SESSION

DATA CENTERS + BIG DATA ANALYTICS

Chair(s)

TBA

Using Random Neural Network for Load Balancing in Data Centers

Peixiang Liu

Graduate School of Computer and Information Sciences
Nova Southeastern University
3301 College Avenue, Fort Lauderdale FL 33314, USA
lpei@nova.edu

***Abstract*—A data center which consists of thousands of connected computer servers can be considered as a shared resource of processing capacity (CPU), memory, and disk space etc. The jobs arriving at the cloud data center are distributed to different servers via different paths. In addition, the internal traffic between servers inside the data center needs to be load balanced to multiple paths between them as well. How to select the underutilized or idle paths for the traffic so as to achieve load balancing and throughput optimality is a big challenge. The Random Neural Network (RNN) is a recurrent neural network in which neurons interact with each other by exchanging excitatory and inhibitory spiking signals. The stochastic excitatory and inhibitory interactions in the network makes the RNN an excellent modeling tool for various interacting entities. It has been applied in a number of applications such as optimization, communication systems, simulation pattern recognition and classification. In this paper, we propose to use Random Neural Network (RNN) to solve the load balancing problem in data centers. RNN is able to achieve adaptive load balancing based on the online measurements of path congestion gathered from the network.**

***Index Terms*—Random Neural Network, Reinforcement Learning, Load Balancing, Data Center**

I. Introduction

In the era of cloud computing, the cloud provides services including full software applications and development platforms, infrastructures such as servers, storage, and virtual desktops to corporate and government organizations, and individual users. A data center which consists of thousands of connected computer servers can be considered as a shared resource of processing capacity (CPU), memory, and disk space etc. Traditionally data centers have been built using hierarchical topologies: edge hosts are organized in racks; each edge host is connected to the Top-of-Rack (ToR) switch; these ToR switches then connect to End-of-Row (EoR) switches, which are interconnected with each other via core switches. Such topologies work well if most of the traffic flows into or out of the data center. However, if most of traffic is internal to the data center, the higher levels of the topology can be a bottleneck due to the uneven distribution of bandwidth.

Recently, other topologies such as FatTree [1] which employs commodity network switches using Clos network, PortLand [2], VL2 [3], and BCube [4] were proposed to address the oversubscription and cross section bandwidth problem faced by the legacy three-tier hierarchical topologies. Depending on the traffic pattern, paths can be congested even if the topology offers 1:1 oversubscription ratio. How to select the underutilized or idle paths to carry the network traffic in order to avoid network congestion and improve the data center throughput is still a big challenge.

Different approaches have been used to spread traffic across different paths in data centers. For example, Equal-Cost Multi-Path (ECMP) [5] splits the flows roughly equally across a set of equal length paths based on the hash of some packet header fields that identify a flow. However, for some topologies such as BCube in which paths vary in length, ECMP is not able to access many paths available in the network since it spreads the traffic across the shortest paths only.

Multipath TCP (MPTCP) [6] was used to improve the data center performance and robustness. MPTCP stripes a single TCP connect across multiple network paths. Rather than sending the traffic on one single network path, additional subflows can be opened between the client and the server either using different ports or any additional IP addresses the client or server may have. MPTCP achieves load balancing by moving traffic off more congested paths and placing it on less congested ones based on the congestion control dynamics on those multiple subflows. In other words, the load balancing is implemented at the transport layer using the TCP congestion control mechanism. MPTCP still relies on the routing algorithm of the data center to select the path for each subflow. Further, MPTCP adds more complexity to transport layer which is already burdened by requirements such as low latency and burst tolerance. Data center fabrics, like the internal fabric within large modular switches, behave like a giant switch. The load balancing function in data center should not be bind to the transport layer. For some data center applications such as high performance storage systems, the kernel is bypassed so MPTCP cannot be used at all.

Alizadeh et al proposed a network based distributed congestion aware load balancing mechanism for data centers, which is called CONGA [7]. In CONGA, TCP flows are split into flowlets, which are assigned to different fabric paths based on the estimated real-time congestion on fabric paths. CONGA operates in an overlay network consisting of ”tunnels” between the fabric’s leaf switches. When an endpoint (server or VM)

sends a packet to the fabric, the source leaf switch determines which destination leaf switch the packet needs to be sent using the destination endpoint's address (either MAC or IP). The packet is then tunnelled from the source to the destination leaf switch. Once the destination leaf switch receives the packet, the original packet is decapsulated and delivered to the intended recipient, the destination endpoint. The path congestion metric is stored at the destination leaf switch on a per source leaf, per path basis and is fed back to the source leaf by piggybacking to support the load balancing decision making at the source leaf switch. The path (uplink port) with the least congestion metric is chosen for the arriving new flowlet.

Since CONGA uses piggybacking to feedback the congestion metric back to the source leaf switch, the metrics may be stale if no sufficient traffic exits for piggybacking. To handle this, CONGA gradually reduces the congestion metric of a port to zero if that metric has not been updated for a long time (e.g., 10ms) to force the port being selected. However, this metric aging approach has following issues. On one hand, if the port is actually congested (not being updated for long time doesn't mean the port is not congested), changing the congestion metric to zero and directing more traffic to this already congested path is not a good decision; on the other hand, if the port is actually less congested than what the congestion metric indicates (the stale congestion metric has a high value because no packets were sent to this port to collect the updated lower congestion metric information), it is OK to send more traffic to this port. However, this less congested port has been left unused for a long time because the source leaf switch thought it was congested based on the stale congestion metric, which is not good either. To gather the congestion metric of a path, that path must be used to transmit network packets. If only the best paths are selected by the source leaf switch for the incoming flowlets, then only the congestion information on those selected paths will be measured. Assigning more traffic to any paths will cause them to be more congested than before. Even if the network has sufficient reverse direction traffic to feed that information back to the source leaf switch, it only confirms that those paths are more congested. CONGA assigns traffic to the least congested paths, which will cause those paths to be more congested; then assign traffic to other less congested paths, which will cause them to be more congested too. Since every path selected by CONGA becomes more congested, finally what the source leaf switch learns is a high congestion metric for every path. There is no positive feedback to inform the source leaf switch that some other paths are idle or less congested since those paths were not explored by any traffic at all. Metric aging, which forces the stale congestion metric to be zero and then selects the path with the stale congestion metric, actually acts like round-robin – the path with the oldest congestion metric is selected.

A Random Neural Network (RNN) is an interconnected recurrent network of neurons, which has the following interesting features: i) *Each neuron in a RNN is represented by its potential, which is a non-negative integer;* ii) *A neuron is considered to be in its "firing state" if it has a positive potential;* and iii) *The signal transmitted between any two neurons are in the form of spikes of a certain rate.* Since the RNN was introduced by Gelenbe [8], it motivated a lot of research which generated various RNN extension models. The RNN has been applied in a number of applications such as optimization, image processing, communication systems, simulation pattern recognition and classification [9].

In this paper, we briefly introduce the RNN model and how the RNN with reinforcement learning was successfully used to design the Cognitive Packet Network (CPN) architecture, which offers adaptive QoS driven routing based on on-line measurement and monitoring to address the users' Quality of Service (QoS) requirements [10]. Then we propose to use RNN with reinforcement learning to select the paths based on the path congestion metric gathered from the network to address the load balancing issue in data centers. The rest of the paper is organized as follows. In section II we present related work on load balancing in data center. In Section III, we briefly describe the mathematical model of RNN and its learning capability in Cognitive Packet Network (CPN). Then we discuss the approach of using RNN with reinforcement learning to solve the load balancing problem in data center in section IV. Finally section V concludes this paper.

II. Related Work

Maguluri et al [11] considered a stochastic mode of jobs arriving at a cloud data center and proposed a load balancing and scheduling algorithm that is throughput-optimal without assuming that job sizes are known or upper-bounded. Paiva et al [12] studied how to assign data items to nodes in a distributed system to optimize one or several of a number of performance criteria such as reducing network congestion, improving load balancing, among others. Grandl et al presented Tetris [13], a multi-resource cluster scheduler that packs tasks to machines based on their requirements of all resource types to avoid resource fragmentation as well as over-allocation of the resources.

In Equal-Cost Multi-Path (ECMP) routing [5], the packets are routed along multiple paths of equal cost. Various methods were proposed for the router to decide which next-hop (path) to used when forwarding a packet. In Hash-threshold, the router selects a key by performing a hash (e.g,. CRC16) over the packet header fields that identify a flow. The N next-hops are assigned unique regions in the key space and the router uses the key to determine which region (next-hop) to use. In Modulo-N algorithm, the packet header fields which describe the flow are run through a hash function. A final modulo-N is applied to the output of the hash which directly maps to one of the N next-hops. Another method is Highest Random Weight (HRW). The router generates a weight using a pseudo-random number generator with packet header fields which describe the flow and the next-hop as seeds. The next-hop which receives the highest weight is selected as the routing option. Basically,

ECMP distributes traffic to multiple paths without considering path quality or congestion.

Raiciu et al discussed how to use Multipath TCP (MPTCP) to improve data center performance and robustness in [6]. Dense interconnection data center topologies provides many parallel paths between pair of hosts. MPTCP establishes multiple subflows on different paths between the same pair of endpoints for a single TCP connection. The intuition of MPTCP is that by exploring multiple paths simultaneously and using the congestion response of subflows on different paths to direct traffic away from congested paths and place it on less congested ones, MPTCP is able to achieve higher network utilization and fairer allocation of capacity of flows.

Alizadeh et al [7] presented the design, implementation, and evaluation of CONGA, a network-based distributed congestion-aware load balancing mechanism for data centers. The majority of the functionality of CONGA resides at the leaf switches. The source leaf switch makes load balancing decisions based on the congestion metrics gathered from the network. CONGA leverages the Virtual eXtensible Local Area Network (VXLAN) encapsulation format used for the overlay to carry the congestion metric exchanged between source and destination leaf switches [14]. It uses the Discounting Rate Estimator (DRE), a simple module present at each fabric link, to measure the link congestion. In each CONGA packet, there are 4 related fields: *LBTag* (4 bits) or load balancing tag, which is the port number of the uplink the packet is sent on by the source leaf switch; *CE* (3 bits), which is used to store the extent of path congestion; *FB_LBTag* (4 bits) and *FB_Metric* (3 bits), which are used by the destination leaves to piggyback congestion information back to the source leaves. Apparently, *FB_LBTag* is the port number and *FB_Metric* is its associated congestion metric. Each leaf switch maintains a flowlet Table which consists of a port number, a valid bit and an age bit. When a packet arrives, the flowlet table is lookup for an entry based on the hash of the fields which are used to identify the flow. If the entry is valid, then the port number in that entry is used, otherwise, this incoming starts a new flowlet and a port is assigned to this flowlet based on the local and remote congestion metrics the leaf switch collected. The leaf switch detects flowlets using a timer and the age bit. Each incoming packet reset the age bit and the timer periodically (every T_{fl} seconds) checks and sets the age bit. If the age bit is set when the timer expires, it means no packets used that entry in the last T_{fl} seconds and the entry times out.

III. Random Neural Network Model

In Random Neural Network model, there are N fully connected neurons which exchange *positive* and *negative* impulse signals. At any time t, each neuron i in the network is represented by its signal potential $k_i(t)$, which is a non-negative integer. The neuron potential is increased by 1 if a *positive* signal arrives. The arrival of a *negative* signal causes the neuron potential to be reduced by 1. Note that *negative* signal has no effect on the neuron potential if it is already zero [15]. Neuron i is called excited when its potential is positive ($k_i(t) > 0$). An excited neuron can fire signals to other neurons in the network or send the signals outside the network. After a neuron fires a signal, its potential is reduced by 1. We use p^+_{ij} and p^-_{ij} to represent the probability that neuron i sends a positive/negative signal to neuron j respectively. $d(i)$ is used to represent the probability that the signal departs from the network. Obviously the following equation holds for all $1 \leq i \leq N$.

$$d(i) + \sum_{j=1}^{N} (p^+_{ij} + p^-_{ij}) = 1 \tag{1}$$

The external inputs to neuron i are modeled with Poisson processes of rate $\Lambda(i)$, and $\lambda(i)$. When neuron i is excited, it fires signals at a rate of $r(i)$, which follows the exponential distribution. Therefore, we have the rates at which positive and negative signals are sent out from neuron i to j, w^+_{ij}, and w^-_{ij}, where $w^+_{ij} = r(i)p^+_{ij}$, and $w^-_{ij} = r(i)p^-_{ij}$. $r(i)$, the total firing rate from the neuron i, can then be expressed as follows,

$$r(i) = \frac{\sum_{j=1}^{N} (w^+_{ij} + w^-_{ij})}{1 - d(i)} \tag{2}$$

Fig. 1 shows an example random neuron network with three neurons i, j, and k.

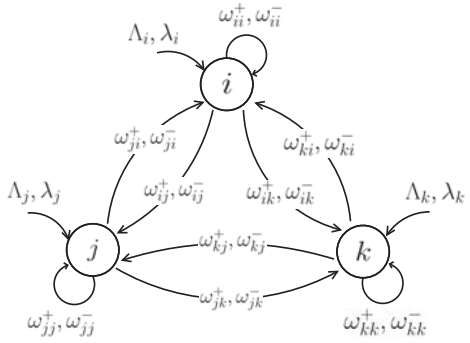


Fig. 1. RNN with 3 neurons

Gelenbe [8] showed that the network's stationary probability distribution can be written as the product of the marginal probabilities of the state of each neuron in the network. A vector of neuron potentials at time t, $k(t) = [k_1(t), k_2(t), ..., k_N(t)]$, is used to describe the network state. Let $k = [k_1, k_2, ..., k_N]$ be a particular value of the vector. The stationary probability distribution can be expressed as $p(k) = \lim_{t\to\infty} Prob[k(t) = k]$ if it exists. Each neuron i in the N-neuron RNN has a state q_i which is the probability that i is excited. The q_i, with $1 \leq i \leq N$, satisfies the following nonlinear equations:

$$q_i = \frac{\lambda^+(i)}{r(i) + \lambda^-(i)} \tag{3}$$

with

$$\lambda^+(i) = \sum_{j=1}^{N} q_j w^+_{ji} + \Lambda_i,$$

$$\lambda^-(i) = \sum_{j=1}^{N} q_j w^-_{ji} + \lambda_i. \tag{4}$$

A. RNN in Cognitive Packet Network (CPN)

RNN with reinforcement learning has been successfully applied in the Cognitive Packet Network (CPN) which provides adaptive QoS driven routing based on online measurements of the network [10], [16]. There are three different types of packets in CPN: Smart Packets (SPs), which explore the network for paths; Dumb Packets (DPs), which carry the payload using the path discovered by SPs; and Acknowledgment Packets (Acks), which bring back the paths discovered by SPs and feed the collected network measurement to RNNs. In CPN, intelligence is introduced to routers in the packet switching network so that the routers are able to learn from their interactions with the network. A RNN is created at each router to make routing decisions. The number of neurons in that RNN is equal to the number of neighbors of the router, with each neighbor represented by a neuron in the RNN. Note that the number of neighbors is actually the number of routing options at the current router. As we discussed earlier, each neuron in a RNN is represented by its signal potential, which is a non-negative integer. The probability that any neuron i is excited, q_i, can be calculated using equation (3). When the RNN is queried for routing suggestion, the output link corresponding to the most excited neuron is returned as the routing decision.

B. Learning process of RNN in CPN

The objective of the learning algorithm is to output a distribution of neuron potential with the neuron corresponding to the desired routing option being most excited so that it can be selected. At each CPN router, we keep a threshold value which denotes the average QoS measurement of the paths from that router to the destination. When an Ack arrives, the instant QoS measurement it carries is compared with the threshold to decide whether the RNN should "reward" or "punish" the routing option the Ack brings back, which is accomplished through adjusting the weight matrices $W^+ = \{w_{ij}^+\}$ and $W^- = \{w_{ij}^-\}$. The routing decisions resulting in higher QoS than the average are rewarded, while the others are punished. This learning process continues until there is no traffic in the network. The new q_i values are calculated using equation (3) every time the weights matrices are modified so that the routing decisions for the following SPs are changed adaptively.

A lot of experiments have been conducted to study how CPN works well to satisfy users' various QoS routing requirements, which include hopcount, delay, and the combination of hopcount and delay [17], [18], [19], [20]. Fig. 2 shows how the average length of paths used by packets changed over time when hopcount was used as the QoS goal. We started the experiments by sending out *Constant Bit Rate* (CBR) of $200pps$ from the source to the destination node in a 26-node testbed for two minutes assuming the source node knew nothing about the network. SPs were generated to explore the network and discover the paths which were then brought back by the ACKs so DPs could use to carry the payload. The same experiments were conducted for 20 times and the average measurements with 95% confidence intervals were reported.

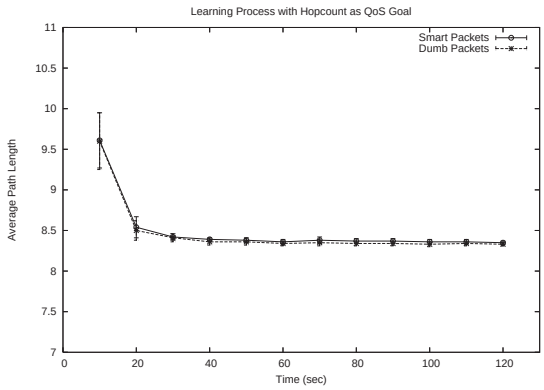


Fig. 2. CPN learning process with hopcount as QoS goal

In Fig. 2, each point represents the average path length used by the packets (SPs or DPs) travelling from the source to the destination in every 10 seconds. We can see that the average length of the paths discovered in the first 10 seconds is about 9.60, and the curve keeps decreasing until it reaches to about 8.34. At the very beginning, the RNNs did not have any knowledge about the network so that the routing decisions they suggested to SPs were not the best options. With the time went on, more SPs were sent out to explore the network and more QoS measurements were fed to the RNNs, which resulted in the weights of RNNs being updated. The more RNNs interact with the network, the more network measurements they learn, and the more reliable were the decisions they made. The decreasing trend of the curves in Fig. 2 clearly shows the learning process of CPN.

In CPN, some SPs do not follow the direction to which the RNNs point, they are routed randomly instead which enables the SPs to explore the network thoroughly rather than stick only to those good paths suggested by RNNs. In our experiments, 5% of the SPs were randomly routed. The length of the shortest paths between the source and destination node in our testbed is 8. Considering the random routing of those 5% smart packets, we can safely draw the conclusion that RNNs have successfully learned the shortest paths when hopcount was used as the QoS goal. Fig. 2 also shows that the average length of the paths used by DPs is almost the same as that of SPs, which is reasonable since DPs use whatever paths the SPs discovers.

IV. RNN for Load Balancing in Data Center

In modern data centers, multiple paths are available for switches to choose from for the same destination. For example, in Fig. 3, a data center with leaf-spine architecture, each leaf switch is fully connected to all four spine switches. In this data center network topology, each leaf switch has 4 switching options to any other leaf switches in the network.

We propose to use Random Neural Network (RNN) to help leaf switches make load balancing decisions. A RNN is created at each leaf switch for every destination leaf switch. The number of neurons in the RNN is equal to the number of uplinks the leaf switch has, with each neuron representing an

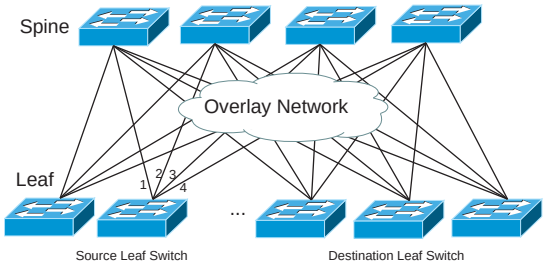


Fig. 3. Data Center Network Topology

outgoing uplink. In Fig. 3, the RNN at each leaf switch has 4 neurons since each leaf switch has 4 uplinks which connect to 4 spine switches. In CONGA [7], the reserved bits in the VXLAN header [14] are used for the source and destination leaf switches to exchange the path congestion information. We have briefly described how the path congestion information can be gathered and fed back to the source leaf switch in section II. In our approach, we use the same method to gather the path congestion metrics from the network. In this section, we focus only on how the gathered path congestion metrics associated with each uplink port are fed to the RNN running at the source leaf switch to help it make adaptive load balancing decisions.

We use Discounting Rate Estimator (DRE) to measure the load of a link. A register, X, is maintained by DRE for each link. X is incremented for every packet sent over the link by the packet size in bytes, and is decremented periodically with a multiplicative factor γ: $X \leftarrow X \times (1-\gamma)$, where $0 < \gamma < 1$. It is easy to show that X is proportional to the rate of traffic over the link, or the congestion level of the link [7]. A lower X indicates the link is less congested. Then

$$X_{path} = \max_{1 \le i \le n} \{X_i\} \tag{5}$$

can be used to represent the congestion level of a path with n links, where X_i is the congestion measurement of the ith link of the path. X_{path} is used as the QoS measurement of a path between the source and destination leaf switch. The QoS goal is the metric which characterizes the quality of each outgoing uplink, obviously, smaller values are preferred for path congestion, X_{path}.

Given the QoS goal X_{path} that the load balancing module must achieve as a function to be minimized, we can formulate a reward function $R = Reward(X_{path})$, which can be seen as the quantitative criterion to measure the QoS of the path, with higher reward value indicating better QoS (the path is less congested). A simple example reward function can be as follows,

$$R = \frac{1}{\beta \cdot X_{path}} \tag{6}$$

where β is a constant parameter.

Obviously, the reward function can be applied to the QoS measurement piggybacked in the packets sent from the destination leaves to the source leaf switch for us to judge the QoS of paths using different outgoing uplinks. The congestion metric carried in *FB_Metric* field represents the QoS of paths using outgoing uplink identified by *FB_LBTag*. The successive calculated values of R based on the QoS measurements piggybacked in different packets received by the source leaf switch are denoted by $R_l, l = 1, 2, \cdots$, which are used to compute a decision threshold T_l,

$$T_l = \alpha T_{l-1} + (1-\alpha) R_l \tag{7}$$

where α is some constant ($0 < \alpha < 1$), which is typically close to 1. R_l is the most recent value of the reward.

The reinforcement learning algorithm uses T_l to keep track of historical value of the reward. T_l can be considered as the average congestion metric (QoS) of paths from the source to the destination leaf switch using any uplinks. Suppose we have made the lth decision which corresponds to an outgoing uplink (neuron) j and that the lth reward calculated for the path congestion received is R_l. We first determine whether R_l is larger than, or equal to, the threshold T_{l-1}; if this is the case, it means the instant measured QoS for outgoing uplink j is better or not worse than the threshold QoS. In other words, it means the paths using outgoing uplink j are less congested or not worse than those paths using the other uplinks to reach destination leaf switch.

Once a neuron in RNN is excited, it sends out "excitation spikes" and "inhibition spikes" to all the other neurons at different firing rates, which are defined as the positive or negative weights: w_{ij}^+ is the rate at which neuron i sends "excitation spikes" to neuron j when neuron i is excited and w_{ij}^- is the rate at which neuron i sends "inhibition spikes" to neuron j when neuron i is excited. Since the paths for the destination leaf switch via uplink j are less congested or not worse than paths using other uplinks, we increase very significantly the excitatory weights going into neuron j and make a small increase of the inhibitory weights leading to other neurons in order to reward it for its success; otherwise, if R_l is less than T_{l-1}(the measured path congestion is worse than the threshold QoS), we simply increase moderately the excitatory weights leading to all neurons other than j and increase significantly the inhibitory weight leading to neuron j in order to punish it for its not being very successful this time.

- If $T_{l-1} \le R_l$

$$\begin{aligned} w_{ij}^+ &\leftarrow w_{ij}^+ + R_l \\ w_{ik}^- &\leftarrow w_{ik}^- + R_l/(N-1), \text{ for } k \ne j \end{aligned} \tag{8}$$

- Else

$$\begin{aligned} w_{ik}^+ &\leftarrow w_{ik}^+ + R_l/(N-1), \text{ for } k \ne j \\ w_{ij}^- &\leftarrow w_{ij}^- + R_l \end{aligned} \tag{9}$$

Once the weights are updated, the probability that each neuron i is excited, q_i, is computed using the nonlinear iterations (3) and (4) presented in section III. In the RNN model for load balancing, parameters $\Lambda(i)$ and $\lambda(i)$ can be set as constant values. The uplink associated with the most excited neuron is the best option when the path congestion is

considered as the QoS goal. We define p_j, the probability that uplink j is selected by the source leaf switch to send traffic flowlet to the destination leaf switch, as follows,

$$p_j = \frac{q_j}{\sum_{i=1}^{N} q_i}, \text{ for } 1 \leq j \leq N \tag{10}$$

where N is the total number of uplinks of the source leaf switch. When a flowlet is detected, the source leaf switch first determines the destination leaf switch based on the fields in the packet header used for flow identification, then it consults the RNN of that destination leaf switch for $p_j, 1 \leq j \leq N$ to help make load balancing decisions, the uplink j with the highest p_j will most probably be selected.

In our approach, the RNN is trained by the congestion metrics of the paths in data center. The congestion metric of a path is gathered by packets using the path. Along its way traveling from the source to its destination, each packet checks the congestion level of every link measured by Discounting Rate Estimator (DRE). The maximum among those link congestion metrics is used to as the path congestion metric (see equation 5). The unused bits in the VXLAN header of the VXLAN frames exchanged between the leaf switches are used to carried the congestion metric of a path. As long as there exists network traffic in data center, the congestion metric of the path being used will be collected by network packets and fed back to the RNN running at the source leaf switch. With reinforcement learning, the RNN is able to tune its parameters based on the online path congestion measurements so as to offer adaptive load balancing for data centers. The uplinks are selected based on the congestion metrics of the paths using each uplink: the port with less congested paths is more likely to be selected than the port with more congested paths to service the incoming traffic.

Unlike CONGA, in which the uplink with the least congestion metric is always chosen to service the incoming flowlet, RNN distributes the incoming traffic flowlets to different uplinks based on the congestion metrics of the paths using each uplink. RNN avoids the undesirable output we described in section I that CONGA has. The introduction of a little bit randomness to RNN's load balancing decision logic will ensure every path in the network be explored so that the congestion metric of the path is brought back to the source leaf switch to support RNN to make better load balancing decisions.

V. Conclusions

In this paper, we proposed to use Random Neural Network (RNN) to address the loading balancing problem in data centers. We briefly introduced the RNN model, discussed the learning process of RNNs in Cognitive Packet Network (CPN) and presented experiment results to demonstrate that RNN was able to provide adaptive, QoS driven routing to network packets based on online monitoring and measurements. We described how to use a similar approach in data centers to achieve adaptive load balancing based on the online measurements of path congestion. In our approach, every path in the network gets the opportunity to be measured therefore more valuable path congestion metrics are fed back to the RNNs located at every leaf switch in data center to make better load balancing decisions.

References

[1] Mohammad Al-Fares, Alexander Loukissas, and Amin Vahdat. A scalable, commodity data center network architecture. *SIGCOMM Comput. Commun. Rev.*, 38(4):63–74, August 2008.

[2] Radhika Niranjan Mysore, Andreas Pamboris, Nathan Farrington, Nelson Huang, Pardis Miri, Sivasankar Radhakrishnan, Vikram Subramanya, and Amin Vahdat. Portland: A scalable fault-tolerant layer 2 data center network fabric. *SIGCOMM Comput. Commun. Rev.*, 39(4):39–50, August 2009.

[3] Albert Greenberg, James R. Hamilton, Navendu Jain, Srikanth Kandula, Changhoon Kim, Parantap Lahiri, David A. Maltz, Parveen Patel, and Sudipta Sengupta. Vl2: A scalable and flexible data center network. *SIGCOMM Comput. Commun. Rev.*, 39(4):51–62, August 2009.

[4] Chuanxiong Guo, Guohan Lu, Dan Li, Haitao Wu, Xuan Zhang, Yunfeng Shi, Chen Tian, Yongguang Zhang, and Songwu Lu. Bcube: A high performance, server-centric network architecture for modular data centers. *SIGCOMM Comput. Commun. Rev.*, 39(4):63–74, August 2009.

[5] C. Hopps. Analysis of an Equal-Cost Multi-Path Algorithm. RFC 2992 (Informational), November 2000.

[6] Costin Raiciu, Sebastien Barre, Christopher Pluntke, Adam Greenhalgh, Damon Wischik, and Mark Handley. Improving datacenter performance and robustness with multipath tcp. *SIGCOMM Comput. Commun. Rev.*, 41(4):266–277, August 2011.

[7] Mohammad Alizadeh, Tom Edsall, Sarang Dharmapurikar, Ramanan Vaidyanathan, Kevin Chu, Andy Fingerhut, Vinh The Lam, Francis Matus, Rong Pan, Navindra Yadav, and George Varghese. Conga: Distributed congestion-aware load balancing for datacenters. In *Proceedings of the 2014 ACM Conference on SIGCOMM*, SIGCOMM '14, pages 503–514, New York, NY, USA, 2014. ACM.

[8] E. Gelenbe. Random neural networks with negative and positive signals and product form solution. *Neural Comput.*, 1(4):502–510, December 1989.

[9] Stelios Timotheou. The random neural network: A survey. *The Computer Journal*, 53(3):251–267, 2010.

[10] Georgia Sakellari. The cognitive packet network: A survey. *The Computer Journal*, 53(3):268–279, 2010.

[11] Siva Theja Maguluri and R. Srikant. Scheduling jobs with unknown duration in clouds. *IEEE/ACM Trans. Netw.*, 22(6):1938–1951, December 2014.

[12] João Paiva and Luís Rodrigues. On data placement in distributed systems. *SIGOPS Oper. Syst. Rev.*, 49(1):126–130, January 2015.

[13] Robert Grandl, Ganesh Ananthanarayanan, Srikanth Kandula, Sriram Rao, and Aditya Akella. Multi-resource packing for cluster schedulers. *SIGCOMM Comput. Commun. Rev.*, 44(4):455–466, August 2014.

[14] M. Mahalingam, D. Dutt, K. Duda, P. Agarwal, L. Kreeger, T. Sridhar, M. Bursell, and C. Wright. Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks. RFC 7348 (Informational), August 2014.

[15] Erol Gelenbe. G-networks with triggered customer movement. *Journal of Applied Probability*, 30(3):pp. 742–748, 1993.

[16] Erol Gelenbe. Cognitive Packet Network. *U.S. Patent No. 6,804,201*, October 2004.

[17] Michael Gellman and Peixiang Liu. Random neural networks for the adaptive control of packet networks. In *Proceedings of the 16th international conference on Artificial Neural Networks - Volume Part I*, ICANN'06, pages 313–320, Berlin, Heidelberg, 2006. Springer-Verlag.

[18] Erol Gelenbe and Peixiang Liu. Qos and routing in the cognitive packet network. In *World of Wireless Mobile and Multimedia Networks, 2005. WoWMoM 2005. Sixth IEEE International Symposium on a*, pages 517–521, 2005.

[19] Peixiang Liu and Erol Gelenbe. Recursive routing in the cognitive packet network. In *Testbeds and Research Infrastructure for the Development of Networks and Communities, 2007. TridentCom 2007. 3rd International Conference on*, pages 1–6, 2007.

[20] Peixiang Liu. The random neural network and its learning process in cognitive packet networks. In *Natural Computation (ICNC), 2013 Ninth International Conference on*, pages 95–100, July 2013.

Generic Architecture for Building Knowledge Base Automatically by Analyzing Big Data Available in the Web Environment

Marcus Vinicius da Silveira and Gwang Jung
Department of Mathematics and Computer Science
Lehman College, the City University of New York
250 Bedford Park Blvd West
Bronx, NY
marcus.silveira@lc.cuny.edu; gwang.jung@lehman.cuny.edu

Abstract-- *In this paper, we present generic architecture for building domain specific knowledge bases that can be automatically acquired by analyzing big data collected from the web environment. As a reference implementation, Nursing Home Application is developed based on our proposed architecture. The initial experiment result shows valuable warrant for future studies.*

Keywords: Big Data Analysis, Automatic Knowledge Acquisition, Web Data Mining, Internet Programming

1. Introduction

From the inception of web space, endless resources of information for users to search for every day are available in the web space, and users seek intelligent search for effectively finding useful information for their daily need. For example, given some medical symptoms, users want to search a list of medical specialists in a specific geographic region. Or users want to find restaurants in a specific cuisine that meet their need in terms of quality, prices, atmosphere, and so forth. In other words, users become more and more interested in getting effective answers from the web as a Knowledge Based System (KBS).

Traditionally, a KBS often termed as an Expert System [1], makes extensive use of specialized domain knowledge to solve problems at the level of human expert. The knowledge base used in such a KBS may be either expertise elicited from human domain expert(s) or knowledge which can be extracted from resources such as professional books or magazines.

In general, as shown in the Figure 1.1, a KBS contains the knowledge base with which the inference engine of the KBS makes a conclusion. These conclusions made by the KBS are answers with respect to the user's query. The answers are formulated as facts or a list of potentially relevant suggestions.

The KBS is generally designed for a specific problem domain. The conventional way of building a knowledge base is carried out by having knowledge engineers repeat the cycle of interviewing the domain experts, constructing a prototype, testing, and interviewing again. Such knowledge acquisition process is very time consuming and labor intensive task. To extract domain knowledge from a big data set often available in the web space is an alternative way to get knowledge from domain experts.

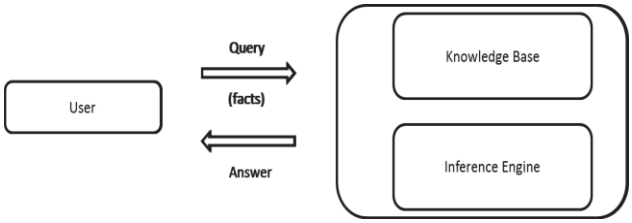


Figure 1.1: Simple View of a Domain KBS

To reduce knowledge acquisition effort, knowledge bases used in KBS are often formulated as heuristics [1]. Heuristics are termed as shallow knowledge or empirical knowledge gained from experience which may help producing potentially useful solutions.

One of the most important and difficult knowledge acquisition tasks is to generalize the target domain to formulate (or discover) its primary concepts (or conceptual categories), and terms (or attributes) that describe a concept, and the relationships between concepts and terms. Knowledge acquisition task should involve a process for identifying important concepts in a specific domain and involve a process of determining the degree of relationships between concepts and terms. Such relationships are heuristics that are the basis of the KBS.

In this paper, we are motivated to develop a generic architecture which automatically builds a domain specific knowledge base by analyzing a big data set collected from the web space. A list of concepts and terms (attributes) that describe concepts are assumed to be given or extracted from a set of web sites in a specific domain.

The proposed method is based on statistical feature of concept and term co-occurrence in a specific domain. If a term co-occurs with a concept in many web documents, the term is considered more relevant to the concept. Such heuristics are basis for our knowledge base.

The concepts in a specific domain in the proposed architecture can be manually derived or extracted from the semantic web based Resource Description Framework (RDF) in a specific domain. For example, as a case study, we developed a prototype which is used to search medical specialties, for nursing home application, based on the knowledge base (heuristics) acquired automatically by the proposed knowledge acquisition system. The prototype is based on the concepts (i.e., specialties) and terms (i.e., symptoms) extracted from the RDF of well-known medical institutions [4, 5].

The salient features of the proposed architecture are summarized below:

1. Knowledge base in a specific domain is automatically acquired by analyzing term co-occurrence between a set of concepts and a set of terms. Term co-occurrence frequencies are obtained by a simple web crawler that makes use of existing search engines such as Bing and Altavista.
2. Knowledge acquisition of the proposed architecture is generic in the sense that the weight (or degree of relationship) between concept and a term can be automatically calculated in any domain. Term frequency and inverse term frequency are used to calculate the degree of relationship more accurately for effective query processing.
3. Knowledge representation is based on simple inverted file, which can be formulated as normalized tables of a simple database schema in a DBMS.
4. User query is a simple vector whose element represents the importance (normalized weight) of a term in the query.
5. Inference engine is based on the cosine similarity measure between user query vector and a concept vector. Inference engine (i.e., query processor) is thus time efficient because of simple query and knowledge representations

In section 2, automatic knowledge acquisition along with overview of the proposed architecture will be explained. In section 3, two different algorithms to extract knowledge base from the collected data set are presented. An algorithm for implementing query processor as an inference engine is also explained.

In section 4, we present the Nursing Home Application (NHA) as a reference implementation of the proposed generic knowledge based search architecture. In section 4, experiment results are shown for validating the utility of the proposed architecture based on NHA as a case study. Conclusion of the paper is presented in section 5.

2. Proposed Architecture and Knowledge Acquisition Framework

Knowledge base is formulated by first identifying the important terms and concepts, and then determining the interrelationships between concepts and terms in a specific domain.

Personal Construct Theory (PCT) has been frequently used to elicit concepts and terms in a specific domain from domain expert [6]. However, PCT based approach for eliciting concepts and terms often requires much longer turnaround time. In our case conceptualization processes are carried out by extracting concepts from well-defined RDF based data set in a specific domain.

Upon completing the conceptualization process, our web brokering architecture will run a web crawler which collects concepts and term co-occurrence statistics from a well-known search engines (such as Altavista, Bing, Yahoo, Google, etc..). As an example, our web crawler will give a query that consists of concept C_i and term t_j pair, and get the frequency of documents where C_i and t_j co-occurs from each search engine. This querying process continues until we exhaustively collect co-occurrence document frequency statistics with respect to all concept and term value pairs. The frequency statistics are collected from multiple search engines to get unbiased statistics (refer to the Figure 2.1). Inverse term frequency is also considered to calculate accurate relationship between a term and a concept.

Initially, co-occurrence frequency matrix between concepts and terms is formulated in our document frequency database. Multiple document frequency matrices are constructed for multiple search engines. The frequency will be analyzed to calculate importance of each term in a concept.

For example, in a specific medical domain, given a concept C_1 (specialty) named "internal medicine", and a set of terms (symptoms) such as "loss of weight", "loss of appetite", and "low blood pressure".

Then the frequency vector corresponding to the concept C_1 will be converted into a weight vector that can be represented as an inverted list, as shown in Figure 2.2. All these inverted lists of all the concepts will be generated and converted into the normalized database tables as the knowledge base of the proposed architecture.

Maintenance of concepts and terms are carried out by the administrator as shown in the Figure 2.1. Term frequency and inverse term frequency are used to calculate the degree of relationship between a concept and a term more accurately for effective query processing. In section 3, the use of inverse term frequency is illustrated when the degree of relationship is calculated.

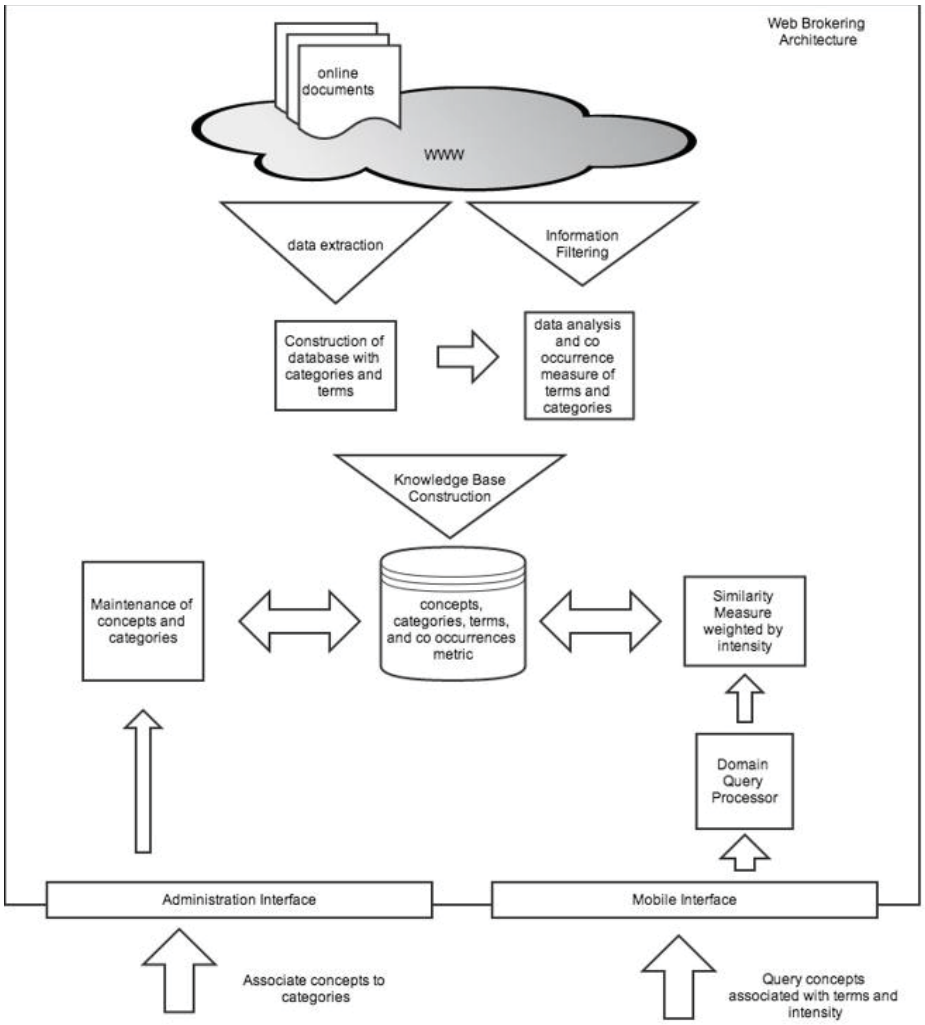


Figure 2.1: Automatic Knowledge Acquisition Framework

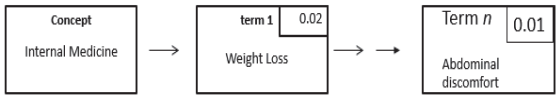


Figure 2.2: An Example of Inverted list of the concept C_1 "Internal Medicine"

Domain query processor is our inference engine for the intelligent web search architecture. Query is formulated by a user as a vector whose element represents an importance of a term in a query. The query processor makes inference to calculate the similarity between a user query and the concepts based on the knowledge base.

3. Algorithms for Building Knowledge Base and Query Processor

In this section, we explain how to calculate the importance (weight) of a term t_j in a concept C_i, which is the central component of our proposed knowledge acquisition system. We then present an inference algorithm that is the basis of the query processor.

3.1 Bucket Algorithm

Given a set of concepts and a set of terms that describe concepts, we start by checking the co-occurrence of both concept and terms from the web document frequency database we collected from the web (refer to the Section 2).

Our novel bucket algorithm can be explained by the following example: if a particular concept C_i, and term t_j pair (C_i,, t_j) co-occurs in one million documents, while the concept C_i, and term t_k (C_i,, t_k) co-occurs in one thousand web documents, then the correlation (or degree of relationship) of (C_i, t_j) is stronger than (C_i,, t_k). We need to consider the co-occurrence frequencies of all (C_i, t_j) pairs in the collected

frequency database, where | C_i | is n and |t| is m. In other words, the number of concepts (i.e., conceptual categories) in a domain is n and the number of terms that describe each concept C_i is m.

There is a need to normalize the data so that multiple search engines use the same scale. For example, search engine Bing could return frequency values lying between 1 thousand and 1 billion, while Altavista could return frequencies between 500 and 500 million. The way to normalize document frequency data is the basis of the bucket algorithm, where min and max results found for any term t_j of a particular concept C_i, and the possible values are divided in 10 buckets with the Equation 1 shown below. Let us define *f*, *max*, and *min* as follows:

f: {{$C_i \cap t_j$ | the number of web documents where C_i and t_j co-occurred}
max: {{$C_i \cap t_j$ | the maximum number of web documents where C_i and t_j co-occurred, where j = 1 to m and m is the number of terms}
min: {{$C_i \cap t_j$ | the minimum number of web documents where C_i and t_j co-occurred, where j = 1 to m and m is the number of terms}

$$W(C_i,t_j) = trunc\left(\frac{f-min}{max-min}\right) - 1 \quad (1)$$

The equation 1 explains that by knowing the difference between the results found for a concept C_i and a term t_j minus the lower bound (min) found for the concept C_i, we can then divide by the range of term co-occurrence frequency for the C_i (max – min). This will give us an integer value between in an interval [0, 9]. We take a coarse granularity of 10 buckets, but this could be changed into more number of buckets if finer grained analysis is needed. Bucket algorithm gives us a set of importance of terms in a set of concepts for different search engines.

3.2 Algorithm based on Jaccard Coefficient

Another algorithm to calculate the relationship between C_i and t_j is based on Jacard Coefficient [8].

Let |C_i| denote the total number of web documents that include conceptual category C_i. Let |T_j| denote the total number of web pages that include a term t_j. Let |C_i,, t_j| denote the number of documents where C_i and T_j co-occurs.

Degree of relationship strength between concept C_i and t_j is calculated by the following equation.

$$W(C_i,t_j) = \frac{|C_i \cap T_j|}{|C_i \cup T_j|} = \frac{|C_i \cap T_j|}{|C_i| + |T_j| - |C_i \cap T_j|} \quad (2)$$

As an example, in a medical domain, let SP_i be a concept $Specialty_i$ and S_j be a term $Symptom_j$. In other words, S_j is a term that describes a concept SP_j in generic knowledge representation. By the Equation 2, we derive

$$W(SP_i,S_j) = \frac{nofdocs\,(SP_i \cap S_j)}{nofdocs\,(S_j) + nofdocs\,(SP_i) - nofdocs\,(SP_i \cap S_j)}$$

, where *noofdocs* represents number of documents co-occurred (3)

Let's assume we have 16 concepts (i.e., specialties) and 850 terms (i.e., symptoms) to be considered. Then,

$$nofdocs\,(S_i) = \sum_{i=1}^{16} nofdocs(S_i)\text{, and}$$
$$nofdocs\,(SP_j) = \sum_{i=1}^{850} nofdocs(SP_j).$$

3.3. Similarity Measure (Used by the Query Processor as the inference engine)

The similarity between a user query q and concept C_i is calculate by the Equation 4.

$$Similarity(C_i,q) = \frac{\vec{C_i} \cdot \vec{q}}{|\vec{C_i}| * |\vec{q}|} \quad (4)$$

Let n be the number of terms (descriptors) used to describe a concept C_j, and q_i be the importance of each term t_i assigned by the user, and w_{ij} be the degree of association between term t_i and concept C_i. Equation 4 can be rewritten as cosine similarity value between concept vector C_i and query vector q, as shown in the Equation 5 [3, 4]. $Sim(C_i,q)$ in the equation 5 represents the similarity value of C_i with respect to q.

$$Sim(C_i,q) = \frac{\sum_{i=1}^{n} w_{ij} * q_i}{\sqrt{\sum_{i=1}^{n} w_{ij}^2} * \sqrt{\sum_{i=1}^{n} q_i^2}} \quad (5)$$

When we calculate the similarity value between concept C_i and a user query q by the Equation 5, we consider inverse concept frequency (ICF) for calculating w_{ij} (the importance of term t_j in concept C_i). ICF is based on the observation that the terms which rarely occur in the Knowledge Base (KB) are considered more informative.

Let |C| be the total number of concepts (e.g., specialties) in the KB, and let $W(C_i,t_j)$ be the importance of term t_j in the concept C_i as we calculated before based on bucket algorithm and Jacaard Coefficient algorithm.

ICF based w_{ij} is calculated by the following two equations:

$$ICF = \log_{10}\frac{|C|}{W(C_i,t_j)} \quad (6)$$
$$w_{ij} = W(C_i,t_j) \times ICF \quad (7)$$

4. Case Study: Nursing Home Application (NHA)

In developed countries, the aging population demand improvements on nursing homes, since as the concentration of elder people become prominent and people are busier than ever, better services need to be provided in order to maintain a good level of care to this aging population when they most need it.

In addition, since most of the nursing homes have limited information infra-structure, finding the most appropriate medical doctor with specialties who can take care of the patient's emergency case is often difficult and not agile.

In this section, we present Nursing Home Application (NHA) as a reference implementation of the proposed generic architecture, on whose basis the knowledge based system (KBS) can be built based on web data analysis, as we explained in the previous sections.

NHA is a specialized application responsible for identifying which doctors (with specific specialties) are required based on a list of symptoms provided by the nursing home staff.

The idea is to quickly type in what symptoms the patient is feeling, and let the NHA decide which doctors should be notified based on that input. Once those are identified, the system broadcasts notifications to all available nursing home affiliated doctors for that specialty, trying to get an appointment as soon as possible. At that point, the system did its job and it's up to the doctors and nursing home staff to arrange when and where the appointment will happen.

4.1 Nursing Home Application (NHA)

The NHA involves the development of two applications. First, a mobile client where nursing home staff can interact with the application, provide patients' symptoms, and notify the emergency to the searched doctors. Second, a backend server that is responsible for providing web-services to be accessed by the client (mobile or desktop PC). The back end server implements the inference engine (refer to the section 3.3) that draws decision effectively and time efficiently by combining medical knowledge base system (KBS) automatically built (see sections 2 and 3). NHA will help nursing home staff members to identify the most appropriate doctors with specialties, and NHA automatically notifies the participating doctors to handle accordingly.

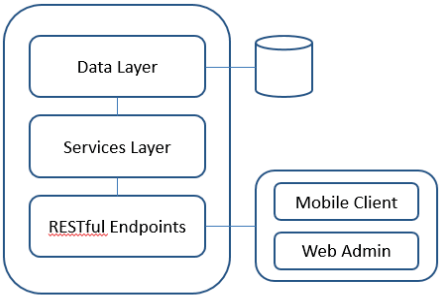


Figure 4.1: Architecture of the NHA

The realm of the application constitutes of a back-end component – responsible for diagnosing, assigning a particular doctor to a patient, and managing the information; a Mobile (or desktop PC) front-end component – responsible for presenting the user interface. An administration front-end component, responsible for presenting the management user interfaces where an administrator manage KBS, inference engine, and other data sets necessary for the back end web-services.

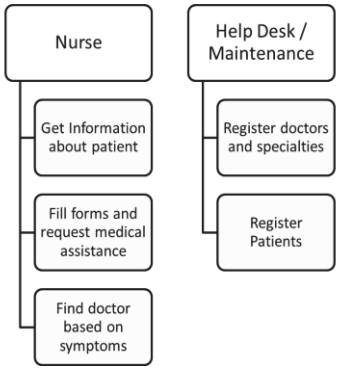


Figure 4.2: Interactions between Users and the NHA System

The core logic of this application relates to how it can search nursing home affiliated doctors based on the symptoms the patients are feeling, as shown in the Figure 4.2 and 4.3.

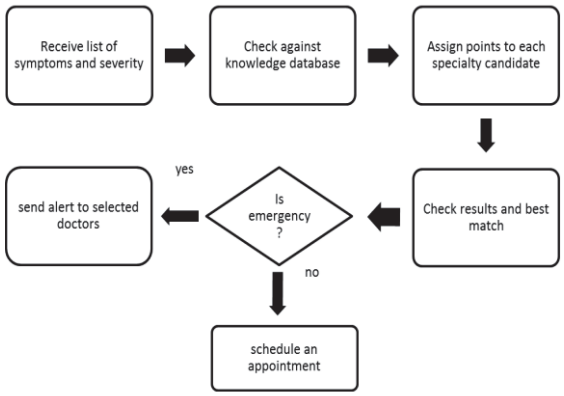


Figure 4.3: Control Flow Diagram for Finding the Doctor

4.2 Building the Knowledge Base for the NHA

Our proposed architecture requires a quality knowledge base. As explained in sections 2 and 3, the most important data that is needed for building the knowledge base for NHA is the correlation between specialties (concepts) and symptoms (terms). The stronger a particular symptom is for a specialty, easier it is to identify the patient's problem to converge into a specialty candidate and appropriate related doctors. At the same time, we consider inverse concept frequency to calculate

the relevant scores by the inference engine (see equation 6 and 7)

In order to construct a knowledge base we need to define what concepts and terms are important to the domain of study, where they could be acquired, and how. For the Nursing Home Application, we identified medical specialties and symptoms as the concepts and terms to be analyzed. Next, we researched on the web for websites that listed both of them, for convenience. Since we weren't successful on trying to find both on the same document, we researched them independently and finally got two well-known resources: the *American Board of Medical Specialties* (ABMS), for our list of specialties, and *Gemina*, a project from the University of Maryland that studies symptoms and genomic sequences. Since the information was available on the website, after evaluating the HTML code of the web pages, we built scripts that parsed the content, saved into files that were then used as input to our knowledge acquisition system that is outlined in the Figure 2.1.

More specifically, first, we discovered a set of websites (Medical Institutions and Universities) that contained either a list of symptoms or a list of specialties. In order to extract data from these websites, we used JavaScript and the JQuery library to analyze the HTML structure, select the list of elements, and output the list to be exported/saved to a file. Using a simple standalone application, this file is then read and saved into the database for later use into building our knowledge base.

We start by checking the co-occurrence frequencies of both specialty and symptom from the well-known search engines from the web. Co-occurrence frequency is the number of documents retrieved with respect to a pair of a specialty and a symptom.

From the results, instead of trying to analyze what each document understands from the query, we rely on what the search engines do best: find documents that contain the terms you specify. That said, if a particular symptom x and specialty y pair (x, y) returns 1 million documents, while symptom u and specialty v (u, v) returns 1 thousand results, that means the correlation of (x, y) is stronger than (u, v). In other words, x and y are common to be seen together, and doctors on the field of y would know better how to treat or refer to a coworker that knows how to treat symptom x.

The idea then is to come up with the product of all symptoms with specialties and perform web searches for each, counting the total number of returned documents. With that number, there is a need to normalize the data so that multiple search engines use the same scale (e.g. Bing could return values between 1 thousand and 1 billion, while AltaVista could return values between 500 and 500 million.) when the correlation between a symptom and specialty pair is calculated.

We could build knowledge bases based on collected symptom and specialty co-occurrence values based on various web search engines. We collected such co-occurrence statistics for more than 1600 symptoms and more than 100 specialties.

For the NHA application, we built User Interfaces to interact with the mobile application, and web-based application.

4.3 Experiment Result for the Nursing Home Application

To evaluate the effectiveness of the knowledge base that was automatically acquired by the proposed architecture for the NHA application, nine queries are formulated as shown in Table 4.1. A query is formulated by extracting key words from the description of each disease from web sources (e.g., Alzheimer's Disease from http://www.wakehealth.edu/Health-Encyclopedia/Health-Topics/Alzheimers-Disease.html)

Table 4.1: **Nine Queries Used for Experiment**

Disease	Relevant Specialties	Symptoms (Query)
Brain Tumor	neurology or neurology with specialization Geriatric Medicine	**Q1:** headache, vomiting, inability to speak, loss of consciousness
Anemia	Hematology Neurology with Special Qualification Neurology Cardiology Diabetes and Metabolism	**Q2:** weakness, breathing problems, dizziness, fluctuation of heart rate, abnormal heart rhythms, headache, change in skin color, chest pain
Alzheimer's	geriatric medicine	**Q3:** memory loss, confusion, memory impairment, inability to form words, inability to think clearly
Osteoporosis	geriatric medicine	**Q4:** pain, joint pain, hip pain
IBS (Irritable Bowel Syndrome)	Gastroenterology	**Q5:** abdominal pain, gas pain, constipation, diarrhea
Pancreatitis	Gastroenterology	**Q6:** lesions in pancreas, loss of weight
Congestive Heart Failure	Cardiology	**Q7:** breathing problems, cough, wheezing, fatigue, leg swelling, abdominal swelling
Asbestosis	Pulmonary disease	**Q8:** breathing problems, chronic cough, chest pain, loss of appetite, abnormal chest sound
Asthma	Pulmonary disease	**Q9:** wheezing, breathing problems, cough, abnormal chest sound

We used three search engines to collect specialty symptom co-occurrence data for building knowledge base from the cyber space: Bing, AltaVista, and Blekko. For the performance evaluation of the effectiveness of the knowledgebase, we select 16 specialties out of 123 specialties for a simple illustration of the utility of the NHA application. The Table 4.2 is Specialty table each specialty is identified with identification number (ID). For each query, relevant specialties are found in the selected set of 16 specialties.

Table 4.2: List of Selected Specialties

ID	SPECIALTY
10	Cardiology
11	Cardiovascular Disease
12	Diabetes and Metabolism
13	Family Medicine
14	Gastroenterology
15	Geriatric Medicine
16	Geriatric Psychiatry
17	Gynecologic Oncology
18	Hematology
19	Internal Medicine
20	Internal Medicine - Critical Care Medicine
21	Neurology
22	Neurology with Special Qualification
23	Orthopedic Surgery
24	Pain Medicine
25	Pulmonary Disease

The average recall and precision measures [7] are used to evaluate the effectiveness of knowledge base that were built by big data analysis based on our generic architecture. Precision is the ratio of the number of relevant specialties retrieved to the total number of irrelevant and relevant specialties retrieved among the selected 16 specialties. Recall is the ratio of the number of relevant specialties retrieved to the total number of relevant specialties in the knowledge base.

The relevant specialties with respect to the disease as a query Q1, as shown in the Table 4.1. The Table 4.2 shows the list of selected specialties to be considered for the experiment.

We calculate the average recall and precision values based on knowledge bases formulated by two different knowledge acquisition algorithms (Bucket algorithm, Jacaard Coefficient-based algorithm; see section 3). When similarity is calculated we used the Equation 5 and term co-occurrence and Jacaard tie-strength with ICF (Inverse Conceptual Frequency) as we explained in the section 3.3.

The following Table 4.3 shows the recall and precision values for the knowledge bases built based on analyzing big data collected from four search engines with respect to nine queries listed in the Table 4.2.

Table 4.3: Retrieval Performance in terms of Average Recall (AR) and Average Precision (AP)

Search Engine and Algorithm	Q1 (AR, AP)	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9
Bing (Bucket)	(0.67, 0.37)	(0.60, 0.29)	(0.63, 0.53)	(0.67, 0.66)	(1.00, 0.14)	(0.75, 0.16)	(0.63, 1.0)	(0.75, 0.42)	(0.75, 0.24)
Bing (Jacaard)	(0.67, 0.63)	(0.60, 0.58)	(0.63, 0.49)	(0.67, 0.58)	(1.0, 0.2)	(0.75, 0.27)	(0.63, 0.50)	(0.75, 1.0)	(0.75, 0.75)
Altavista (Bucket)	(0.67, 0.33)	(0.6, 0.48)	(0.63, 0.54)	(0.67, 0.32)	(1.0, 0.2)	(0.75, 0.13)	(0.63, 1.0)	(0.75, 0.23)	(0.75, 0.26)
Altavista (Jacaard)	(0.67, 0.52)	(0.6, 0.33)	(0.63, 0.33)	(0.67, 0.34)	(1.0, 0.5)	(0.75, 0.67)	(0.63, 0.38)	(0.75, 0.58)	(0.75, 1.0)
Blekko (Bucket)	(0.67, 0.37)	(0.6, 0.58)	(0.63, 0.38)	(0.67, 0.37)	(1.0, 0.09)	(0.75, 0.11)	(0.63, 0.82)	(0.75, 0.12)	(0.75, 0.33)
Blekko (Jacaard)	(0.67, 0.35)	(0.6, 0.61)	(0.63, 0.38)	(0.67, 0.53)	(1.0, 0.2)	(0.75, 0.75)	(0.63, 0.29)	(0.75, 0.21)	(0.75, 0.33)

We could get effective retrieval performance with the automatically generated knowledge bases created based big data analysis based on our proposed knowledge acquisition architecture shown in the Figure 2.1. In general, Jacarrd algorithm slightly outperforms the bucket algorithm for formulating the knowledge bases.

Bing based search performance is the best among the three search engines, which indicates more number of data

collected is more useful for building accurate knowledge base.

We could not use google search engine because of restrictions of using software generated queries. But with proper loyalty payment, we may be able to get more accurate knowledge base from google big document databases.

5. Conclusion

In this paper, we present generic architecture for building domain specific knowledge bases that can be automatically acquired by analyzing big data collected from the web environment. A list of concepts and terms (attributes) that describe concepts are assumed to be given or extracted from a set of web sites in a specific domain. We presented two different algorithms for analyzing big data for building knowledge base from the collected data sets. An algorithm for implementing query processor as an inference engine is also presented.

As a reference implementation of the proposed generic knowledge based search architecture, Nursing Home Application (NHA) was developed. Limited but meaningful experiment result is shown for validating the utility of the proposed architecture based on NHA as a case study.

Among the limitations of the NHA application is the dependency on user perception about their symptoms in order to take decisions, which might not exactly represent the reality, since users may have different perceptions and may omit important symptoms. In addition, the scale used to represent what the user is feeling may vary between patients, so even though they might be feeling the same thing, one might say their symptom seems to be very bad, while another might say it's mild.

Obtaining quality big data is critical for building useful knowledge base for a domain specific application such as the NHA. Future work can be done to better analyze and enrich the knowledge base based on the semantic web analytics available in the Cyber space.

6. References

1. Kendal, S.L. and Creen, M, *An Introduction to Knowledge Engineering*, London, Springer, 2007

2. Ricardo Baeza-Yates, "Soft Computing Applications to Intelligent Information Retrieval on the Internet", International Journal of Approximate Reasoning, Vol. 34, issues 2-3, 2003, pp. 97-114

3. Ricardo Baeza-Yates, and Berthier Ribeiro-Neto, *Modern Information Retrieval*, Addison-Wesley 1999.

4. Octavian Udera et al, "Annotated RDF", ACM transactions on Computational Logic, December 2007, pp. 1 - 35

5. David Booth et al., RDF as a Universal Healthcare Exchange Language, http://dbooth.org/2014/rdf-as-univ/

6. Enrique Carol, et al, "Optimizing Relationships Information Repertory Grids, in Artificial Intelligence and Practices II, Max Bramer Ed., in the proceedings of IFIF 2008 Boston Springer Information Processing 276:163-172, 2008

7. V. Raghavan, P. Bollmann, and G. Jung, "A critical investigation of recall and precision as measures of retrieval system performance", *ACM Transactions on Information Systems (TOIS),* pp. 205-229

8. K.W. Church and P Hanks, "Word Association Norms, Mutual Information and Lexicography", in the proceedings of *ACL 89 Association of Computational Linguistics*, 1989

Applying Scalable Machine Learning Techniques on Big Data using a Computational Cluster

Dev Dua[1], Sujala D. Shetty[2], Lekha R. Nair[3]
Department of Computer Science
BITS Pilani – Dubai Campus
Dubai, U.A.E.
devdua@live.com[1], sujala@dubai.bits-pilani.ac.in[2], lekharnair@gmail.com[3]

Abstract— **Machine Learning is a relatively new avenue in exploring Big Data, and this involves having a working understanding of the commonly used machine learning techniques, and the algorithms that each technique employs. There will be a focus on making the algorithms scalable to utilize large amounts of data, and this will be done using open source machine learning tools and libraries. Since big data resides on the internet, or on a cloud network, the machine learning algorithms studied in this paper will be utilized in applications deployed on a cloud service like Windows Azure or Amazon Web Services, which will carry out compute tasks on big data residing in the cloud.**

Keywords - Big Data, Machine Learning, Cluster Computing

I. INTRODUCTION

The computers of the current year have been improving exponentially in terms of performance as per Moore's Law, and development of fast and efficient computing platforms has significantly helped us to understand computationally and structure-wise complex systems, such as biochemical processes, and sophisticated industrial production facilities and financial markets [7]. The human tendency of thinking and analyzing, and further predicting, arises from the fact that given historical data, we can estimate and model the processes in the system at a level of abstraction that, although not able to provide a complete understanding of the inner workings, is detailed enough to provide useful information about dependencies and interconnections at a higher level. This, in turn, can allow us to classify new patterns or predict the future behavior of the system.

We have been harnessing the processing power of computers to build intelligent systems, systems that, given training data or historical data as mentioned above, can learn from, and as a result give us results when the test data is fed into the system. During the previous few decades, there has been incremental growth in our data generation and storage capabilities [2]. In general, there is a competitive edge in being able to properly use the abundance of data that is being collected in industry and society today. Efficient analysis of collected data can provide significant increases in productivity through better business and production process understanding the highly useful applications for e. g. decision support, surveillance and diagnosis.

The focus of this paper is on exploring and implementing intelligent applications that harness the power of cluster computing (on local machines as well as the cloud) and apply machine learning on big data. However, the concepts that will be explored are by no means specific to these fields, and can be extended/modified for other fields as well.

II. OBJECTIVES

The objective of this paper is to meet the following objectives:

- Explore machine learning techniques, and evaluate the challenges faced when operating on Big Data.
- Explore current machine learning libraries, analyze the feasibility of exploiting them on a cloud platform
- Understand the basics of cluster computing, and how an Apache Spark cluster can be setup on Microsoft Azure.
- Cluster geospatial data, and analyze the performance of the implementation on a cluster.

III. UNDERSTANDING MACHINE LEARNING

To put it simply, one can say that machine learning focuses on designing and implementing algorithms and applications that automatically 'learn' the more they are executed. We will however not be concerned with the deeper philosophical questions here, such as what learning and knowledge actually are and whether they can be interpreted as computation or not. Instead, we will tie machine learning to performance rather than knowledge and the improvement of this performance rather than learning. These are a more objective kind of definitions, and we can test learning by observing a behavior and comparing it to past behaviors. The field of machine learning draws on concepts from a wide variety of fields, such as philosophy, biology, traditional AI, cognitive science, statistics, information theory, control theory and signal processing. This varied background has resulted in a vast array of methods, although their differences quite often are skin-deep and a result of differences in notation and domain. Here we will briefly present a few of the most important approaches and discuss their advantages, drawbacks and differences.

A. Association Rule Learning

ARL is an ML method for discovering relations among attributes in large transactional databases, and is quite popular and well researched. The measures used to discover similarities are varied, and it mainly involves generation of item sets recursively to finally build the rules, based on support count and confidence. This way of learning is often applied in market basket analysis (affinity analysis) where trends that relate products to transaction data are discovered to boost the sales of the organization.

B. Artificial Neural Networks

An ANN learning algorithm is inspired by the structure of the biological computer i.e. the brain, and is structurally designed in a manner similar to biological neural networks. The interconnected group of artificial neurons structure and divide the computation in such a manner that information can be processed in a parallel manner. Applications of NNs include use in tools that model non-linear statistical data. NNs make it easy to model complex relationships and process a large amount of inputs and compute outputs in a massively parallel manner. Other applications include pattern discovery and recognition, and discovering structure in statistical data distributions.

C. Support Vector Machines (SVMs)

SVMs, is a binary learner used for regression and classification, are supervised ML methods. It is applied mostly to categorical data, where the training set of data has records belonging to 1 of 2 categories. The model generated by the SVM training algorithm is then used on the test data to predict which category does each record fall into. Thus it can be seen as a non-probabilistic linear classifier. The data is represented as points in space, mapped so that the 2 categories are divided by a gap that is ideally as far apart as possible. The test records are then fit into the same space so that they fall into a point in space that corresponds to the category they fall into.

D. Clustering

Clustering can be viewed as separating records of data into subsets, called clusters, so that data points lying within the same cluster are highly similar, and this similarity is determined by employing pre-designated criteria. Data points belonging to different clusters are ideally placed as far as possible, i.e. they are highly dissimilar. There are various types of clustering techniques – partitional, hierarchical, and density based clustering being the most common. They are built on the basis of some similarity metric and the result of clustering is scrutinized by looking at the relative placement of members within the same cluster (internal compactness), and also how well separated different clusters are from each other. This ML method is an example of unsupervised learning. Applications of clustering are varied, from spatial data analysis to document clustering.

E. Collaborative Filtering

CF is a recommendation technique being increasingly for generating suggestions/recommendations. Collaborative filtering can be viewed as the process of filtering information to discover patterns involving 'collaboration' among data sources, viewpoints, multiple agents, etc. Collaborative filtering can be applied to very large data sets, and is a commonly applied to social media and entertainment services, like Netflix.

These approaches above are applied to many types of data sets, which vary in size, structure, attributes and complexity. Also, most of these approaches don't work well with all kinds of data, i.e. there is no 'super-algorithm' that can encompass all types of data sets. Therefore this is one problem that connects machine learning with big data. This scenario is better described as scalability [6], where the application/algorithm has to be redesigned to deal with huge sets of data, which are structurally big and complex to be read and operated upon by conventional computers. The structure of the data being used also matters, and impacts the way that it has to be pre-processed before the machine learning application can actually start working on the data.

IV. Big Data and the Challenges to Data Analytics

Big data is a buzz word used to describe the explosive generation and availability of data, mainly on the web [1]. Big Data, going by the name, is so large that traditional software techniques and databases fail to process this exponentially growing structured and unstructured data. It is not only the monolithic structure of big data that makes it a challenge, other factors include its rate of generation (that might be too fast to capture such huge amounts of data successfully without losing the other incoming data) or one may not have the processing prowess to quickly analyze the data. It can be characteristically described by [10] -

- *Volume:* This describes the scale of data being handled. An estimate shows that 40 zettabytes (equivalent to 43 trillion gigabytes) of data will be created by 2020, a 300x increase compared to data generated by 2005. It is also estimated that 2.3 trillion gigabytes of data are generated every day, and is exponentially growing.
- *Variety:* This refers to the different forms of data. It also indicates the various sources that generate structured and unstructured data. Taking healthcare as an example, in 2011 itself, data in healthcare was estimated to be 161 billion gigabytes. On YouTube, more than 4 billion hours are viewed every month.
- *Velocity:* It deals with the rate at which sources like human interaction with things like social media sites, mobile devices, etc., networks, machines and business processes, generate the data. This characteristic is most important when dealing with huge flows of streaming data. Velocity of Big Data can be handled by sampling data from data streams. For example, 1TB of information about stock trades is captured by the

New York Stock Exchange during each trading session. If this is analyzed in an efficient way, businesses can really benefit.

- *Veracity:* Veracity describes the abnormality, biases, noise and inaccuracy in data. The immense flow and size of the data itself is so overwhelming that noise and errors are bound to exist. Thus, to have clean data, filters and other monitoring measures need to be implemented to prevent 'dirty data' from accumulating.

Loosely structured data is often inaccessible and incomplete. Difficulties in being able to create, manipulate, and manage big data are the most common problems organizations face when dealing with large databases. Since standard procedures and tools are not built from the ground up to analyze massive amounts of data, big data particularly poses a problem in business analytics. As can be inferred, the above elicited characteristics of big data make it particularly hard for machine learning tasks to be carried out on it. Sampling such huge data is the first difficulty that is faced. The lack of structure (or poorly defined structure) is another hurdle while preprocessing the data. The performance of the algorithm also suffers because of the sheer volume of the data to be trained. Thus, an efficient platform with high computational prowess and the ability to handle huge sizes of data is required.

V. Current Machine Learning capable Cluster Computing Platforms and their Limitations

Since the 4 V's of big data, as described in the previous section are a hurdle to processing of data at a small scale, a high performance computing solution, or an alternative to high performance computing on a small or distributed scale has to be explored. There are platforms that have been in existence for a long time now, but not all of them currently support applying machine learning on big data, in an explicit and intuitive way, or tradeoff between performance and ease of use.

The key idea behind Hadoop is that instead of having a single juggernaut server that handles the computational and storage task of a very large dataset, Hadoop divides the whole task into a set of many subtasks, using the divide and conquer paradigm. After all the single tasks have been done, Hadoop is responsible for managing and recombining all the single subsets once their computation is over and the output is generated. In this case, it is possible to divide heavy computational tasks into many single node machines even if they are not so powerful, and obtain the results.

The simple programming model of Hadoop provided by the built in software library is basically a framework that enables distributed processing of large datasets across single clusters containing a few worker nodes (as shown in Figure 1), to clusters of computers containing several nodes each. Hadoop can take advantage of the storage and local computation offered by every node in the cluster, and can scale up from single servers to thousands of machines effortlessly.

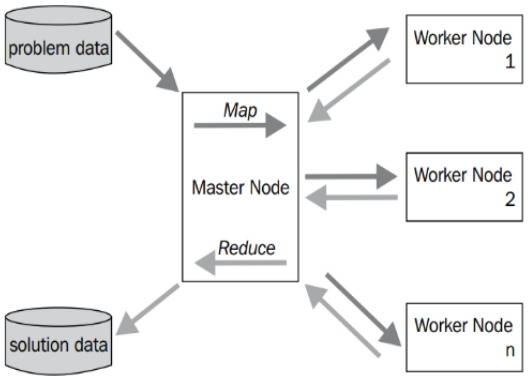


Figure 1. A high level abstraction of Hadoop's MapReduce paradigm.

Users who wished to exploit this great performance model offered by Hadoop to run machine learning tasks, used Apache Mahout, as it was tuned to Hadoop in a very efficient way. Apache Mahout [8][9], another Apache Software Foundation project, is a suite of open source implementations of scalable machine learning algorithms. The library contains algorithms primarily in the areas of classification, clustering and collaborative filtering. To enable parallelized operations, the implementations of the algorithms in Mahout use the Apache Hadoop platform. Like most of the projects in Apache Incubator, Mahout is a work in progress as various machine learning algorithms haven't yet been made available to users, even though the number of implemented algorithms has grown quickly.

Mahout fixes one of the major issues with Machine Learning techniques, which is scalability. Mahout can scale algorithms to large data sets. Since the algorithms implemented in Mahout have been written with Hadoop and MapReduce at their core, the core libraries of machine learning contain code that highly optimized to extract maximum performance out of the available nodes in the cluster. Currently Mahout supports mainly three use cases: collaborative filtering, clustering, and classification.

Even though Mahout on Hadoop are advantageous in many ways, there are some limitations [4][5]. Apache Mahout on Hadoop, although a great platform for data scientists, is not intuitive and easy to learn. The real-time and offline Hadoop backend are not integrated into one system. There exist some performance bottlenecks in the computation of item-item similarities, and finer control needs to be implemented over the sampling rate in most applications. Hadoop tends to convert the Job into a Batch Processing task. Also, since it is iterative in nature, just I/O and serialization of the data during Mapping (in MapReduce) can take up 90% of the processing time. The machine learning task itself runs for only about 10% - 15% of the actual running time. Also, there is no real-time data analysis or data stream analysis for dynamic machine learning applications. This called for development of and even more powerful and fast computing platform, that could take the best of Hadoop's MapReduce, but implement it in a much more optimized and efficient way.

VI. The Apache Spark platform

Apache Spark[11] was an incubator project, and gained a lot of attention from the data science community, regardless of its incubation status. Apache Spark is now a fully supported Apache product, and is out of its incubation status. Apache Spark is an open source computing engine evolved from Hadoop, and built from the ground up to deliver speed, ease of use, and sophisticated analytics as a powerful platform for the computing community

The component of prime interest is MLLib, the Machine Learning library for Apache Spark. It features highly optimized implementations of machine learning algorithms in Scala, and written from the base up to handle big data effectivelySpark give users access to a well-designed library of parallel and scalable machine learning algorithms. MLLib contains high-quality scalable machine learning algorithms as well as unbelievable speed that out performs MapReduce and many other machine learning libraries available publically. Since it is a component of Spark, it is usable through not only Scala, but Python and Java as well. MLlib is a Spark subproject providing machine learning primitives, relevant to mainly classification, regression, clustering, collaborative filtering and gradient descent. Algorithms under each category are:

- classification: logistic regression, linear support vector machine(SVM), naive Bayes
- regression: generalized linear regression (GLM)
- collaborative filtering: alternating least squares (ALS)
- clustering: k-means
- decomposition: singular value decomposition (SVD), principal component analysis (PCA)

A. Experimental Setup

The setup of Spark is fairly simple [12], and it is recommended that the pre-built binaries be download from the Spark website. The results obtained for this paper were collected by running the program on Spark version 0.9.1, when it was still in the incubation state. No substantial changes were made in the MLLib library, so the results obtained using Spark 0.9.1 will be identical to those possible with version Spark 1.0. A Spark cluster was deployed using Cloud Services on Microsoft Azure, and Linux VMs were used as the cluster nodes. Each machine had 4 core processors, with 14GB of memory each. Since the VMs had to be connected to each other in the cluster, a Virtual Network was setup, with RSA secured SSH.

VII. Clustering Geo-Spatial Data using the k-means Clustering implementation of MLLib

Most clustering methods used today either use k-means in conjunction with other clustering techniques, or they modify the algorithm in terms of sampling or partitioning. Given the number of clusters to be formed 'k', and 'n' data points in the data set, the goal is to choose k centers so as to maximize the similarity between each point and its closest center. The similarity measure most commonly used is the total squared distance between the point and the mean. This algorithm, also called the Lloyd's algorithm first initializes k arbitrary "centers" from the data points, typically chosen at random, but using a uniform distribution. Each point is then assigned to the cluster whose center it is nearest to. After this, the centers are re-evaluated, keeping in mind the centers of mass of the points that surround the current center. Until the centers stabilize, the last 2 steps are repeated.

Thus, it can be considered to be one of the simplest unsupervised learning algorithms that can be used to find a definite clustering of a given set of points, even with varied data types. The objective function that this algorithm aims to minimize, is the squared error function. The objective function is given as below:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

Here J is a chosen distance measure between a data point and the cluster center, and thus J is an indicator of the distance of the n data points from their respective cluster centers.

Since there are only a limited number of clustering ways that are possible, this algorithm will always give a definite result, and will always terminate. Also, users who go for the k-means algorithm are interested not in the accuracy of the result it produces, but the simplicity and speed with which it gives the clustering result. It does sometimes generate arbitrarily bad clustering, but the fact that it doesn't rely on how the starting dummy cluster centers were placed with respect to each other makes it a good option when performing clustering tasks. In particular, it can hold with high probability even if the centers are chosen uniformly at random from the data points. The area in which k-means can be improved considerably is the way the initial centers are chosen. If this process is optimized, the algorithm can be considered to be more computationally sound, and overall a good option to go for. In the next section, we look at 2 of the best improvements made to the algorithm to date, both of which are used in the clustering library of Spark.

A. The k-means++ and k-means|| algorithms

As discussed earlier, k-means is relatively not a good clustering algorithm [13] if the quality of clustering or the computational efficiency is considered. Analysis shows that the running time complexity of k-means is exponential in the worst case scenario. K-means aims at locally optimizing the clusters by minimizing distance to the center of the clusters, and thus the results can possibly deviate from the actual globally optimal solution to a considerable extent. Although repeated random initializations can be used to tweak the results a little bit, they prove to be not so effective in improving the results in any way. In spite of all these shortcomings, there are a meagre number of algorithms that can match the simplicity of and speed of the k-means algorithm. Therefore, recent research has focused on optimizing and tweaking how the centers are initialized in the first step. If the initialization method is improved, the performance of the algorithm can be vastly sped up, both in terms of convergence and quality. One of the procedures to improve the initialization is k-means++.

The k-means++ algorithm makes a small change in the original initialization, by choosing just the first mean (center) at

random, uniformly from the data. It also takes into consideration the contribution of a center to the overall error, and each center chosen by the k-means++ algorithm is selected with a probability that is proportional to this contribution. Thus, intuitively, k-means++ exploits the relatively high spread out of a good clustering. The new cluster centers chosen by k-means++ are thus the ones that are preferably further away from the previously selected centers. After analysis, it has been shown that k-means++ initialization improves the original algorithm by serving a constant approximation (O(log k) in some cases, when the data is difficult to cluster) of the optimum solution, if the data is known to be well cluster-able. The evaluation of the practical execution of the k-means++ algorithm and its variants is critical if performance of an actual running implementation is to be optimized. Tests demonstrated that correctly initializing the original k-means algorithm did lead to crucial improvements and lead to a good clustering solution. The k-means++ initialization obtained order of magnitude improvements, using various data sets, when the random initialization was put into effect.

However, its inherent sequential structure is one downside of the k-means++ initialization. Although when looking for a k-clustering of n points in the data set, its total running time is the same as that of a single K-Means iteration, it is not easily parallelizable. The probability with which a point is chosen to be the ith center depends critically on the realization of the previous i-1 centers (it is the previous choices that determine which points are away in the current solution).

A simple bare bones implementation of k-means++ initialization makes k iterations through the data in order to select the initial k centers. This fact is augmented and made clear when big data is brought into picture. As datasets become bigger, as in the case of big data, so does the number of partitions into which the data can be divided. For example, a typical cluster number k = 100 or 1000 is chosen to cluster, say clustering millions of points. But in this case, k-means++ being sequential in nature, proves to be very inefficient and slow. This slowdown is even more noticeable and unfavorable when the rest of the algorithm, i.e. the actual k-means algorithm can be parallelized to run in a parallel environment like MapReduce. For many applications, an initialization algorithm is desirable that guarantees efficient parallelizability, while providing the same or similar optimality to k-means++.

To make k-means++ even better, and to formulate a parallel implementation, Bahmani et al. developed k-means||. the k-means|| algorithm, instead of sampling a single point in each iteration, samples O(k) points and repeat the process for approximately O(log n) rounds. These O(k log(n)) points are then re-clustered into k initial centers for the original k-means. This initialization algorithm, which we call k-means||, is quite simple and lends itself to easy parallel implementations.

B. Description and pre-processing of the dataset

3D Road Network (North Jutland, Denmark) Data Set is essentially geo-coordinates of a road network in North Jutland (spanning over 185x135 sq. km), which has been augmented by adding the altitude (elevation information) of those geo-coordinates to the data set[3]. The Laser Scan Point Cloud Elevation technology was used to achieve this. This 3D road network was eventually used for benchmarking various fuel and CO2 estimation algorithms. For the data mining and machine learning community, this dataset can be used as 'ground-truth' validation in spatial mining techniques and satellite image processing.

Attribute Information:

1. OSM_ID: OpenStreetMap ID for each road segment or edge in the graph.
2. LONGITUDE: Web Mercaptor (Google format) longitude
3. LATITUDE: Web Mercaptor (Google format) latitude
4. ALTITUDE: Height in meters.

Since the first attribute is not significant in clustering the points, only the other 3 relevant attributes had to be extracted for the actual clustering step. The data set file was loaded into GNU Octave, and extraction was achieved by initializing a matrix of dimensions 434874X4 and then slicing off the first attribute using the built in slicing implementation of Octave. The resulting matrix was a 434874X3 matrix, which was then written to disk as a TXT file. This file was then used in the next step, which is dividing the data into training and test data sets.

The next step to preparing the data for training the K-Means model was to sample the data into a training data set, and a test data set. Different proportions of test and train data were tested - 40% of training data and 60% of test data, 50% of training data and 50% of test data, 60% of training data and 40% of test data, 70% of training data and 30% of test data. The best results were found in the last sample, as a good and robust model was built. At the end of pre-processing two files were created, train_70.txt (304412 records) and test_30.txt (134062 records).

C. Explanation of the program

In the program, we use the KMeans object of the MLLib library to cluster the data into clusters. The number of desired clusters is passed to the algorithm, which after performing numerous rounds of clustering, computes the Within Set Sum of Squared Error (WSSSE). WSSSE is the sum of the squared distance between each point in the cluster and the center of the cluster, and is used as a measure of variation within a cluster. You can reduce this error measure by increasing k. In fact the optimal k is usually one where there is an “elbow” in the WSSSE graph.

The parameters accepted by the train() method of the KMeans object are –

i. Data: The training data in the form of and RDD (Resilient Distributed Dataset) is fed into the train method, which will be iterated through to build the KMeans model.
ii. No. of clusters: specifies the number of clusters that the data is to be partitioned into.
iii. Max iterations: maximum number of iterations of the initialization algorithm (random, k-means++ or k-means||) is to be run.
iv. No. of runs: number of times the k-means algorithm has to be run, and this is a crucial parameter as k-means does not guarantee a globally optimal solution.

Increasing the number of runs can give some surety that the best clustering result would be obtained.

v. Initialization mode: initializationMode specifies either random initialization or initialization via k-means||.

VIII. Test Cases and Analysis of Results

The test cases were formulated in a way that could help analyze how the implementation of the clustering algorithms included with the MLLib library of Apache Spark performed with a relatively dense, yet simple data set. The data set used, due to its spatial nature is inherently suitable for clustering. Yet, the points that have been recorded as part of the 3D Road Network, are at really close proximity of each other, and thus the data is very dense. The data, being dense, is a challenge for k-means as k-means goes for a partitional approach rather than a density based clustering approach. This would lead to understandable errors in clustering, and that would be an interesting point to observe. Also, since there are 434874 lines containing 3 floating point numbers each, performance of the algorithms with respect to the parameters specified for the clustering would be a crucial point to observe.

The test cases were designed to range from less computationally intensive to highly computationally intensive tasks. The tests cases have been described below –

i. Cluster count k = 10, maxIterations = 10, Runs = 10

A relatively low number of clusters specified guarantees that the algorithm will take a short amount of time to run. Also, because the runs are limited to 10, the algorithm will produce a high error of clustering. Since this is the first test case, it serves to be a placeholder for designing the next few test cases.

ii. Cluster count k = 20, maxIterations = 50, Runs = 100

Increasing the cluster count guarantees a lower WSSE, but since the number of maxIterations have been increased, along with the number of runs, it will be interesting to note the effect this change in parameters has on the performance as well as running time of the clustering.

iii. Cluster count k = 30, maxIterations = 15, Runs = 50

iv. Cluster count k = 40, maxIterations = 15, Runs = 50

v. Cluster count k = 50, maxIterations = 15, Runs = 50

vi. Cluster count k = 100, maxIterations = 15, Runs = 50

The above 4 runs simply increase the number of clusters, and this is done to observe trends in performance when only the cluster count is increased.

The results obtained exhibited interesting patterns, and helped infer that performance of the clustering is directly linked to the cluster count parameter. The legend is a triple, (k,m,r) which stands for cluster count k, maxIterations m and runs r. The results were measured in seconds, and since the magnitude of the results obtained when changing the no. of slave nodes ranged from 100s of seconds to 1000s, the results had to be normalized to have a clearer and more intuitive insight into the patterns in performance. The normalization was carried out using the z-score method, which transforms data into a range of [-2,2]. It uses the standard deviation and mean of the data to calculate the values. Also, this method proves useful to easily identify outliers, as any value that has a z-score > 2 or z-score < -2 doesn't fall in the normal range of the data being normalized. After z-score normalization, the runtime in seconds was plotted against the number of slave nodes (Worker nodes) being used by the algorithm. The resulting graph is shown in the following figure.

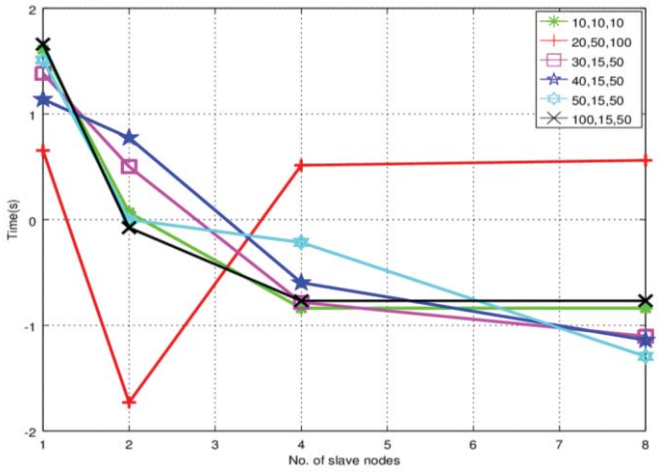


Figure 2. Clustering time vs. Number of Slave Nodes

As can be seen, in the first case, the time to cluster data decreases as number of slave nodes are increased. The performance doesn't change much when the number of slave nodes is increased from 4 to 8, as most of the slaves are scheduled randomly, and the rest remain idle while the jobs are running on the other nodes.

In the 2nd case, the max iterations and runs are increased, and the unnecessary stress on the computation is apparent. This case completely stands out from the rest of the cases as time complexity shoots up due to the relatively more extreme parameters. The 100 runs take longer on 4 and 8 slave nodes, which is unexpected according to the trend. This could be explained by arguing that scheduling and distribution process would be easier on 2 slaves as compared to that on 4 and 8 slaves, and more so when there is just one file being operated upon. This case helps infer that the number of runs increases the complexity and causes unnecessary fluctuations in running time, when accuracy of the model is not favorable over the speed (as in the case of big data). So, in further cases the runs are reduced to 50, and max iterations reduced to 15, as it was observed that the k-means++ converged in not more than 15 iterations every time.

In the consecutive cases, only the cluster count was increased by 10 with each case, and the number of slave nodes were varied as before. The trend remained the same across the last 4 cases – the running time decreased, with run times almost the

same in the case of 4 and 8 slave nodes. This is due to idle states of the nodes when they're not needed, mostly in the case of the 8 slave nodes.

The result of clustering is however more understandable in terms of the Average WSSE (Within Set Squared Errors) which dropped considerably across all 6 cases. This is attributed solely to the number of clusters being created, and has no relation with the other parameters of the KMeans model. As the number of clusters are increased, the WSSE decreases. Here, the values plotted are the average of the WSSE calculated in each case where the number of slave nodes was calculated.

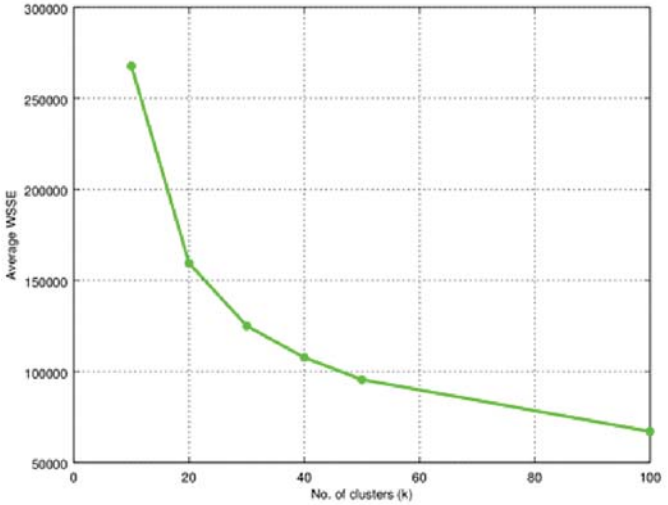


Figure 3. Average WSSE vs. Number of clusters

IX. CONCLUSION

The focus of this paper was to explore platforms that can be used to implement intelligent applications that harness the power of cluster computing (on local machines as well as the cloud) and apply machine learning on big data. Current cluster computing platforms like Google Cloud Engine and Apache Hadoop, and their corresponding machine learning libraries – Prediction API, and Apache Mahout were studied, and compared against Apache Spark and MLLib.

A cluster was created on Windows Azure, and each node in the cluster had a quad core processor with 14 GB of RAM, running Ubuntu Server 12.04 LTS. Apache Spark was downloaded and built on each machine. The program was written in Python, and interfaced with Apache Spark using Pyspark. A simple clustering task was run on a relatively large and complex data set, and the run times were recorded. Varying the configuration of the cluster with every run showed some interesting trends in the results. As compared to traditional iterative implementations of k-means clustering, running it on Apache Spark on a cloud cluster definitely gave it an advantage on run time.

With the rise of diverse, flexible and economical cloud service, users from both research and business backgrounds can harness the power of Spark on a cloud cluster, and apply data mining and machine learning concepts to their everyday tasks. It is even more suited for big data, as Spark features excellent parallelization of data, and optimized code libraries so that jobs can be processed quickly. Big data and machine learning are essentially a very good combination of areas to work upon, and research carried out in these areas are definitely going to influence the development of intelligent and computationally powerful platforms for the ever growing domain of Big Data.

REFERENCES

[1] NG DATA, *"Machine learning and Big Data analytics: the perfect marriage"*, Internet: http://www.ngdata.com/machine-learning-and-big-data-analytics-the-perfect-marriage/

[2] Daniel Gillblad, Doctoral Thesis, Swedish Institute of Computer Science, SE–164 29 Kista, Sweden, 2008, "*On practical machine learning and data analysis*", Internet: http://soda.swedish-ict.se/3535/1/thesis-kth.pdf

[3] 3D Road Network (North Jutland, Denmark) Data Set, UCI Machine Learning Repository, Internet: http://archive.ics.uci.edu/ml/datasets/3D+Road+Network+%28North+Jutland%2C+Denmark%29

[4] Sean Owen, Contributor at Quora**, "***What are the pros/cons of using Apache Mahout when creating a real time recommender system?***",** Internet: http://www.quora.com/Apache-Mahout/What-are-the-pros-cons-of-using-Apache-Mahout-when-creating-a-real-time-recommender-system

[5] Nick Wilson, BigML, "Machine Learning Throwdown", Internet: http://blog.bigml.com/2012/08/02/machine-learning-throwdown-part-1-introduction/

[6] Georgios Paliouras, Department of Computer Science, University of Manchester, Thesis on "*Scalability of Machine Learning Algorithms"*, Internet: http://users.iit.demokritos.gr/~paliourg/papers/MSc.pdf

[7] Tom M.Mitchell, School of Compter Science, Carnegie Mellon Univesity, Pittsburgh, July 2006, "*The Discipline of Machine Learning"*, Internet: http://www.cs.cmu.edu/~tom/pubs/MachineLearning.pdf

[8] Gaston Hillar, "*Machine Learning with Apache Mahout: The Lay of the Land*", Internet: http://www.drdobbs.com/open-source/machine-learning-with-apache-mahout-the/240163272

[9] Apache Mahout, Apache Foundation, Internet: https://mahout.apache.org/

[10] IBM, Articles on Big Data, Internet: http://www.ibm.com/developerworks/bigdata/

[11] Apache Spark, Apache Foundation, Internet: http://spark.apache.org/

[12] Mbonaci, "Spark standalone cluster tutorial", Internet : http://mbonaci.github.io/mbo-spark/

[13] Songma, S.; Chimphlee, W.; Maichalernnukul, K.; Sanguansat, P., "Classification via k-means clustering and distance-based outlier detection," *ICT and Knowledge Engineering (ICT & Knowledge Engineering), 2012 10th International Conference on* , vol., no., pp.125,128, 21-23 Nov. 2012

SESSION

INTERNET AND WEB COMPUTING, INFRASTRUCTURES + SECURITY, PRIVACY AND RELATED ALGORITHMS

Chair(s)

TBA

Energy Efficient Opportunistic Wireless Networks as Digital Inclusion Support and Easy Access to Information in Southeast of Goiás-Brazil

Waldir Moreira*, Antonio Oliveira-Jr, Rosario Ribeiro, Tercio Filho, Dalton Matsuo
Claudio Souza, Roberto Piedade, Marcos Rabelo*, Marcos Batista*
Federal University of Goiás (UFG), Brazil
*Graduate Program in Modeling and Optimization, Federal University of Goiás (UFG), Brazil
Email: {waldir.moreira, antoniojr, rosarioribeiro, tercioas, dalton_tavares, clsouza, roberto.piedade, rabelo, marcos.batista}@ufg.br

***Abstract*—The lack of network infrastructure and communication technologies and services results in a high rate of digital exclusion in the rural areas of southeast of Goiás state in Brazil. Despite the public and private investments on different sectors of society, the population still suffers from a lack of Internet connectivity which generates immeasurable social, cultural, and economic losses for the entire state of Goiás. This work presents a project that aims to study, validate, optimize and implement technologies and infrastructure based on energy efficient opportunistic wireless networks (within the context of Delay Tolerant Networks - DTN). The goal is to facilitate the exchange of information and knowledge in the rural communities of southeast of Goiás state. Thus, applications related to medicine, education, agriculture and environmental protection can be easily deployed and used to improve the life quality of the population in such isolated rural areas. Potential solutions are expected to be evaluated and validated through simulators (Opportunistic Network Environment - ONE, Network Simulator - NS-2/NS-3) and real test bed (prototype based on Arduino hardware). In addition to bringing different benefits (e.g., social and digital inclusion) for the target community, the presented project aims at exchanging experience among undergraduate students, postgraduate researchers and institutions involved, resulting in increased qualification of such knowledge in this area.**

***Index Terms*—wireless communication; opportunistic networks; social networks; energy efficiency; social and digital inclusion; rural communities.**

I. Introduction

Despite of the investments made to encourage the use of Information and Communication Technologies (ICT), there are many Brazilians who are still digitally excluded. This problem is further increased in rural areas that are characterized by little (or even absent) existence of the communication infrastructure and ICT centers.

The public switched telephone and mobile cellular networks are the most common forms used to minimize this gap. However, needn't to say that the cost associated to these types of infrastructure is very high for local rural communities that normally lives off a very low income. Another characteristic of such rural areas is the density of the population that is generally distributed in several acres of land and people sometimes have to travel several hours to reach other communities or urban centers. These characteristics result in the isolation of rural communities, where the exchange of information between government agencies, health centers, schools and the people themselves is very difficult.

It is known that the development of a specific region and the intellectual performance of individuals within such region are the products of how information is exposed to these individuals. As ICT is an important factor in this development process, we can see the reason why rural communities suffer with the digital exclusion and the difficult access to information.

The process of social inclusion, its advantages and challenges, especially in isolated rural areas, have been discussed by researchers, teachers, and companies around the world. However, due to lack of large-scale deployment of opportunistic wireless communication solutions at these rural areas, many services are still not available to the population.

Along with these difficulties of accessing information, there is the problem of global warming that is currently increasing everyday. Hence, it is imperative to reduce carbon emissions of ICT solutions independently of the regions (rural or urban) where such solutions are being deployed.

Studies show that ICT solutions account for 2% to 3% of total carbon emissions on Earth and that 50% of all energy used in the ICT sector is consumed by wireless access networks [1]. Thus, reducing energy consumption through energy efficient mechanisms/routing contributes substantially to the reduction of gas emissions, helping to protect the environment from global warming.

With this in mind, this project focuses on sporadic contacts that occur between the different elements (e.g., vehicles, people, schools, health centers) that are present in the rural areas and that can be used to establish an asynchronous form of communication with others elements in urban centers. This means taking advantage of the flow of vehicles (e.g., buses, ambulances, trucks) used for the daily transporting of passengers, goods and patients between rural and urban areas to facilitate the exchange of information and increase digital inclusion index of the local rural communities. At the same time, the project has a concern with the environment proposing solutions that are energy efficient.

Along these lines, the Store-Carry-Forward paradigm (SFC) found in Energy Efficient Opportunistic Wireless Networks

fits this project. This is due to the fact that the devices based on such paradigm store and carry the information up to another device (intermediate or final) to forward it. Since this type of network uses low-cost equipment, this communication alternative for rural areas becomes very attractive, easily meeting the needs regarding the exchange of information in such communities.

By employing the SFC paradigm, the exchange of information between ICT centers located in rural and urban areas (html requests, asynchronous access to email), health centers and hospitals (diagnoses, test results), schools and education departments (transcripts, performance of teachers), and local and state government agencies (population density) can be done in a simple and inexpensively manner. Most importantly, this paradigm contributes to the economic and intellectual development as well as the improvement of the life quality of individuals of these rural communities.

So, this project aims to study, validate and implement technologies and infrastructure based on Energy Efficient Opportunistic Wireless Networks (within the context of Delay Tolerant Networks - DTN) to improve the life quality of the population in the southeast of Goiás state, facilitating the exchange of information, access to knowledge and digital inclusion. The aim is to develop solutions with low operating costs and that take advantage of opportunistic contacts that occur between the elements (e.g., schools, health centers, public transport) that are part of the rural communities and that can be used to send/deliver information between these areas and urban centers which have easy and readily available access to connectivity.

Opportunistic routing is a critical factor and the use of social aspects can be employed as a deciding factor to generate a social network for the efficient transmission of information. This work is inserted in the scope of a project which initially intended to expose concepts related to opportunistic networks and information exchange focusing on routing protocols based on social aspects.

This paper addresses the social factor as a support to help routing protocols decide on how information shall travel in opportunistic networks deployed in such rural communities. It also briefly describes the concepts related to opportunistic networks and routing protocols based on social aspects (e.g., concept of community and daily routines of users). To illustrate their potential, the Opportunistic Network Environment (ONE) simulator is considered, providing a comparative study between two protocols that use the social factor for routing, dLife [2] and Bubble Rap [3].

II. State-of-the-Art Literature

There are different energy efficient and social-aware opportunistic forwarding solutions, which are based on metrics that rely on contacts between the users' devices to decide how/when to forward information, increasing delivery probability.

Regarding social-aware solutions, they may vary according to the employed forwarding mechanism [2], [4], [5]: flooding the network with multiple copies of information (e.g., Epidemic [6]), using the contact history (e.g., PROPHET), predicting future contacts (e.g., EBR [7]). Additionally, there have been solutions (e.g., Label [8], SimBet [9], SocialCast [10], Bubble Rap [3], PeopleRank [11]) incorporating social characteristics into the forwarding mechanisms and showing efficient delivery capabilities.

As for the energy efficient routing mechanisms, they can be classified according to the employed energy-aware mechanisms [12], [13], [14], [15], [16], [17], [18]: minimizing power consumed while transmitting a packet (e.g., MTPR [19]), selecting paths comprising devices with a larger amount of available battery (e.g., MREP [20], MMBCR [21], CMMBCR [22]), selecting path based on the battery consumption of devices (e.g., MDR [23]), estimating power consumption on an end-to-end path (e.g., LPR [24]).

Upon this vast number of solutions, we intend to consider those that present low consumption of device resources (e.g., battery, processing, etc.)and that present high delivery rate in rural areas which are the main target of this project.

Moreover, the specification and development of efficient solutions for forwarding/routing relevant to the project begin with the development process of improved solutions that will be used to improve the life quality of the target rural communities, helping to reduce the effect the digital divide and facilitating the exchange of information in these regions. The aim is to develop different energy efficient opportunistic routing solutions based (or not) in social aspects that help to improve the exchange of information. Thus, it is expected solutions that take advantage of any contact between the elements present in these rural areas to increase the delivery probability of information.

Finally, the project considers the implementation and validation of the outcome solutions. The validation process of the resulting solutions of this project will be done in two parts: through simulations and then prototypes for testing in a real-world environment. With the simulations, the objective is to reach a stable version of the solutions, which shall show the advantages and points to be improved. Once a stable version of the solutions is reached, the implementation of prototypes is initiated, aiming at an evaluation of these solutions in the real world, that is, in rural area of the southeast of Goiás state.

A. Social Opportunistic Routing

Forwarding in opportunistic networks is a crucial factor for its performance. Currently, the use of social networks (such as Facebook and Twitter) reached a global scale, associated with the use of mobile devices, often used to access any social network, also significantly increased due to factors such as lower cost and greater ease of access to technology compared to some years ago. According to Cisco's report on the growth forecast of global mobile traffic by 2015, there will be a smartphone per capita on the planet.

In the case of opportunistic routing, we can take advantage of social interaction and increased use of mobile devices to create or improve routing protocols in opportunistic networks through social aspects. Considering a network where nodes represent mobile devices carried by users, we can generate a

social graph of the network and use the social relationships of each user to determine the best route to opportunistically send information to a given destination node.

As mentioned earlier, there ate different social-aware forwarding solutions, able to efficiently deliver information in opportunistic environments. This sections briefly goes over few social-aware strategies, namely dLife, PeopleRank [11] and the BubbleRap [3].

dLife [2] considers the daily routine of users and implements two additional utility functions: the Time-Evolving Contact Duration (TECD, which determines the social weight among users based on their social interaction during their daily routines); and the Time-Evolving Importance (TECDi, that measures the importance of the node considering its neighbors and the social weights towards them).

Social interactions between users can change constantly. When a node meets another one, its personal network changes and therefore the entire social structure which that node belongs also changes. Considering that. dLife with TECD captures the dynamism of the social behavior of users, reflecting the users routines more efficiently than those solutions that rely solely on history of contacts.

Thus, forwarding performed by dLife considers the social weight of the node that carries the message towards its destination and the social weight of the intermediate node to this same destination. In cases where the intermediate node has a social weight (i.e., a strong social relationship with the destination), the source node sends a copy of the message to the intermediate node. Otherwise, the importances (TECDi) of node that carries the message and of the intermediate are taken into account when replicating information. That is, the intermediate node receives a copy if it is more important than the node that holds the information at that time.

Inspired by Google's page rank algorithm, Peoplerank [11] presents a success rate close to the epidemic routing, but without the disadvantages of large replication rates and network overhead. In networks with intermittent connectivity, we do not have a full knowledge of the topology for mobility and availability of nodes, but we have information on social interactions between their users. While opportunistic contact information changes constantly, relationships in a social network tend to remain stable. So, Peoplerank uses this information to make more stable and efficient routing. Peoplerank ranks each node according to its importance in the social graph and uses that rank as a guide for making decisions and nodes with a higher rank have more importance on that network. The idea is that the more sociable a node is, the greater the chance of a message reaching its destination using that node as an intermediary.

Bubble rap [3] focuses on the social structure of a community and its central nodes to optimize routing. Bubble rap is based on community aspects and centrality, which are present throughout society, that is structurally divided into communities and sub-communities. Within these communities some nodes interact more than others, are more popular in specific environments, thus defining the between centrality algorithm. Bubble rap also uses K-Clique and cumulative window algorithms i) to identify the communities that nodes belong to; and, ii) to determine local (i.e. within the community) and global centralities of nodes. Based on community information (when the intermediate belongs to the same community of the destination) and centrality (when the intermediate nodes has centrality greater than the node that holds the message), Bubble rap decides when to create a copy of the message.

III. Methodology

The methodology adopted in this project includes a detailed analysis of the state of the art and the definition of rural scenario and its requirements. At this stage, it is important to interact with industry and society/community in order to align the scenario and project application requirements. This project is expected to also use a real test environment for validation of it outcomes. Thus, two complementary approaches will be used: (i) discrete-event simulation and (ii) actual implementation of prototypes to validate the developed solutions.

Regarding the validation of mechanisms and proposed solutions, simulation is an approach based on development models based on abstractions of reality. The simulations have the advantage of being easy to make extensive tests with low cost prior to an actual real-world implementation, since the models are specified through programming languages and executed over simulated scheduled events. For this project, different simulator may be used, such as Opportunistic Network Environment (ONE) and Network Simulator (NS-2 / NS-3), which are widely known by the scientific community and industry for simulation of computer networks, sensors and various applications.

Regarding the real-world implementation, this project includes the development of a real, low-cost test environment (test-bed) for deployment of a communication networks in rural area of the southeast of Goiás state. This real-world environment is composed of current mobile devices (smartphones, tablets, laptops) and the use of prototypes based on the Arduino hardware.

The project will closely analyze the existing technologies and their applications in the scenarios and requirements defined together with the community, as well as the new solutions concerning the communication technologies that can be proposed to achieve the project's objectives. Integration with existing communications networks (if any) such as Internet access near the rural communities covered by the project, will be considered as complementary alternatives to the process of opportunistic transmission of content in the rural context.

Regarding the wireless communication network technology, standard IEEE 802.11 (Wi-Fi), we intend to use a programmable wireless router to allow the change and the addition of new low-cost control mechanisms that are proposed in this project. Thus, the project will be using OpenWRT as the operating system, an open source Linux distribution, which appears flexible enough for the installation of new solutions.

For testing, the project will use free and open source test and network management tools to measure the capacity, reach, and the communication stability in the defined scenario. With respect to applications to be made available to the rural population (multimedia applications, medicine,

education, agriculture, environmental protection, among many others), they will be evaluated following the recommendations of the agencies and national and international standardization institutions such as the National Agency Telecommunications (Anatel), Brazilian Association of Technical Standards (ABNT), Telecommunication Standardization Sector (ITU-T), Institute of Electrical and Electronics Engineers (IEEE) and the Internet Engineering Task Force (IETF).

The results dissemination methodology is through technical reports and scientific papers as well as the organization of meetings with industry and academia partners, and especially with the community/society involved in this project. This methodology is related to the documentation, dissemination of information, protection of solutions and ideas resulting from this project. Technical reports will be written along with papers and articles (to be published in journals, conferences, symposia and national and international events) for the dissemination of results and increasing the visibility of this solution deployed in the state of Goiás.

A. Expected Results

The results of the project will be described in scientific papers and submitted to workshops / conferences / national and international journals, seeking greater visibility to research centers and universities involved, strengthening their graduate programs and encouraging involved students and researchers in improving their carriers. Articles shall provide visibility to FAPEG as it is the funding entity of this project. The resources of this project will provide resources for the development of undergraduate and graduate works. Through the department of Intellectual Property and Start ups of UFG, this project will seek partners from the public and private sectors with regards to technology transfer and stimulate the creation of technology-based companies to the region. Successful experiences of technology transfer involving the participants in this project with the telecommunication company NTT DoCoMo of Japan resulted in products, articles, and patents, and shall have economic return to the university and the state of Goiás.

The expected scientific contributions shall serve as indicators of success of this project since it is intended to support digital inclusion in rural communities and to provide easy access to information to many different elements of these communities (e.g., schools, health centers, etc.). As the resulting solutions are based on the contacts that occur among the different elements found in the scope of this project and aim to be the most energy efficient, they end up not limited to the rural context. This is because such solutions can be applied to urban scenarios where the contacts between portable / mobile devices are also high (further increasing the delivery probability of information), and limitations on the use of energy also applies.

IV. Performance Evaluation

This section presents an initial analysis of results involving the dLife and Bubble Rap protocols within the context of the project. The performance evaluation is carried out over the ONE simulator. As performance metrics, we considered: delivery probability (i.e., ratio between the total number of messages and the total number of successfully delivered messages), cost (i.e., number of replications for each message) and latency (i.e., the time between the creation and delivery of the message).

Just to illustrate the potential of the considered social-aware opportunistic routing solutions, we simulate them over a synthetic mobility scenario, considering the city of Helsinky. The scenario comprises 150 nodes divided into 8 groups of people and 9 groups of vehicles. The car groups follow the Shortest Path Map-Based Movement model available in ONE, where the nodes (i.e., vehicles) randomly choose a target destination and move with speed between 7 m/s and 10 m/s.

The groups of people follow Working Day Movement model also available in ONE. The represented people move with speeds between 0.8 and 1.4 m/s. Additionally, each group has different points of interests such gym, bar, office, etc. With this model, it is assumed that people spend daily 8 hours at work and and are set with a 50% probability of having some activity after work (e.g., gym, bar, restaurant, park, etc.). The messages have TTL values of 1, 2 and 3 days, and their sizes vary between 1 to 90 KB as to represent different applications (e.g., asynchronous chat, email) . Finally, the buffer set in each node is limited to 2 MB, since it is expected that users are not willing to share all nodes' resources (i.e., storage) with others.

Results aim to simply provide a comparison between social-aware protocols considering the effect of different TTL values on the considered performance metrics. Also these results help understanding the potential of these solutions in a scenario close to the one targeted in this project, rural areas with mobile nodes carrying information on behalf of others.

Figure 1 shows the delivery probability of the two social-aware protocols in the evaluated scenario with TTL values set at 1, 2 and 3 days. The dLife protocol has a higher probability of delivery compared to the Bubble Rap. This is due to the fact that Bubble rap depend on the identification of communities and computations of node centrality. Since in the scenario evaluated the communities among node are not initially formed and most of the nodes have a low centrality, this introduces a burden to the solution which spends time computing community and centrality information, negatively affecting its delivery capability. The advantage of dLife is due to the fact that it can indeed capture the dynamics of the network which considers daily routine of users.

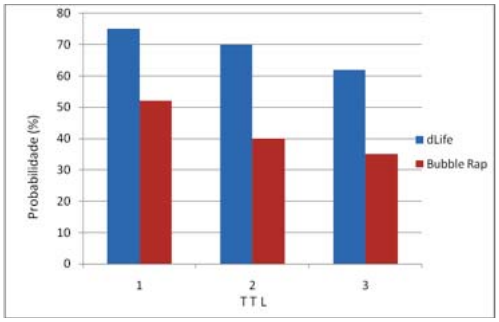


Figure 1. Delivery probability.

Figure 2 shows the number of replicas created for each

message with TTL values set at 1, 2 and 3 days. The results show that the dLife protocol has a performance approximately 38% better than Bubble Rap. This difference is due to the fact that dLife gets a better view of the social graph of the network and its most important nodes in specific time periods. The higher cost of Bubble rap is due to the existence of low popularity of intermediate nodes. Thus, to reach the desired destinations the proposal rely on the available centrality levels, which generates a higher rate of replication.

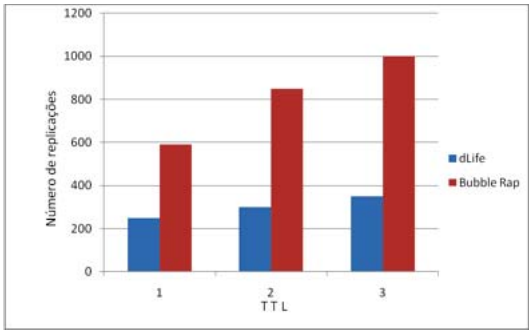


Figure 2. Cost.

Figure 3 shows the latency experienced by messages with TTL values set at 1, 2 and 3 days. The dLife protocol clearly presents a latency of approximately 45% less than Bubble Rap. In this case, dLife takes forwarding decisions independently of the notion of community, relying exclusively on the strength of the social ties with the desired destinations to increase its delivery probability. As the scenario is highly dynamic, Bubble rap ultimately takes longer to have a more accurate view of communities and node centrality, which results in many replicas being created to nodes that are not socially well connected to the destinations, consequently affecting the latency experienced by the messages.

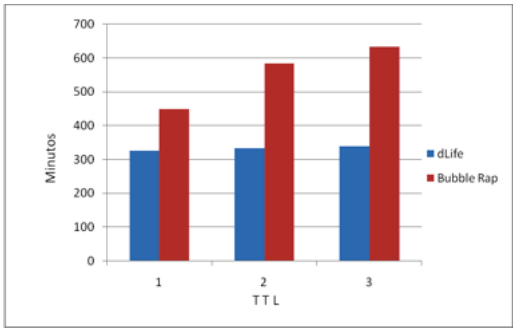


Figure 3. Latency.

The obtained results indicate that opportunistic forwarding based on social aspects is the direction to be followed in the project. Still, social-aware solutions based on metrics that reflect the notion of communities and notion centrality must be carefully considered as such metrics may introduce unwanted overhead, leading to negative performance behavior. Nevertheless, social-based solutions have shown great potential and can be employed in the rural scenario, subjected of the presented project.

Despite of not measuring the energy efficiency of the solutions, our view is that such feature is inherent to these solutions. By relaying on relevant social relationships, a social graph, over which opportunistic routing shall operate, can be defined. Consequently, less transmissions are required which spare network and nodes resources, thus reducing energy consumption.

V. Conclusions and Future Work

This project aims at innovative solutions to facilitate access to information and promote digital inclusion. These solutions are expected to be validated and published, contributing to the development of applications related to medicine, education, agriculture and environmental protection to be employed in the rural communities of the southeastern of the Goiás state. It is important to note that the application of the proposed mechanisms is not limited to rural areas of Goiás and can be further extended to work efficiently in urban settings, requiring very few modifications due to the use of open interfaces and modular components.

In addition, this project shall i) provide an exchange of experience and technology between the involved partners and interested parties (industry, government agencies); ii) result in skilled undergraduate and graduate students; iii) create new lines of research; iv) increase regional scientific production and product development; and v) tighten relations between technicians, students, teachers and researchers of the involved institutions.

Besides showcasing the main aspects of the project, this paper presents the first analysis on the usage of opportunistic networks as an alternative for rural areas when conventional Internet model does not seem the best option. Within this context, the paper addresses the application of social factors to support opportunistic forwarding protocols. Initial results on a performance comparison evaluation between two opportunistic routing protocols based on social aspects show that the project is in the right direction of providing energy efficient solutions. To support our claims, we considered a simple scenario, close to the targeted rural communities, where the dLife protocol performed better than the Bubble Rap.

As future work, we intend to thoroughly evaluate all opportunistic routing protocols based on social aspects also considering other scenarios, more specifically rural environment. In addition to simulations, the project targets the development of a real-world test environment as to help understanding how we can improve the life quality of the rural communities of the southeastern of the Goiás state.

Acknowledgment

This work is supported by Fundação de Amparo à Pesquisa do Estado de Goiás (FAPEG) in the context of the project entitled "Redes Oportunistas Sem Fio Energeticamente Eficientes como Suporte a Inclusão Digital e Fácil Acesso a Informação no Sudeste Goiano" number 201210267001156. We would like to also acknowledge CAPES for the PNPD grant provided do Dr. Waldir Moreira.

REFERENCES

[1] H.-O. Scheck, "Ict and wireless networks and their impact on global warming," in *Wireless Conference (EW), 2010 European*, pp. 911 –915, april 2010.

[2] W. Moreira, P. Mendes, and S. Sargento, "Opportunistic routing based on daily routines," in *Proceedings of WoWMoM*, San Francisco, USA, June, 2012.

[3] P. Hui, J. Crowcroft, and E. Yoneki, "Bubble rap: Social-based forwarding in delay-tolerant networks," *IEEE Transactions on Mobile Computing*, vol. 10, no. 11, pp. 1576–1589, November, 2011.

[4] W. Moreira and P. Mendes, "Social-aware Opportunistic Routing: The New Trend," in *Routing in Opportunistic Networks* (I. Woungang, S. Dhurandher, A. Anpalagan, and A. V. Vasilakos, eds.), Springer Verlag, May, 2013.

[5] W. Moreira, P. Mendes, and S. Sargento, "Social-aware opportunistic routing protocol based on user's interactions and interests," in *Proceedings of AdHocNets*, Barcelona, Spain, October, 2013.

[6] A. Vahdat and D. Becker, "Epidemic routing for partially-connected ad hoc networks," tech. rep., 2000.

[7] S. Nelson, M. Bakht, and R. Kravets, "Encounter-based routing in dtns," in *INFOCOM 2009, IEEE*, pp. 846–854, April 2009.

[8] P. Hui and J. Crowcroft, "How small labels create big improvements," in *Pervasive Computing and Communications Workshops, 2007. PerCom Workshops '07. Fifth Annual IEEE International Conference on*, pp. 65–70, March 2007.

[9] E. M. Daly and M. Haahr, "Social network analysis for routing in disconnected delay-tolerant manets," in *Proceedings of the 8th ACM international symposium on Mobile ad hoc networking and computing*, MobiHoc '07, (New York, NY, USA), pp. 32–40, ACM, 2007.

[10] P. Costa, C. Mascolo, M. Musolesi, and G. P. Picco, "Socially-aware routing for publish- subscribe in delay-tolerant mobile ad hoc networks," *IEEE J.Sel. A. Commun.*, vol. 26, no. 5, pp. 748–760, June, 2008.

[11] A. Mtibaa, M. May, C. Diot, and M. Ammar, "Peoplerank: social opportunistic forwarding," in *Proceedings of INFOCOM*, pp. 111–115, 2010.

[12] A. Oliveira-Jr, R. Ribeiro, W. Moreira, A. Neto, and E. Cerqueira, "A comparative analysis of green routing metrics in user-centric networks," in *XXXIII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC 2015) - XX Workshop de Gerência e Operacão de Redes e Servicos (WGRS 2015)*, (Vitória-ES-Brazil), May 2015.

[13] A. Oliveira-Jr and R. Sofia, "Energy-awareness in multihop routing," in *Wireless Networking for Moving Objects* (I. Ganchev, M. Curado, and A. Kassler, eds.), vol. 8611 of *Lecture Notes in Computer Science*, pp. 137–156, Springer International Publishing, 2014.

[14] C. E. Jones, K. M. Sivalingam, P. Agrawal, and J. C. Chen, "A survey of energy efficient network protocols for wireless networks," *Wirel. Netw.*, vol. 7, no. 4, pp. 343–358, July, 2001.

[15] A. Junior, R. Sofia, and A. Costa, "Energy-awareness metrics for multihop wireless user-centric routing," in *The 2012 International Conference on Wireless Networks (ICWN'12)*, July 2012.

[16] A. Junior, R. Sofia, and A. Costa, "Energy-awareness in multihop routing," in *2012 IFIP Wireless Days (WD)*, pp. 1–6, November 2012.

[17] A. Oliveira-Jr, R. Ribeiro, D. Matsuo, T. Filho, W. Moreira, A. Neto, and E. Cerqueira, "Green routing metrics for multi-hop wireless people-centric networks," in *Proceedings of the Latin America Networking Conference on LANC 2014*, LANC '14, (New York, NY, USA), pp. 11:1–11:4, ACM, 2014.

[18] A. Junior and R. Sofia, "Energy-awareness metrics global applicability guidelines." IETF Internet Draft, draft-ajunior-roll-energy-awareness-01 (working in progress), January 2014.

[19] K. Scott and N. Bambos, "Routing and channel assignment for low power transmission in pcs," in *Universal Personal Communications, 1996. Record., 1996 5th IEEE International Conference on*, vol. 2, pp. 498–502 vol.2, September 1996.

[20] Q. Xie, C.-T. Lea, M. Golin, and R. Fleischer, "Maximum residual energy routing with reverse energy cost," in *Global Telecommunications Conference, 2003. GLOBECOM '03. IEEE*, vol. 1, pp. 564–569 Vol.1, December 2003.

[21] S. Singh, M. Woo, and C. S. Raghavendra, "Power-aware routing in mobile ad hoc networks," in *MobiCom '98: Proceedings of the 4th annual ACM/IEEE international conference on Mobile computing and networking*, (New York, NY, USA), pp. 181–190, ACM, October 1998.

[22] C.-K. Toh, "Maximum battery life routing to support ubiquitous mobile computing in wireless ad hoc networks," *Communications Magazine, IEEE*, vol. 39, pp. 138–147, June 2001.

[23] D. Kim, J. Garcia-Luna-Aceves, K. Obraczka, J. Cano, and P. Manzoni, "Power-aware routing based on the energy drain rate for mobile ad hoc networks," in *Computer Communications and Networks, 2002. Proceedings. Eleventh International Conference on*, pp. 565–569, October 2002.

[24] M. Maleki, K. Dantu, and M. Pedram, "Lifetime prediction routing in mobile ad hoc networks," in *Wireless Communications and Networking, 2003. WCNC 2003. 2003 IEEE*, vol. 2, pp. 1185–1190 vol.2, March 2003.

Topic Templates

A clustering technique to categorize web based searches

Prathamesh R Divekar
Department of Computer Science
The University of Georgia
Athens, Georgia, USA
pdivekar@uga.edu

Dr. Hamid R. Arabnia
Department of Computer Science
The University of Georgia
Athens, Georgia, USA
hra@cs.uga.edu

***Abstract*— Topic based classification and searches have always been a hefty challenge along the corridors of data mining. Reading a large amount of articles and indentifying them of to be the same genre or precisely one subject matter is nearly impossible. With the ever popular need for refinement and quick results we have cropped up a technique to apply graph clustering and probabilistic theory along with known data mining concepts to develop a relationship between words that present high instances of existing together across a majority of documents. These words or topics as we call them form a "Topic Graph". A Graph is thus a set of words with a high frequent, high probabilistic relationship amongst them. In more technical theory, it is a highly connected graph with words as nodes and relationships between these words as edges. We can apply these concepts of Topic Graphs to refine and categorize search result along with creating new Graphs if the need arises. One of the possible resulting applications should be able to provide precise and specific search answers satisfying user's requests.**

***Index Terms*— Topic Maps, Graph Theory, Word Relations**

I. Introduction

The invention of World Wide Web (www) has ushered in the era of search engines and information retrieval. Although, the ascension of internet along with its many diversities and interests provided a near unlimited area of storage space for information, it's just too huge to search and thus makes its more and more difficult to find information. Popular web search engines line Google, Yahoo, AltaVista, Infoseek and MSN do exist to help people find information on the web. Most of these systems return a ranked list of web pages in response to a user's search request. Web pages on different topics or different aspects of the same topic are mixed together in the returned list. The user has to sift through a long list to locate pages of interest [18]. Some believe this to be an annoying issue.

Most internet search engines of the present perform a phenomenal task of providing a linear list of sorted or ranked results for a query. For known-item queries, users often find the site they are looking for in the first page of results. However, a list may not suffice for more sophisticated exploratory tasks, such as learning about a new topic or surveying the literature of an unfamiliar field of research, or when information needs are imprecise or evolving [19][20]. Many a times a single word even though a proper noun, may have complete different meanings. For example: 'S5' may refer to a 'Samsung Galaxy S5' or an 'Audi S5'. The former is the newest phone offered by the Samsung Galaxy series, while the latter refers to a luxury sedan offered by the ever popular German car manufacturer 'Audi'. Queries having ambiguous terms may retrieve documents which are not what users are searching for [22]. No search engine can predict what the user wants to search for at any given moment of time. Although most search engines provide possible related searches or search suggestion, but can never know for sure what the user wants to search.

Another issue is the "why"! The "why" of user search behavior is actually essential to satisfying the user's information need. After all, users don't sit down at their computer and say to themselves, "I think I'll do some searches." Searching is merely a means to an end – a way to satisfy an underlying goal that the user is trying to achieve. (By "underlying goal," we mean how the user might answer the question "why are you performing that search?")[21]. That goal may range from buying the grocery to the newest video game. Or it may range from finding the latest election results to finding what his next door neighbor is up to. Or it even may be to find if some famous celebrity said something controversial to voicing his own opinion about a certain pertaining issue or a plethora of possibilities. In fact, in some cases the same query might be used to convey different goals - For instance, an user searching for 'Samsung Galaxy S5' might get results ranging from the technical knowhow, price to possible outlet stores that sell the product. He may only want to know about the technical issues of the phone rather than the best place to purchase it and if the search results produced somehow tend to be more inclined towards 'possible places to purchase' type, that itself may annoy him enough to produce a negative impression about it.

Perhaps, no technology of the present or the near future may have the capability to completely solve such issues, but techniques like related searches and suggestion may resolve them to some extent. This paper presents a possible solution or at least a way to reduce user's annoyance with search engines called as "Topic Templates" - a system to compartmentalize

search results based on sets of closely associated words. We have devised a technique to formulate bands of words together, having highly logical relations in the real world. By real world I mean the human universe as we know it. The common issue of unnecessary, ambiguous and redundant search results can be reduced to some extent.

What we attempt is to provide search engine, users or any possible searching techniques a hierarchy to search for. Instead for directly searching for the user query, a system may use our topic sets for searching a query thus returning results associated directly with these sets and organized categorically based on these sets. This paper introduces this concept of topic graphs and topic sets, their benefit for searching and a process through which they are forged from any available collection of documents. The final result that is produced is a compilation of groups of words that can be then used as templates for searching as mentioned before. For example: Our previous example of 'Samsung Galaxy S5' could be associated with searching for price, outlet stores, tech specs or comparison with competitors. For that follows topic sets can be available: [Galaxy, S5, Target, At&t], [S5, Verizon, At&t, T-Mobile, ...], [Samsung, S5, Galaxy, PC Magazine, Chips, Amoled, …] or [Galaxy, S5, HTC, One, M8, ….] etc.

A. The Concept

The core theory of this document is based on the simple fundamental belief that "no word is ever alone". Thus we have personified words and phrases to have relationships with other words. Any definition of any topic is a group of words semantically arranged together to make sense. Thus, every single term is any sort of search query can be believed to be associated with a set of words that define them and rather add character to them. We establish relationships between keywords by using a graphical approach. Why Graphs? Because graphs are an efficient data structure to represent hierarchical information. This question gets answered more clearly in later chapters.

The application presented in this paper is a four stage structure in technical terms but we can introduce it in a three step perception to put forth a foundation for the concept as follows:

- **The Corpus:** A plethora of documents exist in every format in the universe. Every known information is represented in a written format so that it can reach every corner of the world to be distributed. The first step is to take a bunch of such documents at random and formulate their base topic based on devising key words from written paragraphs.
- **The application:** The application will establish associations between available keywords by calculating a probabilistic weight between those words that frequently appear together in a host of documents. The result will be a graph with keywords as vertices and their relationships with other words would be the reason for its edges and their weights. The weights help to establish a relationship score between these words.
- **Clusterize:** The final step is to group together those words which have the maximum or near maximum relationship score between them. So we eliminate weaker edges and form several sub graphs. We then establish a hierarchical tree structures for these sub graphs which gives us topic graphs which can be further categorized and organized into topic sets.

The further sections in this paper will present in depth the whole process of how an assemblage of random documents result in the formation of topic templates. We will describe the core system architecture - which essentially is four stages, a working on-paper demonstration, an evaluation of our tests that provide an evidence for the theory and possible enhancements to the proposed concept.

II. BACKGROUND AND MOTIVATION

The need to categorize has always been evident in human nature. Categories present a systematical approach in any organizational approach. Simple examples are evident in our day to day lives starting from our personal bedrooms to kitchens, from our TV guides to restaurant menus, from groceries to books. Categorical and systematic organization of information has always been and will always be the prime need and expectation of any venture.

Three general techniques have been used to organize documents into topical contexts. The first one uses structural information (Meta data) associated with each document. The DynaCat system by Pratt [23] used Meta data from the UMLS medical thesaurus to organize search results. In the SuperBook project [24], paragraphs of texts were organized into an author-created hierarchical table of contents. Others have used the link structure of Web pages to automatically generate structured views of Web sites. Maarek et al.'s WebCutter system [25] displayed a site map tailored to the user's search query. Manually-created systems are quite useful but require a lot of initial effort to create and are difficult to maintain. Automatically derived structures often result in heterogeneous criteria for category membership and can be difficult to understand [18]. A second way to organize documents is by clustering. Documents are organized into groups based their overall similarity to one another. Zamir et al. [26, 27] grouped Web search results using suffix tree clustering. Hearst et al. [28, 29] used the scatter/gather technique to organize and browse documents. Clusters are usually labeled by common phrases extracted from member documents [18]. A third way to organize documents is by classification. In this approach, statistical techniques are used to learn a model based on a labeled set of training documents (documents with category labels). The model is then applied to new documents (documents without category labels) to determine their categories [18].

Our concept uses a bit of each three methods described above. The first method of organizing in a hierarchical format based on the structural information of the document is the final result of our whole application. The final output is a file containing terms organized as trees and we formulate our topic sets based on these trees. The second method is applied to create the topic trees. We clusterize one major graph obtained by establishing relationships between all the terms and then organize these clusters into a hierarchical tree structure. The whole idea and its introduction are based on the third method. We formulate these topic trees and graphs along with topic sets and then propose an application to categorize searches based on these topic sets we have obtained.

Search Engines like Google, Bing, and Yahoo now-a-days deliver a customized search result. This leads to an effect that has been called a filter bubble. Thus, the user has information retrieval process based on his past experiences and searches rather than the present ongoing. News articles produce new results on a daily basis which can never be found in any user's search history because they are new. Thus it becomes imperative to categorize searches based on inter topic relationships as well. According to Eli Pariser, who coined the term, users get less exposure to conflicting viewpoints and are isolated intellectually in their own informational bubble. Pariser related an example in which one user searched Google for "BP" and got investment news about British Petroleum while another searcher got information about the Deepwater Horizon oil spill and that the two search results pages were "strikingly different"[1][2][9]. The bubble effect may have negative implications for civic discourse, according to Pariser. Since this problem has been identified, competing search engines have emerged that seek to avoid this problem by not tracking or "bubbling users [10].

A filter bubble is a result state in which a website algorithm selectively guesses what information a user would like to see based on information about the user (such as location, past click behavior and search history) and, as a result, users become separated from information that disagrees with their viewpoints, effectively isolating them in their own cultural or ideological bubbles. Prime examples are Google's personalized search results and Face book's personalized news stream [1][2][9].

We argue although these personalized searches present the user with results that adhere to their interests and liking rather than presenting the data which is more complementary to the updated happenings on the planet. We thus take this into account and present a more public knowledge based categorizations rather to help the user negate this filter bubble. Instead we try and filter out those redundant and unnecessary results that prove more of a nuisance.

A. Why Graphs?

A month ago a search for 'Malaysia Airlines' would have returned the web address of the chief website of the airlines company or a schedule depicting the coming and goings of certain flights operated by the same airline organization. A search for most now would be expected to return results concerning foremost the words "MH370, Flight 370, disappearance, mystery etc." – all concerned with the latest tragedy that occurred over heaven and earth. Such incidents that happen over time change the complexion of the world as we believe it. Such changes not only affect our mind set but automatically set its tone on the cyber world as well. Among all one thing that remains consistent is the ability or the human nature to associate relationships between what we call words no matter how much situations change. Times change we roll on with it and learn to adapt the fact that words will be associated with each other for an era.

Every search provides with some related search suggestions yet there are some redundant results or those that contain the searched query as a mere formality. Every term – or word – has some associations – related words – which have a certain high probability of occurring together in a host of documents. Such word association can be assumed to have a relation between them and can be assigned certain weights - to signify the strength of their bond – which would ultimately mean that they appear together in most documents and searching for one along with its siblings can return more specifically categorized results. Also classifying documents (books, articles, web pages etc) based on these words and association can help users to have more content specific searches.

So where does the graph theory come in? Consider each term as a node or a vertex of the graph. It's a vast ocean of words out there and the only thing we know about them is that they belong to some document. We then establish some relationships between some of these words and create edges amongst them. Thus we have a huge graph where words are connected to each other. Next we contract this graph. We remove some unneeded edges and with them some isolated vertices. Thus we have particularly important words with a certain calculated weights and each of these has some associations with other words – strong or weak. After we are done with the contractions, we need to have only those word associations that have enough to be together forever. Thus we form clusters from the main graph. Each of these clusters is based on the fact that the terms in them a certain high probability of appearing together in different document. The thing to note here is that not all the words a cluster necessarily appear in the same document. The probability that a document would contain all the terms of the cluster in probably pretty low but the certain group of terms in the cluster do appear together with a high probability. In more technical terms we have highly connected graphs – not complete graphs – which signify the close associations between words and gives us more than one topic sets out of a topic graph.

For example: staying with the Malaysia Airlines disaster. Suppose we have over ten articles concerned with the latest incident that occurred over the Indian Ocean waters. Each of these documents filters out words that occur in them with a certain high frequency. We can have the following words – *Malaysia, Flight, MH, and 370, Kuala Lumpur, Indian, Ocean, India, Australia, Asia, Beijing, Boeing, Thailand, gulf, Malay,* accident and some more. Now the actual incident happened when "**Malaysia Airline Flight MH 370**, travelling from **Kuala Lumpur** International airport to **Beijing** International airport went missing less than an hour after its take off". All the bold words in the above sentence are the key words discussed or mentioned when people speak about the incident. To be more precise, these words appear together with high probability in a group of documents related to this incident. Thus these words could be clustered together since they have such a strong association. Thus we have a topic graph G (V, E) where *V = {Malaysia, Airline, Flight, MH 370, Kuala Lumpur, Beijing}.*

The above example thus implements the theory of graph clustering to string together terms which have a high certainty of appearing together in a corpus of document.

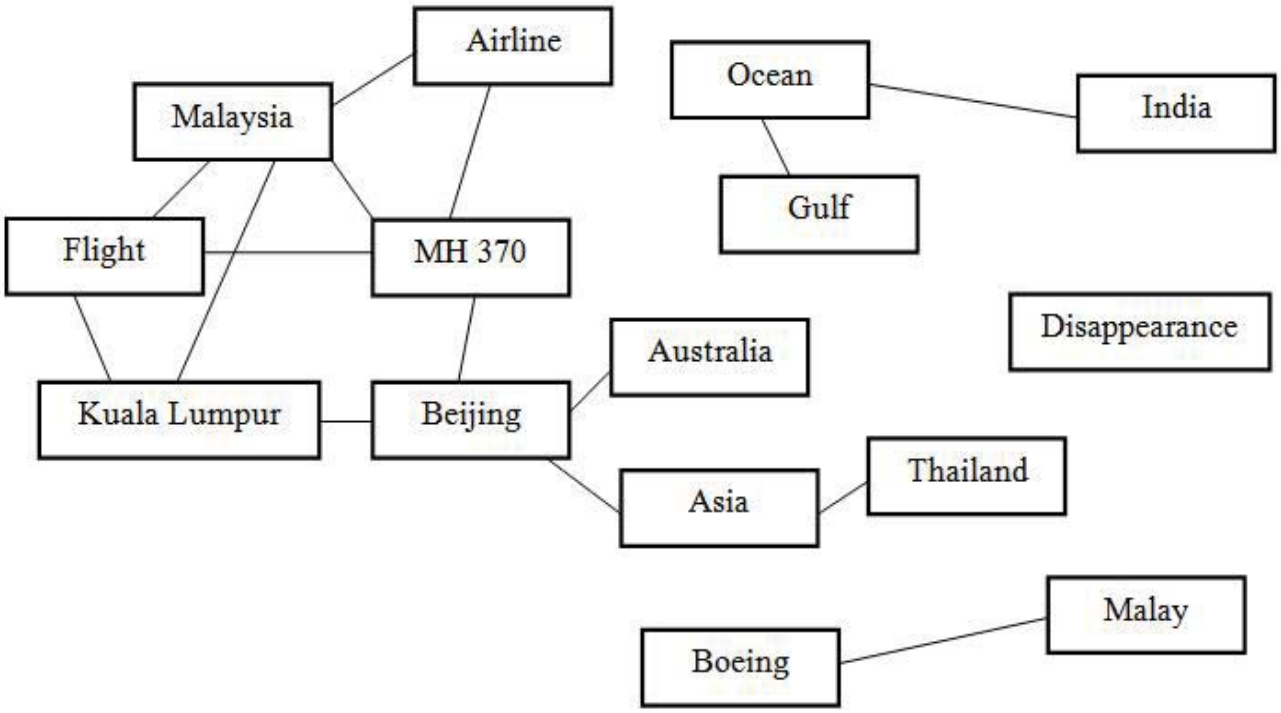


Figure 1: A graph of related terms based on the example mentioned above.

B. Graph Representation

In computer science, a graph is an abstract data type that is meant to implement the graph and hyper graph concepts from mathematics. A graph data structure consists of a finite (and possibly mutable) set of ordered pairs, called edges or arcs, of certain entities called nodes or vertices. As in mathematics, an edge (x, y) is said to point or go from x to y. The nodes may be part of the graph structure, or may be external entities represented by integer indices or references. A graph data structure may also associate to each edge some edge value, such as a symbolic label or a numeric attribute (cost, capacity, length, etc.) [11]. We exploit this of nodes and edges since our prime objective is to establish a relationship between words. The terms themselves become the nodes and the edges between them become a reason to define the existence of a relationship between them.

Various ways to implement graphs exist in programming terms. The two most basic ways are Adjacency Lists and Adjacency Matrix. Operations with a graph represented by an adjacency matrix are faster. But if a graph is large we can't use such big matrix to represent a graph, so we should use collection of adjacency lists, which is more compact. Using adjacency lists is preferable, when a graph is sparse, i.e. $|E|$ is much less than $|V|^2$, but if $|E|$ is close to $|V|^2$, choose adjacency matrix, because in any case we should use O ($|V|^2$) memory

[12]. Adjacency matrix and adjacency lists can be used for both directed and undirected graphs [12]. We however use a combination of these two basic representations. We use Hash Maps and Linked List. Both these Data structure are used for different phases of the system architecture which has been described in the paper.

We use Hash Maps to define the term graph (graphical representation of all key words and their relationships with other key words in the corpus). The key is the term as the node of the graph and the value is a List of terms that and its connections. Linked list are utilized during the Topic Graph creation phase where we arrange all terms in a hierarchical tree formulate topic sets later. Each node of the tree/graph is linked to its child in the structure thus having a kind of a unidirectional traversal – our graphs are not directed.

	Hash Map
Memory complexity (optimal – O(|E|)	**O(|E|)**
Add new term (optimal – O(1))	**O(1)**
Remove term (optimal – O(1))	**O(1)**
Search for a term (optimal – O(1))	**O(1)**
Enumeration of vertices (term) adjacent to 't (term in question)' (optimal – O(|K|))	**O(|K|)**

Table 1: Memories and complexities for a HashMap. We consider each term to a vertex in terms of a graph. Thus K is the number of adjacent terms to a term t [12].

III. THE SYSTEM ARCHITECTURE

In this section we describe the multi stage architecture of our topic-set maker and provide a detail insight into each component of the system. We have discussed in the prior sections the challenges that we encounter in creating these graphs and categorized topic sets. Having a multi stage system is essential to prevent redundancy along with preserving the precision that does not yield false results.

The journey from a group of documents to creating a graph and a more precise topic graphs to topic sets, is a fourfold. Each stage outputs a distinct set of files with more precise and simplified information than its predecessor stage. Our input files start as text files. Each input file is an article or paragraphed sentence as defined in legible English language. By legible English we do not mean, they are random words just typed for the sake of typing or a computer language program – which though English do not meet the legible criteria – that is they can be successfully parsed by an English language parser. The end result is one single file though called as '.tt' file – tt stands for topic template, which has essentially topic graphs and topic sets associated with each graph. The end result also contains two more files: 1) Set of edges and their weights. 2) A hash map file where the key is a term and value is a list of all its true connection. We will explain further what true connections are.

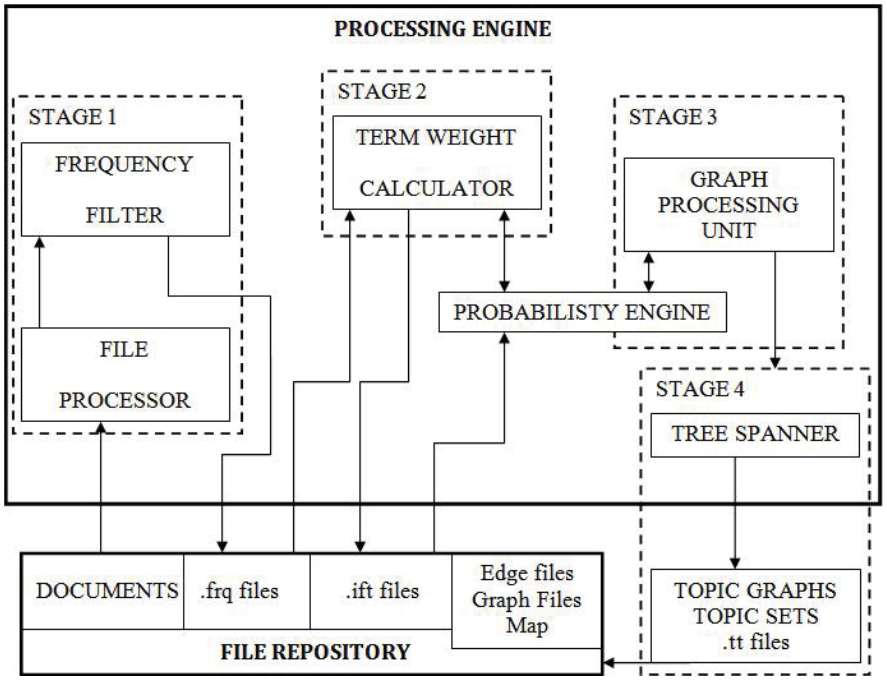


Figure 2: TGS System Architecture

Figure 2 is the system architecture of a simple topic template creator. The system architecture is divided into two parts – The Processing Engine and the Repository.

The Processing Engine as the name specifies does all the work starting from filtering files and driving the topic sets from the input corpus. The processing engine holds the four stage architecture which gives the final output. Both input and the output are used from and stored in the file repository of the system.

All files – input documents, .frq, .ift, Graph, Edge, Hash Map and .tt – are stored in the repository. The repository or the File Repository could be any kind of a storage space like a database, online files repository – GitHub, SVN, Bit Bucket etc, or just a folder on the local host or some common server. Use of some kind of a personal digital library could be encouraged as it presents multiple advantages – no physical library, round the clock availability, multiple access and uses, information retrieval, finer preservation and conservation, possibly infinite space (for a considerably low cost)[13].

A. Stage 1: File Filtering

Stage 1 of the architecture involves the base input corpus. The corpus is a group of documents that could be any of the following:

1) Newspaper and/or magazine Articles
2) Wikipedia blogs
3) Online blogs

Apart from the above (which we used for testing) the corpus could include any legible English language articles popularly talking about some base topic and its constituents but with proper sentence construction. Documents written in modern urban slang or popular short hand abbreviations are discouraged so as not to get unnatural results.

The output of the file filtering stage is the '.frq' files. These assign an identity – some integer id – to every file and contain a list of keywords and their term frequencies.

Term Frequency: Tf(t, d) is the raw frequency of a term in its document, i.e. the number of times the term appears in the document[15].

File Processor: The file processor stems down the available text to words. During the stemming process we cut down plurals, verb forms etc to their base forms and we get rid of high frequency stop words like articles, prepositions [14].

Frequency Filter: We calculate the 'Tf' for every keyword and using a certain threshold we filter out those terms that pass a certain threshold frequency. For example the average frequency of a term in a document is 15; we remove all those key words that have 'Tf' less than 15.

B. Stage 2: Term Weight Processing

Stage 2 gives a certain popularity score to every word based on the following quantities. The output of the term weight processing engine is an '.ift' file. The .ift files have a table with each term its Tf, Idf, term weight and the Df. Term Weight Calculator is the only component of stage 2 that achieves this target.

Inverse Document Frequency: The inverse document frequency is a measure of whether the term is common or rare across all documents. It is obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient [15].

$$Idf\ (t, D) = \log \frac{N}{|\{d \in D : t \in d\}|} \quad (1)$$

N: The total number of documents in the corpus.

Documents Frequency: It is the number of documents where the term t appears. In equation (.1) $|\{d \in D : t \in d\}|$ is the Document Frequency or Df.

TfIdf: The term weight of the TfIdf is calculated as the product of Tf and Idf($\mathrm{Tf}(t, d) * \mathrm{Idf}(t, D)$). A high weight in Tf * Idf is reached by a high term frequency (in the given document) and a low document frequency of the term in the whole collection of documents; the weights hence tend to filter out common terms. Since the ratio inside the Idf's log function is always greater than or equal to 1, the value of Idf (and TfIdf) is greater than or equal to 0. As a term appears in more documents; the ratio inside the logarithm approaches 1, bringing the Idf and TfIdf closer to 0 [15].

C. Stage 3: Term Graph Arrangement

Stage 3 assembles the term graph – important key words based on certain Tf and Term Weight thresholds – and formulates edges between them considering that each term is a node for these edges. Creating edge at this stage is just based on the fact that two words of an edge occur together in multiple documents. Based on the number of occurrences we calculate $p(E)$ that is the probability of the edge which is essentially a ratio - O/U – the ratio of the occurrences over the union of sets that contain the terms that bind the said edge E.

Stage 3 produces three files:
1) Graph: A file containing terms and all its connection.
2) Edge: A file that keeps track of all edges and its probabilistic weights.
3) Hash Map: A file that keeps track of all terms and the documents it appears.

Graph Processing Unit: Not to be confused with a GPU (Graphical Processing Unit), a graph processing unit in this particular system gives the term graph. It assembles edges together between vertices and assigns weights to these edges.

Probability Engine: The probability engine calculates the p(E) for every edge, filters out those below a certain threshold and creates "TRUE EDGES" out of all those available edges. Thus a true edge is and Edge between two terms in a term graph, whose probabilistic weight surpasses a certain threshold set by the creator.

D. Stage 4: Topic Templates

The fourth and the ultimate stage of the system give us a file with comprehensive topic graphs and topic sets. The output is '.tt – topic templates' files as mentioned before. The files presents a hierarchical tree structures of topics arranged together establishing a more systematic parent child relationship between terms. We analyze these relationships further to give topic sets. We define the two subjects of our final output as follows:

Topic Graphs: *Topic graph or topic trees are inter-related term derived from our base corpus, arranged in a hierarchical fashion much similar to a family tree thus establishing a parent child relationship between words that appear in a term graph.*

The thing to remember about topic trees is that, direct siblings and parent – child definitely appear together with high probability in documents together in the corpus. A parent and its direct child will be present together across certain high number of documents, but parent its grandchildren may or may not. Similarly two siblings of the same parent will appear together in significant number of documents but cousins may not.

Topic Sets: *Topic sets are a set of terms such that, each term in the set has a high probabilistic relationship with more than two terms in the set.*

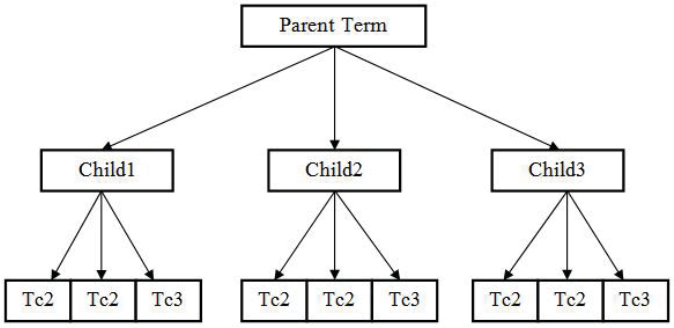


Figure 3: Topic Tree

Figure 3 is a one such topic graph. It is not essentially a tree because even we do not depict it; an edge definitely exists between two siblings of a parent. A tree structure just helps in a hierarchical arrangement. A topic set thus will have parents, its children and grandchildren. Cousins won't be a part of the same topic template thus giving categorical divisions when it comes to every term.

Tree Spanner: The tree spanner clusters the term graph into topic graphs and further divides the topic graph into topic sets.

TGS Repository: The TGS repository is a child repository of the File repository that finally stores all the topic graphs and topic sets that we would get from a base corpus.

IV. TOPIC TEMPLATES

In this section we elaborate what Topics templates are; how topic sets are formulated from a topic graphs aided by a basic logical example – in this sense logical adheres more to obvious than sensible. We start with no more than four documents. Explain how and why we do what we do at every process and at the end present a pseudo code algorithm to implement our theory. To make the application more apparent in layman terms, we will try to establish the similarity between that family tree structure and our topic graph formulations as we mentioned in our introduction section.

A. The Corpus

For the sake our non computerized human implementation of the system that we propose, we have chosen four paragraphs rather than entire documents to make the working of the system apparent and simple to a layman. These for paragraphs – called as documents henceforth – are summarized descriptions of the last four *Harry Potter* books. They are follows:

Doc 1: ***Harry Potter*** *and the* ***Goblet*** *of* ***Fire.***

Book four of the ***Harry Potter*** *tests* ***Harry*** *in the most unusual way not only his abilities to cope with life threatening challenges but also his friendship with* ***Ron Weasely*** *and* ***Hermoine Granger.*** ***Hogwarts*** *and the* ***ministry*** *of* ***magic*** *after a hiatus of almost a century organize the* ***tri wizard tournament*** *between three best known magical schools –* ***Drumstrangs*** *and* ***Beauxbatons.*** ***Albus Dumbledore*** *who has sensed the signs of the eminent return of Lord* ***Voldemort,*** *has employed ex-auror* ***Alastor Madeye Moody*** *as the new* ***defense*** *against the* ***dark arts teacher*** *with a view to protect* ***harry.*** *New characters and plots are introduced in the fourth with most awaited of the main villain of the series in this book. Things will change for* ***Harry*** *and his friends.*

Doc 2: ***Harry Potter*** *and the* ***Order*** *of the* ***Phoenix.***

Lord ***Voldemort*** *has returned and though the* ***ministry*** *of* ***magic*** *is arrogant enough to ignore it,* ***Dumbledore*** *has summoned* ***Sirius Black*** *and the*

Order *of the* ***Phoenix*** *to organize a resistance against the* ***dark arts****.* ***Harry, Ron*** *and* ***Hermoine*** *return to* ***Hogwarts*** *where the* ***ministry's*** *motivation to curb* ***Dumbledore's*** *so called lies has taken an unexpected stand.* ***Dolores Umbridge*** *has been appointed the new* ***Defense*** *against the* ***Dark arts teacher*** *and the High* ***Inquisitor*** *to inculcate some discipline among the failing standards of the school. If that's not enough* ***Harry*** *is constantly facing nightmares which actually is a direct connection to* ***Voldemort's*** *mind. Tough times await* ***Harry*** *as the* ***Hogwarts*** *he knows will never be the same.*

Doc 3: ***Harry Potter*** *and the* ***Half Blood Prince***

The death of ***Sirius Black*** *and the revelation of Lord* ***Voldemort's*** *return have sparked the same panic and unrest in the magical world as it was fifteen years ago.* ***Dumbledore*** *and* ***Harry*** *embark on a mission to discover the life and lies of* ***Tom Riddle*** *a.k.a Lord* ***Voldemort****. Their journey leads them into the world of* ***Horcruxes*** *–* ***Dark*** *objects that store a* ***wizard's/witch's*** *soul making him undefeatable. Life in* ***Hogwarts*** *is back to its usual self though with the imminent danger of the death eaters apart from the fact the* ***Severus Snape*** *as the new* ***Defense*** *against the* ***dark arts teacher****.* ***Harry*** *thus struggles to come with terms* ***Snape's*** *latest victory. New challenges await* ***Harry, Ron*** *and* ***Hermoine****, some not associated with the dangers of the real world as they come of edge. This book will indeed prove to a cliff hanger.*

Doc 4: ***Harry Potter*** *and the* ***Deathly Hallows***

The last battle, the final war. ***Snape's*** *betrayal which led to the death of* ***Albus Dumbledore*** *has sparked fall of* ***ministry*** *and* ***Hogwarts*** *into the hands of the death eaters.* ***Harry, Ron*** *and* ***Hermoine*** *embark on the mission set by* ***Dumbledore*** *to find and destroy Lord* ***Voldemort's Horcruxes*** *which will lead to his defeat. On their journeys they discover the existence of the* ***deathly hallows*** *which are believed to be objects that would make the owner a master of death. A race issues between the good and the bad over the possession of these* ***deathly hallows*** *which lead to the biggest war* ***Hogwarts*** *has ever seen.*

B. File Filtering

We filter these four documents in stage 1. We remove all the high frequency stop words, and create a table with each document and its keywords and their term frequencies.

For the sake of easing our non computerized calculations we have only considered proper nouns that appear in the documents mentioned above. Our system does not classify between nouns, adjectives, pronouns etc. since we do not want to enter into the realms of Natural Language Processing. So we remove all stop words and unnecessary words. We keep the obvious key words in the documents. Then we assign the term frequencies to each term. We can set a certain frequency threshold o filter out low frequency words from the document. In this case we set it to 2.

$$\textbf{\textit{Filter documents}}\ (t, d)$$
$$\quad if\ (tf(t,d) < 2)$$
$$\quad\quad remove\ t\ from\ d$$
$$\textbf{\textit{end}}$$

Algorithm 1**:** Filter Documents

Doc 1 - 01	**Doc 2 - 02**	**Doc 3 - 03**	**Doc 4 – 04**
Harry 5	Harry 4	Harry 4	Harry 2
Potter 2	Order 2	Voldemort 2	Deathly 3
	Phoenix 2	Dark 2	Hallows 3
	Voldemort 2	Snape 2	Hogwarts 2
	Ministry 2		Dumbledore 2
	Dumbledore 2		
	Hogwarts 2		
	Dark 2		

Table 2**:** contents of a .frq file: Terms and frequency

Table 2 is a typical '.frq' file which contains all the term that qualify a certain set frequency threshold and their frequencies. Thing to note is that each column of the table is a separate file each accompanied by the document id.

C. Term Weight Processing

We get four '.frq' files from the first stage with terms and their frequencies. We apply and calculate the Document Frequency (Df), Inverse Documents Frequency (Idf), Tf*Idf of each term from '.frq' files. Thus we have a popularity quotient for which term which signifies how important the term is in its document.

	Tf	**Df**	**Idf**	**Tf * Idf**	**Doc Id**
Harry	5	4	0	0	1
Potter	2	1	0.602	1.204	1

Table 3**:** '.ift' file for Doc 1

	Tf	**Df**	**Idf**	**Tf * Idf**	**Doc Id**
Harry	4	4	0	0	2
Order	2	1	0.602	1.204	2
Phoenix	2	1	0.602	1.204	2
Voldemort	2	2	0.301	0.602	2
Ministry	2	1	0.602	1.204	2
Dumbledore	2	2	0.301	0.602	2
Hogwarts	2	2	0.301	0.602	2
Dark	2	2	0.301	0.602	2

Table 4**:** '.ift' file for Doc 2

	Tf	Df	Idf	Tf * Idf	Doc Id
Harry	4	4	0	0	3
Voldemort	2	2	0.301	0.602	3
Dark	2	2	0.301	0.602	3
Snape	2	1	0.602	1.204	3

Table 5: '.ift' file for Doc 3

	Tf	Df	Idf	Tf * Idf	Doc Id
Harry	4	4	0	0	4
Deathly	3	1	0.602	1.806	4
Hallows	3	1	0.602	1.806	4
Hogwarts	2	2	0.301	0.602	4
Dumbledore	2	2	0.301	0.602	4

Table 5: '.ift' file for Doc 4

D. Term Graph Arrangement

We process all of the '.ift' files from stage 2 and prepare term graph. Term graphs are formulated using three files – The Edge file, the hash map and the Graph. The edge file contains all the edges with their associated probabilistic weights. The Hash map has all the terms and a list of all documents it is contained in. And the Graph contains all the vertices and its adjacency lists.

The first file created is the Hash Map which contains all the term and each term is associated with a Document Set – Ds. Ds of term t contain the Ids of all the documents the term t belongs to. We formulate the edges as follows:

E (u, v) is an Edge if u and v are terms from processed '.ift' files and the size of the intersection of the Document Sets of u and v is greater than a certain predefined threshold which is mandatorily more than or equal to 2. We calculate the probability weight of each edge as:

$$Pwt\big(E(u,v)\big) = |Df(u) \cap Df(v)| \,/\, |Df(u) \cup Df(v)| \quad (2)$$

The contents of the three files of this stage for our corpus are as follows:

Hash Map:
Harry: [1, 2, 3, 4]
Voldemort: [2, 3]
Dumbledore: [2, 4]
Hogwarts: [2, 4]
Dark: [2, 3]

Edge:
$Harry, Voldemort: wt = 2, Pwt = 0.5$
$Harry, Dumbledore: wt = 2, Pwt = 0.5$
$Harry, Hogwarts: wt = 2, Pwt = 0.5$
$Harry, Dark: wt = 2, pwt = 0.5$
$Voldemort, Dark: wt = 2, pwt = 1$
$Hogwarts, Dumbledor: wt = 2, Pwt = 1$

Graph:
$Harry \rightarrow Voldemort, Dumbledore, Hogwarts, Dark$
$Voldemort \rightarrow Dark$
$Dumbledore \rightarrow Hogwarts$

E. Topic Templates

The last stage produces the Topic Graph/Tree and the Topic sets. We start by traversing every vertex in the graph and looking at every edge in the term graph. For every vertex in the graph we see if those present in its adjacency list belong to the lists of each other. Those thus that have connections between them become the children of the first vertex we started from. Next we check the 'Pwt' of each edge that we have chosen. If the Pwt (Child1, Child2) is more than that of Pwt (Parent, Child2) then Child2 becomes the child of Child 2 in the tree.

CreateTree (G, t, t_i, t_j)
For every $t \in$ term graph
  For every $t_i \in$ Adjacency list (t)
    if $t_j \in t_i$
      Add t_i, t_j to child list of t
    if $(Pwt(t_i, t_j) > Pwt(t, t_i)$
      Add t_j to child list of t_i
  end
end

Algorithm 2: Create tree Algorithm

For every tree that we get from the term graph every LHS and RHS of the parent is a topic set. To limit the number of words in a topic set, our tree are limited to no more than four generations. The '.tt' files of our corpus yield the following results:

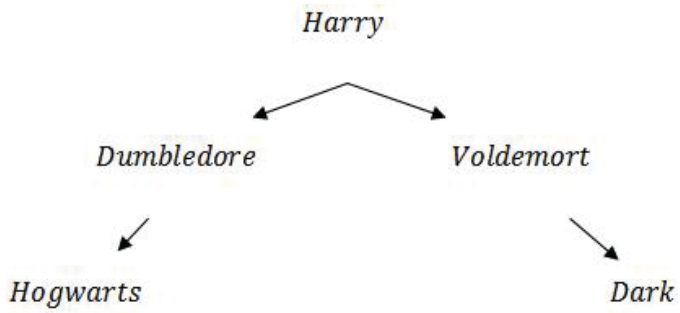


Figure 4: Topic Tree for Corpus

We get two topic sets from the above tree. Our main parent 'Harry' has two children and two grand children. Each pair of child – grandchild along with the parent is a topic set as follows:

1. $[Harry, Dumbledore, Hogwarts]$
2. $[Harry, Voldemort, Dark]$

Now we can arrange any document related to harry potter based on these two topic sets. Each set will return a specific set of documents related to the subjects in the sets. For example Document 3 will not be a search result for set 1 and Document 4 will not be returned when the search is concentrated towards set 2. On a more fantasy note, students at Hogwarts school of Witchcraft and Wizardry can have a more categorized search in the Hogwarts Library based whether their interest lies in Albus Dumbledore and Hogwarts or Lord Voldemort and the Dark arts.

V. EXPERIMENTAL EVALUATION

The evidence to prove the success of our enterprise, we designed and implemented a thorough application adhering to our System Architecture. The application was programmed using the objected oriented concepts where each of the important units of the system like Graph, Edges, vertices, terms and words were systematically organized as classes. Java was the OOTP language used for the purpose of demonstration with Eclipse Kepler being the programming tool.

We took assistance of certain predefined 'JARS' for the purpose of information and data handling. 'Lucene' [17] being the key framework for deriving quantities like Tf, Idf for terms. We also used existing classes of java to arrange and organize information that suited best to needs. The use of Hash Map, Array List, Hash Sets, stacks and queues is the best example for this.

Our experimental corpus consisted of primarily news articles in '.txt' format. The corpus was a host of documents ranging from 100 words to 2000 words. Our results constitute the findings of topics sets that exists for 10, 25, 50, 100, 150, 200 and 300 articles. We assigned six different test cases for every value of size of the corpus. Thus for seven values of 'N – size of the corpus' we has six test cases, thus we performed 42 distinct test for our application. For part one of the testing process we manually categorized the documents based on their subject matter. We tested our finding across each category. Part two of the process included testing some of the topic sets we obtained and using them as search queries for google. Test series one, provided a base to prove the existence of the topic template and test series two was helpful in quality assessment.

To observe the changes in the number of topic sets produced we set different threshold values for the following quantities: Tf, Tf * Idf, Cardinality of intersection of adjacency lists and the Pwt of edges. Some thresholds like corpus size and term frequency were effective for the entire application. Others like Tf*Idf were only stage specific. For setting and resetting every threshold value we found different sets of topic graphs.

Based on the different quantities we used as Thresholds, we had the following six test cases (for each Tf is the term frequency and Pwt is the probabilistic weight and N is the size of the corpus:

Test 1: Tf >= minimum, Pwt >= 3/N
Test 2: Tf >= minimum, Pwt >= 5/N
Test 3: Tf >= minimum, Pwt >= 7/N
Test 4: Tf >= (minimum + average)/2, Pwt >= 4/N
Test 5: Tf >= (minimum + average)/2, Pwt >= 6/N
Test 6: Tf >= average, Pwt >= 3/N

N	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6
10	3	2	0	5	0	2
25	7	7	4	7	4	4
50	14	11	7	10	7	11
100	36	21	14	21	11	13
150	40	26	25	44	22	29
200	64	41	30	48	34	37
300	78	56	36	67	45	60

Table 6: Number of distinct graphs/trees for every test subject

N	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6
10	31	6	0	12	0	1
25	76	24	2	47	23	5
50	126	45	26	80	33	12
100	244	96	53	171	38	26
150	377	165	107	207	65	63
200	497	210	128	269	128	93
300	585	262	165	372	190	135

Table 7: Number of Topic Sets for every Test Subjects

The tests presented the following observation:

1. The 'number of Topic Trees' and "number of Topic Sets' for each set increased with the increase in the number of documents. We set certain lower limits on the Term Frequencies and the probabilistic weights. This made the term graphs obtained in stage 3 was varying in size thus varying the number of sets and trees for every case.
2. An increase was observed in top to bottom manner. Which meant, the larger the value of N, more were the sets and trees produced? A decrease was observed going left to right when pwt threshold was increased. Test cases 1, 2 and 3 followed this pattern for Tf = minimum and Test cases 4 and 5 also followed suit for Tf = (minimum+average)/2. In most cases it was observed that for Tf >= average as a threshold, least number of trees and sets were produced.
3. Most of the cases produced multiple sets for each topic tree constructed. Our observations showed that 90% of the tests had trees to sets ratio greater than 1. Also the cases where the ratio was less than or equal to 1, did not have more than 25 documents in the corpus. Most had only 10. One thing to note here is that, size of the corpus does not directly affect the outcome as much as how many documents are of the same base subject. For example in a set of 10 documents if all documents

mutually exclusive of each other when it comes to similarity of topic, subject or genre, the set and trees produced are likely to be zero.

4. Topic sets of trees corresponding to different base subjects did not mingle with each other thus keeping the distinct nature of the results produced.

These sets of test cases helped in formulating the number of sets and trees that could be potentially produced by setting certain thresholds. Thus based on our observations we were able to conclude that the best results could be obtained by keeping the Term Frequency threshold to the lower minimum average [(minmum+average)/2] along with the pwt threshold somewhere between 4/N and 6/N. This particular threshold could be increased for a larger number of documents. But again, the fact that specificity of base subject is more important than the actual corpus size will actually decide the quality of results.

We tested and searched some of the results on www.google.com. Some of the sets we tested for were as follows:

[Samsungs S5 Galaxy Samsung HTC M8]
[Winter Soldier Marvel Captain Iron Man]
[Bahrain Williams Ferrari Alonso]
[Chelsea PSG Ibrahimovic]

For each of the sets above we observed that the results produced were query specific. The first had results concerning results that compared Samsung Galaxy S5 and HTC one M8. The second displayed results specific only to second Captain America movie and its relations to the marvel universe. The third had results concerning the F1 Grand Prix of Baharain with respect to Ferrari, Fernando Alonso and Williams F1 rather than actual results as a whole. The last set had results specific to the match between Chelsea and Paris Saint Germain rather than description of the two soccer clubs.

VI. RELATED WORD AND FUTURE PROSPECTS

Haphazardly arranged information is not information but just data of no importance. Information retrieval is an abundantly improving commodity today especially over the web. With the rise of social networking people expect more from the internet more than ever before. Ranking is another evolving aspect to organize information over the web. A set various standards exists to rank and index information. These theories are not limited to the internet but our aspects of our materialistic lives as well. Our introduction of topic graphs and topic sets can be used as complementary to both ranking and indexing techniques. Ranking and indexing documents by categorizing them on the basis of the topic sets that we would provide would enhance the information retrieval process. The same can be done to improvise the ranking and the indexing of social micro-blogging web-services like twitter and facebook. Visual information retrieval can be applied the same theory as well. Images, videos and other multimedia searches can divided based on the same concept of topic graphs. Any information that has some subject based classification associated with it can be ranked, indexed, categorized and classified basis of some sets of topic graphs.

Topic Maps: Topic Maps is a standard for the representation and interchange of knowledge, with an emphasis on the find ability of information. Topic maps were originally developed in the late 1990s as a way to represent back-of-the-book index structures so that multiple indexes from different sources could be merged. However, the developers quickly realized that with a little additional generalization, they could create a meta-model with potentially far wider application [16].

A topic map represents information using

- Topics, representing any concept, from people, countries, and organizations to software modules, individual files, and events,
- Associations, representing hyper graph relationships between topics, and
- Occurrences representing information resources relevant to a particular topic.

Topic Maps are similar to concept maps and mind maps in many respects, though only Topic Maps are ISO standards.

Ontology: In computer science and information science, an ontology formally represents knowledge as a hierarchy of concepts within a domain, using a shared vocabulary to denote the types, properties and interrelationships of those concepts [7][8].

Ontologies are the structural frameworks that are used in information organization. Their utilization ranges from various fields artificial intelligence, the Semantic Web, systems engineering, software engineering to biomedical informatics, library science, enterprise bookmarking, and information architecture. Ontologies can be used for knowledge representation about a host of topics or just a part of them. The creation of domain ontologies is also fundamental to the definition and use of an enterprise architecture framework [7] [8].

Substantial work has been performed on the translation of natural language questions to formal queries using ontology or a database [5] [3] [4] [6]. While these approaches have been shown to yield remarkable results, it is not clear if users always want to specify a full natural language question. In fact, the success of commercial search engines shows that users are quite comfortable with using keywords. Thus, it seems important to also develop approaches which are able to interpret keywords

Future aspects of Topic graphs can include an assortment of uses ranging from basic everyday uses to the realms of the World Wide Web. We wish to pursue the use of topic sets to develop a new indexing scheme for any web based search. Based on the parent child relationship of the words, we can present a ranking scheme for topic graphs. We can use our topic sets to urge a crawler to find more documents that subject to a particular topic sets. These in turn can be ranked and index with the terms of the graphs to provide a more categorically based ranking and help in better information retrieval.

VII. CONCLUSION

We present "Topic Graphs" and "Topic Sets", a probabilistic relationship based association words to cluster and categorize topics together into a hierarchical tree based format. This tree format can be further used to create topic sets which present templates of words that have high associations with each other. These sets contain words that have a high probability of appearing together across a number of documents in the world. These words thus are strongly related to each other.

To prove the existence of these topic sets we began by processing a certain corpus of documents. We filtered out unnecessary and unwanted words out to keeps the more common but popular and important words in each document. We then used graph theory to formulate relationships between these more popular terms. We treated every term as a vertex and the weighted edges between them defined the strength of their relationships. Based on this relationship we created topic graphs which essentially are topic trees. The parent child relationship between these topic graphs helped us to formulate the required topic sets.

We used a well defined and categorized corpus to test our theory. We created a through application in Java for every stage of the system to test our theory. The application yielded comprehensive topic graphs and topic sets. To prove the logical existence of the theory we also presented an on paper example for a small corpus of four documents each with an average of 120 words. We extracted results that provided the evidence for the nature of the thesis.

Towards the end we presented various applications along with related and future aspirations for our theory to grow further in the web and non technical world. We believe that such kind of categorizing will help precise and efficient searching of information.

REFERENCES

[1] Lewis, Chris. "Information Patterns." *Irresistible Apps*. Apress, 2014. 81-97.

[2] Weisberg, Jacob. "Bubble trouble: Is web personalization turning us into solipsistic twits." (2011).

[3] Lopez Lopez, Vanessa, Michele Pasin, and Enrico Motta. "Aqualog: An ontology-portable question answering system for the semantic web." *The Semantic Web: Research and Applications*. Springer Berlin Heidelberg, 2005. 546-562.

[4] Cimiano, Philipp, Peter Haase, and Jörg Heizmann. "Porting natural language interfaces between domains: an experimental user study with the orakel system." *Proceedings of the 12th international conference on Intelligent user interfaces*. ACM, 2007.

[5] Popescu, Ana-Maria, Oren Etzioni, and Henry Kautz. "Towards a theory of natural language interfaces to databases." *Proceedings of the 8th international conference on Intelligent user interfaces*. ACM, 2003.

[6] Bernstein, Abraham, and Esther Kaufmann. "GINO–a guided input natural language ontology editor." *The Semantic Web-ISWC 2006*. Springer Berlin Heidelberg, 2006. 144-157.

[7] Gruber, Thomas R. "A translation approach to portable ontology specifications." *Knowledge acquisition* 5.2 (1993): 199-220.

[8] Fensel, Dieter. *Ontologies*. Springer Berlin Heidelberg, 2001.

[9] Pariser, Eli. *The filter bubble: What the Internet is hiding from you*. Penguin UK, 2011.

[10] Zhang, Yuan Cao, et al. "Auralist: introducing serendipity into music recommendation." *Proceedings of the fifth ACM international conference on Web search and data mining*. ACM, 2012.

[11] Cormen, Thomas H., et al. *Introduction to algorithms*. Vol. 2. Cambridge: MIT press, 2001.

[12] Kolosovskiy, Maxim A. "Data structure for representing a graph: combination of linked list and hash table." *arXiv preprint arXiv:0908.3089* (2009).

[13] Gertz, Janet. "Selection for preservation in the digital age." *Library Resources & Technical Services* 44.2 (2000): 97-104.

[14] Rajaraman, Anand, and Jeffrey David Ullman. Data Mining. Cambridge University Press, 2011.

[15] Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*. Vol. 1. Cambridge: Cambridge university press, 2008.

[16] Pepper, Steve, and Graham Moore. "XML topic maps (XTM) 1.0." *TopicMaps. Org Specification xtm1-20010806* (2001).

[17] Gao, Rujia, et al. "Application of Full Text Search Engine Based on Lucene." *Advances in Internet of Things* 2 (2012): 106.

[18] Chen, Chun, et al. "Ti: an efficient indexing mechanism for real-time search on tweets." *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*. ACM, 2011.

[19] Kules, Bill, Jack Kustanowitz, and Ben Shneiderman. "Categorizing web search results into meaningful and stable categories using fast-feature techniques." *Digital Libraries, 2006. JCDL'06. Proceedings of the 6th*

ACM/IEEE-CS Joint Conference on. IEEE, 2006.

[20] White, Ryen W., Bill Kules, and Steven M. Drucker. "Supporting exploratory search, introduction, special issue, communications of the ACM." *Communications of the ACM* 49.4 (2006): 36-39.

[21] Rose, Daniel E., and Danny Levinson. "Understanding user goals in web search." *Proceedings of the 13th international conference on World Wide Web*. ACM, 2004.

[22] Baeza-Yates, Ricardo, Carlos Hurtado, and Marcelo Mendoza. "Query recommendation using query logs in search engines." *Current Trends in Database Technology-EDBT 2004 Workshops*. Springer Berlin Heidelberg, 2005.

[23] Pratt, Wanda. "Dynamic organization of search results using the UMLS." *Proceedings of the AMIA Annual Fall Symposium*. American Medical Informatics Association, 1997.

[24] Landauer, Thomas, et al. "Enhancing the usability of text through computer delivery and formative evaluation: the SuperBook project." *Hypertext: A psychological perspective* (1993): 71-136.

[25] Maarek, Yoelle S., et al. "WebCutter: a system for dynamic and tailorable site mapping." *Computer networks and ISDN systems* 29.8 (1997): 1269-1279.

[26] Zamir, Oren, and Oren Etzioni. "Grouper: a dynamic clustering interface to Web search results." *Computer Networks* 31.11 (1999): 1361-1374.

[27] Zamir, Oren, and Oren Etzioni. "Web document clustering: A feasibility demonstration." *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1998.

[28] Hearst, Marti A., and Jan O. Pedersen. "Reexamining the cluster hypothesis: scatter/gather on retrieval results." *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1996.

[29] Hearst, M., J. Pedersen, and D. Karger. "Scatter/gather as a tool for the analysis of retrieval results." *Working notes of the AAAI fall symposium on AI applications in knowledge navigation*. 1995.

Polymorphic Worms Detection Using Longest Common Substring

Mohssen M. Z. E. Mohammed, Eisa Aleisa
College of Computer and Information Sciences
Al-Imam Muhammad Ibn Saud Islamic University,
Riyadh, Saudi Arabia
m_zin44@hotmail.com; aleisa@ccis.imamu.edu.sa

Neco Ventura
Dept. of Electrical Engineering, University of Cape Town
Rondebosch, South Africa
neco@crg.ee.uct.ac.za

Abstract— Polymorphic worms are considered as the most dangerous threats to the Internet security, and the danger lies in changing their payloads in every infection attempt to avoid the security systems. We have designed a novel double-honeynet system, which is able to detect new worms that have not been seen before. To generate signatures for polymorphic worms we have two steps. The first step is the polymorphic worms sample collection which is done by a Double-honeynet system. The second step is the signature generation for the collected samples which is done by using Longest common substring algorithm. The system is able to generate accurate signatures for single and multiple worms.

Keywords-honeynet; worms; String matching.

1. Introduction

An Internet worm is a self-propagated program that automatically replicates itself to vulnerable systems and spreads across the Internet. Worms take the attack process one step further by self-replicating. Once a worm has compromised and taken over a system, it begins scanning again, looking for new victims. Therefore a single infected system can compromise one hundred systems, each of which can compromise another one hundred more systems, and so on. The worm continues to attack systems this way and grows exponentially. This propagation method can spread extremely fast, giving administrators little time to react and ravaging entire organizations. Although only a small percentage of individuals can identify and develop code for worms, but once the code of a worm is accessible on the Internet, anyone can apply it. The very randomness of these tools is what makes them so dangerous. A polymorphic worm is a worm that changes its appearance with every instance [1].

It has been shown that multiple invariant substrings must often be present in all variants of worm payload. These substrings typically correspond to protocol framing, return addresses, and in some cases, poorly obfuscated code [8].

Intrusion detection systems serve three essential security functions: they monitor, detect, and respond to unauthorized activities. There are two basic types of intrusion detection: host-based and network-based. Host-based IDSs examine data held on individual computers that serve as hosts, while network-based IDSs examine data exchanged between computers [15, 16].

Our research is based on Honeypot technique. Developed in recent years, honeypot is a monitored system on the Internet serving the purpose of attracting and trapping attackers who attempt to penetrate the protected servers on a network. Honeypots fall into two categories. A high-interaction honeypot such as (Honeynet) operates a real operating system and one or multiple applications. A low-interaction honeypot such as (Honyed) simulates one or multiple real systems. In general, any network activities observed at honeypots are considered suspicious [1, 9].

This paper is organized as follows: Section 2 approximate string matching algorithms. Section 3 discusses the related work regarding automated signature generation systems. Section 4 introduces the proposed system architecture to address the problems faced by current automated signature systems. Signature generation algorithm for Polymorphic Worm will be discussed in section 5. Section 6 concludes the paper.

2. Approximate String Matching Algorithms

In this section, we give an overview on the approximate string matching algorithms.

2.1 Preliminaries

The problem of string matching is very simply stated in [17]. Given a body of text T[1…n], we try to find a pattern P[1…m] where $m \leq n$. This can be used to search bodies of texts for specific patterns, or say for example in biology, can be used to search strands of DNA (Deoxyribonucleic acid) for

specific sequences of genes. The issue of exact string matching has been extensively worked on. However, approximate string matching is a much more complicated problem to solve which has many more real world applications. The truth is that in real world applications, the issue is not so systematic. This is where approximate string matching is needed. Instead of searching for the exact string, approximate string matching searches for patterns that are close to P. In other words, approximate string matching allows for a certain amount of errors between the two strings being compared. One of the earliest applications of approximate string matching was in text searching. The approximate string matching algorithms can be applied to account for errors in typing. Internet searching is particularly difficult because there is so much information and much of it has errors in it. Also, since the Internet spans many different languages, errors frequently arise in comparing words across language barriers. Also, text editors have to use approximate string matching when performing spell checks. Additionally, spell checkers have to generate a list of "*suggested words*" that are close in spelling to the misspelled word. Exact string matching is efficient to generate signatures for polymorphic worms.

2.2 Dynamic Programming

Approximate string matching algorithms use Dynamic programming method. In mathematics and computer science, dynamic programming is a method for solving complex problems by breaking them down into simpler subproblems [17]. It is applicable to problems exhibiting the properties of overlapping subproblems, which are only slightly smaller and optimal substructure (which is described below). When applicable, the method takes far less time than naive methods.

The key idea behind dynamic programming is quite simple. In general, to solve a given problem, we need to solve different parts of the problem (subproblems), then combine the solutions of the subproblems to reach an overall solution. Often, many of these subproblems are really the same. The dynamic programming approach seeks to solve each subproblem only once, thus reducing the number of computations. This is especially useful when the number of repeating subproblems grows exponentially as a function of the size of the input.

Top-down dynamic programming simply means storing the results of certain calculations, which are later used again since the completed calculation is a sub-problem of a larger calculation. Bottom-up dynamic programming involves formulating a complex calculation as a recursive series of simpler calculations.

2.3 History of dynamic programming

Richard Bellman first used the term 'dynamic programming' in the 1940s for describing the process of solving problems where one needs to find the best decisions one after another. By 1953, he refined this to the modern meaning, referring specifically to nesting smaller decision problems inside larger decisions, and the field was thereafter recognized by the IEEE (Institute of Electrical and Electronics Engineers) as a systems analysis and engineering topic. Bellman's contribution is remembered in the name of the Bellman equation which is a central result of dynamic programming that restates an optimization problem in recursive form.

The term dynamic was chosen by Bellman to capture the time-varying aspect of the problems. The word programming referred to the use of the method to find an optimal program, in the sense of a military schedule for training or logistics.

2.4 Overview of dynamic programming

Dynamic programming is both a mathematical optimization method and a computer programming method. In both contexts, it refers to simplifying a complicated problem by breaking it down into simpler subproblems in a recursive manner. While some decision problems cannot be taken apart this way, decisions that span several points in time do often break apart recursively; Bellman called this the "Principle of Optimality". Likewise, in computer science, a problem that can be broken down recursively is said to have optimal substructure.

If subproblems can be nested recursively inside larger problems, so that dynamic programming methods are applicable, then there is a relation between the value of the larger problem and the values of the subproblems. In the optimization literature, this relationship is called the Bellman equation.

2.5 Dynamic Programming in Mathematical Optimization

When we talk about mathematical optimization, dynamic programming usually refers to simplifying a decision by breaking it down into a sequence of decision steps over time. This is done by defining a sequence of value functions V_1, V_2, ..., V_n, with an argument y representing the state of the system at times i from 1 to n.. The definition of $V_n(y)$ is the value obtained in state y at the last time n. The values V_i at earlier times i times $i = n-1, n-2, ..., 2, 1$ can be found by working backwards, using a recursive relationship called the Bellman equation. For $i = 2, ..., n$, V_{i-1} at any state y is calculated from V_i by maximizing a simple function (usually the sum) of the gain from decision $i - 1$ and the function V_i at the new state of the system if this decision is made. Since V_i has already been calculated for the needed states, the above operation yields V_{i-1} for those states. Finally, V1 at the initial state of the system is the value of the optimal solution. The optimal values of the decision variables can be recovered one-by-one by tracking back the calculations that are already performed.

3. Related Work

Honeypots are an excellent source of data for intrusion and attack analysis. Levin et al. described how honeypot extracts details of worm exploits that can be analyzed to generate detection signatures [4]. The signatures are generated manually.

One of the first systems proposed was Honeycomb developed by Kreibich and Crowcroft. Honeycomb generates signatures from traffic observed at a honeypot via its

implementation as a Honeyd [5] plugin. The longest common substring (LCS) algorithm, which looks for the longest shared byte sequences across pairs of connections, is at the heart of Honeycomb. Honeycomb generates signatures consisting of a single, contiguous substring of a worm's payload to match all worm instances. These signatures, however, fail to match all polymorphic worm instances with low false positives and low false negatives.

Kim and Karp [6] described the Autograph system for automated generation of signatures to detect worms. Unlike Honeycomb, Autograph's inputs are packet traces from a DMZ that includes benign traffic. Content blocks that match "enough" suspicious flows are used as input to COPP, an algorithm based on Rabin fingerprints that searches for repeated byte sequences by partitioning the payload into content blocks. Similar to Honeycomb, Auto-graph generates signatures consisting of a single, contiguous substring of a worm's payload to match all worm instances. These signatures, unfortunately, fail to match all polymorphic worm instances with low false positives and low false negatives.

S. Singh, C. Estan, G. Varghese, and S. Savage [7] described the Earlybird system for generating signatures to detect worms. This system measures packet-content prevalence at a single monitoring point such as a network DMZ. By counting the number of distinct sources and destinations associated with strings that repeat often in the payload, Earlybird distinguishes benign repetitions from epidemic content. Earlybird, also like Honeycomb and Autograph, generates signatures consisting of a single, contiguous substring of a worm's payload to match all worm instances. These signatures, however, fail to match all polymorphic worm instances with low false positives and low false negatives.

New content-based systems like Polygraph, Hamsa and LISABETH [8, 10 and 11] have been deployed. All these systems, similar to our system, generate automated signatures for polymorphic worms based on the following fact: there are multiple invariant substrings that must often be present in all variants of polymorphic worm payloads even if the payload changes in every infection. All these systems capture the packet payloads from a router, so in the worst case, these systems may find multiple polymorphic worms but each of them exploits a different vulnerability from each other. So, in this case, it may be difficult for the above systems to find invariant contents shared between these polymorphic worms because they exploit different vulnerabilities. The attacker sends one instance of a polymorphic worm to a network, and this worm in every infection automatically attempts to change its payload to generate other instances. So, if we need to capture all polymorphic worm instances, we need to give a polymorphic worm chance to interact with hosts without affecting their performance. So, we propose new detection method "Double-honeynet" to interact with polymorphic worms and collect all their instances. The proposed method makes it possible to capture all worm instances and then forward these instances to the Signature Generator which generates signatures, using a particular algorithm.

An Automated Signature-Based Approach against Polymorphic Internet Worms by Yong Tang and Shigang Chen[9] described a system to detect new worms and generate signatures automatically. This system implemented a double-honeypots (inbound honeypot and outbound honeypot) to capture worms payloads. The inbound honeypot is implemented as a high-interaction honeypot, whereas the outbound honeypot is implemented as a low-interaction honeypot. This system has limitation. The outbound honeypot is not able to make outbound connections because it is implemented as low-interaction honeypot which is not able to capture all polymorphic worm instances. Our system overcomes this disadvantage by using double-honeynet (high-interaction honeypot), which enables us to make unlimited outbound connections between them, so we can capture all polymorphic worm instances.

4. Double- Honeynet System

We propose a double-honeynet system to detect new worms automatically. A key contribution of this system is the ability to distinguish worm activities from normal activities without the involvement of experts.

Figure 2 shows the main components of the double-honeybet system. Firstly, the incoming traffic goes through the Gate Translator which samples the unwanted inbound connections and redirects the samples connections to Ho-neynet 1.

The gate translator is configured with publicly-accessible addresses, which represent wanted services. Connections made to other addresses are considered unwanted and redirected to Honeynet 1 by the Gate Translator.

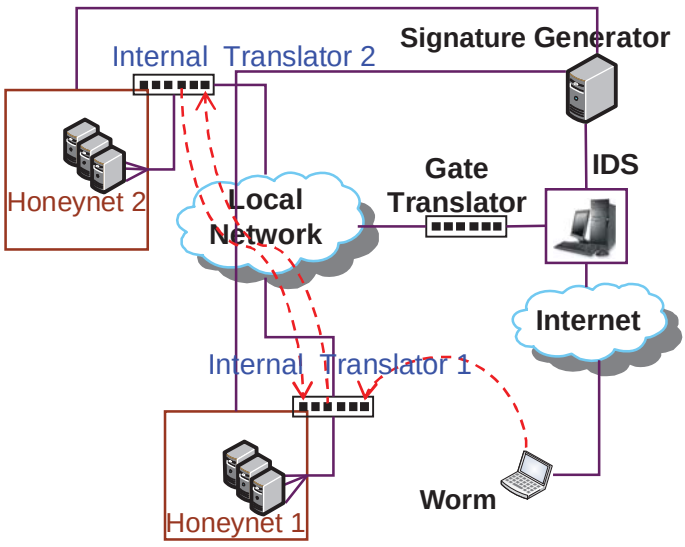


Figure 1. System architecture.

Secondly, Once Honeynet 1 is compromised, the worm will attempt to make outbound connections. Each honeynet is associated with an Internal Translator implemented in router that separates the honeynet from the rest of the network. The Internal Translator 1 intercepts all outbound connections from

honeynet 1 and redirects them to honeynet 2 which does the same forming a loop.

Only packets that make outbound connections are considered malicious, and hence the Double-honeynet forwards only packets that make outbound connections. This policy is due to the fact that benign users do not try to make outbound connections if they are faced with non-existing addresses.

Lastly, when enough instances of worm payloads are collected by Honeynet 1 and Honeynet 2, they are forwarded to the Siganture Generator component which generates signatures automatically using specific algorithms that will be discussed in the next section. Afterwards, the Siganture Generator component updates the IDS database automatically by using a module that converts the signatures into Bro or pseudo-Snort format. The above proposed system implemented by using Vmware Server 2. The implementation results are out of the scope of this paper.

For further details on the double-honeynet architecture the reader is advised to refer to our published works [13].

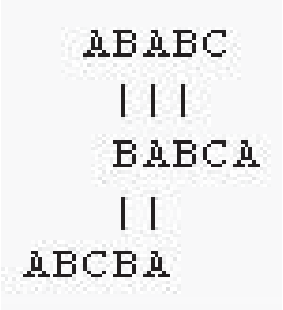
5. Signature Generation Algorithms

In this section, we describe the Longest common substring algorithm which we use it to generates signatures for polymorphic worms.

The longest common substring problem is to find the longest string (or strings) that is a substring (or are substrings) of two or more strings [17].

Example:

The longest common substring of the strings "ABABC", "BABCA" and "ABCBA" is string "ABC" of length 3. Other common substrings are "AB", "BC" and "BA".



Problem definition

Given two strings, *S* of length *m* and *T* of length *n*, find the longest strings which are substrings of both *S* and *T*.

A generalisation is the **k-common substring problem**. Given the set of strings =

$$\{S_1, \ldots, S_K\}, \text{ where } |S_i| = n_i \text{ and } \sum n_i = N.$$

Find for each $2 \leq k \leq k$, the longest strings which occur as substrings of at least k strings.

6. Conclusion

We have proposed automated detection for Zero day polymorphic worms using double-honeynet. We have proposed new detection method "Double-honeynet" to detect new worms that have not been seen before. The system is based on the Longest common substring algorithm that used to generate signatures for polymorphic worms. The main objectives of this research are to reduce false alarm rates and generate high quality signatures for polymorphic worms.

7. References

[1] L. Spitzner, "Honeypots: Tracking Hackers," Addison Wesley Pearson Education: Boston, 2002.
[2] Hossein Bidgoli, "Handbook of Information Security," John Wiley & Sons, Inc., Hoboken, New Jersey.
[3] D. Gusfield, "Algorithms on Strings, Trees and Sequences,", Cambridge University Press: Cambridge, 1997.
[4] J. Levine, R. La Bella, H. Owen, D. Contis ,and B. Culver, "The use of honeynets to detect exploited systems across large enterprise networks," Proc. of 2003 IEEE Workshops on Information Assurance, New York, Jun. 2003, pp. 92- 99.
[5] C. Kreibich and J. Crowcroft, "Honeycomb–creating intrusion detection signatures using honeypots," Workshop on Hot Topics in Networks (Hotnets-II), Cambridge, Massachusetts, Nov. 2003.
[6] H.-A. Kim and B. Karp, "Autograph: Toward automated, distributed worm signature detection," Proc. of 13 USENIX Security Symposium, San Diego, CA, Aug., 2004.
[7] S. Singh, C. Estan, G. Varghese, and S. Savage, "Automated worm fingerprinting," Proc. Of the 6th conference on Symposium on Operating Systems Design and Implementation (OSDI), Dec. 2004.
[8] James Newsome, Brad Karp, and Dawn Song," Polygraph: Automatically generating signatures for polymorphic worms," Proc. of the 2005 IEEE Symposium on Security and Privacy, pp. 226 – 241, May 2005.
[9] Yong Tang, Shigang Chen," An Automated Signature-Based Approach against Polymorphic Internet Worms," IEEE Transaction on Parallel and Distributed Systems, pp. 879-892 July 2007.
[10] Zhichun Li, Manan Sanghi, Yan Chen, Ming-Yang Kao and Brian Chavez. Hamsa, "Fast Signature Generation for Zero-day Polymorphic Worms with Provable Attack Resilience," Proc. of the IEEE Symposium on Security and Privacy, Oakland, CA, May 2006.
[11] Lorenzo Cavallaro, Andrea Lanzi, Luca Mayer, and Mattia Monga, "LISABETH: Automated Content-Based Signature Generator for Zero-day Polymorphic Worms," Proc. of the fourth international workshop on Software engineering for secure systems, Leipzig, Germany, May 2008.

[12] J. Nazario. “Defense and Detection Strategies against Internet Worms “. Artech House Publishers (October 2003).
[13] Mohssen M. Z. E. Mohammed, H. Anthony Chan, Neco Ventura. “Honeycyber: Automated signature generation for zero-day polymorphic worms“; Proc. of the IEEE Military Communications Conference, MILCOM, 2008.
[14] C. C. Aggarwal and P. S. Yu, " Outliner Detection for High Dimensional Data," Proceedings of the ACM SIGMOD Conference, Santa Barbara, CA, May 21-24, 2001.
[15] Snort – The de facto Standard for Intrusion Detection/Prevention. Available: http://www.snort.org, 1 March 2012.
[16] Bro Intrusion Detection System. Available: http://www.bro-ids.org/, 5 March 2012.
[17] Mohssen Mohammed, Sakib Pathan " Automatic Deference Against Zero-day Polymorphic worms in Communication networks", CRC press, USA.

[18] Haykin, Simon, " Neural Networks: A Comprehensive Foundation (2 ed.),". Prentice Hall. ISBN 0132733501.

[19] MacQueen, J. B. (1967), "Some Methods for classification and Analysis of Multivariate Observations," Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press. pp. 281–297.

[20] Keerthi, S. & Gilbert, E. (2002). Convergence of a Generalized SMO Algorithm for SVM Classifier Design. Machine Learning 46: 351–360.

[21] Genton, M. (2001). Classes of Kernels for Machine Learning: A Statistics Perspective. Journal of Machine Learning Research 2: 299-312.

[22] Hodge, V., Austin, J. (2004), A Survey of Outlier Detection Methodologies, Artificial Intelligence Review, Volume 22, Issue 2, pp. 85-126.

SESSION

COMMUNICATION AND NETWORKS + WEB SERVICES

Chair(s)

TBA

Local Trend Detection from Network Traffic Using a Topic Model and Network Router

Kenichi Takagiwa* and Hiroaki Nishi*
*Department of Science and Technology, Keio University, Japan
takagiwa@west.sd.keio.ac.jp, west@sd.keio.ac.jp

Abstract*—In this study, a novel trend detection application from network traffic is proposed, which can find a trend in the a specific region. In this application, a special router is used to capture the packet stream directly. Trend detection is based on a topic model, latent Dirichlet allocation (LDA). This model considers clustering between different web pages and counts the appearance frequency of web browsing history, data that can only be captured by the network router. In order to achieve effective clustering, we propose to categorize relevant URLs and non-relevant URLs from network traffic by using Deep Packet Inspection (DPI). Identifying relevant URLs is a key component for analyzing the trend of web pages because a web request contains a variety of non-relevant URLs such as advertisements, CSS, and JavaScript. This classification is accomplished by using IP address and HTTP headers. To provide this service, we use HTTP rather than HTTPS to perform DPI. Hence we also evaluate HTTPS usage to prove the effectiveness of DPI for current communication. The results of our evaluation shows a 17.4% usage of HTTPS and that HTTPS streams contribute only 1.2% to WIDE traffic, one of Japan's major ISPs. Our proposed application is evaluated using real traffic. The evaluation results prove that our proposed method can detect both major and minor trends in real world traffic.

***Keywords*—trend detection, topic model, network traffic analysis, Service-oriented Router**

I. INTRODUCTION

Understanding customer trends has become important, especially for companies to understand the preferences of users in order to gain profits. User preferences are becoming more diversified as the Internet evolves. However, the network is still able to represent the trend of user behavior. For example, users explore the Internet to learn about items that they are interested in buying. Users spend increasingly more time browsing websites on the Internet. The purchase behavior of users may be influenced during netsurfing.

Trend detection technology has been introduced and can be used to find valuable trends in data. In particular, many companies use this technology to analyze data, and they can use this mined data to improve their business results. Trend detection technology enables companies to understand the trends of the market, the potential interests of customers, and customer feedback. From the perspective of users, trend detection is also important because their interests influence the products made by the company, which improves their satisfaction. This technology has received much attention due to the dynamic nature of the market. Companies plan agile marketing strategies based on dynamic user demand.

In recent years, microblogs such as Twitter, Facebook, and Google+ have emerged as a popular social media, and these share a variety of information, ranging from personal life to the latest news. As microblogging services gain popularity, many trend detection applications for microblogging have appeared. A common method to detect trends is counting hash tags in a microblog. The popular hash tags can be regarded as user trends. Some tweets also contain geo-location information and trends in specific data can be identified. Companies can measure trends based on hash tag popularity. Another method detects trends from popular search engines such as Google and Bing. Trends are estimated by frequent search queries. Popular search queries are regarded as a trend. The geo-location information of the query can be estimated by using the IP, allowing local trends to be estimated.

The methods mentioned above used for gathering information for trend detection have some limitations. Microblogging services cannot obtain data from users who do not use these services. Geo-location information cannot be collected from all users depending on their configurations. If this data is gathered for trend detection, the accuracy of trend detections could be improved because the data coverage would be increased. Because user preferences are diversified, simply counting hash tags and search queries is not enough data to accurately detect trends. Trend detection should focus on the context of user behavior more intensively by applying other methods.

In this study, we propose a new trend detection system that applies the features of the Service-oriented Router (SoR) [1]. The SoR is a router that not only routes packets, but also collects the payloads of the packets. The SoR can capture all of the information transferred over the Internet. When packets pass through the SoR, it analyzes the packets, searches packet payload and stores the matched contents of the packets, the timestamp, and the IP address.

The data captured by the SoR includes the information about who did what, when they did it, and where there they did it. This information can be used for providing high-quality services to Internet users. Because routers are geographically distributed, as shown in Fig. 1, our system can detect the local trends for each area. Specifically, an edge router knows local communication because all communications go through edge routers, as shown in Fig. 2. When local information is gathered

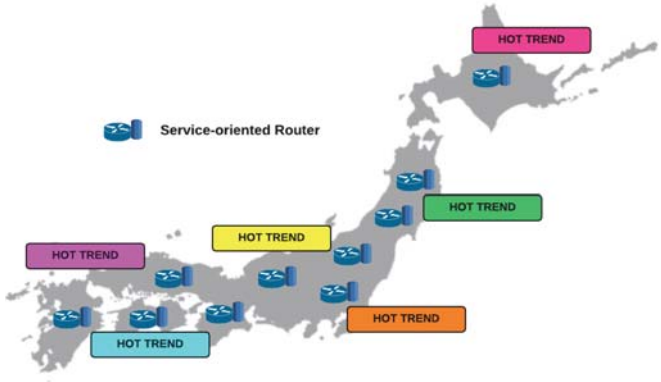


Fig. 1. Trend detection by SoR

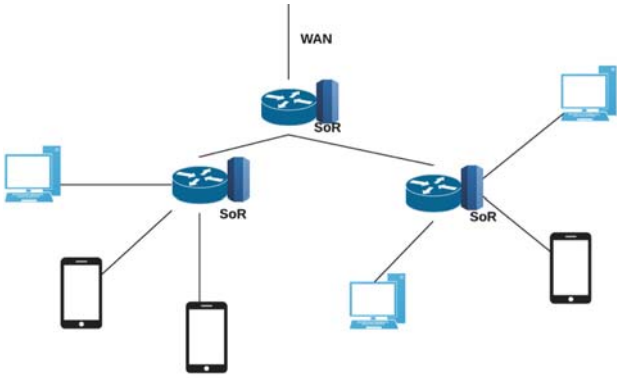


Fig. 2. Local communication thorugh SoR

in the edge router, a local area trend can be detected. This architecture has advantages in terms of computation and location detection, such as fog computing [13].

SoRs can be distributed at the edge of network as computing resources. A SoR can ensure where users are located because user access concentrates at the nearest access point. This location information can be obtained only because SoRs are used as network routers. Although there are alternative methods for gathering location information and web usage, these methods have limitations in information gathering.

A SoR can measure any kind of webpage regardless of access logs and type of services by DPI, except HTTPS. The HTTPS ratio to HTTP has been increased according to the penetration of HTTP/2 [9]. The evaluation of HTTPS is necessary to prove the effectiveness of DPI. To improve the accuracy of trend detection, our proposed application uses network traffic, which contains location information and user communication history.

II. Related Work

As related work concerning DPI up to layer 7, Masuda et al. propose a webpage recommendation application that draws its recommendations from network traffic [2]. This research proposes a method to extract individual web browsing histories from Internet traffic. This method utilizes the viewing time of a webpage as an index for the recommendation. The recommendation for webpages is estimated by a collaborate filter weighted by browsing time. This research analyzes traffic payload, and the URL and browsing time are used to make the recommendation. This method does not analyze the entire content of webpages.

In web data mining, various methods have been proposed for a clustering method of web-content mining [4]. Most of the existing web mining algorithms concentrate on finding frequent patterns, while neglecting less frequent ones in the domain.

Latent Dirichlet allocation (LDA) was originally developed by Blei et al. to find topics in collections of documents [14]. However, LDA has been applied to many different fields other than natural language processing. Cramer et al. propose network analytics that detect significant co-occurrences in the type of network traffic by using time-varying LDA [3]. They show that LDA can be applied to describe the status of a network. They find co-occurrence in user behavior by analyzing well-known ports with LDA. Distinguished topics, such as web traffic, email clients and instant messaging, Microsoft file access, and email servers, are detected from network traffic. Chu et al. propose a method to find web service orchestration patterns from sparse web service logs by using a biterm topic model (BTM) in conjunction with LDA [5]. Web orchestration is an interaction between the internal and external communication in a web service. Trends in user behavior can be estimated from web orchestration by using topic models. This method demonstrates that topic models can be used to detect patterns from sparse web information. Noor et al. propose DWSemClust in order to cluster deep web databases. This approach uses LDA to model content that is representative of deep web databases [6]. Because topics on the web have sparse distribution, an unsupervised probabilistic model is a suitable solution to the problem of clustering. This research demonstrates that LDA can successfully assign a probability for clustering sparse web content. A topic model is applied to analyze usage on the network and the web. However, a topic model for DPI up to L7 has not been explored.

Remotely related to our research, is the work on trend detection in microblogging. Lau et al. propose topic modeling based on the online trend analysis of Twitter [7]. The discovered trends give insight to popular culture and events by analyzing the short messages using an online LDA. This work does not consider location information.

III. Layer 7 Analyzer: NEGI

As described in the Related Work section, most of the previous work does not use multi-domain information. There is a possibility to improve trend detection by using the information extracted from cross-domain content. Hence, a special middleware is required to achieve the extraction of web content and usage from cross-domain.

We used NEGI, a middleware for capturing and reconstructing a packet stream, designed in C/C++ [8]. An overview of NEGI is shown in Fig. 3. NEGI uses libpcap, a network traffic capture library. NEGI can directly monitor the Ethernet port on a server and can provide real-time analysis. More specifically, the router can obtain all of the

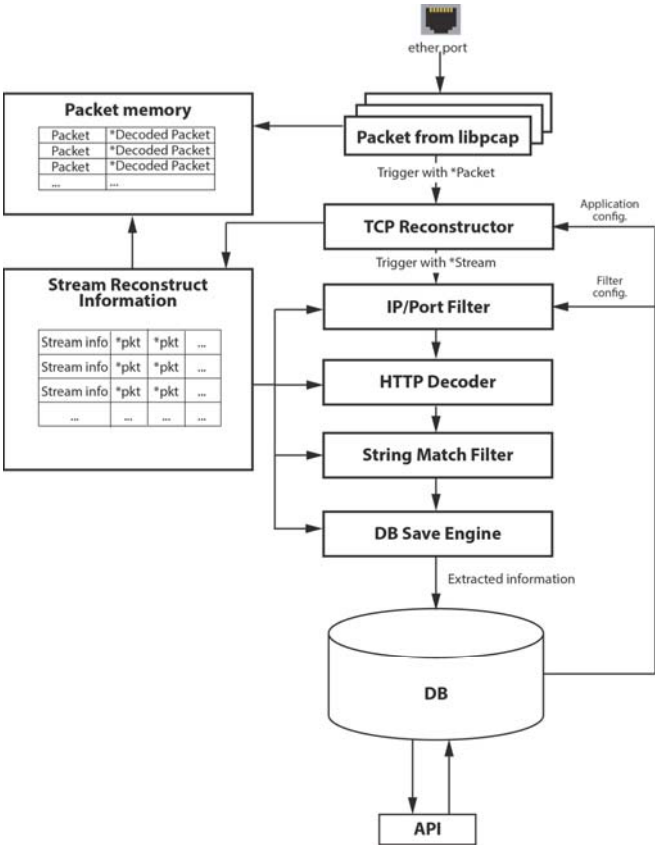


Fig. 3. Architecture of a SoR

communication information. SoR requires reconstructing the packet stream in order to obtain the contents from each TCP/IP stream. This reconstruction enables a router to provide cross-domain information to a trend detection application because the contents are captured in the router, and information in the router is independent of specific servers. Packets are captured from the Ethernet port. The IP/Port Filter shown in Fig. 3 distinguishes relevant and non-relevant packets in order to determine whether the packets should be processed further. This decision is based on five tuples: source IP address, destination IP address, source port, destination port, and the protocol of each packet.

Many streams exist in the network and they arrive at a router in a perfectly mixed state, with some being unordered. Therefore, packets that belong to a specified stream do not always arrive at a router sequentially. NEGI can reconstruct a stream from mixed packets by applying a context switch [12].

As a result of the reconstruction process of the TCP stream, packets with HTTP 1.1 protocol are decoded in the L7 decoder module, as shown in Fig. 3. An L7 decoder can decode contentious gzip and chunked encoding packet-by-packet. After HTTP 1.1 decoding, the packets are passed to the next extraction process. In this process, the relevant part of the stream is extracted using the string match filter shown in Fig. 3, in accordance with the user's query request. If the content matches the user's query, the matching content is stored in an on-memory database. A user can configure the size of the data

Table 1. Results of the URL Filter

Variable	Description
$\beta_{1:K}$	Topic where each β_k is a distribution over the vocabulary
θ_d	Topic proportions for document *d*
$\theta_{d:k}$	Topic proportion for topic *k* in document *d*
z_d	Topic assignments for document *d*
$z_{d,n}$	Topic assignment for word *n* in document *d*
w_d	Observed words for documents *d*
α	Parameter of the respective Dirichlet distributions
η	Parameter of the respective Dirichlet distributions

extraction. Stored data is periodically flushed to an external storage device in order to keep memory requirements low in the SoR. This insertion process to the database is accomplished without waiting the end of the entire reconstruction of a TCP stream in the DB save engine. When the last packet in a stream has been processed, or the stream is incomplete after a user defined timeout, the dedicated memory entry for the packet stream is discarded. Our proposed trend application utilizes the database constructed by NEGI.

NEGI is currently available on an ALAXALA router by using a service module card. It is also available on a Juniper router using the JunosV App Engine.

IV. TOPIC MODEL

A topic model is a statistical model useful for discovering abstract topics that occur in a collection of documents. A topic model is used for trend detection in this study. Many types of topic models have been proposed. LDA [14] is a probabilistic topic model and is, therefore, suitable for trend detection from sparse web data collected from web traffic.

LDA and its derivations have been shown to effectively identify topics from a collection of documents. LDA assumes that a document is made up of a mixture of topics, where a topic is defined as a multinomial distribution over words in the vocabulary. The generative process of LDA is as follows:

1. Draw βk ~ Dir(η), for k = 1, 2, …, K
2. For document *d*, where d = 1, 2, …, D:
   a. Draw θ_d ~ Dir(α)
   b. For token *i* in document *d*, where i = 1,2, …, N_d:
      i. Draw $z_{d,i}$ ~ Multi(θ_d)
      ii. Draw $w_{d,i}$ ~ Multi($\beta_{zd,i}$)

The generative model for LDA is shown in Equation 1. The parameters used are described in Table 1.

$$p(\beta_{1:K},\theta_{1:D},z_{1:D},w_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \prod_{n=1}^{N} p(z_{d,n}|\theta_d)p(w_{d,n}|\beta_{1:K},z_{d,n}) \quad (1)$$

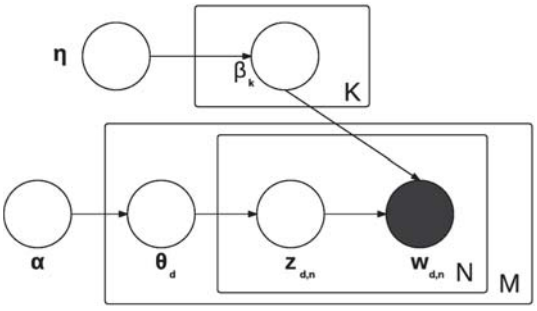


Fig. 4. The graphical representation of LDA

The graphical model for LDA, shown in Fig. 4, defines the joint distribution of random variables.

V. TREND DETECTION APPLICATION

In this section, we describe our proposed trend detection application. It employs three processing steps as shown in Fig. 5 The first step is the TCP reconstruction and the extraction of network streams using NEGI. The second step is the extraction of URLs. The final step is detecting trends from traffic data.

A. *Application design*

The goal of our application is to provide a quick summary of the trends found from network traffic. This problem is close to the problem to find a topic from streamed HTML documents.

We implemented a trend detection application based on the network-stream capture engine in a router. The network traffic captured in the router was processed by NEGI, software for analyzing TCP streams. NEGI is described in detail in Section 3.

B. *URL extraction*

Our trend detection application requires the information about the URLs that people look at. The way in which we extract URLs from traffic is described in [2] as follows. The search condition is "GET," "Host:", "Referer," and "<title". HTTP headers containing the GET request are extracted from network traffic to identify URLs. URLs are reconstructed from each "GET" and "Host:" in the HTTP header. The IP address is used to identify users. For example, in the case where "Host:" is "www.google.com" and "GET" is "/," the obtained URL is "www.google.com/."

Although many HTTP requests are issued to display a web page such as images, scripts, and advertisements, the relevant URL is the first URL among the HTTP requests. This is because the first URL contains text information that a user primarily looks at. In this study, we call the first URL, "Base URL," and the rest of them, "Subordination URL." Filters are implemented to categorize Base and Subordination URLs.

1) *REFERER filter*

The REFERER filter distinguishes Base URLs and Subordination URLs by the fact that the REFERER is different for Base and Subordination URLs. The REFERER of a Base URL is same as the URL that user looked at previously or it is empty. On the other hand, the referrer of a

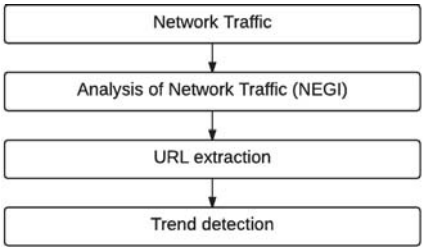


Fig. 5. Trend detection application flow

Subordination URL tends to be same as the Base URL. Therefore, if the REFERER is the same as the URL that user viewed before, or is empty, the URL is classified as Base URL, and if it is the same as the Base URL, the URL is classified as a Subordination URL.

2) *Timestamp filter*

A browser downloads HTML by requesting the Base URL. After the HTML is parsed, the browser downloads images and scripts from the Subordination URLs. This means that the difference of the timestamp between Subordination URLs is shorter than between the Base and a Subordination URL. If timestamps are smaller than the threshold, then these URLs are Subordination URL.

3) *Response filter*

An HTTP response corresponding to a Base URL contains the "<title>" tag because the Base URL is in HTML format. URLs of HTTP responses without a title tag in the payload are categorized as Subordination URLs.

4) *Content-Type filter*

The Content-Type in the HTTP header should be "text/html" because the Base URL is in HTML. If the Content-Type of the HTTP response is not "text/html," then the URL is categorized as a Subordination URL.

5) *URL filter*

A URL filter is a blacklist filter that filters by searching for specific strings in a URL. If a file extension is js, css, mp3, etc., then these URLs are Subordination URL. A static directory like "assets" is also categorized as a Subordination URL.

6) *Title filter*

If some specific string appear in the "<title>" tag, such as "403" or "500", these URL are Subordination URL. This filter is mainly used for detecting error pages.

7) *Integrated Filter*

This filter integrates all of the filters mentioned above. The REFERER and timestamp filter are applied after the other filters because they use before-and-after URLs. When the response filter and timestamp filter are applied, if the HTTP response status is 300, the following processes occur. The filter checks "Location:" that describes the URL and the response filter is omitted if the HTTP header has "Location:".

C. *Trend detection*

Trend detection problem is approximated by what are popular topics in HTML documents captured from traffic. The probabilistic topic model is suitable for trend detection from traffic because webpages that a variety of people look at are

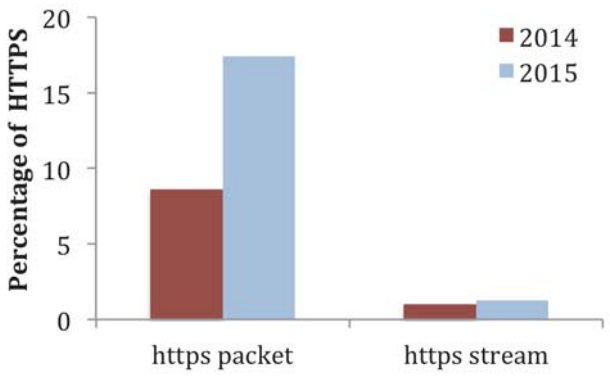


Fig. 6. HTTPS percentage

sparse [4-6]. Unsupervised learning methods can handle such sparse data. Chu et al. compare BTM and LDA and conclude that BTM and LDA are suitable for clustering web data. BTM has an advantage for shorter text [5], but is has the disadvantage of having to use the whole webpage. LDA is sufficient for detecting trends from HTML. For scalability, an online LDA [15] is used for topic estimation because an online algorithm is suitable for processing network streams.

The content of each webpage is obtained from the computation server by requesting the URLs. In order to increase throughput and to keep memory usage low, the SoR focuses on extracted URLs. A particular number of topics are estimated based on a given topic number and batch duration using the online LDA. Before the topics are estimated, the HTML tags, CSS, and stop words are removed from the obtained HTML. Documents generated from the HTML are converted into a bag-of-words model to represent a document. Unlike English, Japanese requires morphological analysis to extract a word from the text. All of the titles in Japanese Wikipedia are used to estimate each word with high accuracy.

VI. EXPERIMENT AND EVALUATION

A. HTTP to HTTPS ratio

The HTTP to HTTPS communication ratio was measured by using WIDE traffic. HTTPS communication cannot be analyzed by NEGI due to the encrypted end-to-end communication. HTTP-HTTPS communication rates provide insight on the impact of our DPI application. The standardization of HTTP/2 was approved in February 2015 and the ratio of HTTPS has been increasing [9]. In HTTP/2, most of the communication is done over SSL/TLS. In this case, it is difficult to operate DPI from communication in the middle of the server and client due to the encryption of the communication. However, it is possible to intervene by using a trusted proxy, bringing back the benefits of value-added services, as described in the IETF proposal [10].

Each dataset is a daily trace, representing 15 minutes of traffic, captured from "sample point F" from a trans-Pacific link between Japan and the United States. The data is publicly available. Packet payloads are omitted and IP addresses are anonymized. [11] Traffic from January 1st to March 28th in 2014 and January 1st to March 28th in 2015 were used for HTTP-HTTPS communication ratio analysis. Traffic with destination port numbers 80 and 443 are regarded as HTTP and HTTPS, respectively. The number of packets and streams are compared in this analysis. The results are shown in Fig. 6. The number of HTTPS packets has increased by 6.7%, reaching a total of 17.4% in 2015. The number of HTTPS streams has increased by 0.2%, reaching a total of 1.2% in 2015. However, HTTPS still does not comprise the majority of web traffic. According to [9], Facebook and YouTube account for majority of the percentage increase in HTTPS traffic. Therefore, only HTTP communication is sufficient to detect trends from traffic.

In our evaluation, we used the stored data set from the network stream captured in our laboratory to evaluate the dataset.

sTable 2. Results of the URL Filter

Threshold[s]	TP	TN	FP	FN	Precision	Recall
0.5	30	2085	2	0	0.94	1.0
0.6	29	2085	2	1	0.94	0.97
0.7	29	2085	2	1	0.97	0.97
0.8	29	2087	0	1	0.97	1.0
0.9	29	2087	0	1	0.97	1.0
1.0	29	2087	0	1	0.97	1.0

B. URL Filter

The integrated filter described in Section 5 was evaluated by traffic captured in our lab from October 7 – 15, 2013. Thirty Base URLs were chosen randomly. Browsing was reproduced by navigating to these URLs from a browser. This browsing was expected to produce 30 Base URLs and 2087 Subordination URLs. Precision and recall were evaluated using the formulas shown in Equations 2 and 3. True positive (TP) is the number of Base URLs successfully categorized as such. False positive (FP) is the number of Subordination URLs wrongly categorized as Base URLs. False negative (FN) is the number of Base URLs wrongly categorized as Subordination URLs.

$$precision = \frac{TP}{TP + FP} \quad (2)$$

$$recall = \frac{TP}{TP + FN} \quad (3)$$

The results of this evaluation are shown in Table 2. Precision and recall were calculated for every timestamp threshold from 0.5 to 1.0 s, in 0.1 s intervals. For values greater than or equal to 0.6 s, FN becomes 1 because its timestamp is below the threshold, and this Base URL is categorized as a Subordination URL. We used 0.8 seconds as the threshold for our experiment because no FP results occur within this threshold.

C. Trend detection

The data set is traffic captured in our lab from October 22, 2013 – January 22, 2014. We extracted 32,262 relevant URLs from this data set by using the URL filters mentioned in

Section 5. There were 21,158 unique URLs. There are 22 people in our lab, including 19 Japanese, 2 Sri Lankans, and 1 Indonesian. Only Japanese and English webpages are used for trend detection. Webpages in other languages are removed. The number of topics to be found by the online LDA was set to 100. The results shown in Table 3 are a mix of Japanese and English. In order to improve readability, Japanese words are translated into English. Some words are 2-gram, which is originally a unigram in Japanese. Data results are divided by week to allow weekly trends to be detected.

As shown in Table 3, the trend detection application discovered many interesting topics. The "words" column in Table 3 lists representative words for each topic. The "probabilities" column indicates the probability that those words would be assigned to a topic. From October 20 – 26, Apple product was grouped. This was caused by the release of the iPad Air 2. From November 24 – 30, a topic relating to the Comet ISON was detected. Comet ISON was popular because it was passing by the Sun. In the week from December 15 – 21, a "daily" topic was discovered. This is related to the fact that Japanese people tend to buy a new daily for the next year at the end of the current year. From January 5 – 11, Puzzle & Dragons, one of the most popular smartphone game applications in Japan, was found. The iPad Air 2 and Puzzle & Dragons were very popular in our lab at that time. Our proposed trend detection application can discover major trends. On the other hand, our method can also detect minor trends. This is illustrated by the fact that the comet ISON was popular with only two of the lab members. Other topics also represent the characters of various lab members. Therefore, both minor and major trends are detected from the traffic using LDA.

VII. CONCLUSIONS

In this paper, we proposed a novel trend detection application that used a topic model and is based on traffic through the Internet router. This method can distinguish relevant URLs from non-relevant ones by applying filters. In order to provide this service, HTTP should be used. Though the volume of HTTPS traffic has been increasing, our evaluation showed that the HTTPS ratio to HTTP is still low. HTTPS usage is at 17.4% and HTTPS streams contribute only 1.2% to WIDE traffic, one of Japan's major ISPs. Our application was able to detect major and minor trends that represent real-world trends from traffic. Our work has advantages in terms of scalability and load distribution because routers are geographically distributed. This research helps consider the importance of value-added service in the HTTP/2 communication.

ACKNOWLEDGEMENTS

This work was partially sponsored by the SECOM Science and Technology Foundation and by MEXT/JSPS KAKENHI Grant (B) Numbers 24360230 and 25280033.

REFERENCES

[1] K. Inoue, D. Akashi, M. Koibuchi, and H. Nishi. "Semantic router using data stream to enrich services", in 3rd International Conference on Future Internet (CFI), pp. 20-23, June 2008.

[2] Kazuki Masuda, Shinichi Ishida, and Hiroaki Nishi., "Cross-site recommendation application based on the viewing time and contents of webpages captured by a Network Router", in ICOMP, Las Vegas, 2013.

[3] C. Cramer and L. Carin, "Bayesian topic models for describing computer network behaviors", in Acoustics, Speech, and Signal Processing (ICASSP), 2011 IEEE International Conference on, pp. 1888–1891, May 2011.

[4] B. Singh, and Singh H. K., "Web data mining research: a survey", Computational Intelligence and Computing Research (ICCIC), 2010 IEEE International Conference on, pp.1-10, 28-29, Dec. 2010

[5] V. W. Chu, R. K. Wong, Chi Hung Chi, P. C. K. Hung, "Web service orchestration topic mining", Web Services (ICWS), 2014 IEEE International Conference on, pp. 225, 232, June 27, 2014 – July 2, 2014

[6] U. Noor, A. Daud, A. Manzoor, "Latent dirichlet allocation based semantic clustering of heterogeneous deep web sources", Intelligent Networking and Collaborative Systems (INCoS), 2013 5th International Conference on, vol., no., pp. 132, 138, 9 – 11 Sept. 2013

[7] J. H. Lau, N. Collier, and T. Baldwin, "On-line trend analysis with topic models: #twitter trends detection topic model online", in COLING, pp. 1519–1534, 2012.

[8] NEGI https://github.com/westlab/negi

[9] D. Naylor et al. "The cost of the S in HTTPS", Proceedings of the 10th ACM International on Conference on Emerging Networking Experiments and Technologies. ACM, 2014.

[10] HTTPBis Working Group. Explicit Trusted Proxy in HTTP/2.0. http://goo.gl/Pbt8G5, February 2014.

[11] K. Cho, K. Mitsuya, and A. Kato. "Traffic data repository at the WIDE project", in USENIX 2000 Annual Technical Conference: FREENIX Track, pp. 263–270, June 2000.

[12] S. Ishida, S. Harashima, M. Koibuchi, H. Kawashima, and H. Nishi. "A memory-efficient method of TCP reconstruction using context switch on internet router", in 190th computer ARChitecture (ARC), August 2010.

[13] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli. "Fog computing and its role in the internet of things", in Workshop on Mobile Cloud Computing, MCC '12, 2012.

[14] D. M. Blei, A. Y. Ng., and M. I. Jordan. Latent dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022. 2003

[15] M. D. Hoffman, D. M. Blei, and F. Bach, "Online learning for latent dirichlet allocation", in Neural Information Processing Systems, 2010.

Table 3. Results of the URL filter

Oct/20 – Oct/26		Oct/27 – Nov/2		Nov/3 – Nov/9		Nov/10 – Nov/16		Nov/17 – Nov/23	
words	probabilities	words	probabilities	words	probabilities	words	probabilities	words	probabilities
ipad	0.010	arduino	0.011	english	0.009	student	0.010	twitter	0.020
apple	0.004	serial	0.008	time	0.006	international	0.006	original	0.014
st	0.004	print	0.007	weblio	0.006	program	0.005	miku hatsune	0.012
air	0.004	others	0.006	englishjapanese	0.005	scholarship	0.005	movie	0.011

Nov/24 – Nov/30		Dec/1 – Dec/7		Dec/8 – Dec/14		Dec/15 – Dec/21		Dec/22 – Dec/28	
words	probabilities	words	probabilities	words	probabilities	words	probabilities	words	probabilities
comet ison	0.016	pururazi	0.013	shield	0.016	daily	0.114	job hunting	0.010
comet	0.012	voice actor	0.012	per month	0.006	plan	0.019	question	0.009
yahoo	0.007	blazblue	0.011	arduino	0.005	size	0.009	information	0.008
nasa	0.007	every hour	0.011	xb	0.005	page	0.009	me	0.008

Dec/29 – Jan/4		Jan/5 – Jan/11		Jan/12 – Jan/18	
words	probabilities	words	probabilities	words	probabilities
privacy	0.034	puzzle dragon	0.057	co2	0.008
ieee	0.016	important	0.034	lang	0.005
data	0.014	collaboration	0.018	display	0.005
public	0.011	advent	0.017	page	0.005

A new TF-IDF Formulation for Web Service Business Protocol Discovery

Abdelkader Moudjari[1], **Salim Chikhi**[2], **and Amer Draa**[2]
[1 2 3]MISC Laboratoory, Department of Fundamental Informatics, and its Applications,
Abdelhamid Mehri University, Algeria
[1]moudjariabdelkader@gmail.com, [2]slchikhi@yahoo.com , [3]draa.amer@gmail.com

Abstract— *The formula of the IT-IDF term relevance measure is adapted in this paper to discover business protocols of web services from the history of their execution. A graph-base representation of the BP is adopted, the nodes represent the states of the system and the edges represent the messages exchanged between the client and the BP in question. This new formula considers the total number of edges that constitute the graph and the relative presence of each edge with regard to the clients that used it. The proposed formulation is used in a probabilistic framework to decide about the probable occurrence of edges in the graph representing the BP being discovered. The proposed approach has been validated using graphs of different degrees of complexity. The obtained results prove the efficiency of this new formula.*

Keywords: TF-IDF, Business protocol discovery, Information retrieval , Log files.

1. Introduction

Adapting web services for new needs, extensions and constraints necessitates more than the static description of the functionality of it, this functionality is generally offered by the WSDL language [1]. In fact, the behaviour of a web service is not always offered by its designer. Fortunately, the trace of execution of a given web service can bee observed in the log file recording the sequence of message exchange between the web service and its clients. However, the log files corresponding to web services generally contain the history of many conversations overlapped. This makes it difficult to discover the correct sequences of messages presenting the different conversations, the sum of these conversations is the business protocol in question[2], [3]. One could think that a simple solution is just using an identifier distinguishing between clients, and so distinguishing the conversations one from the other [2], [3], [4]. However, this solution is quite impossible in real-world applications for the following two reasons. First, many classical web services may have been the result of migrating classical application for which the behaviour description is unavailable. Second, it may be the choice of the designer of the web service not to give any identifying information [4]. Consequently, new BP discovery approaches appeared in recent years. The latter, mainly exploit statistical techniques to give an approached description of the BP being discovered [5], [6], [7].

In this line of thought, this paper uses a famous information retrieval technique, the TF-IDF for business protocol discovery from corresponding log file lacking any information related to the ID of the conversations. The TF-IDF measure offers an estimation resulting from the combination of the relevance of a document with regard to the term being looked for, TF (for Term Frequency), and its relevance within a collection of documents, IDF (for Inverse Document Frequency).

An adaptation of the TF-IDF [8] for the context of BP discovery from log files is needed. To do so, we consider the graphical representation of the web service whose BP is being discovered, in which the edges represent the messages exchanged between the client and the BP in question. In addition, instead of considering the importance of documents, as the basic measure does, we focus on the importance of the edges, to distinguish the most probable edges from the less ones, generally presenting noise.

The rest of this paper is organized as follows. In Section 2, previous works related to business protocols discovery are presented. Section 3 summarizes the basic concepts linked to business protocol discovery and the TF-IDF. The proposed approach is presented in Section 4. Section 5 discusses the obtained experimental results. The paper is concluded in Section 6.

2. Related works

The static behaviour of a web service is generally described through providing the set of operations offered by the web service. The WSDL [1] is known for its power for describing this static aspect. On the other hand, the dynamic behaviour of any web service cannot be known just from its WSDL description; the order in which the web service operations are invoked is not offered. For this reason, business protocols are used; they describe the ordered sequences of operations, called *conversations*, of the web service [2], [3].

As far as BP discovery is concerned, two main categories of approaches exist. In the first class of approaches, the discovery process uses a conversation identifier (CID) which

must be present in the log file, while the second class does not exploit the CID, the latter does not exist in the log files used in this case. It is clear that the approaches based on the use of conversation ID are far to be implemented in some real life contexts. In the context of BP discovery using CID, we cite the work of Benatallah *et al.* [2], where the authors initiated this research field. Later on, the general scheme of discovering business protocols, also dealing with imperfect log files, has been proposed in [3]. The notion of episode has been first exploited in [4].

In the second set of approaches, those not using CIDs, we cite [5], [9], [7]. In [5], the authors exploited the idea of using some attributes as correlating tools and composing attributes to look for correlated messages belonging to the same conversation. This correlation, is generally expressed in the form of correlation rules using logical operators such as the AND and the OR operators. Decomposing the log files into sub-logs according to the sender and receiver of messages and exploiting graph theory to represent the business protocol are the main ideas of the work in [9]. In [7], the authors present a new approach based on linear algebra and linear regression for conversation protocol discovery. They have also solved the case of implicit states (temporized transitions). Recently, the work of Moudjari *et al.* [10] solved the problem of BP discovery using latent semantic analysis exploiting a micro/macro relationship between sets of messages. The approach presented in this paper belongs to the second category. As will be shown in the following sections, it does not need the CID to be recorded in the log files.

An information retrieval system typically searches in collections of unstructured or semi-structured data (e.g. web pages, documents, images, video, etc.) [11]. The authors in [12] gave a classification of different models of information retrieval. These models are: boolean models [13], [14] , probabilistic models [15], [16] and vectors space models [8]. The TF-IDF technique used in this paper is one of the techniques belonging to the latter category of models.

3. Basic Concepts

In this section, the concepts used in this work are defined, namely the notions related to BP discovery and TF-IDF measure.

3.1 Business Protocol Discovery

The definition presented in the following subsections are mainly extracted from [3], [5].

3.1.1 Message

Messages are units of information exchange between clients and servers in web service communication. A message is composed of a set of attributes. Messages are recorded in a specific file called the *log file*, which contains many entries representing each a message. Messages contain many attributes. The attribute we are interested in have been chosen to be: the type of message *Msgtype*, the sender *Snd*, the receiver *Rec* and the time *T* of the occurrence of the event, and so the time of its recording in the log file. An example of time would be 2014/02/22/23/12/12 for expressing: the date, hour, minute and second respectively.

3.1.2 Conversation

A conversation $C = (M_1, M_2, \ldots, M_n)$ is a sequence of messages exchanged between a web service and a client in the purpose of fulfilling a given goal.

3.1.3 Business Protocol

A business protocol is the specification of *all possible conversations* between a web service and its partners [3], [5]. Formally speaking, a BP is defined by the tuple $P = (S, S_0, F, M, T)$; where: S is a finite set of states, S_0 is the initial state, M is the set of messages and T is a set of transitions.

3.1.4 Log files

A log file is a text file containing the events of an application the programmer wants to record. In the present work, the events to be recorded are all the sending and reception events sent or received by a web service.

3.2 TF-IDF Measure

In the context of information retrieval, the relevance (or importance) of a given document with regard to a specific term is characterised by two aspects: the relevance of the document with regard to the term being looked for(TF for Term Frequency) and its relevance within a collection of documents (IDF for Inverse Document Frequency), measured too with regard to the same term. These two indicators have been summed in on measure called the TF-IDF [17], [8].

3.2.1 Term Frequency

The TF (Term-Frequency) is a measure of the importance of a given document with respect to a specific term (being looked for). There are several formulas to calculate this term weight [8], [18], [19]. The most common one is *normalisation* or simply *term frequency*, Equation (1).

$$TF = \frac{Number\,of\,the\,term\,occurrences}{Total\,number\,of\,words\,in\,the\,document} \quad (1)$$

3.2.2 Inverse Document Frequency

The Inverse Document Frequency (IDF) [8], [17], Formula 2, permits reducing the importance of terms in documents; if a term occurs in many documents, it is likely to be less important in the whole request [8], [13], [19]. Equation (2) shows how this term is calculated, where: Nbr_all_doc is the number of all documents in the collection and

Nbr_term_doc is the number of documents containing the term being looked for.

$$IDF = log(\frac{Nbr\_all\_doc}{Nbr\_term\_doc}) \quad (2)$$

4. The Proposed Approach

In this paper, a new algorithm for discovering business protocols of web services is proposed. It is based on the construction of graphs representing these protocols. In this section, details the phases constituting our approach.

The key idea of our contribution is to consider pairs of message instead of single separate messages. Pairs of messages represent messages and their successors in the sub-logs *i.e.*; they represent the edges of the graph corresponding to the business protocol to be discovered. Using the TF-IDF measure seen above, we will look for the most important edges in the log files, where log files recording communications done by all clients over a period of time are exploited.

4.1 An Adaptation of the TF-IDF

The TF-IDF formula is adapted to accomplish the new objective. The new IDF term, called the *Couple Client Frequency* (CCF) here, is given in 3. It is worth to mention, that instead of considering documents, we consider messages in our work.

In the equation, $N\_clients$ presents the total number of clients that interacted with the web service. $Pres[i]$ is the amount of presence of the edge i (couple of successive messages) with respect to all clients. The Log function is used to brake the resulting measure, to make the value of the IDF term as influential as the TF term. In simple words, to allow both terms (TF term and IDF term) to act equally on the final measure.

On the other hand the TF term is called in the following CF (for *Couple Frequency*), presented in Equation (4), where $occ_{i,j}$ is the number of times Edge *i* was chosen by client *j*. It is to be noted that the principle remains the same: an edge (a couple of messages) is considered to be present in the graph if its TF-IDF measure (computed with the new formula) is greater than a given threshold, permitting eliminating noise and incompleteness in the log file.

$$CCF[i] = Log(\frac{N\_Clients - Pres[i] + (\frac{Pres[i]}{N\_Clients})}{N\_Clients}) \quad (3)$$

$$CF_{ij} = \frac{occ_{ij}}{\sum_{i=1}^{m} occ_{ij}} \quad (4)$$

Phase 1, Log File Partitioning: In this phase, the log file describing the web service to be discovered is partitioned to 'sub-logs' with regard to the senders and receivers of messages (per clients). Figure 2 illustrate this step, the log file on the top is partitioned into the sub-logs in the bottom part.

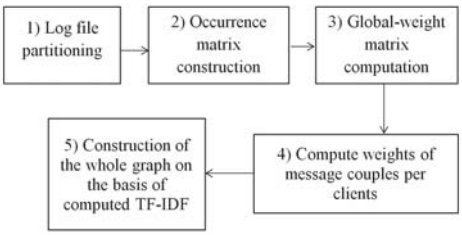


Fig. 1: Steps of the proposed approach

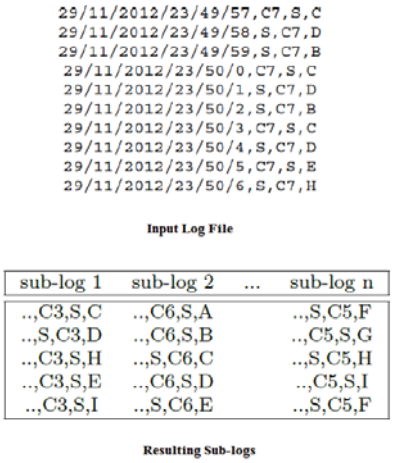


Fig. 2: Partitioning of the log file

Phase 2, Occurrence matrix construction: In this step, the occurrence matrix is built. It has as rows the different pairs of messages and as columns the clients. Pairs of messages are constructed from the sub-logs by considering the message and its successor in the form:*(m_1, m_2)*.

Phase 3, Global-weight matrix computation: After having created the occurrence matrix, the weight matrix is constructed. This is done through two stages; one is focusing on the local weight and the other on the global weight. Local weight is calculated according to Formula (4). It provides the importance of each edge in the sub-log.

On the other hand, the global weight relationship between couples and clients is expressed as a vector that represents the respective weights of edges with respect to different clients. This vector is called the *Couple Client Frequency* (CCF). Its elements are calculated as given in Formula (3).

Once the two weights are calculated, the whole weight matrix is built according to Equation (5). This matrix gives the weight of each edge taking by considering its importance through a local and global view vision perspective.

$$W_{ij} = CF_{ij} * CCF[i] \quad (5)$$

Phase 4, Computing weights of message couples per client: In this step, we calculate the sum of weights of

each pair of messages with all clients using Formula (6). This vector indicates the importance of each edge compared to all other edges and its importance with regard to all clients. Edges having high weight values are actually the edges invoked by most clients.

$$Pi = \sum_{j=1}^{n} Wij \quad (6)$$

Phase 5, Construction of the whole graph by thresholding: To build the whole graph that represents the business protocol taken as input, the weight vector computed in the previous step is used.

Two types of threshold are used and compared in our experiments. The firs threshold is chosen to be the mean of the final vector of weights (Equation 6). The second threshold is the result of the division of the variance of these weight by the number of clients.

The process starts with computing the threshold in question (mean-based/variance-based) to be used for accepting or rejecting the inclusion of some edge in the graph. The threshold is used as follows. The value of the threshold is calculated from the weight vector and edges with weights above this threshold are accepted, but those below it are not directly rejected. This process (calculating a threshold value and thresholding weights) is repeated till no weight is above the threshold. Hence, many threshold values may exist, which allows discovering low frequent edges; less frequently occurring edges would be discovered, since they significantly differ from noise with regard to a local threshold.

After this, initial and final states are discovered. The edges connecting final to initial states are then deleted. The process accomplishing these two tasks is the following. In typical cases, recording log files starts just with the beginning of communication; *i.e.* the first message to be recorded is the initial state. The final state is then not hard to be deduced. If a client executes a BP only once, the same policy used in the case of initial states is adopted: a final state is a state that occurs in final position in the majority of sub-logs. In the case of having a BP executed many times by the same client, we will have a graph in which every message has a successor. In this case, the final state is determined by looking for the immediate predecessor of the initial state.

5. Experimental results

To validate the proposed formula, and so the whole approach, the assessment method presented in Figure 3 used in many works in the literature [20], [9], [10] is adopted in our work.

Some numerical results are then presented to illustrate the efficiency of our approach. Mainly, we start by generating a synthetic log file corresponding to a predetermined (test) web service, then its graphical representation is used. As a quality measure, we assess the degree of similarity between the input and output graphs in terms of the well-discovered edges, the missing edges and the wrongly added edges. Formula (7) summarizes this metric, where: Nbr_edges is the number of edges in the input graph, $missing$ is the number of missing edged in the output graph and $incorrect$ is the number of incorrectly added edges.

$$Q = \frac{Nbr\_edges - (missing + incorrect)}{Nbr\_edges} \quad (7)$$

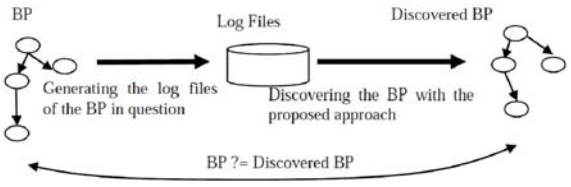


Fig. 3: Validation scheme

5.1 Synthetically Generated Graphs

As mentioned above, we applied our approach on some web services of different degrees of complexity. In the following, we introduce the results obtained from applying the proposed approach ten independent times on the web services presented by the graphs given in Figure 4. Tables 1, 2 and 3 present the numerical results obtained from applying the proposed approach on the graphs (a), (b) and (c) of Figure 4, respectively.

Table 1: Results of application on Graph (a)

Size of the log file	Missing / Added edges	Nbre of iterations TH_1	Missing / Added edges	Nbre of iteration TH_2	Perf by TH_1	Perf by TH_2
1209	4/0	2	1/0	1	0.82	0.95
5755	3/0	3	2/0	1	0.86	0.91
11449	4/0	3	2/0	1	0.82	0.91
22106	2/0	3	2/0	1	0.91	0.91
22449	2/0	2	1/0	1	0.91	0.95
73561	0/0	3	2/0	1	1.00	0.91
117112	4/0	2	2/0	1	0.82	0.91
202107	0/3	4	2/0	1	0.86	0.91
206477	0/3	3	1/0	1	0.86	0.95
246148	0/6	4	2/0	1	0.73	0.91
Averages performance					0.86	0.92

Table 2: Results of application on Graph (b)

Size of the log file	Missing / Added edges	Nbre of tries TH_1	Missing / Added edges	Nbre of tries TH_2	Perf at TH_1	Perf at TH_2
2897	7/0	2	2/0	1	0.84	0.95
4190	7/0	2	4/0	1	0.84	0.91
12812	8/0	2	3/0	1	0.81	0.93
21893	7/0	3	3/0	1	0.84	0.93
27199	5/0	2	3/0	1	0.88	0.93
41665	3/0	3	3/0	1	0.93	0.93
127765	3/0	3	3/0	1	0.93	0.93
198246	3/0	3	3/0	1	0.93	0.93
296113	3/0	3	3/0	1	0.93	0.93
342391	0/11	5	3/0	1	0.74	0.93
Averages performance					0.87	0.93

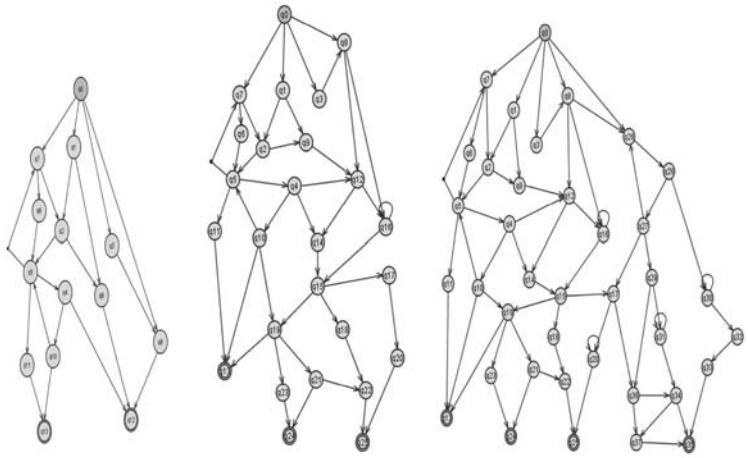


Fig. 4: The graphs used for validation: (a) simple, (b) medium and (c) complex

Table 3: Results of application on Graph (c)

Size of the log file	Missing / Added edges	Nbre of tries TH_1	Missing / Added edges	Nbre of tries TH_2	Perf at TH_1	Perf at TH_2
4229	12/0	4	1/0	1	0.82	0.98
13293	9/0	3	2/0	1	0.86	0.97
27428	8/0	3	3/0	1	0.88	0.95
42807	9/0	3	2/0	1	0.86	0.97
54342	9/0	3	3/0	1	0.86	0.95
140902	5/0	3	3/0	1	0.92	0.95
198246	3/0	3	3/0	1	0.95	0.95
244468	3/0	3	2/0	1	0.95	0.97
319392	0/8	4	2/0	1	0.88	0.97
357559	0 /0	4	0/0	1	1.00	1.00
Averages performance					0.90	0.97

As seen from the tables, the proposed approach succeeds in discovering the most of edges iof the business protocols in question. The system gave an average performance (accuracy) of 0.86 using Threshold 1 and 0.92 using Threshold 2 in the case of Graph (a), which has 22 edges, 0.87 using Threshold 1 and 0.93 using Threshold 2 in the case of Graph (b), having 43 edges, and 0.9 with Threshold 1 and 0.97 with the use of Threshold 2 in the case of Graph (c), with 66 edges, respectively. This proves the efficiency of our algorithm. Hence, scalability does not matter for discovering business protocols with the proposed method.

Another plus of the proposed approach is its tolerance to noise, the used log files contained a noise generated in a random manner. As seen from the results, our algorithm could resist noise which is likely to occur in genuine log files.

In addition, the obtained performances show the superiority of the second threshold, the variance-based compared to the one, it gave better performances than the mean-based one. It is also worth to mention that the last value of this threshold is critical; very small values would not prevent noise from causing the appearance of wrongly added edges, while high values would cause less frequently used edges to disappear, which leads to missing edges.

5.2 A Real-world Case Study

In this section, a graph representing a real-word web service is used to validate our approach. It is extracted from the works in [10], [21]. As used above, in the graph, the vertices represent the messages exchanged between the web service ans its clients. The edges represents the succession that exists between these messages.

A log file is generated starting from this graph. It contains the sequence of events described above. Table 4 shows the obtained results obtained from applying our algorithm (ten times) for BP discovery on this graph, of course with generating ten different log files. In the table, different sizes of the log file are considered. It is clear, from the table that proposed approach succeeded in discovering the BP of the web service at hand. which proves again the applicability of the proposed approach on real-world applications.

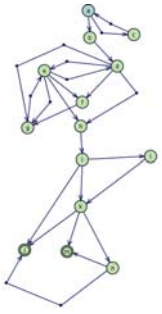


Fig. 5: The real-world web service graph used for further validation

Table 4: Results of application on a real-world web service graph

		(Missing /Added) edges		Performance using	
Size of the log file	Total edges	TH_1	TH_2	TH_1	TH_2
4735	24	3	0	0.88	1.00
47747	24	3	0	0.88	1.00
100744	24	0	0	1.00	1.00
161023	24	0	0	1.00	1.00
237207	24	6	0	0.75	1.00
278055	24	5	0	0.79	1.00
314744	24	5	0	0.79	1.00
346777	24	6	0	0.75	1.00
352141	24	4	0	0.83	1.00
363224	24	6	0	0.75	1.00
Averages performance				0.84	1.00

6. Conclusion

This paper proposed and used a new formulation of the TF-IDF metric for business protocol discovery of web services. The new formula considers the total number of edges that constitute the graph and the relative presence of each edge with regard to the clients that used it. The proposed formulation has been used in a probabilistic framework to decide about the probable occurrence of edges in the graph representing the BP being discovered. The proposed approach has been validated using graphs of different degrees of complexity, in addition to a real world web service graph. The obtained results have proven the efficiency of this new formula. Finally, two ways of thresholding the edge existence probabilities are used, the first is a mean-based threshold, the second is a variance-based one. The variance-based threshold gave better results.

As future perspective of research we want also to apply this business protocol discovery on a bigger systems of a real life application and use other information retrieval metrics to evaluate the importance of edges in log files. In addition, an advanced study of the thresholds efficiency is planned.

References

[1] C. Roberto, M. Jean-Jacques, and A. Ryman, *Web Services Description Language (WSDL) Version 2.0 World Wide Web Consortium (W3C)*, 2002. http://www.w3.org/TR/wsdl12.

[2] B. Benatallah, F. Casati, and F. Toumani, "Analysis and management of web service protocols," in *Conceptual Modeling–ER 2004*, pp. 524–541, Springer, 2004.

[3] B. Benatallah and H. R. Motahari-Nezhad, "Servicemosaic project: Modeling, analysis and management of web services interactions," in *Third Asia-Pacific Conference on Conceptual Modelling (APCCM2006)* (M. Stumptner, S. Hartmann, and Y. Kiyoki, eds.), vol. 53 of *CRPIT*, (Hobart, Australia), pp. 7–9, ACS, 2006.

[4] D. Devaurs, F. De Marchi, and M. S. Hacid, "Caractérisation des transitions temporisées dans les logs de conversation de services web," *Revue des Nouvelles Technologies de l'Information*, vol. E, no. 9, pp. 45–56, 2007.

[5] H. R. Motahari-Nezhad, R. Saint-Paul, B. Benatallah, and F. Casati, "Protocol discovery from imperfect service interaction logs," in *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, pp. 1405–1409, IEEE, 2007.

[6] B. Serrour, H. Kheddouci, *et al.*, "Une méthode à base de graphes pour la corrélation de messages dans les logs," in *JDIR*, 2010.

[7] K. Musaraj, T. Yoshida, F. Daniel, M.-S. Hacid, F. Casati, and B. Benatallah, "Message correlation and web service protocol mining from inaccurate logs," in *Web Services (ICWS), 2010 IEEE International Conference on*, pp. 259–266, IEEE, 2010.

[8] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing and Management*, vol. 24, no. 5, pp. 513–523, 1988.

[9] B. Serrour, D. P. Gasparotto, H. Kheddouci, and B. B., "Message correlation and business protocol discovery in service interaction logs," in *In the Proceedings of the 20th international conference on Advanced Information Systems Engineering*, pp. 405–419, 2008.

[10] A. Moudjari, S. Chikhi, and H. Kheddouci, "Latent semantic analysis for business protocol discovery using log files," *International Journal of Web Engineering and Technology*, vol. 9, no. 4, pp. 365–396, 2014.

[11] M. Sanderson and W. B. Croft, "The history of information retrieval research," *Proceedings of the IEEE*, vol. 100, no. Special Centennial Issue, pp. 1444–1451, 2012.

[12] R. Baeza-Yates, B. Ribeiro-Neto, *et al.*, *Modern information retrieval*, vol. 463. ACM press New York, 1999.

[13] N. Fuhr and C. Buckley, "A probabilistic learning approach for document indexing," *ACM Transactions on Information Systems (TOIS)*, vol. 9, no. 3, pp. 223–248, 1991.

[14] J. v. R. C, "A non-classical logic for information retrieval," *The Computer Journal*, vol. 29, no. 6, pp. 481–485, 1986.

[15] K. S. Jones, S. Walker, and S. E. Robertson, "A probabilistic model of information retrieval: development and comparative experiments: Part 1," *Information Processing & Management*, vol. 36, no. 6, pp. 779–808, 2000.

[16] K. S. Jones, S. Walker, and S. E. Robertson, "A probabilistic model of information retrieval: development and comparative experiments: Part 2," *Information Processing & Management*, vol. 36, no. 6, pp. 809–840, 2000.

[17] K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[18] W. B. Croft, "Experiments with representation in a document-retrieval system," *INFORMATION TECHNOLOGY-RESEARCH DEVELOPMENT APPLICATIONS*, vol. 2, no. 1, pp. 1–21, 1983.

[19] E. Chisholm and T. G. Kolda, "New term weighting formulas for the vector space method in information retrieval," 1999.

[20] W. M. Van der Aalst, B. F. van Dongen, J. Herbst, L. Maruster, G. Schimm, and A. J. Weijters, "Workflow mining: A survey of issues and approaches," *Data & knowledge engineering*, vol. 47, no. 2, pp. 237–267, 2003.

[21] K. Musaraj, *Extraction automatique de protocoles de communication pour la composition de services Web*. PhD thesis, Lyon1 University, French, 2010.

Expansion of Region for Redundant Traffic Reduction with Packet Cache of Network Node

Shunsuke Furuta, Michihiro Okamoto, Kenji Ichijo, and Akiko Narita
Graduate school of Science and Technology, Hirosaki University, Hirosaki, Aomori, Japan

Abstract - *In recent computer networks, a large amount of data of the same contents are transferred repeatedly. They are often delivered concurrently, as represented by live broadcast. We have developed network nodes with packet caches in order to reduce such redundant traffic in TCP/IP network. We call the node TR node. We have obtained successful reduction rates experimentally in limited network topology in our studies. In this paper, we propose improved TR node with which region of redundant traffic reduction is extended. Improvements consist of two modifications. One is to quit renting a space in the IP header for a cache control argument and take the argument out to the header allowing fragmentation of a packet. The other is cache synchronization generating a synchronization packet. We show results of implementation of the proposed modifications of the TR node and advantages using them.*

Keywords: traffic reduction; network node; packet cache; cache synchronization; fragment

1 Introduction

In recent computer networks, the same contents are often transferred repeatedly. If the contents are transmitted through the same route, transmission is wasteful spending of resources for communications. It is desirable to reduce redundancy from the network traffic to utilize limited resources for computer networks efficiently.

There are two dominant methods to reduce the redundancy from network traffic. One is multicast, and the other is caching. Multicast is a method to eliminate redundancy in traffic to let a packet own multiple destinations substantially. It is appropriate for concurrent transmission of redundant data. IP multicast, with which routers at branch points duplicate datagrams, is considered the most powerful technique. However, some requirements hinder utilizing IP multicast. All equipment on transmission route must be acceptable for IP multicast. Furthermore, only datagram type is acceptable for a transport layer protocol. Meanwhile, application level multicast is relived from such limitations since end hosts manage transmission iterations. At the same time, efficiency of this method is lower than that of IP multicast. The same data are passed through the same route unlike IP multicast in principle. Efficiency of it depends on quality of multicast tree with the hosts and overhead for constructing the trees is not negligible. Caching is a method to shift traffic redundancy to access to storage devices. Caching has been employed for elimination of redundancy produced by repeated request for the same content. Proxy server that sends data instead of the original sender is popular. A proxy server that stands nearer than the original server does shortens length of transmission route and number of network segments containing redundant traffic decreases. Caching device of the proxy server is usually secondary storage so that a file is unit of caching. Hence, it is difficult to reduce redundant traffic with concurrency such as live broadcast or video conference. Packet caching using primary storage is another type of caching that enable to reduce concurrent redundant traffic.

We have developed the network nodes for reducing redundant traffic with packet cache [1]-[6]. We call it TR (traffic reduction) node. Two or more TR nodes cooperatively reduce traffic in a service region. Advantage of the TR node over multicast is transparency for endpoint hosts. There have been several problems to resolve in basic methodology of the TR node in [1]. Speed up of processing in [2] [3] is important to gain enough throughput with inexpensive resources. Methods for efficient memory utilization in [4] [5] also contribute to establish system with the TR nodes in reasonable cost. Another important problem is expansion of the service region. In the original TR node network, an encoding TR node and a decoding TR node must be placed alternately. We introduced forwarding TR node in [6] to decrease number of TR nodes and lower cost in service region consisted with multiple network segments. To enlarge the service region drastically, two modifications have been necessary. One is to replace a pure forwarding TR node with ordinary router, and the other is to adapt the TR node to branch topology of computer networks. It is complicated to adapt branch topology completely so that we focused on the situation with a sender and multiple receivers. In this paper, we propose methods to realize them and show benefit.

2 Basic Method for traffic reduction

We outline operation of the TR node for eliminating redundancy in traffic with Fig.1 and Fig. 2. The TR node is a functional router. We have implemented its functions with programs that work on Linux operating system. We designed the TR node for TCP/IP environment and implemented it assuming Ethernet for link layer protocol to perform experiments easily. There are three roles in traffic reduction

with TR nodes. They are encoding, decoding, and forwarding. In Fig. 1, there are two TR nodes. TR node E is situated on the upstream side and receives the same data repeatedly in a short time. It works as an encoder. When it receives a packet, it searches the same data in its packet cache as transported by the received packet. If it is the first time for TR node E to receive data A, the node cannot find it in the cache. TR node E records the data in the cache and transmits data A to TR node D with an instruction code of requesting for recording the data. The instruction code is put in IP header. When TR node E receives data A for the second time, it can find the data in its cache unless the data has been already removed. TR node E encodes data A and sends it with an instruction code to request decoding. Fig. 2 shows format of encoded data. Encoded data is smaller than received one. In general, TCP does not always divide data streams at the same byte offset into segments. Therefore, the data received by TR node E may partially match with one of the records in the cache. The node often represents the data using several cache records or portion of raw data in this case. TR node E sends original data if result of encoding has longer size than received one. TR node D lay on the downstream side works as a decoder. TR node D reconstructs data A referring its cache. If any TR nodes are placed between TR node E and TR node D, they merely forward the received data.

We define the reduction rate Rs_j in a network segment j as follows.

$$Rs_j = \frac{D_j - C_j}{D_j} \qquad (1),$$

where D_j is transmission rate in the case that the TR nodes do not reduce redundant traffic at all, and C_j is transmission rate obtained with the TR node operations for redundant traffic reduction. We define the reduction rate of target network R_n as weighted average with D_j over traffic reduction region as follows.

$$R_n = \frac{\sum_j Rs_j \cdot D_j}{\sum_j D_j} \qquad (2).$$

We can obtain ideal reduction rate when all redundant streams are divided at the same offset and transported with the largest frames having the shortest headers. In this case, the encoding TR node builds the shortest packets from the largest data. The encoded packet contains only one block of type 0. Packets with block type 1 or block type 0 containing multiple blocks decline traffic reduction rate. Under TCP/IP/Ethernet environment, a packet with 14-byte header, 8-byte preamble and 4-byte FCS for Ethernet, 20-byte IP header, and 20-byte TCP header has the largest data. The ideal value of Rs_j is 1440(M-1)/(1526M) = 0.94(M-1)/M with number of redundant streams M if we set 4 bytes for each the sizes of the fields of number of blocks, block type, record number, offset, and length in Fig. 2.

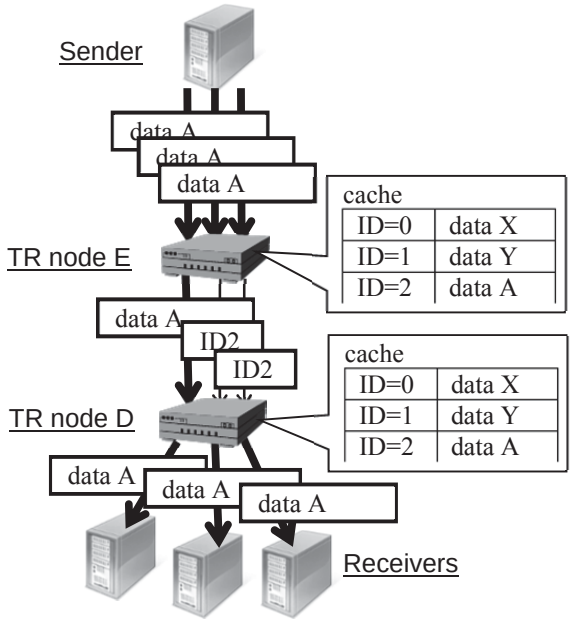
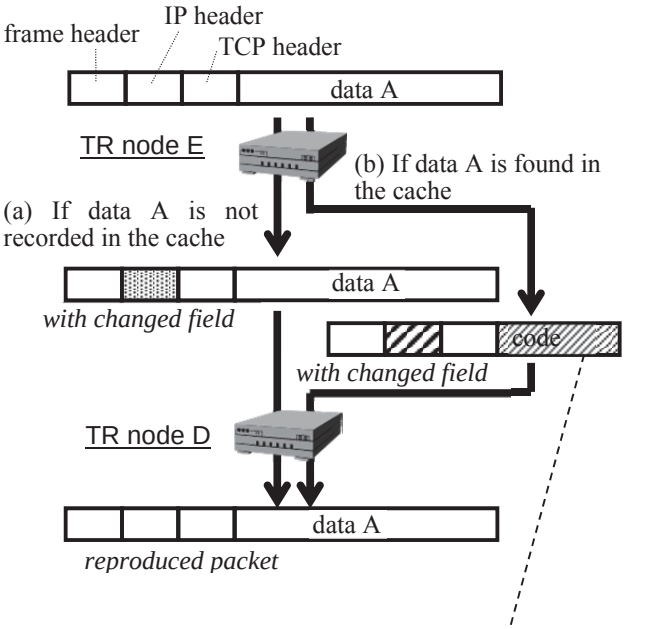


Figure 1. Basic operation of the TR node for eliminating redundancy in traffic.



number of blocks	data body			
	block 1	block 2	block 3...	block k

block type = 0	type	record #	offset	length
block type = 1	type	length	portion of data	

Figure 2. Reduction of packet size with the TR nodes.

3 Expansion of service region

3.1 Restriction of service region with the previous TR node

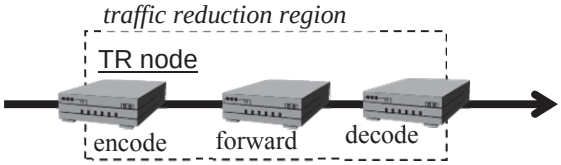
There were confinements of service region with the previous TR node as shown in Fig. 3. Arrows in the figure mean routes of TCP originated packets. We had to arrange TR nodes continuously without branch, as in case (a) provided that number of the forwarding TR node may be 0 or more than 1. There were roughly two reasons for restriction of service region of the TR nodes. One was illegal change of IP header and the other was flaw in method for cache synchronization.

The previous TR node used total length field in the IP header for record number in the packet cache. While network nodes could know packet length seeing frame size and header length, usage of the total length field by the TR node was illegal for IP. If an ordinary router was situated in the TR node network like case (b), the router discarded an encoded packet containing illegal values in the total length field. If we would reduce redundant traffic at both upstream side and downstream side of the ordinary router, we had to set pairs of an encoding TR node and a decoding TR node on both sides.

Flaw in the method for cache synchronization caused trouble with such a TR node network that involved branches in a transmission route as shown in case (c) and (d). In case (c), branches of transmission route shoot from a TR node toward downstream side. The previous TR node did not guarantee cache coherency if branch point existed. The packet with instruction for recording for cache synchronization was passed only one route of the branches. If the fastest data of redundant streams was sent to the direction to X, the data was recorded in the caches of TR node A and B. When successive data of the same content was sent to TR node C, TR node A encoded it, and TR node B forwarded it to TR node C despite TR node C did not have the original data. TR node C constructed data using information in the received packet and different source in its cache. Then an incorrect packet arrived at a receiver. This trouble was not brought if only one route of transmission was used for each group of redundant streams. Avoiding the trouble, TR node B sent encoded packets only one direction and decoded packets to others.

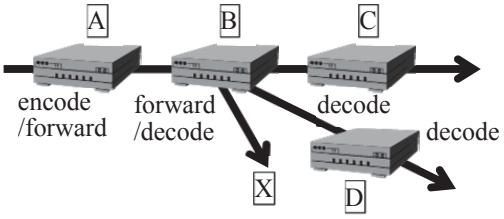
The decoding TR node also failed to make up a proper packet in case (d), in which several branches came together from upstream to one. In this case, packets with instruction for recording from multiple encoders might overwrite records in the cache of the gathering spot and caused conflict. When TR node E sent an encoded packet assuming that TR node G could decode properly, the corresponding record in the cache of TR node G might have been already replaced with a data sent from TR node F. This disorder was caused by any concurrent flow of TCP streams from two or more encoding TR nodes to one forwarding or decoding TR node.

Nevertheless, a destination host can detect error with checksum in TCP header and discard incorrect packets. A sender host executes retransmission after expiration of retransmission timer. The encoding TR node can know retransmission comparing IP addresses, port numbers, and sequence number of a received packet to list of traces for packets that have come by. The encoding TR node does not encode data on the retransmitted packet. The data arrives at the destination without recasting in turn. The most consequential problem is not delay of transmission but decline of throughput. TCP retransmission is accompanied with suppression of transmission rate because of mechanism for congestion avoidance.

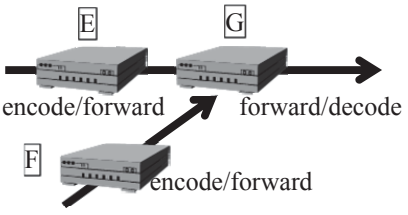


(a) Case of successful traffic reduction.

(b) Case of failure in transmission at the ordinary router.



(c) Cache synchronization failure of un-sent record.



(d) Cache synchronization failure due to competition.

Figure 3. Restriction of service regions by the previous TR nodes.

Fig. 4 shows a target network topology. This is typical for server-client type service. Service regions with the TR nodes are expanded drastically. Resolving illegal change of IP header, we can add unknown network constructed with ordinary routers to the service region. Improving cache

synchronization method for branch of routes toward downstream side, we can apply service of the TR node to branch topology of a local area network. We can obtain significant benefit with these resolutions so that we tentatively postpone the problem of cache synchronization of branches coming from upstream.

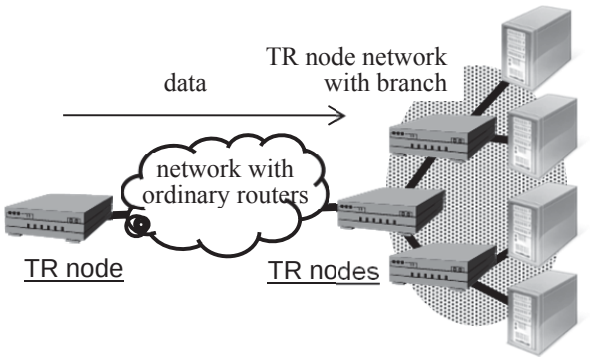


Figure 4. Service region enlarged.

3.2 Remediation of illegal change of the IP header

The encoder TR node puts an instruction code in the protocol field of the IP header of the received packets. The instructions are given in TABLE I. They are replaced with the protocol number of TCP by the decoder TR node. Codes F2 and FB are introduced in this study as explained later. The instructions require arguments in TABLE I and II. The previous encoding TR node wrote argument N on total length field in the IP header. Disagreement between true length and the value in the field means error for an ordinary router obeying IP. The router discards the packets encoded by the TR node with instruction FC and FE. The reason why we adopted the illegal change was to avoid increase of traffic in any case. If we stop using the total length field in the IP header for argument N, we must extend the length of a packet to send putting the argument in the other place. In this study, we placed emphasis on advantage to reduce traffic in many network segments with a few TR nodes.

The present TR node at the edge of upstream side in the service region makes a temporary packet adding the argument N to the end of data field for instruction FC and FE. If length of it exceeds the upper limit of length of the packet in the network segment, the temporary packet is fragmented as shown in Fig. 5. We introduced a new instruction F2 to notify the fragmentation for decoding TR node. The fragmentation is carried out with manner of IP except packet reassembly node. The decoding TR node reconstructs the original data. Assuming the same condition as we show in the section 2 for ideal reduction rate, we obtain Rs_j as (1440(M-1)-48)/(1526M), which is only 0.02 smaller than that obtained by the previous TR node even if M = 2. The proposed method increases traffic if redundancy is not contained. Ideal reduction rate is 1438*(M-1)/(1524M) that is accomplished if fragment does not occur, that is, the length of data in the original packet is always 2 bytes shorter than the maximum length.

TABLE I. INSTRUCTIONS EMBEDDED IN THE PROTOCOL FIELD BY THE ENCODER NODE

code	***Request***	***argument (See TABLE II.)***
F2	reconstruct and record	D,N,F
FA	no operation	
FB	synchronize	D,N
FC	record	D,N
FD	decode	B
FE	decode and partially record	D,N,B

TABLE II. ARGUMENTS FOR THE INSTRUCTIONS.

argument	***attribution(position)***
D	raw data (data field)
N	record number (IP header in the previous TR node, see text)
F	fragment control information (IP header)
B	encoded block (data field)

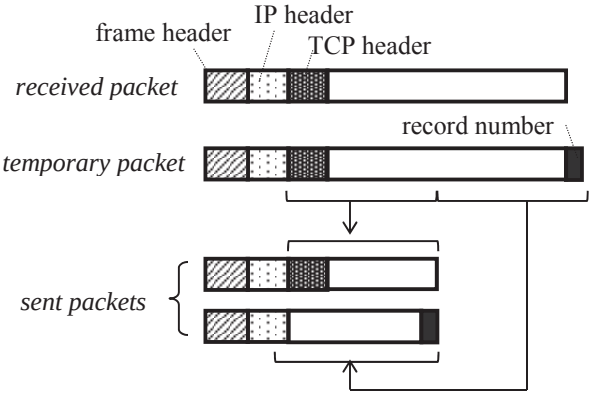


Figure 5. Packet fragmentation of a long packet with instruction FC and FE

3.3 Cache Synchronization

Cache synchronization over arbitrary network is unrealistic. We tried to realize cache synchronization for the TR node sequence with branch toward downstream and without ordinary router. Data transmission between nodes is inevitable for cache coherency. There are roughly two methods. One is always to keep coherency. With this method, when the TR node sends an encoded packet with instruction FC or FE, it transmits the same data to all succeeding TR nodes. This is a simple method to implement. However, the transmission may be wasteful because there may be TR nodes that are not on the route of the redundant streams. Nevertheless, the transmission causes momentary increase of traffic. The other method is to delay synchronization until the

data becomes required to decode. The TR nodes guarantees cache coherency when it sends an encoded packet with instruction FD or FE for which its descender TR node must reconstruct an original data.

We introduced a new instruction FB for cache synchronization as shown in TABLE I. Destination of a packet with this instruction is the neighboring TR nodes so that TCP header is unnecessary. In this study, we assume branch topology is only in local network. We can allow the TR node to borrow total packet length field in IP header for record number. Note that the packet with FB instruction escape fragmentation if we put the record number in the data field since the packet length is shorter by size of eliminated TCP header.

The packet cache in the TR node is a large data table with several management fields. We appended fields of synchronization flag to the packet cache. The node on upstream side is responsible for cache synchronization between two neighboring nodes. The fields are arranged for all TR nodes on downstream side. When the TR node makes new record, it clears the fields of the record. As the TR node sends the data in the record to one of next hop nodes, it sets flag corresponding to the current next hop node. When the TR node sends a packet with instruction FD or FE, the node confirms whether the involved record in the encoded data is already sent to the next hop. If the corresponding field is unset, the node transmits the data. The most favorable case is that the received packet contains only one encoded block. The TR node reconstructs a packet with instruction FC and sends it only. If the received packet has more than two encoded blocks or instruction FE, the TR node generates synchronization packets for unsent data. It is also necessary to send the received packet in this case. Fig. 6 shows an example for generation of synchronization packets. The received packet has two encoded blocks of un-synchronized cache records in this example.

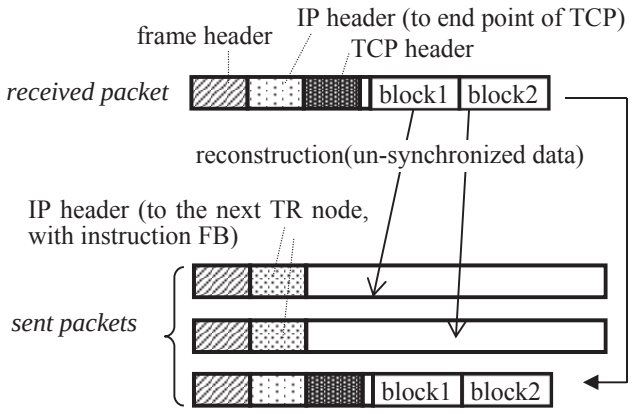


Figure 6. An example for generation of synchronization packet.

4 Evaluations and discussion

We built a computer network for evaluation of the proposed methods. The functions of the TR node was implemented on computers with AMD Opteron 1210 (1.8 GHz) CPU, 1GB main memory, Debian 4.0 (Linux 2.6.18-6-486) operating system. The redundant contents consisted of random numbers. Line speed was 100 Mbps. In the present measurements, sender sent packets using bandwidth fully.

4.1 Passing an ordinary router

To confirm validity of the proposed method described in the section 3.2, we constructed a computer network shown in Fig. 1, and set an ordinary router between two TR nodes. We emulated 5 receivers with 5 processes working on a machine. We counted number of packets with each instruction at upstream side and downstream side of the router. Measurements were carried out a few seconds and 300 seconds. Fig. 7 shows the result this measurement.

With the present modification, packets went through the ordinary router successfully. If all redundant streams are divided at the same offset and data length is always maximum, the encoding TR node sends 4 packets with instruction FD for 5 received packets, and 2 for the 5 with instruction F2. Number of sent packets is larger than that of received packets since fragmentation occurs. In the condition of the present experiment, packet sizes were mostly maximum frame size so we observed packets with instruction F2 frequently. As mentioned before, this is not a condition that gives the best reduction rate. Furthermore, packets with instruction code FE were generated for stream division at different position. We obtained $R_n = 0.74$. The maximum reduction rate is 0.75 for N = 5 as described in the former section.

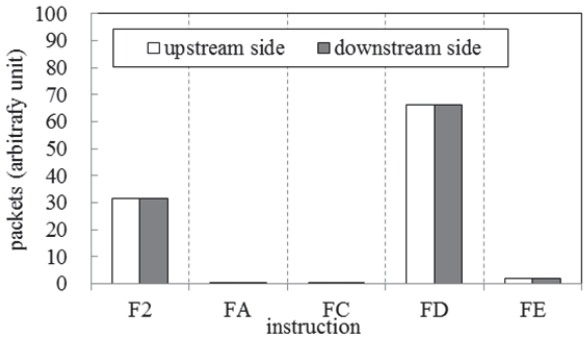


Figure 7. Number of packets for each type and behavior of an ordinary router for encoded packets.

We can obtain benefit of redundant traffic reduction with less TR nodes with the present modification. For example, we consider such network that contains 2 consecutive network segments separated by 3 network nodes. We can achieve $R_n = 0.74$ with 2 TR nodes at the ends of

considered region under the same condition of redundant traffic as the present experiment. If we use the previous TR nodes, 3 TR nodes are required to obtain nearly equivalent efficiency to $R_n = 0.74$. When we can prepare only 2 TR nodes of the previous type, we must place an ordinary router before or after sequence of the 2 TR nodes and elimination of redundant traffic is performed only in one network segment. Then $R_n = 0.38$ over the region. Advantage of the present specification becomes larger in the network with more ordinary routers.

4.2 Validity of cache synchronization

We carried out measurement for evaluation of the present cache synchronization using the computer network shown by Fig. 8. The sender dealt 6 redundant streams, that is, 3 streams to the receiver Y and A, respectively. Transmitted data by the sender split on halves toward TR node C and D. In ideal case, one for 6 packets of received packets by TR node B is of instruction FC and 5 for 6 are of FD and have the smallest data field. Then the node sends 2 packets of FC and 4 of FD for 6 received packets. Packets with instruction FB are never generated. $R_n = 71$ in this case.

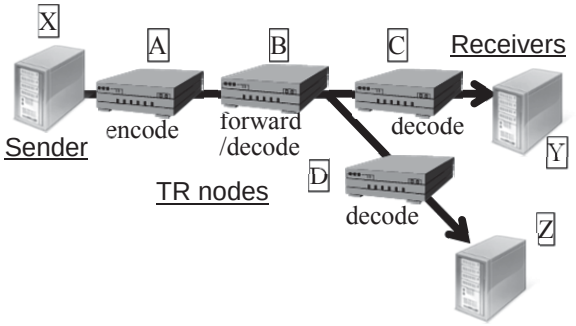


Figure 8. An example of TR node network with branch constructed for evaluation

Experimental results are given in Fig. 9 and Fig. 10. The former presents transmission rates and the latter shows rate of packets with each instruction sent by TR node B. Segment A-B means zone between TR node A and B in Fig. 8. $R_n = 0.68$ in the experiment. Fig. 10 demonstrates generation of packets with instruction FB and denotes existence of packets with instruction FD having multiple encoded blocks. Hence, R_n was lower than ideal value. If we construct the network of Fig. 8 with the previous TR nodes, TR node B must forward encoded packets to only TR node C (or D), regarding TR node D (or C) as an ordinary router and decodes packets for it (or C). Otherwise, throughput significantly drops because of congestion avoidance by TCP. TR node B sends one packet with instruction FC and two with FD to TR node C, and three reconstructed packets to TR node D for six received packets. Then $R_n = 0.31$ at most. Thus, the obtained value of R_n with the proposed cache synchronization method was much better than that with the previous TR node. Advantage of the present TR node increases for larger TR node network.

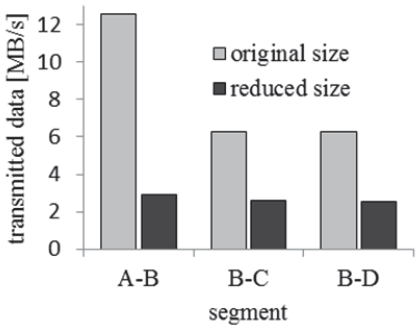


Figure 9. Reduction of redundant traffic using the present cache synchronization method.

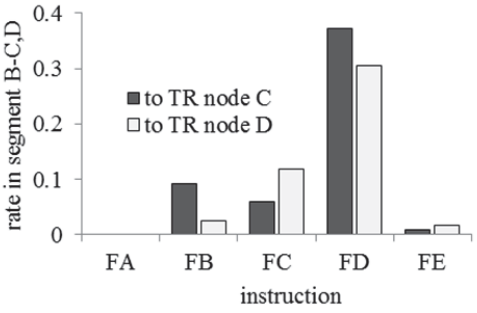


Figure 10. Rates of instructions

5 Related works

Concept of shared cache in [7] is very similar to the TR node. Rabin fingerprint [8] is used to generate identifier of cached data. Cache coherency is assumed in the discussion of it. [9] and [10] also use Rabin fingerprint and routing is optimized for redundant traffic elimination. In [9], all caches on routers that are involved in traffic reduction hold the same data. On the other hand, a central module to manage status of caches is introduced in [10]. It offers deployment of redundant data over caches along route of packets in order to save memory space. [7], [9], and [10] emphasize independency of protocol or application. Procedure proposed in [11] is advantageous in memory saving. Cache of ingress router holds only identifier. Policy of synchronization that all egress routers must synchronize with corresponding ingress routers completely is the same as [7] or [8]. Study of [11] is focusing on P2P traffic exchanged by a particular application BitTorrent. Data identifiers are obtained by the hash function MD5. Traffic reduction routers in [11] confirm cache

coherency with message sent by egress routers to the ingress ones as data updated.

We lay weight independency of protocol in the same way as [7], [9], [10], while region for traffic reduction is different from those studies. Their targets are ISP network. On the other hand, our target is traffic traversed WAN and LAN as shown in Fig. 3. The TR nodes do not suppose optimal routing for redundancy elimination whereas it must know whether if there is another TR node on downstream. This characteristic can cut cost of introducing functional router and routing management. Furthermore, server-client type service is our target so that assigning responsibility of cache synchronization to the upstream node is appropriate rather than the other method. Some differences between related works and our study are trade off. One of them is method to obtain identifier. We use so simpler method than the other related works do that load of the nodes is lighter. However, the identifier does not reflect characteristic of whole data. Therefore, the TR node is weak for accidental matching. The next problem is memory saving. Memory saving in [10] is substitution between resource and reduction rate, or memory and bandwidth. Similarly, saving copy of data in the encoding TR node is memory consuming but free from collision.

6 Conclusions and Future works

In this paper, we proposed procedures to apply service with the TR node to wider network region. One is to enable to employ its service over network including ordinary routers. It is achieved by shifting an argument for cache management from the field in IP header to data field, with fragmentation if required. The other is to adapt the TR node for branch topology. It is accomplished with cache synchronization control sending synchronization packets. Therese methods are implemented to the computer network constructed for experiments. Validity of our proposed method was shown by experiments and estimations.

The most important advantage of the TR nodes is redundant traffic reduction of real time transmission. TV meeting is one of the most significant application as well as live broadcast with sever-client type delivery. In the case of TV meeting, two or more sender exists and cause competition of cache utilization. In future work, correspondence to branch topology with multiple encoder TR nodes is desirable.

7 References

[1] Yasuyuki Saito, Kenji Ichijo, Akiko Narita, and Yoshio Yoshioka. “Data reduction with cache in TCP/IP network”; Tohoku-Section Joint Convention of Institutes of Electrical and Information Engineers, p.110, Aug. 2007.

[2] Tomoya Shikanai, Kengo Kimura, Sayuri Yamamoto, Yasuyuki Saito, Kenji Ichijo, Akiko Narita, and Yoshio Yoshioka. “Modification and evaluation of a network node for traffic reduction”; Tohoku-Section Joint Convention of Institutes of Electrical and Information Engineers 2010, p. 138, Aug. 2010.

[3] Tomohiro Yoshida, Yuki Otaka, Akiko Narita, “Implementation of Function for Redundant Traffic Reduction on Kernel of Network Node”; Tohoku-Section Joint Convention of Institutes of Electrical and Information Engineers 2014, 2G14, Aug. 2014.

[4] Yuki Otaka and Akiko Narita. "Efficient Packet Cache Utilization Of A Network Node For Traffic Reduction”: Proceedings of the 2013 International Conference on Internet Computing and Big Data (ICOMP '13), pp. 109-112, Jul. 2013.

[5] Yuki Otaka and Akiko Narita, "Efficient Assignment of Packet Cache Region for Traffic Reduction of Multiple Redundant Contents”; Proceedings of the 2014 International Conference on Internet Computing and Big Data (ICOMP '14), pp. 117-123, Jul. 2014.

[6] Shunsuke Furuta, Yuki Otaka, Akiko Narita. “Design and Evaluation of a Network Node for Traffic Reduction”; The Special Interest Group Technical Reports of IPSJ 2013-CSEC-61, No. 15, pp.-7, May. 2013.

[7] N. Spring and D. Wetherall. “A protocol-independent technique for eliminating redundant network traffic”; Proceedings of the ACM SIGCOMM 2000 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication, pp. 87-95, Aug. 2000.

[8] M. Rabin. “Fingerprinting by random polynomials”; Harvard University Technical Report, TR-15-81, 1981.

[9] Ashok Anand, Archit Gupta, Aditya Akella. “Packet Caches on Routers: The Implications of Universal Redundant Traffic Elimination”; Proceedings of the ACM SIGCOMM 2008 conference on Data communication, pp. 219-230, Aug. 2008.

[10] Ashok Anand, Vyas Sekar, and Aditya Akella. “SmartRE: an architecture for co-ordinated network-wide redundancy elimination”; SIGCOMM Computer Communication Review, 39 (4), pp. 87-98, Sep. 2009.

[11] Shu Yamamoto, Akihiro Nakao. “P2P packet cache router for network-wide traffic redundancy elimination”; Proceedings of International Conference on Computing, Networking and Communications (ICNC), 2012, pp. 830-834, Jan. 2012.

SESSION

POSTER PAPERS

Chair(s)

TBA

Big Data on Performance Logs – A Collaborative Monitoring Cloud for ERP Systems

H. Müller[1] **and K. Turowski**[1]
[1]Very Large Business Applications Lab, Otto von Guericke University, Magdeburg, Germany

Abstract - *Although outsourcing is a viable instrument to save operational costs, the majority of ERP systems is still operated in-house due to privacy, security and dependency concerns coupled with ERP's exceptional significance for business continuity. In this abstract paper, we propose a research artefact that is planned to become an alternative option to classical in-house or off-promises operation models and enables fully controlled in-house operation with cloud-supported performance analyses. Therefore, we started to analyze 230 million performance log entries of about 8,700 standard SAP ERP systems and evaluate its suitability for a value creating Big Data scenario. Integrating performance data and hardware information of ERP systems enables cross-system and cross-customer analyses and, potentially, to deliver additional knowledge to ERP operating IT departments through a cloud service*.

Keywords: ERP; Performance; Big Data; Cloud.

1 Introduction

Outsourcing of information technology has attracted significant interest from both research communities and industries [1]. Firms benefit from economies of scale by combining their operational costs for hardware, software and staff through external service providers. Particularly in cases of planned greenfield projects or major architecture changes, outsourcing is a viable instrument to save investment costs by spreading them over the entire contract period. Besides, information management has transformed from a technical perspective of "plan-build-run" to a business perspective of "source-make-deliver". Driven by market orientation, product orientation and product lifecycle management [2], modern business departments consume IT products delivered by both internal and external providers, while quality attributes are ensured by means of service level agreements. While firms make use of these effects for various business applications, outsourcing, on the other hand, involves risks that still dominate decisions for certain core applications. Bryson and Sullivan discussed reasons for and against outsourcing which include concerns about security, privacy and application service provider (ASP) dependency [1], [3]. Olsen states that the biggest risks of outsourcing are downtime and loss of operational data. Therefore, companies tend to outsource applications which are not business critical in the effect that business continuity can still be ensured. According to the senior director of IT applications at Informatica Corp., the human resources (HR) module of their business suite could be outsourced, because the business will still continue to run if HR goes down [4]. Olsen argues that companies view ERP as too mission-critical to yield control. The CIO at Federal-Modul Corp., Mike Gaynor, clearly states a lack of control and trust when it comes to ERP outsourcing. He illustrates the dominating attitude of IT executives by a metaphor advising to never hand off the brains of operation, although companies might farm out for extra arms and legs [4]. One of the most widely used ERP system is developed by the German SAP AG. A recent study among German SAP customers from October 2014 [5] states that 75% of the surveyed companies do not have any cloud strategy regarding their SAP application landscape, whereby 62% of these evaluate SAP clouds as 'not relevant'. For an average of 56% of the companies across all surveyed industries, any software-as-a-service (SaaS) offerings for SAP are neither used nor planned or discussed. Only 6% of the surveyed companies are discussing infrastructure-as-a-service (IaaS) alternatives to their SAP in-house operation and 72% of those companies who make use of SAP IaaS are extending their existing SAP solution [5]. Thomas and Chirania discuss on whether or not to outsource ERP and explain existing restraints by the lack of implementing unique business processes [6].

Based on [7], Olsen summarizes seven options of ERP operation from in-house till ASP and states that each specific organization might generate variants of these that suit their particular needs [1]. Our research project seeks to develop and evaluate an additional option, which includes in-house operation with cloud-supported performance analyses. Therefore, additional knowledge derived from performance data of various ERP systems is going to be extracted by means of Big Data techniques.

2 State of the Art

For ERP systems that can be distributed across multiple application servers and serve hundreds of users simultaneously, performance monitoring becomes a complex task and response times need to be measured and analyzed in various dimensions. Therefore, IT departments make use of consultancy services, e.g. offered by hardware partners. Those services include ERP system performance monitoring during

productive operation and subsequent analyses of log records [8], they are requested and delivered whenever necessity is assumed. Generated performance reports serve as decision support for system administrators regarding sizing requirements, load balancing strategies, job scheduling and other points of action which affect the system performance. Within SAP systems, log entries are called statistical records and created automatically after each dialog step, performed by any user. These records are used to assess, e.g., system performance on business transaction level. When adding information about utilized hardware components, statistical records can be used for benchmarking, too. Cloud-based benchmark tools for hardware components already exist in other domains where performance matters, e.g., the gaming community [9], [10]. For SAP ERP systems, standard application benchmarks exist, e.g., the Sales & Distribution (SD) benchmark, which simulates a defined amount of simultaneously working users and measures response times within a given load interval [11]. In the following, we provide an overview of an SAP ERP performance cloud, which we are going to evaluate with respect to its ability to provide dedicated transaction performance analyses for each customer, performance comparisons across systems, hardware and customers, as well as benchmarking capabilities.

3 Proposed Research Artefact

Taking into account the existing concerns and demands regarding the operation of core business functionality, we developed an operational option that supports in-house operation but cloud based performance analyses of ERP systems. In that manner, only monitored performance data will need to be shared with the service provider while master and transaction data as well as operational control remains at operator site. Therefore, customers benefit from tools, data models and analysis techniques implemented once by the service provider and utilized for multiple systems and customers. In addition, integration of monitored data from different systems enables cross-system and cross-customer analytics. Therefore, customers are given the opportunity to assess their ERP system landscape by comparing own system performance with, e.g., mean values derived from empirical distribution functions across various systems on similar hardware. Hence, ERP performance logs contain information that can be extracted on a global scale leading into a Big Data use case, which we are going to investigate and implement as appropriate. During our research, we will focus on both operator's and provider's perspective and their different objectives.

4 Research Design and Outlook

We planned our research based on the information systems research framework provided by [12]. Therefore, we design an option of collaborative ERP in-house operation as a research artefact in multiple iterations, considering stated business needs from the environment and contributing results from planned analytics to the research knowledge base. Together with a major infrastructure partner of SAP customers, we started to clean and preprocess more than 230 million statistical records of about 8,700 SAP systems. An entity relationship (ER) model for a common database schema was developed and performance data has been imported into one in-memory database, which integrates a web server that can be used for providing the customer's user interface. Using statistics, we started preliminary work on identified use cases for response time predictions and hardware comparisons on business transaction level. In the current project phase, we collect further use cases of data analyses and seek feedback from both scientific community and industry. During the whole research, we plan to focus on both data analytics and the technical infrastructure required for the proposed research artefact.

5 References

[1] D. L. Olson, "Evaluation of ERP outsourcing," *Computers & Operations Research*, vol. 34, no. 12, pp. 3715–3724, 2007.

[2] R. Zarnekow, W. Brenner, and U. Pilgram, *Integrated information management*. Springer, 2006.

[3] K.-M. Bryson and W. E. Sullivan, "Designing effective incentive-oriented contracts for application service provider hosting of ERP systems," *Business Process Management Journal*, vol. 9, no. 6, pp. 705–721, 2003.

[4] B. DePompa, "Time to outsource ERP?" Computerworld, 2003 [Online]. Available: http://www.computerworld.com/article/2571205/it-outsourcing/time-to-outsource-erp-.html

[5] "SAP goes Cloud." Pierre Audoin Consultants, 2014 [Online]. Available: https://www.pac-online.com/download/13330/144294

[6] E. Thomas and V. Chirania, "Whether or not to outsource ERP?," 2005.

[7] D. L. Olson, *Managerial issues of enterprise resource planning systems*. McGraw-Hill, Inc., 2003.

[8] "Data Sheet - Software OPTIMIZATION Services - SAP SystemInspection Service." Fujitsu, 2013 [Online]. Available: http://globalsp.ts.fujitsu.com/dmsp/Publications/public/ds-optimization-services-sap-systeminspection-en.pdf

[9] "Mobile Phones, Tablets, Graphics Cards, Processors and Motherboard Performance." 2015 [Online]. Available: http://www.futuremark.com/hardware

[10] "How does your computer rank against millions of others?" 2015 [Online]. Available: http://www.systemrequirementslab.com/rank-my-computer

[11] "Sales and Distribution (SD and SD-Parallel)." [Online]. Available: http://global.sap.com/campaigns/benchmark/appbm_sd.epx

[12] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design Science in Information Systems Research," *MIS quarterly*, vol. 28, no. 1, pp. 75–105, 2004.

SESSION

LATE BREAKING PAPERS: INTERNET COMPUTING AND WWW

Chair(s)

TBA

World Wide Web: A Survey of its Development and Possible Future Trends

Abdulelah A. Algosaibi, Saleh Albahli and Austin Melton

Department of Computer Science
Kent State University
Kent, OH, USA
{aalgosai,salbahli,amelton}@kent.edu

Abstract- *The World Wide Web is considered one of the main sources in accessing information. Over the last decades, a number of improvements have been achieved that helped the Web reach its current state. For many, the World Wide Web has become indispensable to their daily lives. There are a number of research projects going on to enhance the current status and develop the future of the Web. It is therefore important to look into a new version of the Web in order to improve the way that information is expressed to make more intelligent choices and obtain a better meaning of the information over the Web. In this survey, we study the evolution of the Web from Web 1.0, Web 2.0, Web 3.0, Web 4.0, to Web 5.0. We are pointing out document types and technologies employed to understand the changes from Web 1.0 to Web 3.0 and to predicate the future of the Web (Web 4.0 and Web 5.0). Also, we present the current status and concerns about the Web as an information source and communication channel.*

Keywords- Web Generations; Web 1.0; Web 2.0; Web 3.0; Web 4.0; Web 5.0; World Wide Web; WWW; Semantic Web; WebOS; Intelligent Agent.

1. Introduction

The ways of communicating and accessing information have changed, and more people are relying on the Web as a primary source of information. This information can be obtained from different places such as web sites, blogs, online publications, social networks, databases and much more. Indeed, the Web is considered as one of the main sources of information. It is a massive information exchange platform that was introduced by Tim Burners-Lee [1]. Basically, the idea was to link documents over the internet. Now, with the evolution of the technologies, not only can we connect documents, but we also can understand documents. Documents, in general, come in three categories [2]. The first type is structured documents. Here the formation of the document and the inner data are structured in a way that for each piece of information, it is explicitly known how that piece of information fits with the other data. This leads to the retrieval of more relevant data. Thus, the semantic information extracted from these kinds of documents is rich. Examples of structured documents include databases or spreadsheets. The second type of documents is semi-structured documents. In this type of document, the data are structured, at least in part, based on semantics, but the underlying structure and the semantics are not explicitly given. The structure, however, is still helpful in extracting the semantics from the document. This kind of document, if carefully handled, can produce relatively rich semantic information because the semantics is embedded in data and the data's structure. Examples of semi-structured documents are HTML documents, WordNet [3], and XML documents. The third type of documents is unstructured documents where the knowledge is available only in the data and not in the structure of the documents. In this freely structured format, the documents do not follow any predefined or representation structure. Examples are standard text files. The Web, as an information source, is holding enormous amounts of information. In its first generation (Web 1.0) and early stages of second generation (Web 2.0), the Web was limited by the available technologies at the time. Also, most of information on the Web was not understandable by the machine. This makes the vision of the Semantic Web [4] (Web 3.0) an urgent task. Web 3.0 heavily depends on the structured documents type. We explain how the Web (Web 1.0 to 3.0) has evolved over time with respect to documents types and technologies, show the basis of Semantic Web concepts and technologies that form the foundation and expectations for the next Web (Web 4.0 and Web 5.0), and outline the Web's current status and some concerns.

The rest of this paper is organized as follows: section 2 reviews the related work of Web generations. In section 3, we discuss our viewpoint of Web 1.0 to Web 3.0. We present future trends in section 4. Finally, the conclusion is in section 5.

2. Related Work

Research [5] has surveyed the Web generations' background evolution from Web 1.0 to Web 4.0. They studied the characteristics of these generations and provided some comparisons. Research paper [6] has mentioned the fifth generation of the Web (Web 5.0). Web generations also were described by Weber and Rech [7]. They proposed a definition of each generation. Key differences between the first two generations were studied in [8]. Diana, Marta, Carlos and Alberto [9] have discussed the possibilities for the fifth generation. This survey is distinguished from the above surveys by

studying the evolution of the Web with respect to the technologies and document types of each generation. Also, we are pointing out the current Web status and concerns.

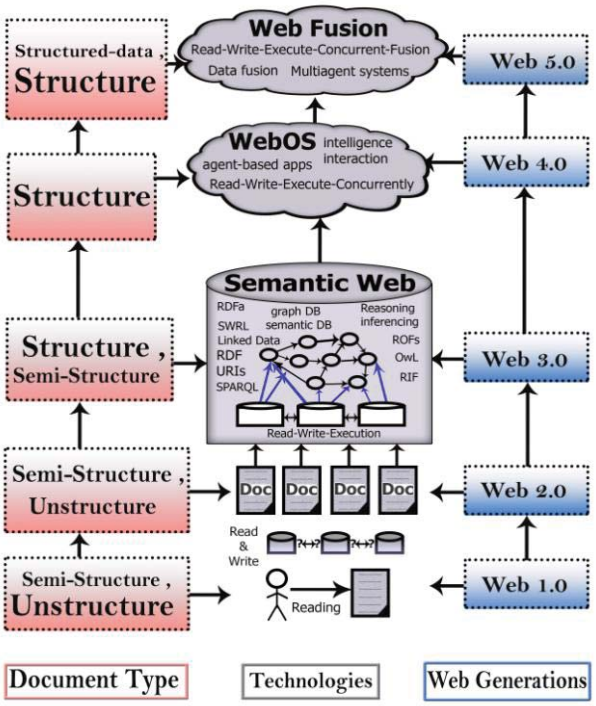


Figure 1: Web generations with respect to attention on technologies and document types

3. Web Generation

As mentioned earlier, documents come in three categories. Each one of these categories has received attention at certain generations. Table 1 below shows when each kind of document received higher attention. These documents are processed by given technologies. The processing of these documents was limited by the availability of Web technologies at particular Web generations. Since progress was made to improve each generation, the related technologies have improved simultaneously within certain periods [5].

	Unstructured Documents	Semi-Structured Documents	Structured Documents
Web1.0	X	X	
Web2.0	X	X	
Web3.0		X	X

Table 1: Document types in each Web generation

A. Web 1.0

Web 1.0 was simple in terms of information and how it was represented. In fact, it was considered a "static web" on the Web. Web 1.0 was limited to the features it provided and did not exceed the layout representation of a website. Most of the information on Web 1.0's websites was on the webpage itself. It was about a websites that do not interact with visitors to offer implicit functions such as an online sound-clips converter. A good example of a Web 1.0 website is a professor's homepage that has only information about courses, publications or pictures. Web 1.0 was not implemented to offer a service that required further configuration.

1) Document Type

Most of the attention was on unstructured documents. As a result, Web technologies were simple in terms of the power of processing these documents (e.g., a web directory for navigating through a list of websites). Semi-structured documents had limited availability and did not exceed Hyper Text Markup Language (HTML) documents or data describing EXtensible Markup Language (XML) documents.

2) Technologies in Place

Technology was focused on how to process these understructure documents. HTML, Cascading Style Sheets (CSS), XML, or web browser technologies were a clear attempt to enhance the limitations of functionality that users experiences. HTML is the publishing language of the World Wide Web [10]. HTML tags on data give presentation capability that is processed by a web browser. These tags structure a webpage. This structure is used for representing information in a webpage. HTML5 proposed semantic tags (e.g., `<section>`). However, it is still on structural stages. A webpage, in Web 1.0 that was a user-end interface consisted of HTML tags along with text. This version of the Web interfaced with a web directory [11] that created a methodology to browse websites. Later, technologies such as "crawler" and "spider," which store text within webpages was have proposed [12][13]. Thus, processing Web 1.0 was limited by the technologies available.

B. Web 2.0

Web 2.0 is an extension of Web 1.0. It is more enhanced in terms of the features, functions, services and usefulness than Web 1.0. Web 2.0 is considered as offering "dynamic web" (e.g., social networking sites, wikis, video sharing sites, online shopping, and web applications). It is a web-as-participation-platform [14]. If desktop-application software is available, then with Web 2.0 it is possible to have another version be a web-based application. Web 2.0 is bidirectional communication [8]. All of these features have a number of back configurations to make them work. Some configuration processes, for example, can start with choosing the domain (e.g., health, sports, news), then building the necessary database, then writing the query code within web page code, then designing the layout of the page (e.g., drop-down menu, radio check), and ending with the presentation of the results of a query. Web 2.0 often follows software engineering principles in order to build scalable web applications. Following these principles will help in maintaining and enhancing web applications.

1) Document Type

Unstructured documents have a massive amount of information that increases the number of services done over the Web. Semi-structured documents allow for more precise processing. Also, semi-structured documents get increased attention. The technology can more easily process these kinds of documents to include Natural Language Processing [3] and HTML documents parsing [15].

2) Technologies in Place

Technologies in this Web generation enhanced the processing functionality of Web documents. The technologies drove Web 1.0 to become more interactive. The technologies allowed the Web to be more dynamic (e.g., Web forums). The Web technologies allowed Web applications to be in place (e.g., JavaScript, PHP, Python, JSP, ASP.NET and JAVA). Also, Web 2.0 became more mature to give the user the ability to choose the architecture of the Web application whether it was client-side such as JavaScript or server-side as JSP. There are some web editors that help make the creation of Web 2.0 much easier and simpler (e.g., Microsoft Visual Studio that works with ASP.NET and C#). In fact, with the evolution of Web 2.0, allowed for richer Web content. As a result, navigation through the Web, particularly within the website, became harder. Navigation methodologies (e.g., sitemap and mature search engines such as Google) were invented to help speed up the process of finding a desired content.

C. Web 3.0

It is also called the Semantic Web (SW). SW is another type of Web that builds above the existing version of the Web as shown in figure 1 and is the vision for the coming World Wide Web [4]. It needs structured documents that a machine can process by querying or inferencing to derive more precise information. Tim Berners Lee, the inventor of the Web, coined the term Semantic Web for a Web of meaning that makes it understandable to machines rather than just readable by machines. As such, the Web has been developing toward this vision by embedding huge quantities of machine-processable metadata, structure and different semantic Web technologies into the current Web. Basically, the Semantic Web tries to shift the thinking of published data in the form of Webpages (i.e., HTML documents) to allow machines to understand the contents. The content of Web 1.0 and Web 2.0 suffered from a number of issues, including the amount of information and how to access it and enable delegation [16]. They were provided for humans rather than for comprehension by machines. It therefore was not easy to automate data across the Web. To this end, the key idea behind the Semantic Web is to identify and link the content of the Web in a way that allows machines to understand and derive meaning from the data. Recognizing this vision requires new approaches, languages, technologies and data representation models to be built. For this reason, a variety of semantic languages and standards are maturing, and different applications, tools, and services are developing as shown in figure 2 below. In the case of implementing the Semantic Web for a single website, this implementation does not help much. Rather, the multiple of sources should be semantically structured to continue the chain of multiple websites in order to reach the required information that serves the ultimate goal of the Semantic Web. Data on Web 1.0 and Web 2.0 are about connecting information. However, with Web 3.0, it is about connecting knowledge and semantically structuring documents.

1) Document Type

The most attention on Web 3.0 was devoted to work on structured documents. Structuring documents includes machine-processable format such Resource Descriptive Language (RDF) or Web Ontology Language (OWL). RDF acts as a data model used to manage, structure and reason about the data found on the Web, and to show how the data relate in reality [17]. In this manner, RDF is a way to represent a small chunks of knowledge and how they are related to each other something lacking from the Web 2.0 and XML particularly. RDF is a graph data model (fig 2). In RDF, the graph serves to define the massive collection of triples in graph form. It is labeled directed graphs that represent statements as a set of nodes forming a network of information and how they are related to each other. Ontology languages such as RDF Schema (RDFS) and OWL are semantic Web languages that provide semantic meaning for RDF data. From the document type points of view, OWL represents semantics in the same way as RDF (structured way) but has more classes and properties to add semantic richness to RDF. Therefore, RDFS and OWL are data modeling for displaying RDF data in a structure ontology-based format.

2) Technologies in Place

The existing Web 2.0 utilizes natural language, multimedia, files, graphics, and much more so that it is easy for people to read and process information; however, with Web 2.0 it is difficult for machines to derive meaning from the information. It is not an easy task for computers to traverse the data meaningfully and even for humans to locate related information. In this context, technologies must be developed to realize the vision of Web 3.0. Some important technologies in the Semantic Web era are SPARQL Protocol and RDF Query Language (SPARQL), LINKED DATA, RDFaCE (RDFa Content Editor), JSON, and DBpedia Spotlight. SPARQL, Along with previous ontology languages (RDFS and OWL), is one of the three main semantic technologies. It is a query language for semantic datasets that works via pattern matching. Moreover, SPARQL not only provides facilities to query against data, but it also queries against the semantic schema. This query language can work with disconnected datasets (e.g., Linked Data) that are already mapped by RDF in order to retrieve data and obtain results. In this way, Linked Data, although it is still being developed, links different RDF datasets.

Furthermore, one of the important Web 3.0 technologies is RDFa (RDF in Attributes). It is a

technology for embedding and serializing data between XHTML tags. Therefore, it links the technologies from Web 2.0 with Web 3.0. Recently, big search engines like Google, Yahoo and Bing have been using RDFa to collect and integrate data from web sites. Hence, RDFa helps, for example, Google to provide rich semantics data to express concepts to the user in a few lines under every search result (Google Rich Snippets). Another company using RDFa is Facebook. It uses Open Graph Protocol to display information contained on web sites. Having considered RDFa, it is also reasonable to look at JSON (JavaScript Object Notation). JSON is used to interchange and process data and appears to be an alternative to XML. Consequently, it is another technology that can combine the two eras (Web 2.0 and Web 3.0) by loading data that is represented in Web 2.0 (e.g., Ajax, XML, XHTML) more quickly to avoid delays in site rendering.

4. Future Trends: Can the future be predicated?

Ten years from now, "semantic technology could be as ubiquitous as the Web is today, and combined with the capabilities of Web scalability, real-time reasoning, and world-scale knowledge management, there are both exciting new possibilities ahead as well as brave new challenges." [16].

A. Web 4.0

Although as of yet there is no clear technology for Web 4.0, but it is widely believed that Web 4.0 will be a "symbiotic Web" or "WebOS" that is a read-write-execution-concurrency Web [5] [18] that will handle structured documents. Research in [5] mentions that the Web is moving into artificial intelligence. Personalized agents can work smoothly with users to improve user experiences. These agents will be able to make a decision based on what a user wants, based on current conditions, along with other information to create an ideal response.

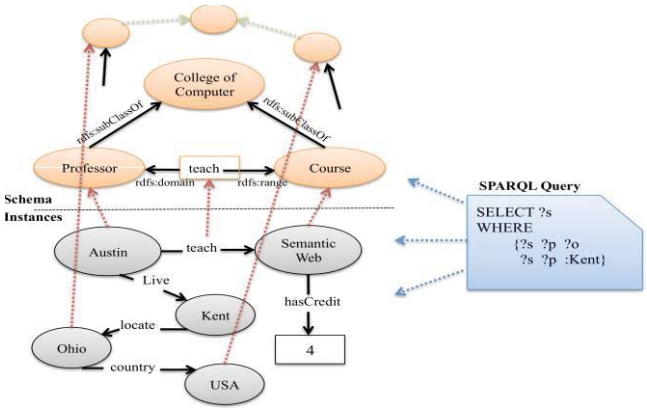


Figure 2: RDF/S graph

This all can be done with the power of data modeling, such as RDF/OWL, that allows reuse, integration, and reasoning. Web 4.0 is expected to handle complex intelligent interactions over the Web. As defined by [6], one of the most critical developments of Web 4.0 will be the migration of online functionality into the physical world. Daniel Burrus [20] mentioned that Web 4.0 is beginning already; it is about an "ultra-intelligent electronic agent" and personalized intelligent agent anticipated to be on every device. For example, according to Burrus, this personal intelligent agent says: "Good morning. You're flying to Boston today; take a raincoat, it's raining. By the way, that fight you were taking, it's already been canceled. Don't worry about it. There was a mechanical problem. I've already booked you on a new one, I'll tell you about on the way to the airport. But remember you're going to exercise everyday and I'm here to remind you that you're going to exercise." And you might say, "I don't know if I want to exercise today, and it'll show you a nude profile of yourself. And you'll say, 'You know what, I think I'm going to exercise today.'" The agent would be able to tell you what you have not asked for but what you should have asked for.

Having considered the rich ontologies and knowledge sharing in Web 4.0, it is also reasonable to look at automated and on-the-fly reasoning. It is expected to be an essential part of the ubiquitous Web to enhance the scalability of reasoning whether using heavyweight or lightweight reasoning. This advance will lead to more flexibility and effectiveness in cloud computing with new operating systems (OS), called WebOS or OS in the cloud [16]. WebOS with the OS and all its functionalities and contents such as data, applications and documents in the cloud, will all be accessible in one place. For example, Google bought Nest Labs, a company for home automation, and recently introduced the Brillo Operating System that is an extension to the physical world [21]. We are expecting that it would open the prospects for the Web 4.0 era. Therefore, some challenges, such as large-scale Web, will be decreased as a result of applying a heavyweight reasoning with real-time reasoning over the WebOS.

It is important to understand that Web evolution is based on needs. That is, Web 2.0 complemented Web 1.0 by adding more functionality. Web 3.0 solves major issues in Web 2.0. Thus, Web 4.0 should enhance the user experience of Web 3.0. As Web 2.0 enhanced the functionality of Web 1.0, Web 4.0 is expected to enhance the functionality of Web 3.0. The main reason is that Web 4.0 will more affectively process semantic structure. Web 4.0 will merge semantic Web technologies to produce agent-based applications that provide products that work as operating systems to facilitate the interaction between humans and machines. Web 4.0 is expected to introduce intelligent applications based on Semantic Web technologies. We think that Web 3.0 is providing the infrastructure to develop Web 4.0.

B. Web 5.0

Paper [9] thinks of the fifth generation (Web 5.0) as a "Sensory Emotive Web." Web 5.0 takes into account the feelings of the user. It is guided by technologies that already exist to measure feelings and their effects. As an example, a company called Emotiv Systems [22] works in the field of neurotechnology. With headphones, the human brain can communicate with a machine. The machine can read conscious thoughts, emotions, facial

expressions and head rotations. Research in [6] describes Web 5.0 as a "Symbionet Web.". They mention it is "emotionally" neutral and notes that it does not count on a user's feelings. We think it would consider structured data along with structured documents. Also, Web 5.0 may take advantage of data fusion algorithms and applications to merge with pervious Web generations. The data fusion field has already proven successful in a number of domains, including discovery science and business intelligence, and it may be able to work with multiagent systems [24]. We think that Web 5.0 is about "Fusion Web" where machines and people will process data in forms that they can deal with, interact with, and make decisions with. Web 5.0 could be a read-write-execute-concurrent-fusion Web. We may see Web 5.0 or "Fusion Web" join Web For All [25] to support people with special needs. Web 5.0 is expected to be built on the power of Semantic Web "Web 3.0" and Symbolic Web "Web 4.0."

C. Web Current Status and Concerns

At the last W3C 20th anniversary symposium [26], a number of topics were covered, including a long-term view of the World Wide Web and Access for All, and a panel discussion session titled "The Future of the Web and How It Is Run." On this panel, Web fragmentation was a hot topic. Web fragmentation is a topic that concerns Web leaders, including Sir Tim (Web inventor) and Vint Cerf (Internet co-father) and other field pioneers. Technically, Internet fragmentation is about having more than one Internet that results in Web fragmentation. In business, The World Web Consortium (W3C) has considered the Digital Rights Management (DRM) in HTML, which includes Encrypted Media Extension (EME) [27] specifications to prevent fragmenting the Web into free-Web and charged-Web. Jeff Jaffe mentioned that "It is W3C's overwhelming responsibility to pursue broad interoperability, so that people can share information, whether content is protected or available at no charge." Fragmenting the Web also was mentioned at a government level when German Chancellor Angela Merkel called for the European Union to create its own regional Internet for security concerns [28]. Eugene Kaspersky, chairman and CEO of Kaspersky Lab, mentioned that Internet fragmentation means the end of the Web by saying, "But I fear that we are at a turning point for the internet, and may even be going into reverse. The utopia of a borderless digital global village may be coming to an end. Fragmentation of the world wide Web is already taking place—along national borders" [29].

Also, it is worth mentioning that the Semantic Web is growing exponentially, and its technologies are refining and developing fast from many different sources every day. Some visionaries [16] of the Semantic Web predict that in the next few years, Semantic technologies will be matured enough to act as valuable sources of structured documents for different applications and wider integration of Web content. In addition to do inference on the given knowledge for those applications, semantic technologies also will have the ability to provide data provenance, which will allow for tracing and verifying the sources of information. Consequently, we foresee that different types of application domains will be enhanced by semantic research, including health care, smart life, mobile technology, energy, and knowledge discovery.

As a reason for developing the Semantic Web technologies and for increasing the amount of structured document is, we expect that most Web content will be semantically marked up so that metadata will become increasingly important to reach the vision of the Semantic Web. Therefore, applying ontologies over these structured data will be a new gold to do better reasoning.

Finally, the Semantic Web raises some new issues and challenges in research, such as the availability of content, scalability and the stability of Semantic Web languages [31]. Some progress has been made in overcoming them as in ontology learning [43][44][19][30], storing RDF graph [17][32][33][34][35][36] and standardizing necessary technology [37]. However, some remaining challenges are still open for research, such as maintaining machine processable data and providing some mechanism that supports engineering tasks for ontologies [31][38]. Also, some challenges are mentioned in [39], such as vastness with data redundancy and inconsistency when ontologies from different parties are combined. Also, some researchers [40][41][42] have showed that some semantic security policies are needed for the evolving Semantic Web. Last, we think that in order to rebound the growth of the Semantic Web, current and new Websites need to seriously consider publishing the semantic part of the Website along with its deployment.

5. Conclusion

This paper surveyed the Web generations: approaches, technologies, and general trends. Conversion from Web 1.0 to Web 2.0 is optional, as it depends on the need to have an interactive website. Rather, we think conversion from Web 1.0 or Web 2.0 to Web 3.0 is necessary. Some solutions have been proposed to help in reaching Webs (e.g., RDFaCE). Websites can be part of many Web generations [8]. Basically, the Semantic Web tries to shift the thinking of published data in the form of webpages (i.e., HTML documents) to allow machines to understand the contents [16]. That is, computers are able to interoperate and think on our behalf.

In the next Web (Web 4.0) or symbiotic Web, knowledge is expected to be structured well. Therefore, ontologies will be the new gold; namely, ending up with an enriched and efficient conceptualization analysis of a specific area of interest. This will lead to high knowledge sharing over databases and documents on the Web. Therefore, agent-based applications will be able to intelligently reason and perform tasks based on metadata. Moreover, experts in the semantic Web foresee that intelligent software agents will be interacting with humans in symbiosis [16]. Thus, the interplay between Web 3.0 and Web 4.0 will be a key technology for the growth of knowledge and for making the Web more useful for humanity.

REFERENCES

[1] T. Berners Lee, R. Cailliau, J. Groff, and B. Pollermann, "World wide web: the information universe," *Internet Res.*, vol. 20, no. 4, pp. 461–471, Aug. 2010.

[2] A. Barforush, A. A., & Rahnama, "ONTOLOGY LEARNING: REVISTED," *J. Web Eng.*, vol. 11, no. 4, pp. 269–289, 2012.

[3] G. A. Miller, "WordNet: a lexical database for English," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, Nov. 1995.

[4] O. Berners-Lee, T., Hendler, J., & Lassila, "The Semantic Web," *Sci. Am.*, vol. 21, pp. 28–37, 2001.

[5] S. Aghaei, M. A. Nematbakhsh, and H. K. Farsani, "EVOLUTION OF THE WORLD WIDE WEB : FROM Web 1.0 to Web 4.0," vol. 3, no. 1, pp. 1–10, 2012.

[6] C. Science and S. Engineering, "Incremental Journey for World Wide Web : Introduced with Web 1 . 0 to Recent Web 5 . 0 – A Survey Paper," *Int. J. Adv. Res. Comput. Sci. Softw. Eng. Res.*, vol. 3, no. 10, pp. 410–417, 2013.

[7] J. Weber, S., & Rech, "An Overview and Differentiation of the Evolutionary Steps of the Web XY Movement: The Web Before and Beyond 2.0," *Handb. Res. Web*, 2009.

[8] G. Cormode and B. Krishnamurthy, "Key differences between Web 1.0 and Web 2.0," *First Monday*, vol. 13, no. 6, Apr. 2009.

[9] A. Benito-Osorio, D., Peris-Ortiz, M., Armengot, C. R., & Colino, "Web 5.0: the future of emotional competences in higher education," *Futur. Emot. competences High. Educ.*, no. Global Business Perspectives, pp. 274–287, 2013.

[10] "W3C HTML, The Web's Core Language." [Online]. Available: http://www.w3.org/html/. [Accessed: 28-May-2015].

[11] "The WWW Virtual Library." [Online]. Available: http://vlib.org/. [Accessed: 28-May-2015].

[12] B. Pinkerton, "Finding What People Want: Experiences with the WebCrawler," *Proc. Second Int. World Wide Web Conf.*, 1994.

[13] D. Eichmann, "The RBSE spider-balancing effective search against web load," *Proc. 1st WWW Conf*, 1994.

[14] D. N. Shah, *A Complete Guide to Internet And Web Programming - Computer Science*. Dreamtech Press, 2009.

[15] "Html Agility Pack." [Online]. Available: http://htmlagilitypack.codeplex.com/. [Accessed: 28-May-2015].

[16] J. A. Fensel, D., Domingue, J., & Hendler, *Handbook of Semantic Web Technologies: Foundations and technologies*. Springer Berlin Heidelberg, 2011.

[17] S. Albahli and A. Melton, "ohStore : Ontology hierarchy solution to improve RDF data management," *9th Int. Conf. Internet Technol. Secur. Trans. ICITST-2014 IEEE*, 2014.

[18] "How the WebOS Evolves? | Nova Spivack." [Online]. Available: http://www.novaspivack.com/technology/how-the-webos-evolves. [Accessed: 28-May-2015].

[19] M. Hazman, S. El-Beltagy, and A. Rafea, "Ontology learning from domain specific web documents," *Int. J. Metadata, Semant. Ontol.*, vol. 4, no. 1, pp. 24–33, 2009.

[20] "From Web 3.0 to Web 4.0 | Big Think." [Online]. Available: http://bigthink.com/videos/from-web-30-to-web-40. [Accessed: 28-May-2015].

[21] "Google takes aim at the internet of things with new Brillo operating system." [Online]. Available: http://www.theguardian.com/technology/2015/may/28/google-brillo-operating-system-internet-of-things. [Accessed: 30-May-2015].

[22] "Emotiv | EEG System ." [Online]. Available: http://emotiv.com/. [Accessed: 28-May-2015].

[23] L. A. Klein, *Sensor and data fusion: a tool for information assessment and decision making*. 2004.

[24] M. Wooldridge, *An Introduction to MultiAgent Systems*. John Wiley & Sons, 2009.

[25] "Accessibility - W3C." [Online]. Available: http://www.w3.org/standards/webdesign/accessibility. [Accessed: 28-May-2015].

[26] "W3C20 Anniversary Symposium." [Online]. Available: http://www.w3.org/20/. [Accessed: 28-May-2015].

[27] "Encrypted Media Extensions." [Online]. Available: http://www.w3.org/TR/encrypted-media/. [Accessed: 28-May-2015].

[28] "The End of the Internet? ." [Online]. Available: http://www.theatlantic.com/magazine/archive/2014/07/the-end-of-the-internet/372301/?single_page=true. [Accessed: 28-May-2015].

[29] "What will happen if countries carve up the internet?" [Online]. Available: http://www.theguardian.com/media-network/media-network-blog/2013/dec/17/internet-fragmentation-eugene-kaspersky. [Accessed: 28-May-2015].

[30] Nicolas Weber and Paul Buitelaar, "Web-based ontology learning with ISOLDE," Processing of the Workshop on Web Content Mining with Human Language at International Semantic Web Conf., USA, 2006.

[31] Richard, J. Contreras, O. Corcho, and A. Gómez-Pérez, "Six Challenges for the Semantic Web," 2002.

[32] D. J. Abadi, S. R. Madden, and K. Hollenbach, "Scalable Semantic Web Data Management Using Vertical Partitioning," 2007.

[33] C. Weiss and A. Bernstein, "Hexastore : Sextuple Indexing for Semantic Web Data Management," pp. 1008–1019, 2008.

[34] D. Wilkinson, K., Sayers, C., Kuno, H. A., & Reynolds, "Efficient RDF Storage and Retrieval in Jena2," *SWDB*, pp. 131–150, 2003.

[35] M. Wylot, P. Cudré-Mauroux, and P. Groth, "TripleProv: Efficient Processing of Lineage Queries in a Native RDF Store," *Int. World Wide Web Conf. Steer. Comm.*, pp. 455–466, 2014.

[36] M. Atre, V. Chaoji, J. Weaver, and G. T. Williams, "BitMat : An In-core RDF Graph Store for Join Query Processing," *Rensselaer Polytech. Inst. Tech. Rep.*, 2009.

[37] "Standards - W3C." [Online]. Available: http://www.w3.org/standards/. [Accessed: 28-May-2015].

[38] G. Antoniou and F. Van Harmelen, *A Semantic Web Primer*. the MIT Press., 2012.

[39] N. Choudhury, "World Wide Web and Its Journey from Web 1 . 0 to Web 4.0," *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 6, pp. 8096–8100, 2014.

[40] B. Thuraisingham, "Security Issues for the Semantic Web," in *37th Annual Computer Software and Applications Conference*, 2013.

[41] L. Kagal, T. Finin, M. Paolucci, K. Sycara, and G. Denker, "Authorization and privacy for semantic Web services," *IEEE Intell. Syst.*, vol. 19, no. 4, pp. 50–56, Jul. 2004.

[42] G. Denker, L. Kagal, and T. Finin, "Security in the Semantic Web using OWL," *Inf. Secur. Tech. Rep.*, pp. 51–58, Jan. 2005.

[43] A. A. Algosaibi and A. C. Melton, "Using the Semantics Inherent in Sitemaps to Learn Ontologies," in *2014 IEEE 38th International Computer Software and Applications Conference Workshops*, 2014, pp. 360–365.

[44] R. Zarrad, N. Doggaz, and E. Zagrouba, "Concepts Extraction based on HTML Documents Structure.," *ICAART (1)*, pp. 503–506, 2012.

Web Accessibility, Litigation and Complaint Mechanisms

Justin Brown[1], **Scott Hollier**[2]
[1] School of Computer & Security Science, Edith Cowan University, Perth, Western Australia
Email: j.brown@ecu.edu.au
[2] School of Computer & Security Science, Edith Cowan University, Perth, Western Australia
Email: s.hollier@ecu.edu.au

Abstract – *This paper examines three well known web accessibility litigation cases, their causes and outcomes. Web accessibility focuses on how websites can be designed to be usable by all netizens, including those with disabilities. The Web Content Accessibility Guidelines and assistive technologies work together to allow websites and content to be access by people with a wide range of disabilities, however such technologies only work when content and technology are in sync. Whilst accessible technologies have become highly available, development of websites according to accessibility standards is still fragmentary, and in too many cases litigation is required to bring about change. After examining three such cases of litigation the authors discuss issues around the accessibility of complaint mechanisms and conclude with figures indicating that the number of people with disabilities should make corporate and government website owners consider the value of an accessible web.*

Keywords: web accessibility, litigation, disability, assistive technology, wcag,

1 Introduction

The provision of a global World Wide Web (WWW) which can be used by all is hardly a new concept. As noted by WWW inventor Sir Tim Berners-Lee, "The power of the Web is in its universality. Access by everyone regardless of disability is an essential aspect." [1]. While the desire to ensure an effective web is present, and both consumers and developers have the mechanisms needed to address access issues, there appears to be a reluctance to fully embrace accessibility solutions, resulting in ongoing complaints and legal action undertaken by individuals who continue to struggle with the basics of online services due to web accessibility barriers.

This paper reflects on the common themes in significant legal precedents relating to web accessibility, the implications for business and how the effective implementation of an internationally-recognised web accessibility standard, in partnership with an effective complaints process, can improve the access of the web today and its availability to all users.

2 The Web for People with Disabilities

For people with disabilities, the web requires two elements to work in harmony together. The first is the assistive technology required by a person with a disability to use a computer or mobile device, the second is for the content being used by such technologies to be built to relevant web accessibility standards. Traditionally assistive technologies were expensive products and this was in itself a barrier, but in recent years developers of mainstream devices such as Microsoft operating systems and Surface tablets, Apples iPhone and iPad products and Google's Android devices, have included accessibility features into their products, including screen readers, on-screen keyboards, switch keys, magnifiers, high contrast settings and captioned video support [2]. While the provision of mainstream accessibility features has significantly improved the availability and affordability of equipment for people with disabilities to engage with the web, the effectiveness of such tools relies on web, and more recently app content, being created and deployed in an accessible way [3-6].

The current standard created by the World Wide Web Consortium (W3C) and recognised by the International Standards Organisation (ISO) are the Web Content Accessibility Guidelines (WCAG) 2.0 [7], or in their standards recognised form, ISO/IEC 40500 [8]. WCAG 2.0 consists of 12 guidelines designed to assist developers in ensuring that their web content will work effectively for people with disabilities and their assistive technologies, featuring key requirements such as the need to ensure that alternative text is present for visual content, that videos feature captions, keyboard shortcuts can be used, content is easy to find and clear guidance is provided to correct mistakes entered into web pages such as through the use of web forms [9]. The guidelines have been widely adopted across international policy and legislative frameworks [10-13] and the technology-neutral approach of the guidelines, combined with the specific success criteria and supporting techniques documents have allowed the standard to remain current and beneficial since their introduction in 2008. However, as the works of Harper and Chen indicate, whilst new web technologies can be implemented by large organisations in short timeframes, accessibility uptake rates

can be much lower over considerably longer periods of time [14].

3 Legal Challenges and Complaints

While the knowledge base on how to create accessible websites based on standards compliance is clear and ever expanding, people with disabilities have continued to face significant accessibility challenges which has led in some cases to legal action [15-19]. Arguably the first notable instance of web accessibility litigation was the case of Maguire v Sydney Organizing Committee for the Olympic Games (SOCOG) [16, 20]. The case focused on Bruce Maguire, a blind person who required ticketing and event information for the Sydney 2000 Olympic Games. However, the information on the Olympic Games website was inaccessible, primarily due to the use of images without alternative text. As a result, Maguire was unable to navigate or access much of the content on the website using screen reader software.

While Maguire raised his concerns with SOCOG he was informed that there were no plans to rectify the issue. As a result, the case was taken to the Human Rights and Equal Opportunity Commission (HREOC), with the argument being put forward that the Olympic website needed to be compliant with the WCAG standard (WCAG 1.0 at the time), and failure to do so would put SOCOG in breach of Section 24 of the Disability Discrimination Act 1992 (DDA). SOCOG initially stated that making all the webpages accessible was too onerous with the Olympics approaching but this argument was rejected as most of the webpages were based on a few key templates which had not been used due to the Olympics not having commenced at the time of the complaint. SOCOG then tried to blame IBM as their web development firm but this was also rejected as it was viewed as SOCOG's responsibility. The ruling found that SOCOG had engaged in conduct that is unlawful under Australia's DDA. However, the website was not fixed, with SOCOG opting to provide Maguire with monetary compensation. It did, however, lead to a notable improvement in the accessibility of future Olympic Games websites.

While the Maguire v SOCOG case received much attention due to the Olympics, it was the case against a large corporation in the USA that is considered particularly relevant due to the country having specific legislation in the form of the Rehabilitation Act of 1973, Section 508 [21] in which major components of the original WCAG 1.0 are legislated as to their applicability within procurement procedures of the US Federal government. While this case, the Target.com v National Federation of the Blind was not specifically relevant to Section 508, it did demonstrate the importance of web accessibility as a fundamental requirement in relation to the Americans with Disabilities Act 1990 (ADA) [22]. The case began in 2006 when the National Federation of the Blind (NFB) took on Target [23] arguing that the inaccessibility of target.com was in breach of the ADA. Unlike the Maguire V SOCOG case, the initial strategy was to discredit the complainant by suggesting that others could use the website without issue. This ultimately led to the case being moved from state to Federal courts and focused on how a court could determine whether a website was accessible or not beyond just user opinion.

Again the result was a combination of issues raised as part of WCAG, and real-world demonstrations of people trying to complete tasks but failing due to issues such as missing alternative text in images and the need to specifically use a mouse to access the 'continue' option [24] As a result, Target addressed the accessibility issue of their website and had to pay $US6 million to compensate to members of the original Californian subclass suit and nearly $US4 million in the plaintiff's legal fees [17].

In Canada another legal case, and arguably the most significant case since WCAG was updated to version 2.0 in 2008 is Donna Jodhan V Attorney General of Canada [25-27]. The primary focus of this case was the inability of a blind woman to apply for government jobs due to the inaccessibility of the websites that hosted the job application information. The arguments were based around the Canadian government's 2001 Common Look and Feel (CLF) requirement that specify that all government information services must have the same look and feel, and as part of that process Internet materials must be available to people with vision impairment. While Jodhan won the case in 2010, the Federal government appealed [28] arguing that while the web content may not have been accessible, the government was only obligated to provide the information through alternatives, and that Jodhan had the opportunity to receive the information via alternative means. This argument was ultimately rejected, with the court finding that Jodhan had been discriminated against as the CLF requirements should ensure that all web users can access the content and that non-web alternatives should not be necessary for web content if WCAG 2.0 were correctly applied. This last point is particular salient in the views of these authors as for an ever increasing number of web-only businesses, there may not be non-web alternatives to online processes.

There are a number of websites which report web accessibility related legal cases and the outcomes [29, 30], many of which have been settled out of court such as an Australian case against the supermarket giant Coles around the accessibility of its website [31]. In a most cases where accessibility complaints end up in a court of law the defending organisation has been initially requested to fix accessibility issues and subsequently refused, typically citing cost, hardship or a perceived lack of legal obligation as the reason for their lack of accessibility compliance [32]. Such behavior seems shortsighted considering the size and profile of the organisations involved, especially if one takes into account the perceived reputational damage such accessibility

obstinacy may cause in the long term versus the short term costs of fixing selected parts of a website.

4 Common Themes

The legal cases discussed here, and other cases settled out of course, highlight two notable themes. The first is the importance of web standards, the second relates to the complaints process. In all cases the legal issues highlighted that the websites involved did not conform to the established web accessibility standard of the time, either WCAG 1.0 or 2.0, meaning that people with disabilities were unable to use their assistive technologies to access the web content. While it's not uncommon for websites to have some accessibility issues, what distinguishes these examples is that the inaccessibility of the web content did not only prevent access to information, but prevented user from completing a particular task. In the case of the Maguire case it was the purchasing of Olympics tickets, for Target USA it was the ability to buy products from their online store, and in the case of Jodhan v Canadian government it was the ability to apply for a job. While web accessibility is an important issue, its implementation becomes much more polarized if the accessibility issues do not allow for a work around of some kind. This is particularly apparent when a business exists only online (for example eBay, Amazon) with no bricks and mortar equivalent, or where formerly manual processes move online, and online only, meaning that a service which might have been partially or fully accessible to a person with a disability becomes completely inaccessible. Telephone directories, job advertisements, ticket sales and bill payments are examples of services which are more and more moving to the web as their primary interface with clients. In the cases outlined above, if those websites were built to comply with WCAG it's unlikely these issues would have occurred as guidelines 1.3 and 4.1 of the current WCAG 2.0 guidelines specifically focus on the need to ensure that processes work for assistive technology users and that coding is correctly implemented to ensure its success.

The second part relates to the complaints process. In these instances when a complaint was made, the cases highlighted that it was often initially difficult to make a complaint due to no one in an organization being designated to provide support for web accessibility issues. Once a complainant had persevered and managed to raise their concerns, the issue was either largely dismissed, arguments were made that other users could access the information and thus was an isolated issue or in the case of the Jodhan case, non-web solutions were deemed to be adequate. The commonality among all these cases is that none of them acknowledged the right for people with disabilities to access web content, and the vital adherence to web accessibility standards that would ensure the web content could be ubiquitously accessed.

It is of note that once WCAG 2.0 is implemented, attitudes can change quickly and what was once seen as a barrier can be promoted as best practice. In the case of Target USA, The NFB provided an award to Target for its excellence in accessibility and its screen reader compatibility, changes which were implemented as a direct result of legal action. This demonstrates that the need to ensure accessible web content is in place is less about the desire of users to make complaints and far more about the need to ensure effective access for all by the content provider.

5 Barriers to Complaints

Thus far this far this paper has examined the outcomes of accessibility complaints not being dealt with speed or empathy. However, sometimes how a complaint is dealt with can be less pressing that the actual ability to submit a complaint or raise a website concern in the first place. Some very large web based organisations do not provide general 'contact us' email addresses or web forms to which users can raise their concerns as to a website's accessibility. To demonstrate this issue, Figures 1 and 2 illustrate the process of contacting eBay (Figure 1) and Amazon.com (Figure 2) in order to raise an issue. These examples are not a reflection on the accessibility of these websites, but rather examples of how very large web-only organisations can use highly structured, drill-down processes in order to filter user communication or complaints. It is perhaps worth noting that Amazon services power the websites of a number of large corporate e-commerce sites, including target.com [33]. In Figure 1 we can see that in terms of eBay users wishing to contact that organisation they must choose from a list of categories, and in this case the author was attempting to select an option which would allow an email to be sent. In terms of categories, the only one which seemed even slightly relevant to a possible website usability concern was 'Reporting other problems' under the 'Reporting another user' category, which in itself was not very informative. Selecting this option brought up a list of headings indicating what a user might do to resolve an issue, at the bottom of which lies a query asking the user if their issue had been resolved, with a Yes or No button for input. If the user selects No, then Other from the pop-up menu which then appears, they then receive a tiny text box into which they can enter 100 characters describing their issue. Given that the key terms of 'web accessibility wcag 2.0' take up 27 characters on their own, entering a meaningful accessibility concern using such a process seems unlikely to produce a positive outcome.

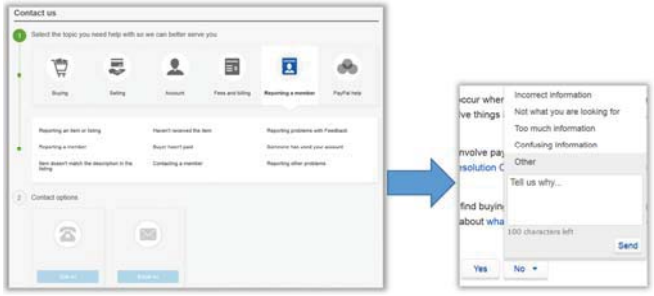


Figure 1: Contact process in eBay

Similarly, the Amazon.com 'contact us' process is also highly structured with categorised contact subjects and a series of drill-down form fields, which does lead to an actual email contact form for website related concerns if the correct combination of options are selected in the correct order.

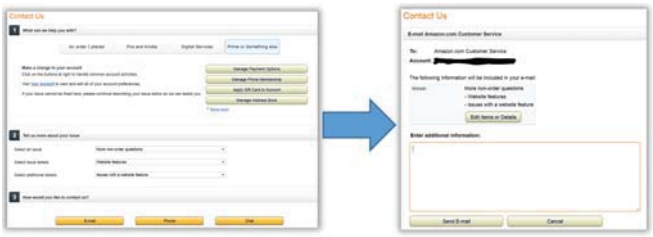


Figure 2: Contact process in Amazon.com

These two examples demonstrate that in some websites the process of making contact with the site owner can be either convoluted or not related to website usability, or both. In other sites there may be no point of contact whatsoever [33], or if there are, they can be form-based and not designed with accessibility in mind, or worse, use CAPTCHA technologies [34-38] to verify the veracity of the human user. If users of assistive technologies find it difficult if not impossible to locate or utilise an avenue of communication with a website owner, then the likelihood of that user contacting external bodies, such as members of the legal fraternity or disability advocacy bodies increases in relation to the users level of frustration and perception of being discriminated against. What started as a desire to seek further assistance or register a complaint can rapidly escalate into a rallying cry for change, as seen in the target.com and Jodhan cases.

6 Conclusion

The issues covered in this paper are complex in that they do not focus purely on the accessibility and usability of websites, but also on the ability of site users to have their accessibility issues heard and recognised if and when they occur. In the legal cases discussed above, site users encountered problems in both communicating their accessibility concerns and having those concerns taken seriously. Arguments made by site owners that they cannot change websites just to suit a single person or even a small group of individuals holds no weight in the eyes of the law, and ultimately, is not reflective of figures for disability in the wider community. In the U.S. figures indicate that between 12-22% of the population report some kind of disability [39], whilst internationally the figure is thought to be on the order of 15% of the global population [40]. These figures indicate that conservatively, one in six people are living with some type of disability, and that in order to access web services, they may require the use of one or more assistive technologies.

It is the view of these authors that the case for accessible web design is self-evident, including the attraction of new users, the retention of existing users, the preservation of corporate reputation [42] and above all, responding to user needs in a compassionate and equitable manner. All too frequently in the disability space, litigation becomes the channel of communication between those with specialist needs and governmental or corporate entities who are not listening. If accessible thinking informs website design and complaints handling processes, perhaps in the future litigation will no longer be the primary driver of change, as the leaders will lead and the followers will follow [43].

7 References

[1] World Wide Web Consortium. (2014, 17/5/2015). Web Accessibility Initiative. Available: http://www.w3.org/WAI/

[2] S. Hollier. (2013, 1/5/2015). 10 milstones in the mainstreaming of accessibility. Available: http://www.creativebloq.com/netmag/10-milestones-mainstreaming-accessibility-7135541

[3] A. Chuter, "Web Accessible and Mobile: The Relationship between Mobile Web Best Practices and Web Content Accessibility Guidelines." vol. 5105, ed Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 498-501.

[4] L. Knudsen and H. Holone, "A multimodal approach to accessible web content on smartphones," 2012, pp. 1-8.

[5] R. Mireia, P. Merce, B. Marc, T. Miquel, S. Andreu, and P. Pilar, "Web Content Accessibility Guidelines 2.0: A further step towards accessible digital information," Program, vol. 43, pp. 392-406, 2009.

[6] L. Möbus, "Making web content accessible for the deaf via sign language," Library Hi Tech, vol. 28, pp. 569-576, 2010.

[7] World Wide Web Consortium. (2008, 5/2/2015). Web Content Accessibility Guidelines. Available: http://www.w3.org/TR/WCAG20/

[8] International Standards Organisation. (2012, 14/2/2015). ISO/IEC 40500:2012 Information technology -- W3C Web Content Accessibility Guidelines (WCAG) 2.0. Available: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=58625

[9] World Wide Web Consortium. (2008). WCAG 2 at a Glance. Available: http://www.w3.org/WAI/WCAG20/glance/

[10]M. Rogers. (2010, 4/3/2015). Government Accessibility Standards and WCAG 2.0. Available: http://blog.powermapper.com/blog/post/Government-Accessibility-Standards.aspx

[11]L. Basdekis, L. Klironomos, L. Metaxas, and C. Stephanidis, "An overview of web accessibility in Greece: a comparative study 2004–2008," Universal Access in the Information Society, vol. 9, pp. 185-190, 2010.

[12]K. Faouzi, M. Basel, and B. Emad, "E-GOVERNMENT WEB ACCESSIBILITY: WCAG 1.0 VERSUS WCAG 2.0 COMPLIANCE," International Journal of Digital Information and Wireless Communications, vol. 3, pp. 56-65, 2014.

[13]F. Kamoun and M. Basel Almourad, "Accessibility as an integral factor in e-government web site evaluation," Information Technology & People, vol. 27, pp. 208-228, 2014.

[14]S. Harper and A. Chen, "Web accessibility guidelines: A lesson from the evolving Web," World Wide Web, vol. 15, pp. 61-88, 2012.

[15]S. Hollier. (2012, 11/2/2015). Do legal precedents help the accessibiltiy cause? Available: http://www.accessiq.org/news/w3c-column/2012/09/do-legal-precedents-help-the-accessibility-cause

[16]A. Arch and O. Burmeister, "Australian experiences with accessibility policies post the Sydney Olympic games," Information Technology and Disabilities, vol. 9, 2003.

[17]J. Grubbs, B. Brice, and S. Jennings, "AMERICANS WITH DISABILITIES ACT AND E-COMMERCE: TARGET CORPORATION AND BEYOND," Southern Law Journal, vol. 22, p. 89, 2012.

[18]S. Kretchmer and R. Carveth, "Analyzing recent Americans with disabilities act-based accessible information technology court challenges," Information Technology and Disabilities, vol. 9, 2003.

[19]B. Parmanto and S. Hackett, "A case study examination of the impact of lawsuits on website accessibility," Disability & Rehabilitation: Assistive Technology, vol. 6, pp. 157-168, 2011.

[20]Australian Human Rights Commission. (1999, 22/10/2014). Maguire v SOCOG. Available: https://www.humanrights.gov.au/maguire-v-socog

[21]United States General Services Administration. (2015, 3/2/2015). Section 508.gov: Opening Doors to IT. Available: http://www.section508.gov/

[22]U. S. D. o. o. Justice. (2015, 4/2/2015). Information and Technical Assistance on the Americans with Disabilities Act. Available: http://www.ada.gov/

[23]World Wide Web Consortium. (2009, 24/2/2015). A Cautionary Tale of Inaccessibility: Target Corporation. Available: http://www.w3.org/WAI/bcase/target-case-study

[24]J. Hatcher. (2011, 4/4/2015). Accessibility, Law and Target.com. Available: http://jimthatcher.com/law-target.htm

[25]O. o. t. C. f. F. J. A. Canada. (2011, 6/5/2015). Jodhan v. Canada (Attorney General). Available: http://recueil.cmf.gc.ca/eng/2011/2010fc1197.html

[26]Marketwired. (2014, Elections Canada Launches Advisory Group for Disability Issues. (Journal, Electronic). Available: http://ecu.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwrV1NSwMxEB20Xjyp-FW_mD9Q3N2syeaiVG3x4NGDt5KYWTwsW7VV6L_vJJuUivTmORACE-a9TGbeA0DBGK4oc2UtK-dnlkVRSZI2o6LOKDerRsxOKDr5tcbQpowY0rSbvvkK-TXTloKJrdT53cfnwHtG-b_VZKBhorGC_59hArwNOznjmr_16vV-nSyIm1L-SboBScZ7sKorxw6S2ECYpnF-KTT-21n3oc-g5AjXqnQ47C7NAWxRewi3oyZ0ZbUzDIoFBp8NY947zXDofrxG5wJDoQqZ5uJjlOadL7Dz7zuCq_Ho5eFpkA47cU0zEfz0YCbC7FAcQ6-dtnQK6LSxsmbSopglKaq0zbU2nASkNbYk6sPJhk3ONq6cwy7HogxT4OoCevOvb7oMzlBLMvemZg

[27]T. Nicki, "Blind must be able to access sites, judge tells Ottawa," in Toronto Star, ed. Toronto, Ont: Torstar Syndication Services, a Division of Toronto Star Newspapers Limited, 2010, p. A.8.

[28]Council of Canadians with Disabilities. (2013, 4/5/2015). Jodhan Decision Advances Access to Web Sites for Persons with Vision Impairment. Available: http://www.ccdonline.ca/en/blog/jodhan

[29]I. Accessibility. (22/5/2015). Lawsuites and Settlement Agreements. Available: http://www.interactiveaccessibility.com/lawsuits-settlement-agreements?page=1

[30]K. Grovers. (2011, 3/4/2015). List of Web Accessibility-Related Litigation and Settlements. Available: http://www.karlgroves.com/2011/11/15/list-of-web-accessibility-related-litigation-and-settlements/

[31]S. Hollier. (2015, 4/5/2015). The Coles web accessibility case - two important lessons for corporate Australia. Available: http://www.accessiq.org/news/w3ccolumn/2015/02/the-coles-web-accessibility-case-two-important-lessons-for-corporate

[32]J. Palazzolo. (2014, 2/4/2015). Disabled Sue Over Web Shopping. Available:

http://www.wsj.com/articles/SB10001424127887324373204578374483679498140

[33]C. Peters and D. Bradbard, "Web accessibility: an introduction and ethical implications," Journal of Information, Communication and Ethics in Society, vol. 8, pp. 206-232, 2010.

[34]M. Evans, "Mark Evans: About Us & Contact Us: Website Workhorses," ed. Chatham: Newstex, 2013.

[35]B. Kelly, L. Nevile, D. Sloan, S. Fanou, R. Ellison, and L. Herrod, "From Web accessibility to Web adaptability," Disability & Rehabilitation: Assistive Technology, vol. 4, pp. 212-226, 2009.

[36]C. Pribeanu, P. Fogarassy-Neszly, and A. Pătru, "Municipal web sites accessibility and usability for blind users: preliminary results from a pilot study," Universal Access in the Information Society, vol. 13, pp. 339-349, 2014.

[37]M. Ribera, M. Porras, M. Boldu, M. Termens, A. Sule, and P. Paris, "Web Content Accessibility Guidelines 2.0," Program, vol. 43, pp. 392-406, 2009.

[38]E. Valdez, O. Martinez, G. Fernandez, L. Aguilar, and J. Lovelle, "Security guidelines for the development of accessible web applications through the implementation of intelligent systems," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 1, pp. 79-86, 2009.

[39]B. Wentz, H. Hochheiser, and J. Lazar, "A survey of blind users on the usability of email applications," Universal Access in the Information Society, vol. 12, pp. 327-336, 2013.

[40]D. Brucker and A. Houtenville, "People with disabilities in the United States," Archives of physical medicine and rehabilitation, vol. 96, pp. 771-774, 2015.

[41]The Lancet, "World Report on Disability," The Lancet, vol. 377, pp. 1977-1977, 2011.

[42]T. Coombs and S. Holladay, "Amazon.com's Orwellian nightmare: exploring apology in an online environment," Journal of communication management, vol. 16, pp. 280-295, 2012.

[43]R. Gonçalves, J. Martins, J. Pereira, M. Oliveira, and J. Ferreira, "Enterprise web accessibility levels amongst the Forbes 250: where art thou o virtuous leader?," Journal of business ethics, vol. 113, pp. 363-375, 2013.

Customized Multimedia Google form for Education – A User Friendly Approach

Chandani Shrestha and Hong Jiang
Benedict College, Columbia, SC

Abstract - *Google form is an extensive and cost free platform to create online homework / quiz and manage educational work. However, the basic form only allows the user to edit text information or embed pictures. It will be a problem if the user wants to include mathematic formula, table, hyperlink, YouTube video, etc. To provide a solution to this problem, this undergraduate research uses HTML code to represent customized multimedia content; uses Google Apps Script to code the related functions to generate an updated multimedia Google form, and save the form in a publically shared Google drive. Considering that not all users are familiar with HTML code for their desired multimedia content, free HTML online editor is recommended to generate desired multimedia HTML code, and the users can simply copy / paste the generated HTML code to the basic Google form in a predefined symbol such as "[[" and "]]". User will get the updated Google form and a link to share by clicking the designed web-service in the menu. This approach creates a tailored Google form and provides a user-friendly free / low cost solution to enable the use of graph, links, videos, images, tables and many other formats in online quiz and homework system.*

Keywords: Google Docs, Apps Script, HTML, Multimedia Form, Online Homework/Quiz System

1 Introduction

The traditional technique of giving out works by teachers and submitting class works by students is usually in papers. It has long started fading away. With the advancement in technology, the education system has become more dependent on technological innovations than ever.

We were searching for a platform, which allows us to create a low cost online homework system without encumbering the educator with the burden of being an expert programmer or technology savvy. And it brought us to the use of Google Docs. Google Docs is one medium that can be used to create an effective Online Homework / Quiz System in little time with reasonable effort. However, it comes with a drawback: the basic form only allows the user to edit text information or embed pictures. It will be a problem if the user wants to include mathematic formula, table, hyper-link, YouTube video, etc.

This research provides a solution for educators to be able to include customized multimedia, for an easy addition of mathematic formula, table, hyper-link, YouTube video, etc. in a Google form.

2 Related Background

Integrating technology with education has some reason. Statistics as of 2008 shows that more than one fourth of the students in higher education level take at least one course online [1]. Even for the classes that are taught in a live classroom environment, the exchanges of notes, homework and quizzes between instructors and students is more convenient through online means rather than hard copies. Some advantages include: it prevents human errors like losing the works without the convenience of easy backup copy, and is environment friendly as well. Studies have shown that e-Learning consumes 90% less energy than traditional courses [8]. Apart from that, students do benefit from using computers and other technological means while doing coursework [9]. When prompted to use computers as a media, rather than paper works, students tend to take the coursework more seriously [3].

Google Docs, unlike many other resources available, the supplementary package that comes with it is free of cost. It provides script editor for coding. It consists of a package including forms, spreadsheet, documents, and Google Drive for free online storage and efficient management of data. These free packages can be used together to create a system that can save an educator great amount of time. For instance, using Google Docs for an Online Homework System a teacher can save time sending confirmation emails to each student by adopting the automated email sending feature [6], or organize the graded and not graded Homework [7]. The use of Google Docs for creating the Online Homework/Quiz System is chiefly advantageous for it includes free online tutorials in texts as well as videos [2].

While using the Google package for an online homework system, Google form is used as the main

platform for the end user to send and receive homework. The problem we encountered is that it lacks the way to present multimedia content that most of the homework layout requires. An efficient Online Homework System has to be able comprise homework for courses like Math, Art, Physics, Biology among others which uses graphs, equations, tables, images and others. Generally only text or multiple choices can be used as an input in the Google Form. This drawback is prompted by the absence of a multimedia selection in a homework assignment or quiz, which makes the Google Form interface highly impractical. Thus the lack of multimedia in a Google Form limits its use as an effective platform. This limitation gives rise to a need to create an interface within Google Form that can accommodate graphs, equations, images, table and various other multimedia selections.

A few years back, when Google has not enabled the image for Google form yet, James Eichmiller [4, 5] provided an idea to include image in Google Form by modifying the source HTML code of the original Google Form to achieve a modified Google form which allows a user to add images [10]. Inspired by his idea, we further developed a user-friendly approach to use HTML code to generate multimedia Google form with vast options for graphics, tables, videos, graphs, equations and etc.

3 Customize Multimedia Form

3.1 Service Design

The overall idea of this research is to allow the users to input the desired multimedia content through HTML code in a predefined symbol such as "[[" and "]]" in the "Help Text" box of a regular Google Form editor. The symbols, "[[" and "]]" are used by the source code to identify the beginning and the end of the HTML code input by the user. Then, the input HTML code for the Google Form is fetched and the substring in "[[" and "]]" is replaced with the desired user defined HTML code. For the users who are not familiar with HTML, some 3rd party free online service is introduced to provide an editor similar with word processor and convert the desired multimedia content to HTML code.

This research uses Google Apps Script editor as the programming tool. The designed Google Apps Script code focuses on: (1) fetching the original source code; (2) searching for the substring with the user input html code in "[[" and "]]"; (3) replacing the substring with the user defined HTML code; (4) and generating the link to the updated multimedia Google form which can be used and shared by the user. The design process is described in the following Figure 1.

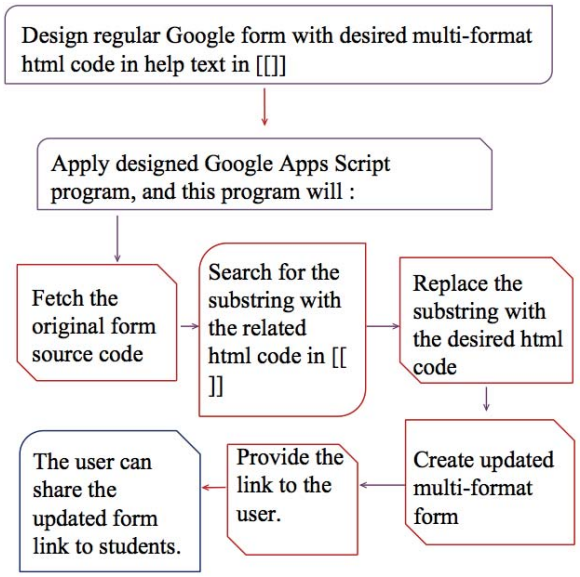


Figure 1: Design Process of Multimedia Form Service

The further designed code with the key components are demonstrated as in Figure 2 – 4.

```
//Get the html source of the form
var response = UrlFetchApp.fetch(formUrl);
source = response.getContentText();

ssName = ss.getName();

htmlText = source;
htmlText = htmlText.replace(/\baria-label="(.)+"/g, "aria-label=\"\"");
index1 = htmlText.indexOf("[[");
index2 = htmlText.indexOf("]]");

while (index1!=-1) {
   subStr = htmlText.substring(index1,index2+2);
   subStrUpdate = subStr.replace(/\[\[/g,'');
   subStrUpdate = subStrUpdate.replace(/\]\]/g,'');
   subStrUpdate = subStrUpdate.replace(/&lt;/g,'\<');
   subStrUpdate = subStrUpdate.replace(/&gt;/g,'\>');
   subStrUpdate = subStrUpdate.replace(/&amp;nbsp;/g,' ');
   subStrUpdate = subStrUpdate.replace(/&quot;<a href="(.)*">/g,'\"');
   subStrUpdate = subStrUpdate.replace(/<\/a>&quot;/g,'\"');
   subStrUpdate = subStrUpdate.replace(/&quot;/g,'\"');
   htmlText = htmlText.replace(subStr,subStrUpdate);
   index1 = htmlText.indexOf("[[");
   index2 = htmlText.indexOf("]]");
}
```

Figure 2: Code to Fetch and Replace HTML Block

In the code of Figure 2, we fetch the source HTML code based on the URL link of the original Google form; then we identify the substring marked with the predefined symbol; and replace the related symbols with desired HTML code. This piece of code generates an updated HTML code to include the original function of Google form and a new looking with multimedia and multi-format.

Figure 3 shows the code to generate and store the updated HTML file in a public folder.

```
htmlName = ssName+'.html';

//Get the folder instance of Public. If it does not exist, create it
try {
 folder = DriveApp.getFoldersByName('Public').next();
}
catch(exception){
 folder = DriveApp.createFolder('Public');
}

folder.setSharing(DriveApp.Access.ANYONE, DriveApp.Permission.VIEW);
//get folder ID this is used to establish the link
folderID = folder.getId();

try{
  file = folder.getFilesByName(htmlName).next();
  file.setContent(htmlText);
}
catch(exception){
  file = folder.createFile(htmlName, htmlText, MimeType.HTML);
}
```

Figure 3: Code to Store the Updated HTML File

We also design a menu in the corresponding spreadsheet, to provide a user-friendly interface. Thus, user can simply click the generated menu item to update the form and show the link, as in Figure 4.

```
var menuEntries =
 [{name:"Create Multimedia Form", functionName: "getHTMLWithImageTags"},
  {name: "Show Link", functionName: "showLink"}];
ss.addMenu("UpdateForm", menuEntries);
```

Figure 4: Code to Add Menu Entries

In the designed web service, we use the apps script to program the desired functions, and it creates a more structurally flexible Google Form and provides a low cost solution to enable the use of graph, links, videos, images, tables and many other formats in Online Homework/Quiz System. This arguably is one of the most effective solutions to create multimedia form as the user is only required to input the HTML code for the desired multimedia content, and the service itself is coded in a way to generate an automated link to the customized Google form. To simplify the application, we also recommend some free online HTML editors. The user can use the existing online service to generate HTML code for desired multimedia content, input the generated HTML code in the "Help Text" box of the original form. By applying the designed web service, a new multimedia Google Form is created. Thus, the user is not expected to be an expert programmer or have knowledge of any HTML coding.

3.2 Designed System Outcome

The above Google Apps Script based Multimedia form service is the core of the system. To make the service easy to use, we included some 3rd party online HTML Editor. And the whole system includes the following four steps, as in Figure 5.

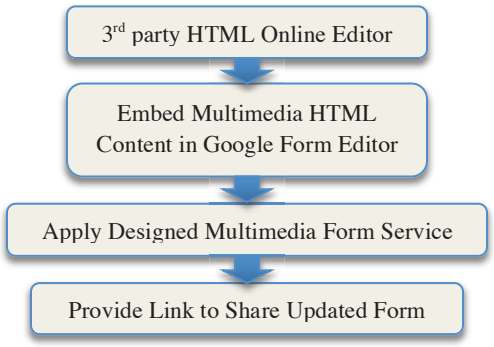


Figure 5: Designed User Friendly Multimedia Form System

For users, they can simply copy the sample spreadsheet to their own Google Drive account, and then they can start to edit their own form. Here we recommend some 3rd party HTML Online Editor, as in Figure 6, to provide the desired multimedia content and convert them to HTML code. The HTML Online Editor provides an environment similar with word processor, and users can edit their multimedia content in it. "Source" button is used to get the related HTML code. Through Google search, users can find many similar free HTML Online Editors. For users' convenience, we list some links below as references:

- http://bestonlinehtmleditor.com
- http://html-color-codes.info/html-editor/
- http://www.html.am/html-editors/
- http://www.quackit.com/html/online-html-editor/

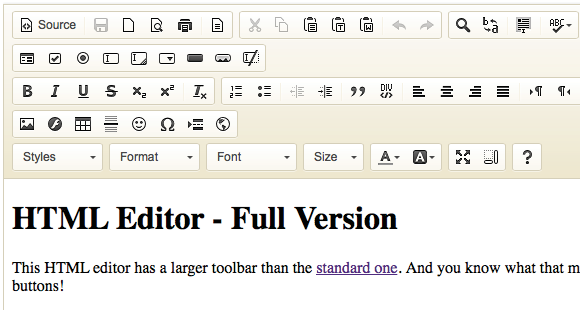


Figure 6: Sample Free HTML Online Editor

Then users can use above HTML Editor to get related HTML code and copy the HTML code to Google Form Editor. In the original Google Form editor, there is a label called "Help Text" followed by a text box, where a user can input any text. Users can simply paste it in "Help Text" part and mark it in [[and]], as in Figure 7.

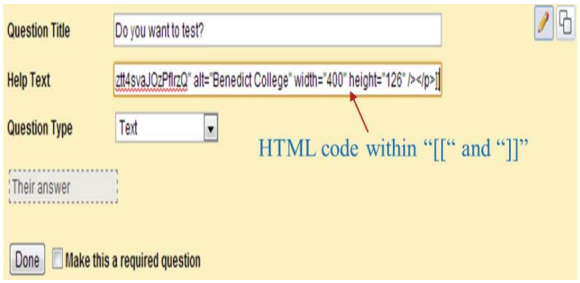


Figure 7: Copy HTML Code in Google Form Editor

Once we finish designing the form, we can go back to the corresponding spreadsheet. As in Figure 8, by clicking the menu "Create Multimedia Form" under "UpdateForm" to active the designed service to generate an updated Multimedia Form.

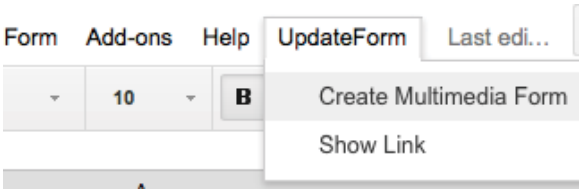


Figure 8: Activate Designed Service

After that, users can click "Show Link" to get the link of the new form and share with others. Some resulted sample form is in Figure 9.

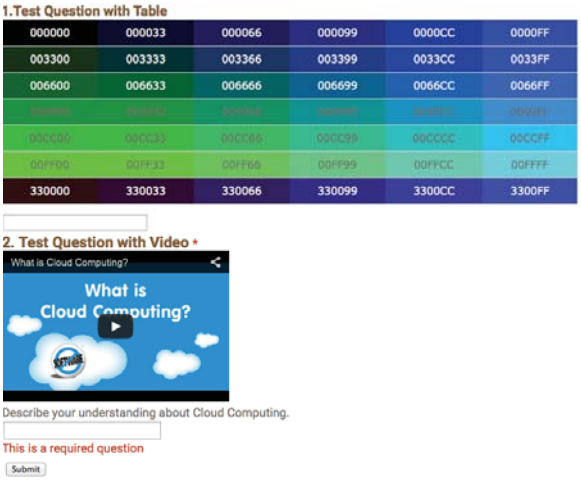


Figure 9: Resulted Sample Form with Table & Video

4 Conclusion

This approach creates a tailored Google form and provides a user-friendly free / low cost solution to enable the use of graph, links, videos, images, tables and many other. Furthermore the steps can be modified and utilized to generate Google Forms with any type of multimedia content questions. Hence, the outcome of this research project enables the use of the free resource, Google Docs, to create an Online Education System without having to compromise the format of questions needed to be asked on quizzes and / or homework. This leads to a more user friendly interface Forms suitable for the inclusion of all course questions such as Math, Chemistry, Physics or any other subject area in a free Online Homework / Quiz System.

* This project is supported by NSF Grant # HRD-1436222.

5 References

[1] I. Elaine Allen and Jeff Seaman. "Learning on demand online education in the United States, 2009". Needham, Mass.: Sloan Consortium. 2010. http://files.eric.ed.gov/fulltext/ED529931.pdf

[2] Google Developers. "Apps Script". 2014 https://developers.google.com/apps-script/

[3] Scott Bonham, Robert Beichner and Duane Deardorff. "Online homework: Does it make a difference?" The Physics Teacher Phys. Teach., 39, 293-293. 2001

[4] James Eichmiller. "Google Forms with Images". Making Technology Work at School. Dec. 7, 2012. http://tech-in-school.blogspot.ca/2012/12/google-forms-with-images.html

[5] James Eichmiller. "Update to Google Forms with Images". Feb 24, 2013. http://tech-in-school.blogspot.com/2013/02/update-to-google-forms-with-images.html

[6] Margaret Heller. "Taking Google Forms to the Next Level". ACRL TechConnect Blog. Nov. 26, 2012. http://acrl.ala.org/techconnect/?p=2343

[7] Alice Keeler. "Google Docs: Mark As Graded". Teacher Tech. Jan. 4, 2015. http://www.alicekeeler.com/teachertech/2015/01/04/google-docs-mark-as-graded/

[8] Christopher Pappas. "Top 10 e-Learning Statistics for 2014 You Need To Know", eLearning Industry, Dec. 1, 2013. http://elearningindustry.com/top-10-e-learning-statistics-for-2014-you-need-to-know

[9] Maryellen Weimer. "Online Homework Systems Can Boost Student Achievement". Jan. 22, 2013. http://www.facultyfocus.com/articles/instructional-design/online-homework-systems-can-boost-student-achievement/

[10] Bryan Weinert. "Adding Images to Google Forms". Dec. 7, 2012. https://youtu.be/Z3uPBe3rh2g

Web accessibility and security: an analysis of online security mechanisms and the Web Content Accessibility Guidelines 2.0

Justin Brown[1], **Scott Hollier**[2]
[1]School of Computer & Security Science, Edith Cowan University, Perth, Western Australia
Email: j.brown@ecu.edu.au
[2] School of Computer & Security Science, Edith Cowan University, Perth, Western Australia
Email: s.hollier@ecu.edu.au

Abstract - *This paper examines a number of common security mechanisms utilised in modern web applications such as CAPTCHAs, timeouts and visual reporting of errors or alerts and the impacts these can have on disabled web users. The authors explore the problems that security mechanisms can cause disabled users and the assistive technologies they utilise for web browsing, and map the relevant issues to the Web Content Accessibility Guidelines 2.0. Whilst proposing a number of possible options for aligning security concerns with those of accessibility guidelines, the authors do conclude by stating that this is a currently underdeveloped area of research and much work remains to be done.*

Keywords: WCAG, web accessibility, web security, CAPTCHA

1 Introduction

The issue of web security has received significant attention in recent times when United States President Barack Obama was questioned regarding the surveillance of online data by the United States National Security Agency, Obama stated "What I've said, and I continue to believe, is that we don't have to sacrifice our freedom in order to achieve security." [1]. For people with disabilities the freedoms the web represents are profound, though these freedoms come hand in hand with potential security issues faced by all web users, and the techniques for minimizing these risks can be extremely detrimental when they collide with the tenets of web accessibility. Currently people with disabilities require two issues to be addressed in equal measure to ensure effective online access: the provision of effective assistive technology tools and the creation of web content designed in a way that adheres to accessibility criteria which in turn allows for the assistive technology products to work in an optimal manner.

While the World Wide Web Consortium (W3C) and the International Standards Organisation (ISO) provide effective guidance as to how the web can be made accessible through the Web Content Accessibility Guidelines (WCAG) 2.0 and ISO 40500 standard, what is rarely acknowledged is the potential for accessibility issues to go beyond the inability to access information, but also to create potential security risks for users where web systems fail to take WCAG 2.0 standards into account. Issues related to CAPTHCAs, forms, images indicating password strength and the use of accessible video in informing the public can all be addressed through the use of WCAG 2.0, yet many websites remain inaccessible, raising the possibility of people with disabilities being particularly susceptible to security risks online or conversely, being hindered by security mechanisms designed to protect users.

This paper explores the elements of the WCAG 2.0 standard most relevant to security risks, the online dangers people with disabilities are likely to face if web based security mechanisms do not conform to WCAG 2.0, and the future implications of such risks as ever more traditional services, such as banking and commerce move inexorably towards a primarily cloud based delivery model.

2 Significance of web accessibility standards

The creation of web standards to support developers in creating accessible websites began with the creation of the World Wide Web Consortium (W3C) Web Accessibility Initiative conceived in 1996 and formalised in1997 [2]. The result was the creation of the Web Content Accessibility Guidelines (WCAG) 1.0 in 1999 [3] with the purpose to "explain how to make Web content accessible to people with disabilities. The guidelines are intended for all Web content developers (page authors and site designers) and for developers of authoring tools. The primary goal of these guidelines is to promote accessibility." [3].

While no specific mention of security or security-related elements were included in these early attempts towards web accessibility, the release of WCAG 2.0 in 2008 [4] marked a turning point in its acknowledgement that some security measures may result in people with disabilities facing security challenges such as completing a CAPTCHA or whether an error in a form [5] could be effectively conveyed in an

accessible manner without compromising security.

While the WCAG 2.0 standard provides some insight as to the potential issues that may occur when web security and accessibility collide, many accessibility issues are not framed in terms of their relevance to security issues and therefore it is the view of these authors that some additional consideration are required.

3 The CAPTCHA conundrum

The Completely Automated Public Turing test to tell Computers and Humans Apart (CAPTCHA) is a web based mechanism designed to be "tough on bots, easy on humans" [6] so as to ensure that websites do not receive spam during signup and registration processes, and can provide a useful barrier against automated cyber-attacks. However, for people with disabilities, the 'easy on humans' part of the equation is often very challenging [7-11]. The CAPTCHA generally contains difficult to read text (visually jumbled) or difficult to hear audio, which can prevent people with vision or hearing impairments from effectively entering the characters presented. For people with low vision, the characters often appear jumbled and illegible, while the audio CAPTCHA can be equally confusing when the word 'seven' is read out over garbled noises resulting in difficulty hearing initial word, and then further difficulty in understanding if the word 'seven' or the number '7' needs to be entered by the user. Work by Bigham and Cavender indicated "that more than a third of blind participants said they had never solved a visual CAPTCHA"[12].

People who are blind or vision impaired rely on the use of a screen reader to access web pages, with these text-to-speech software applications being designed to provide full audio description of any functions or content that can be accessed via keyboard or touch screen on desktop of mobile devices. The challenge of a CAPTCHA in this regard is that a screen reader is designed to use the computing power of the device to process a web page and computer-readable content for the text-to-speech engine, yet this is the exact task that a CAPTCHA is designed to prevent in light of malicious cyber activity [13]. Whilst more accessible alternates to traditional CAPTCHA technologies have been proposed, with a focus on audio challenge rather than visual puzzle [14-17] the emphasis on visual over audio seems to remain the current norm. The WCAG 2.0 standard highlights the CAPTCHA in the first of its 121 guidelines, stating in Guideline 1.1 that "Text Alternatives: Provide text alternatives for any non-text content so that it can be changed into other forms people need, such as large print, braille, speech, symbols or simpler language." [18]. Specifically, the success criteria 1.1.1 for this guideline states that "If the purpose of non-text content is to confirm that content is being accessed by a person rather than a computer, then text alternatives that identify and describe the purpose of the non-text content are provided, and alternative forms of CAPTCHA using output modes for different types of sensory perception are provided to accommodate different disabilities." [19]. In essence, this guideline requires that CAPTCHA elements are labelled as such so that the user understands that information is required to be input to the CAPTCHA, though at the time of writing there is no access solution proposed by WCAG 2.0 to address the accessibility issues associated with completing the CAPTCHA task. The Australian Communications Consumer Action Network (ACCAN) has advocated often on the issue of CAPTCHAs and their impact on disabled web users [20-22], going so far as to launch the 'kill CAPTCHA' campaign [23], explaining that alternative processes such as e-mail verification of the credentials entered into a website can be just as effective in preventing bot attacks whilst providing a more accessible solution. However, this option is rarely preferred by the ICT industry [24] due to it requiring a two-stage verification process. While CAPTCHAs remain a significant aspect of online security despite the accessibility overhead they create, the ACCAN campaign did have some success within the Australia context, with Australia's biggest telecommunications company Telstra removing CAPTCHAs from its website, arguing that people (its clients) with disabilities should be identified as human too.

4 Cyber safety awareness

Outside of the technical mechanisms for addressing security concerns, it is also imperative that information is provided in an accessible way so as to educate the public in regards to their personal security when entering information into websites and applications on connected devices. With dating scams, phishing and computer lockouts on the rise [25, 26] it is important that online information providing guidance on how to protect against these issues is available to all web users, including those with disabilities. In this instance, WCAG 2.0 proves helpful with Guideline 1.2 stating that provisions should be made for alternatives to time-based media. Specifically, the guidelines refers in its success criteria to the accessibility techniques of captions, audio description and sign language as mechanisms to ensure accessible audio-visual content. While the guidelines view captioning as a high priority and sign language less so, the importance of ensuring that the deaf community in particular was able to effectively understand cybersafety messages was highlighted by a campaign run by the Deaf Society of Western Australia [27, 28] specifically to meet a need that the Deaf community were facing in understanding security issues of scam e-mails relating to phishing, fake lotteries and online dating. In this instance, following the WCAG 2.0 requirements to ensure audio-visual materials featured alternative information presentation (such as captions) enabled an effective conveyance of cybersecurity information protecting a community whose disability may have resulted in them being more susceptible to security issues. Another WCAG 2.0 guideline, namely 3.1 states that content should be readable and understandable. This is particularly relevant in ensuring that any security-related discussions are simple and easy to follow for the lay person and those with limited literacy.

Security instructions can often be technical in nature, and the implementation of the success criteria contained in this guideline can ensure that any messages that need to be delivered to the user are structured in a simple way and clearly identified in the expected language of the user, or as Stajano and Wilson put it "Users care only about what they want to access and are essentially blind to the fact that 'the annoying security gobbledygook' is there to protect them. Smart crooks exploit this mismatch to their advantage; a lock that is inconvenient to use is often left open" [26]. Even something as fundamental as declaring the language within the html content of a webpage is particularly important in this regard as incorrect or missing language declarations can cause assistive technologies such as screen readers to mispronounce instructions which could then have flow on security implications.

4.1 Visual presentation of security information

A common online security feature is requiring users to change their password on a periodic basis. Part of this process can involve a visual image displayed to the user to indicate the proposed password's strength using imagery such as a coloured bar or circle with graduated fills used to represent low through high password strength. However, for a number of disability groups the visual information will be inaccessible, particularly for people who rely on screen readers or people with limited cognitive functionality who may not be able to effectively comprehend the visual representation and its implications for their level of security. WCAG 2.0 Guideline 1.3 can provide assistance through ensuring that the relationship of information represented in an accessible way. The guideline states that authors can "Create content that can be presented in different ways (for example simpler layout) without losing information or structure" [29].

Common solutions to this issue are to provide text that describe the information that is also presented visually. A related issue is colour contrast, with many password-related errors being represented by a change of colour such as red. For users with colour blindness or who are blind, this change is often difficult or impossible to assess. Support for this issue is also identified in WCAG 2.0 in Guideline 1.4 which states that authors should make content "Distinguishable: Make it easier for users to see and hear content including separating foreground from background" [30].

Other mechanisms that could be used instead of colour to indicate a change includes the use of additional visual content such as putting a box around the form element (ie focus on the element reporting the error) and providing a text explanation as to what the user needs to do in order to address the security issue. While these WCAG 2.0 guidelines do not specifically mention security implications, following 1.3 and 1.4 would go a long way towards ensuring that security issues associated with visual representations and colour changes are addressed for people with disabilities.

4.2 Entering secure information using forms

Most websites that require users to enter secure information such as credit card details or personal data will do so using a web form. However, there can be a number of challenges with this process for people with disabilities in relation to accessibility. Due to the difficulties that forms can present, there are a number of guidelines that can apply in this situation. Firstly, it is imperative that all elements of the form are completely accessible via keyboard. This is identified in Guideline 2.1 which states that they be "Keyboard Accessible: Make all functionality available from a keyboard" [31].

This is particularly beneficial to people with vision or mobility-related disabilities, as such users will generally employ keyboard or touchscreen gestures or pointing devices to navigate through form elements. If an element is not accessible by keyboard, it is likely the element will be skipped and the process of completing the form cannot be achieved. However, the key guideline that relates to security in forms is guideline 3.3 which state that "Input Assistance: Help users avoid and correct mistakes" [32]. In particular, the labelling of forms can have a significant impact on whether correct security credentials are set up. For example, simply stating 'Date of Birth', a common piece of information used in security and authentication systems, can result in the date being entered into a form in a variety of ways. While most forms typically accept only one valid date syntax (ie MM/DD/YYYY), the user may not be aware of the correct format, or that the field is even requesting a date of birth. Another date related issue is where a date picker control has been embedded into the form, allowing the user to click on a calendar icon and select the required date, such as a date of birth. Date picking controls can be used in conjunction with text-entry fields or can be as the sole input mechanism for gathering a specific date from a user. The basic HTML code snippet below (Figure 1) shows a form using a fieldset and legend approach to indicate to users what information the form fields will be collecting alongside label elements associated with each input field to make it clear to the user what data is required for a given field.

```
2  <h1>Register New User</h1>
3  <p>Please enter your personal details</p>
4  <form action="http://wcagsecuritypaperexample.net/register" method="post">
5    <fieldset>
6      <legend>Personal identification details</legend>
7        <label for="fname">Please enter your first name</label>
8        <input type="text" name="fname" id="fname" />
9        <br />
10       <label for="sname">Please enter your surname</label>
11       <input type="text" name="sname" id="sname" />
12       <br />
13       <label for="dob">Please enter your date of birth in MM/DD/YYYY format</label>
14       <input type="text" name="dob" id="dob" />
15       <br />
16   </fieldset>
17   <input type="submit" value="Register Now" />
18 </form>
```

Figure 1: accessible form design

Lines 13-14 in Figure 1 shows a label requesting the user's date of birth, with the required date format appearing as text within the label. If this verbose labelling approach were

deemed inappropriate (due to page space restrictions for example), an alternate implementation look something like that shown below in Figure 2.

```
13      <label for="dob">Please enter your date of birth:</label>
14      <input type="text" name="dob" id="dob" title="date format MM/DD/YYYY" />
```

Figure 2: title attributes in forms

In this scenario the label content is shorter, but the date format requirement is assigned to a title attribute on the date of birth text field. To take this example one step further, the Figure 3 presents a possible solution to the issue of indicating to users with disabilities when a field is defined as 'required', when a password confirmation is called for and what format that password might take;

```
16      <label for="emailAdd">Please enter your email address (required field):</label>
17      <input type="text" name="emailAdd" id="emailAdd" />
18      <br />
19      <label for="password">Please create a password (required field):</label>
20      <input type="text" name="password" id="password" title="must contain a minimum of 6 letters and two numbers" />
21      <br />
22      <label for="confirmpassword">Please confirm your password (required field):</label>
23      <input type="text" name="confirmpassword" id="confirmpassword" title="must contain the same value entered into the
        password field" />
24      <br />
25  </fieldset>
26  <input type="submit" value="Register Now" />
27 </form>
```

Figure 3: clearly stated form requirements

Lines 16, 19 and 22 of Figure 3 indicate that the required field indicator is presented inside the label for each of the fields rather than using the unhelpful * indicator next to each field with an associated '* *this is a required field*' statement elsewhere in the form. Lines 20 and 23 use title attributes on the first and second password fields, the first to indicate the password formatting requirements and the second to indicate that the same value must be re-entered by the user. Using this more informative approach, where input requirements are located within proximity to the fields taking data entry from the user, the occurrences of entering incorrect data or incorrectly formatted data by users with disabilities should be mitigated to a certain degree.

This leads to another aspect of Guideline 3.3 regarding correcting mistakes. It is often the case that when an error occurs, there is no obvious indication as to what error was made on the form, or sometimes the error is marked by a change of colour such as red which leads to a colour contrast issue as highlighted earlier. In order for people with disabilities to effectively enter information into a form, the form field labels must be clearly identified and if errors are made, clear indications as to what the error is and how it can be addressed must be presented to the user in a way that is compatible with assistive technologies. An example of good practice in dealing with errors in forms can be seen in Figure 4 as taken from webaccessibility.com in their login form demonstration in regards to indicating errors with form input (https://www.webaccessibility.com/login.php);

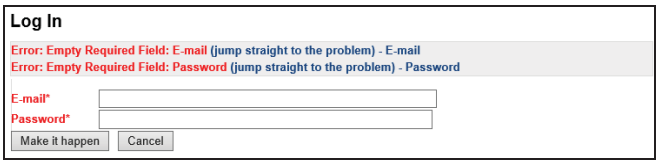


Figure 4: accessible error reporting in forms

In this example, if both the email and password fields are left empty on submission, a user without disabilities can readily identify and remedy the error, whilst a user with a disability using only a keyboard would be taken directly to the hyperlinked '(jump straight to the problem) – Email' link which has a tab index of 0 to attract the first keyboard interaction, which in turn would then send the user straight to the problematic field.

However, despite the importance of users entering information into forms correctly for security related functions, it is this area where WCAG acknowledges that there may be security reasons as to why this cannot be implemented. The only specific reference of security in WCAG 2.0 with success criteria 3.3.3 stating that "Error Suggestion: If an input error is automatically detected and suggestions for correction are known, then the suggestions are provided to the user, unless it would jeopardize the security or purpose of the content" [33].

As with the CAPTCHA, the success criteria endeavors to provide information as to how security relates to the web accessibility standard, but does not provide a solution as to how a person with a disability can effectively navigate such a scenario.

4.3 Timeouts

The final security issue this paper examines in the context of accessibility is situations where security measures are based on some type of time limit for user input, particularly when the website requires the user to complete a specific task such as processing payment information or using email on a public terminal. WCAG 2.0 Guideline 2.2 addresses this by stating "Enough Time: Provide users enough time to read and use content" [34].

The relevant success criteria specifically addresses time-related issues by saying that users should have the ability to either turn off the time restriction, extend the time restriction or adjust the time restriction to ensure that the user has enough time to complete the tasks. While some may view this as a security risk, the ability to extend or adjust time within the web application as they are using it would still require a user to be actively using the web content which in turn should minimize the risk of leaving the content unattended, and also ensure that accessibility requirements are implemented. Whether government agencies, banks or online retailers will provide such customizations on a wide scale remains to be seen, though an ad hoc review of a number of international banking websites revealed a small number of organizations that allowed users to customize their session timeout values, with maximum values ranging from 15 to 60

minutes. It could be argued that if all aspects of a website were designed to be accessible and easy to navigate, that less generous session timeouts would not be as problematic, for users of assistive technologies could accomplish their tasks in a relatively straightforward manner. However, where such users struggle against poor navigation, incomprehensive forms and form validation and complex, multi-step processes, insufficient time can present an insurmountable barrier to website usage.

4.4 Cloud security

While WCAG 2.0 is generally viewed as the definitive guide for web accessibility [35-38] there is both great promise and great concern about the security implications for people with disabilities as service delivery moves inexorably towards cloud-based solutions. Raising the Floor, based in the USA, have created an initiative called the Global Public Inclusive Infrastructure (GPII) which aims to create an environment where "…every device automatically changing into a form you can understand and use, instantly, whenever and wherever you encounter it." [39]. Given that WCAG 2.0 does not currently address all elements of security in its guidelines, a GPII-style approach could incorporate key accessibility features such as those found in WCAG 2.0, integrate it with a universally available set of proven accessibility features which could potentially address accessibility issues, enabling both the developers and the user to interact in an effective and secure environment [40].

However, while security issues are currently focused more around people with disabilities being disadvantaged due to accessibility issues, the cloud could pose a reverse discriminatory environment in that a successful GPII rollout would depend heavily on people with disabilities sending a profile of there needs to the cloud, something that individuals may not comfortable in doing [40]. As such, future accessibility solutions need to consider both the accessibility of information presented to the user, and the user requirements for storing their personal disability-related preferences online and having them retrieved and applied to private and public computing platforms.

5 Conclusions

The WCAG 2.0 standard provides insight as to how creators of websites and content can effectively ensure that potential security issues are addressed whilst maintaining alignment with accessibility goals. In particular, ensuring that users can effectively understand instructions, input information into forms and view security warnings such as password changes and provide guidance relating to timeout techniques.

The authors of this paper feel there is still much work to be done in the design of accessible websites, and even more work remains in the equitable design and implementation of online security mechanisms so that all web users can participate securely and in an informed manner. Whilst this paper raises some of the most common security approaches used in modern websites and how those approaches might impact people with disabilities, any number of more advanced security techniques such as two factor authentication, biometric methods and the use of connection tokens might pose even greater challenges in the near future. It is the hope of these authors that in the pursuit of ever more secure web computing technologies and approaches, web designers and security experts will bear in mind that creating barriers to malicious users can create insurmountable barriers to people with disabilities. This paper has demonstrated that a number of web security issues can be mitigated with website design being aligned with the WCAG 2.0 guidelines, though such alignment does not suit every technology, every use case and above all, every user.

6 References

[1] D. Kerr. (2013, 17/7/2014). Obama: NSA spying doesn't mean 'abandoning freedom'. Available: http://www.cnet.com/news/obama-nsa-spying-doesnt-mean-abandoning-freedom/

[2] D. Dardailler. (2009, 9/9/2014). WAI early days. Available: http://www.w3.org/WAI/history

[3] World Wide Web Consortium. (1999, 1/10/2014). Web Content Accessibility Guidelines 1.0. Available: http://www.w3.org/TR/WCAG10/

[4] World Wide Web Consortium. (2008, 3/10/2014). Web Content Accessibility Guidelines 2.0. Available: http://www.w3.org/TR/WCAG20/

[5] K. Fuglerud and T. Røssvoll, "An evaluation of web-based voting usability and accessibility," Universal Access in the Information Society, vol. 11, pp. 359-373, Nov 2012 2012.

[6] Google Inc. (2014, 11/10/2014). Google RECAPTCHA. Available: http://www.google.com/recaptcha/intro/index.html

[7] K. Brian, "From Web accessibility to Web adaptability," Disability & Rehabilitation: Assistive Technology, vol. 4, pp. 212-226, 2009.

[8] E. Murphy, R. Kuber, G. McAllister, P. Strain, and W. Yu, "An empirical investigation into the difficulties experienced by visually impaired Internet users," Universal Access in the Information Society, vol. 7, pp. 79-91, 2008.

[9] C. Pribeanu, P. Fogarassy-Neszly, and A. Pătru, "Municipal web sites accessibility and usability for blind users: preliminary results from a pilot study," Universal Access in the Information Society, vol. 13, pp. 339-349, 2014.

[10] Ö. Subasi, M. Leitner, N. Hoeller, A. Geven, and M. Tscheligi, "Designing accessible experiences for older users: user requirement analysis for a railway ticketing portal," Universal Access in the Information Society, vol. 10, pp. 391-402, 2011.

[11] S. Shirali-Shahreza and M. H. Shirali-Shahreza, "Accessibility of CAPTCHA methods," presented at the

Proceedings of the 4th ACM workshop on Security and artificial intelligence, Chicago, Illinois, USA, 2011.
[12] J. Bigham and A. Cavender, "Evaluating existing audio CAPTCHAs and an interface optimized for non-visual use," presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Boston, MA, USA, 2009.
[13] S. Ball, "Accessibility in e-assessment," Assessment & Evaluation in Higher Education, vol. 34, pp. 293-303, 2009.
[14] E. Bursztein, A. Moscicki, C. Fabry, S. Bethard, J. Mitchell, and D. Jurafsky, "Easy does it: more usable CAPTCHAs," presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Toronto, Ontario, Canada, 2014.
[15] J. Lazar, J. Feng, T. Brooks, G. Melamed, B. Wentz, J. Holman, et al., "The SoundsRight CAPTCHA: an improved approach to audio human interaction proofs for blind users," presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Austin, Texas, USA, 2012.
[16] S. Shirali-Shahreza, G. Penn, R. Balakrishnan, and Y. Ganjali, "SeeSay and HearSay CAPTCHA for mobile interaction," presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Paris, France, 2013.
[17] K. Fuglerud, I. Tjøstheim, B. Gunnarsson, and M. Tollefsen, "Use of social media by people with visual impairments: Usage levels, attitudes and barriers," pp. 565-572.
[18] World Wide Web Consortium. (2014, 11/10/2014). Text Alternatives: Understanding Guideline 1.1. Available: http://www.w3.org/TR/UNDERSTANDING-WCAG20/text-equiv.html
[19] World Wide Web Consortium. (2014, 11/10/2014). Non-Text Content: Understanding SC 1.1.1. Available: http://www.w3.org/TR/UNDERSTANDING-WCAG20/text-equiv-all.html
[20] Australian Communications Consumer Action Network. (2013, 4/4/2014). Telstra to kill CAPTCHA: ACCAN calls on others to follow. Available: https://accan.org.au/news-items/media-releases/710-telstra-to-kill-captcha-accan-calls-on-others-to-follow
[21] Australian Communications Consumer Action Network. (2014, 4/4/2014). Community position statement on CAPTCHA. Available: https://accan.org.au/our-work/policy/728-community-position-statement-on-captcha
[22] B. Grubb. (2014, 5/4/2014). Google challenged to drop CAPTCHA puzzles. Available: http://www.smh.com.au/digital-life/digital-life-news/google-challenged-to-drop-captcha-puzzles-20140205-321be.html
[23] Australian Communications Consumer Action Network. (2013, 4/4/2014). Consumer advocates unite to kill CAPTCHA. Available: https://accan.org.au/news-items/media-releases/603-consumer-advocates-unite-to-kill-captcha
[24] D. Bushell. (2011, 5/9/2014). In Search Of The Perfect CAPTCHA. Available: http://www.smashingmagazine.com/2011/03/04/in-search-of-the-perfect-captcha/
[25] N. Muscanell, R. Guadagno, and S. Murphy, "Weapons of Influence Misused: A Social Influence Analysis of Why People Fall Prey to Internet Scams," Social and Personality Psychology Compass, vol. 8, pp. 388-396, 2014.
[26] F. Stajano and P. Wilson, "Understanding scam victims: seven principles for systems security," vol. 54, ed. New York: ACM, 2011, pp. 70-75.
[27] Australian Communications Consumer Action Network. (2010, 6/6/2014). ACCAN Grant recipients target Deaf consumers and scams, privacy complaints and culturally diverse consumers Available: https://accan.org.au/news/media-releases/156-accan-grant-recipients-target-deaf-consumers-and-scams-privacy-complaints-and-culturally-diverse-consumers
[28] WA Deaf Society. (2011, 6/7/2014). Internet Scams: How to Protect Yourself. Available: https://www.youtube.com/user/internetscamsprotect
[29] World Wide Web Consortium. (2014, 11/10/2014). Adaptable: Understanding Guideline 1.3. Available: http://www.w3.org/TR/UNDERSTANDING-WCAG20/content-structure-separation.html
[30] World Wide Web Consortium. (2014, 17/10/2014). Distinguishable: Understanding Guideline 1.4. Available: http://www.w3.org/TR/UNDERSTANDING-WCAG20/visual-audio-contrast.html
[31] World Wide Web Consortium. (2014, 12/10/2014). Keyboard Accessible: Understanding Guideline 2.1. Available: http://www.w3.org/TR/UNDERSTANDING-WCAG20/keyboard-operation.html
[32] World Wide Web Consortium. (2014, 9/10/2014). Input Assistance: Understanding Guideline 3.3. Available: http://www.w3.org/TR/UNDERSTANDING-WCAG20/minimize-error.html
[33] World Wide Web Consortium. (2014, 11/10/2014). Error Suggestion: Understanding SC 3.3.3. Available: http://www.w3.org/TR/UNDERSTANDING-WCAG20/minimize-error-suggestions.html
[34] World Wide Web Consortium. (2014, 14/10/2014). Enough Time: Understanding Guideline 2.2. Available: http://www.w3.org/TR/UNDERSTANDING-WCAG20/time-limits.html
[35] V. Conway, J. Brown, S. Hollier, and C. Nicholl, "Website accessibility: a comparative analysis of Australian National and state/territory library websites," The Australian Library Journal, vol. 61, p. 170, 2012.

[36] S. Harper and A. Chen, "Web accessibility guidelines: A lesson from the evolving Web," World Wide Web, vol. 15, pp. 61-88, 2012.
[37] C. Power and H. Jürgensen, "Accessible presentation of information for people with visual disabilities," Universal Access in the Information Society, vol. 9, pp. 97-119, 2010.
[38] S. Hollier. (2012, 9/9/2014). WCAG 2.0 approved by ISO: wjay impact odes it have. Available: http://www.accessiq.org/news/w3c-column/2012/11/wcag-20-approved-by-iso-what-impact-does-it-have
[39] GPII. (2011, 11/10/2014). Global Public Inclusive Infrastructure. Available: http://gpii.net/index.html
[40] S. Hollier. (2014, 22/9/2014). Consumers and cloud accessibility. Available: http://www.mediaaccess.org.au/latest_news/online-media/consumers-and-cloud-accessibility