Final Report

Dynamic Scheduling for Web Monitoring Crawler

Contract Number: FA4869-08-1-4054

AFOSR/AOARD Reference Number: AOARD-08-4504

AFOSR/AOARD Program Manager: Hiroshi Motoda

Period of Performance: 7 March 2008 – 7 March 2009

Submission Date: 27 Feburary 2009

PI: Dr. Byeong Ho Kang/University of Tasmania
CoPI: Professor Paul Compton/University of New South Wales
Professor Hiroshi Motoda/Osaka University
Dr. John Salerno, Air Force Research Laboratory/Information Directorate

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
27 FEB 2009	**FInal**	**07-03-2008 to 07-03-2009**

4. TITLE AND SUBTITLE
Dynamic Scheduling for Web Monitoring System

5a. CONTRACT NUMBER
FA48690814054

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Byeong Ho Kang

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Tasmania,GPO Box 252-100,Hobart TAS 7005,tas,au,7005

8. PERFORMING ORGANIZATION REPORT NUMBER
N/A

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AOARD, UNIT 45002, APO, AP, 96337-5002

10. SPONSOR/MONITOR'S ACRONYM(S)
AOARD

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AOARD-084054

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Web monitoring systems report any changes on the target web pages by revisiting them frequently. As they are operated under significant constrains such as network and computing, it is necessary to minimize revisits with minimal delay and maximum coverage. Various statistical scheduling methods were proposed to resolve this problem. However they are static and cannot easily cope with events in the real world. This paper proposes a new scheduling method that manages unpredictable events. MCRDR (Multiple Classification Ripple-Down Rules) document classification knowledge base was reused to detect events and to initiate a prompt web monitoring process regardless of static monitoring schedule. The experiment demonstrates that the approach proposed improves monitoring efficiency significantly.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**48**	

1 Objectives

One of the main aims of web monitoring system is to collect information from the selected web pages with maximum coverage and minimal delay. To this end, a web monitoring system needs to revisit its monitoring web pages according to its revisit schedule for each monitoring web page. If there are no resource constraints, web monitoring system may obtain high coverage with low delay by as frequent revisits for target web pages as possible. However, revisiting process, technically sending and receiving HTTP messages via internet, is very expensive and is constrained by resource limits, including network and computer capacities. Therefore, it is essential for the Web monitoring system to have an efficient scheduling method that minimizes its revisit frequency while maximizing its coverage and timeliness. Statistical scheduling approaches have provided static solutions for this problem based on the assumption that there exist stable publication patterns. However, real-world web publications are affected by unpredictable events, such as Olympic in sports and Terror in national security domains. Therefore, dynamic scheduling approach is also required to manage real world events that occur during monitoring process effectively.

The followings are our main contributions:

- Firstly, this project proposes an event detection methodology, which reuses document classification knowledge to detect events at the real-time. Our scheduling system detects abnormality of document publications by comparing current classification results with previous one.
- Secondly, this project proposes how to combine event detection methodology with web monitoring scheduling, as the outbreak of an event in a monitoring web page may influence the other monitoring web pages.
- Lastly, this project demonstrates how the proposed event-driven monitoring can operate efficiently compared to the static monitoring methodology.

2 Status of effort

1) Feasibility Study

We conducted various web monitoring studies before this project. A web monitoring system, called WebMon, was created in 2005 and it was used to conduct web monitoring services for the selected domains, such as Australian Government Web pages and health news web pages. We

have conducted web publication analysis and search engine performance analysis using our web monitoring system. The web publication analysis revealed publication patterns of the selected web pages and search engine performance analysis revealed usefulness of the web monitoring system in relation to search engines. As information overload is one of the main problems of web monitoring system, we developed a web document classification system to resolve this problem. The MCRDR knowledge acquisition methodology was employed to facilitate incremental knowledge encoding by the end users. Various knowledge acquisition studies were conducted to evaluate our document classification system and this system was also used to facilitate personalised information delivery and to construct dynamic web portals.

2) System Development

Java program language and MySQL database were used to develop our event-driven monitoring system. Figure 1 illustrates system architecture of the event-based web monitoring system, which consists of five modules.

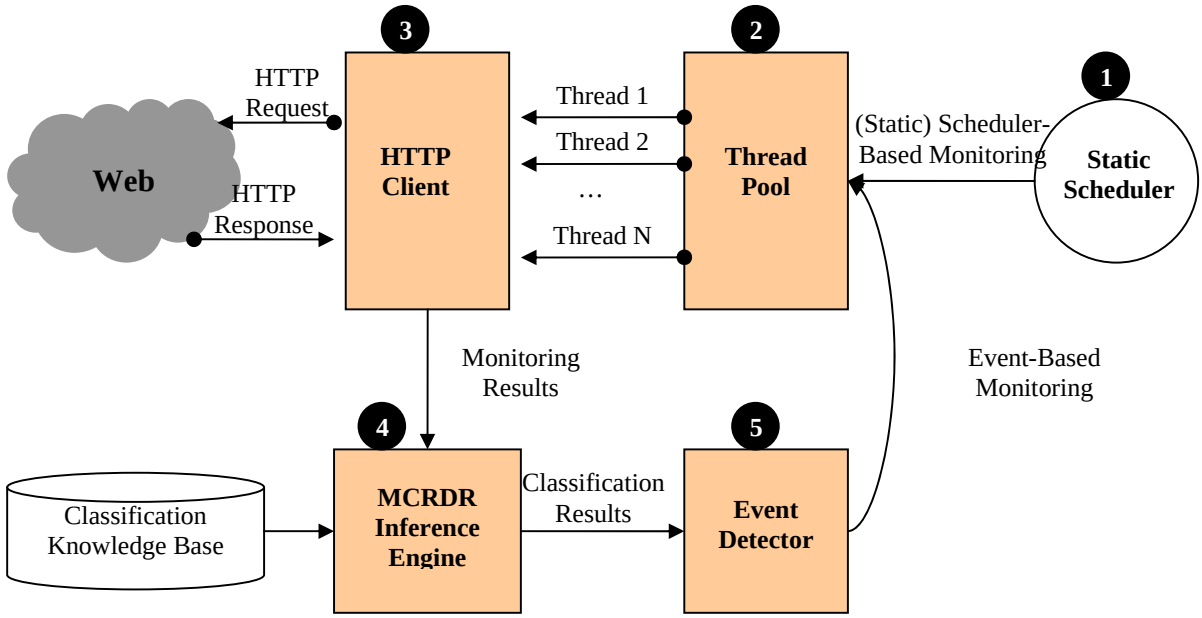


Figure 1. Event-Based Web Monitoring System Architecture

❶ **Static Scheduler**: The static scheduler initiates each static monitoring process. There are many previous researches on static scheduling methods, but they are not included in this project, because this project mainly focuses on the event-driven scheduling. In this system the user can specify revisit intervals such as every 2 hours. This simple static scheduling may be replaced by

more sophisticated methods, but critical results that obtained from this project may not be affected by this substitution.

❷ **Thread Pool**: As the computing resource is generally limited, it is necessary to manage system resources efficiently and thread pool may be used for this purpose. In our system, Java thread pool was implemented to manage large scale of HTTP requests, which prevents run out of resources.

❸ **HTTP Client**: Apache httpclient (http://hc.apache.org/httpclient-3.x/) library was used in this project because this is stable and commonly used in the HTTP network programming. A HTML parser, called htmlparser (http://htmlparser.sourceforge.net/), library was used to manipulate HTML/XHTML documents.

❹ **MCRDR Inference Engine**: The MCRDR (Multiple Classification Ripple-Down Rules) inference engine automatically classified the collected documents. Detailed explanation on the MCRDR classification document system is discussed in [1, 2].

❺ **Event Detector**: At each monitoring session the event detector finds anomaly of document classification for each category. If the number of documents classified into a specific category significantly increased at a specific monitoring session, the system regards this as an advent of an event and initiates an additional event-driven monitoring processes regardless normal monitoring schedule.

3) Event-Driven Scheduling Methodology

Assumptions

Our event-driven scheduling approach is based on the following assumptions:

1. Many publication changes in web pages caused by particular events, such as accidents, financial crisis, and sports games.
2. A group of web pages $P = \{P_1, P_2, P_3 \ldots P_n\}$ may provide "similar" publications, which is classified into the same category, called *C*, with different schedule. Note that all publications from *P* may not classified into the same category, rather most documents classified into the category consists of publications from *P*.
3. It is assumed that average documents classified into *C* at each time window (e.g., from 9:00 am to 10:00 am) is stable over time. Therefore, if the number of classified documents at a certain time is significantly greater than the normal frequency, it is

regarded that an event occurs between the last monitoring time and the current monitoring time.

4. Therefore, if the web monitoring system checks these other pages, then the delay time of the web monitoring will be decreased.

Algorithms

Our monitoring system, which implemented event-driven scheduling algorithm, works as follow:

- Register monitoring web pages of interest
- Set schedule for each monitoring web page using statistic monitoring policies.
- The system generates an hourly average classification table (a 7 days × 24 hours matrix) for each category.
- A_{ij} means average number of publications in the classification at *j* hour on *i* day of week (e.g, 14:00 on Monday)
- Get the current number of publications in the classification of each web site (C_{ij}) (e.g., 14:00 on Monday 12th of January, 2009)
- If $C_{ij} > A_{ij} + T_p$ (classification threshold), then the system finds other monitoring web pages that provide similar contents and execute web monitoring regardless of its original schedule.

Decision Factors

There are three decision factors in the above algorithm. Firstly, it is necessary to determine how to generate an hourly average publication table. Each average value (A_{ij}) can be calculated by overall period or specific time span (e.g. last five week average). Secondly, it is necessary to determine the threshold value (T_p) that is used for event detection. For this project, this value may be determined by heuristics and this value was set by $0.2 \times A_{ij}$. Lastly, it is necessary to determine which web pages are related to a specific web page. This project uses classification history of web pages to determine similarity between web pages. That is, web page similarity is decided by the classification frequency for a specific category.

4) Experiment Design

Experiment Goal

This experiment was designed to evaluate the performance of the event-based scheduling method in comparison with a static scheduling method.

Selection of Experiment Period

We monitored sports news web pages during Beijing Olympic Games. As each game held in the Olympic can be regarded as an event, it is possible to expect articles about the games would be published according to progresses of the games. For example, if some Olympians won or lost in the game, the monitoring web pages would be updated according to the game results. Therefore, Olympic period can be an appropriate environment to examine event-driven scheduling performance.

Selection of Web Pages

First of all, we attempted to choose the countries which tend to be more interested in Olympic game, because web pages from those countries may publish more articles related to Olympic game than other countries. Top 10 countries were selected from the medal table in the official web site of the Olympic 2004 Athens.[1] Generally, it is expected that highly ranked countries in the previous Olympic Game are more likely to make good results in the next Olympic Game. However, the contents in the selected web pages must be understandable, so only English-based countries; Australia, United States and Great Britain have been selected for this experiment. Once the target countries have been selected, we chose 10 generally "well-known" news web sites for each country. Obviously, focusing on the "well-known" news web sites biases our results to a certain degree, but it is believed that this selection was not too extreme case and those web sites are considerably popular in the selected countries. Sports pages as well as home pages of these selected web sites were monitored, because Olympic-related articles may be posted more to the sports section. As a result, 2 types of web pages (homepage and sports page) from each selected web site have been registered and monitored about one month.

Web Monitoring

A web monitoring system module was used to download web pages. It periodically revisits registered web pages to check whether or not there are any hyperlink changes in these web pages. A total of 60 selected web pages were registered to the web mentoring module. As we had no

[1] Seen at http://www.olympic.org/uk/games/past/index_uk.asp?OLGT=1&OLGY=2004

information about publication patterns of each web page, we used a single fixed scheduling strategy. The system continually revisited all the web pages with two hours interval from 7th August 2008 to 3rd September 2008. During this monitoring period, originally total 146,033 Web pages; 50,778 pages from Australia, 45,892 pages from United States and 47,521 pages from Great Britain were downloaded and stored into the database. The size of all collected data was around 580MB. Each data includes the information about its URL (stored in the form of absolute URL), the link text (title), data origin (to indicate which Web site it was published from), its contents (HTML source associated with absolute URL) and the time when it was captured by the system.

Document Classification with MCRDR Classifier

It is important to determine what standard we need to apply to classify the articles. It can vary and be highly fluid because it depends on what people are interested in. In this experiment, however, we focused on Olympic and decided to define two big categories first, which were *Countries* and *Sports*, because these two concepts are the main subjects in Olympic Games. Under the countries category, top 10 countries category in the previous Olympic 2004 were created and categories of 28 summer sports, referred to the official Web site of the Beijing 2008, were created under sports. Table 1 summaries category structure used in this experiment.

Table 1 Category Structure

Countries	Australia, China, France, Germany, Great Britain, Italy, Japan, South Korea, Russia, USA
Sports	Aquatics, Archery, Athletics, Badminton, Baseball, Basketball, Boxing, Canoeing, Cycling, Equestrian, Fencing, Gymnastics, Handball, Hocky, Judo, Morden Pentathlon, Rowing, Sailing, Shooting, Soccer, Softball, Table Tennis, Taekwondo, Tennis, Triathlon, Volleyball, Weightlifting, Wrestling

Rules have been created between 5th September 2008 and 13th October 2008, around 39 days. Total 3,707 rules and 7,747 conditions have been made manually in only a little more than one month. Each rule generally contains average 2 conditions. There are 1,413 refine rules and 2,294 stop rules. The reason for the number of stop rules was much greater than the number of rules was that we made rules which were too wide range and not too specific when the rules were created first. As a result, many articles which were not really related to a particular folder were

classified as well, so we inevitably created many stop rules to classify these articles. A total of 29,714 articles were classified and this is around 22% of the entire articles from the dataset. As the articles are classified into multiple categories, the total number of classification is 47,629 and this means each article was classified into 1.6 categories.

Simulated Web Monitoring

Simulated web monitoring was conducted to evaluate the event-driven web monitoring method. It is assumed that each monitoring web page publishes the collected web page published at the capturing time and the simulated monitoring system revisits each web page with given revisit intervals such as every four hours and captures new information between the last visit time and the current visit time. As the original web monitoring was conducted two hours intervals, but started at different time, there are time gaps between the original capturing time and the simulated capturing time, which was regarded as delay.

For this experiment, five categories (31, 32, 34, 36, and 40) were chosen under the 'countries' category with respect to the amount of the classified documents and the top five web pages that contributed most to these five categories were also chosen for simulated monitoring. Firstly, the simulation program conducted a static web monitoring by revisiting the monitoring web pages periodically (e.g., every two hours). This simulation provided default delay of each monitoring web page. Then the simulation program dynamically monitored using our event-based scheduling method. It is assumed that the document classification knowledge has been already acquired at the beginning of the monitoring process. Monitoring intervals for each web page were set by top-down order, bottom order and random order. The top-down order means that the shortest intervals (2 hours) were assigned to the web page that has the largest number of classified document in a category and the longest intervals (24 hours) were set to the web page that has the smallest number of classified document in a category. The bottom-up order means inversely assigned the interval times and the random order means there is no relationship between monitoring intervals and classified documents count. There setting was considered because the number of classified documents may affect on the performance of the event-driven scheduling. For each monitoring cycle, the simulation system calculated the number of classification for each category and compared it with the average value to find 'events'. If any

event was found, the system performed monitoring immediately for the relevant web pages. Each web page's delay of the event-based simulation was calculated in order to compare them with those of the static monitoring results.

5) Results

Simulated monitoring results are summarised in Table 2 ~Table 5, where the left column shows the Category IDs that were selected for this experiment and each cell represents delay time and its unit is minutes.

Static Web Monitoring Results

As the static web monitoring does not change web monitoring schedule, the number of revisit frequency is also fixed during the monitoring period as illustrated in Table 2. Revisit frequency decreases as the monitoring intervals increase. The results also show that the average delay of each category increase as the monitoring intervals increase.

Table 2 Revisit Frequency and Delay of the Static Web Monitoring

Category ID	Monitoring Intervals (Hour)				
	2	4	8	12	24
31	58.57	122.11	241.12	392.51	694.84
32	61.00	121.90	242.87	396.15	692.19
34	57.26	113.46	252.37	405.12	722.13
36	63.96	124.79	269.72	373.14	761.43
40	63.09	130.73	250.16	354.02	673.31
Grand Total	60.78	122.60	251.25	384.19	708.78
Revisit Frequency	251	125	62	41	20

Event-Based Dynamic Simulated Web Monitoring Results

Event-driven dynamic simulation results are summarised in Table 3 ~ Table 5. Firstly, the results show that delay times are significantly shortened as the event-driven scheduling is employed. Secondly, the improvements are different among different monitoring interval assignment strategy. The top-down approach show better performance compared to other strategies.

Table 3 Delay of the Dynamic Web Monitoring Using Top-Down Policy

Category ID	Monitoring Intervals				
	2	4	8	12	24
31	60.41	78.55	83.32	85.85	97.62
32	64.84	92.49	107.91	110.39	134.79
34	58.91	73.35	76.74	77.46	85.05
36	69.36	119.38	142.29	152.63	193.67
40	61.91	108.70	112.62	127.32	112.62
Grand Total	63.09	94.49	104.58	110.73	124.75

Table 4 Delay of the Dynamic Web Monitoring Using Bottom-up Policy

Category ID	Monitoring Intervals				
	2	4	8	12	24
31	64.10	87.17	95.50	99.59	113.20
32	56.06	94.34	103.15	129.82	136.19
34	60.31	76.94	86.12	89.67	106.26
36	68.00	105.68	157.90	167.95	166.81
40	60.31	102.96	133.37	163.67	124.50
Grand Total	61.76	93.42	115.21	130.14	129.39

Table 5 Delay of the Dynamic Web Monitoring Using Random Policy

Category ID	Monitoring Intervals				
	2	4	8	12	24
31	64.10	80.36	94.61	91.80	104.75
32	61.38	102.07	111.00	124.30	107.36
34	55.72	81.86	70.39	93.78	69.49
36	58.54	122.34	143.95	189.86	252.00
40	60.31	102.96	133.37	163.67	124.50
Grand Total	60.01	97.92	110.66	132.68	131.62

3 Abstract

Web monitoring systems report any changes on the target web pages by revisiting them frequently. As they are operated under significant constrains such as network and computing, it is necessary to minimize revisits with minimal delay and maximum coverage. Various statistical scheduling methods were proposed to resolve this problem. However they are static and cannot easily cope with events in the real world. This paper proposes a new scheduling method that manages unpredictable events. MCRDR (Multiple Classification Ripple-Down Rules) document classification knowledge base was reused to detect events and to initiate a prompt web

monitoring process regardless of static monitoring schedule. Our experiment demonstrates that our approach improves monitoring efficiency significantly.

4 Personnel Supported

Dr. Claire DeEstate
Research Assistant, School of Computing and Information Systems, University of Tasmania
Yang Sok Kim
Ph.D Student, School of Computing, University of Tasmania

5 Publications

Journal

Kim, Y. S., Kang, B.H. (2007). "Tracking Government WebSites for Information Integration" Information Research, 12(4) paper colis09. [Available at http://InformationR.net/ir/12-4/colis/colis09.html] (Note. Short version of this paper was presented at the 6th International Conference on Conceptions of Library and Information Science "Featuring the Future". 2007. Boras, Sweden.

Refereed Conferences / Workshops

Kim, Y.S. and B.H. Kang. *Search Query Generation with MCRDR Document Classification Knowledge*. in EKAW 2008 - 16th International Conference on Knowledge Engineering and Knowledge Management Knowledge Patterns. 2008. Acitrezza, Catania, Italy.

Kim, Y.S. and B.H. Kang. *A Study on Monitoring Web Page Locating Heuristics*. in The 2008 International Conference on Information and Knowledge Engineering (IKE'08). 2008. Monte Carlo Resort, Las Vegas, Nevada, USA.

6 Interactions

Review Meeting at AFRL/AFOSR in ROME (25/09/2008 – 26/09/2008)

Review Meeting at UTAS by the program manager, Prof. Hiroshi Motoda (AOARD) (6/12/2008 – 9/12/2008)

7 Honors/Awards

8 Archival Documentation

9 Software and/or Hardware (if they are specified in the contract as part of final deliverables):

Sample demo web monitoring site is available at this URL (Note this does not include the event driven monitoring function)
http://www.cis.utas.edu.au/iWeb/iwebAOARD2008/index.php

Prototype software is enclosed.

References

[1] Park, S.S., Y.S. Kim, and B.H. Kang. *Web Document Classification: Managing Context Change*. in *IADIS International Conference WWW/Internet 2004*. 2004. Madrid, Spain.

[2] Kim, Y.S., S.S. Park, E. Deards, and B.H. Kang. *Adaptive Web Document Classification with MCRDR*. in *International Conference on Information Technology: Coding and Computing ITCC 2004*. 2004. Orleans, Las Vegas, Nevada, USA.

informationresearch
Vol. 12 No. 4, October, 2007

Proceedings of the Sixth International Conference on Conceptions of Library and Information Science—"Featuring the Future"

Tracking government Websites for information integration

Yang Sok Kim and Byeong Ho Kang
School of Computing, University of Tasmania, Private Bag 100 Hobart TAS 7001 Australia

Abstract

Introduction. Nowadays government policies, laws, and other valuable information are published via the web. However, it is very difficult to ensure the comprehensiveness, accuracy and currency of all Webpublications manually, because there are too many Websites that are usually maintained by different departments and agencies.

Method. We proposed a Web monitoring system based Web information integration method to resolve this problem. It has been utilized for Australian and Tasmanian government Web information integration for the State Library of Tasmania since July 2005.We have been monitoring 249 federal, state and local government Websites for the Web information integration since July 2005. Monitoring Websites consist of federal government homepages (21 sites, 8%), federal government media release pages (108 sites, 44%), Tasmanian government homepages (73 sites, 30%), Tasmanian government media release pages (16 sites, 6%), and Tasmanian local government homepages (31, 12%)

Analysis. We analysed overall, monthly, weekly and daily monitoring trends and usage of the monitoring results.

Results. In total, the Web monitoring system collected 30,279 documents during the monitoring period. The most prolific domain is federal media release pages (16,075 documents, 51%). This is followed by Tasmanian government homepages (6,288 documents, 20%), Tasmanian government media release pages (4,673 documents, 15%), federal government homepages (3,243 documents, 10%), and local council homepages (1,493 documents, 5%). These monitoring results are used to modify the current Web information (62%), to add new information(22%), and to delete current Web information (16%)of two Tasmanian government information portal.

Conclusions. Our reseach shows that Web monitoring based approach is very useful, because it supports Web information

integration without requiring any changes in the current system.

change font

Introduction

Electronic government (e-Government), which refers to governments' use of information and communication technology to exchange information and services with citizens, businesses, and other arms of government, is often regarded as the new way forward for the public sector. According to the UN report on global e-Government readiness, "An increasing number of e-Government initiatives are being employed to improve the delivery of public services to the people, and to tap the potential synergy from the interaction between new technologies, an educated population and an enabling environment for the attainment of knowledge-based economies.(UN,2004). e-Government may be applied by the legislature, judiciary, or bureaucracy, in order to improve internal efficiency, the delivery of public services, or processes of democratic governance. Even though there are various service opportunities in e-Government, information publication via the Webis one of the main services, which enables people to access high quality information that was not accessible in the past. West (2004) reported that 89 percent of websites in 2004 provided access to publications and 62 percent had links to databases.

However, it is very difficult to ensure the accuracy and currency of all Webpublications manually because the Webuses a passive information delivery mechanism, called pull technology. In this case, people should visit the publication Websites to acquire new information, but this is a time consuming and uncertain process because there are too many Websites that are usually maintained by different departments and agencies. Nowadays Websites can send new information to subscribed users by using e-mail or more sophisticated methods such as RSS (Really Simple Syndication), but this is not provided by all Websites because of financial or technical limits.

This project was motivated by the Web information integration of federal, state and local government Websites. The State Library of Tasmania (SLT) operates two Websites to provide integrated government information. Tasmania Online (www.tas.gov.au) is a portal of Tasmania business, community and government Websites. The aim of Service Tasmania Online (www.service.tas.gov.au) is to provide flexible access to a wide range of resources on the Webprimarily for the state government, but also for federal and local government. Additional information on Service Tasmania Online is available at http://www.service.tas.gov.au/ stabout/stabout.asp.

We proposed a monitoring system that continuously monitors new and changed content on selected federal, state and local government

Web pages. The system provides two key services, one monitoring home pages and the other media release pages. This complements the indexing and description of those sites on Service Tasmania Online and Tasmania Online as the SLT must continuously monitor the Websites to ensure the accuracy and currency of the information provided. This process - conducted manually before we proposed this project - was a time consuming repetitive and routine task. SLT cataloguers were also required prior knowledge of the content of the monitored Websites. The Web monitoring system both saves staff time and removes a large component of a repetitive and routine task, and ensures the timely delivery of new and changed Webcontent in one access point to the SLT cataloguers.

This paper consists of the following contents: Section 2 summarizes related research on government Web information integration. Section 3 explains our Web monitoring system implementation details. Section 4 summarizes basic workflow and benefits. Section 5 summarizes Web monitoring and usage results. Conclusions and further studies are discussed in Section 6.

Related work

Government information on the Web

Providing access to government information is the most common e-Government initiative. Many governments have tried to vastly increase the number of interactions and to provide large amounts of online information for efficiency, better services to citizens and improved governmental processes. There are many benefits both for the public and for government of this kind of service: reducing distribution costs for government agencies, ensuring 24/7 access to information, removing the delay between production of and access to information, and more timely update of material.(Pardo,2000) A significant proportion of total Web information comes from governments because they provide large amounts of information via their Websites to exploit the above benefits (Wagner *et al.*).

Integration of government Web information is a significant problem, because it is created by different departments and agencies, and people cannot access all Websites to find relevant information from the government Websites. Governments have used several technologies to provide useful information to citizens through the web, such as portals(Wimer,2001), content management systems, e-mail broadcasting and list serves or discussion forums. All these solutions can help to disseminate and exchange information, but each has its own strengths and weaknesses. However these approaches are limited in their ability to provide an automatic timely and integrated information service. In fact, The realization of the full potential of e-Government depends on the same goals that government agencies have been pursuing for many years: true horizontal and vertical integration of services" (Pardo,2000).

Information integration with the semantic Web

Some researchers used Semantic Webtechnology to integrate government Web information in the semantic level (Wagner *et al.*, 2006,

Drumm, 2006, Gugliotta, 2005). The Semantic Webaims to facilitate semantic interoperability and integration by using XML based machine-processible information (McIlraith, 2001). Although the Semantic Webis regarded as the future of the current web, there are some limitations because large parts of legacy systems do not support a Semantic Webservice. It is not easy to transform legacy information into Semantic Webserviceable information even though some middleware can help legacy systems to provide a Semantic Webservice. We did not use Semantic Webtechnology for this project because our aim is to provide Web information integration without changing legacy systems or information or adding any additional middleware to the existing systems.

Timely information integration with Web monitoring

Timely Web information integration can be accomplished by the Web monitoring technique. Web monitoring systems/services collect new information from selected Websites by continually revisiting these sites. By doing this they can obtain new information in real-time. There are other systems that support real-time information dissemination. Nowadays XML based RSS services are prevalently used to serve real-time information provision (Powers,2005). People can get new information by registering RSS service URLs, called feeds, to the RSS aggregator systems. However, these systems/services have limitations because they only work when the service Websites provide XML based RSS services. In comparison with an XML based RSS service, the most significant benefit of the Web monitoring method is that we can obtain information in real-time without changing current Webcontents. The Web monitoring system has been researched since the beginning of the Web (Pandy *et al.*, 2004, Liu *et al.*, 2000, Tan, 2002) and nowadays the following Web monitoring systems or services are available:

Service/Product Name	URL
WatchThatPage	http://www.watchthatpage.com
Wisdomchange	http://www.wisdomchange.com
ChangeDetection	http://www.changedetection.com
ChangeDetect	http://www.changedetect.com
Track Engine	http://www.trackengine.com
Copernic Tracker	http://www.copernic.comen/products/tracker
WebsiteWatcher	http://aignes.com

Table 1: Web monitoring System

Although these systems and services are available, there is no significant large scale research that focuses on the Web information integration of government Websites. In this research, we focus on this issue and report operation results of 248 Websites over 1.5 years.

Web information monitoring system

We developed a Web information monitoring system to collect newly uploaded publications from selected Web pages. The system monitors specific Web pages that consist of various objects such as text, hyperlink, and images. Hyperlinks are the most important object among them, because they are usually linked to the specific Web document. A hyperlink consists of link text that is located between <a> and tag and a URL that indicates the location of a specific document (linked content).

The system works as follows. The system sends an HTTP request message to the Web server of the registered Web page according to the fixed revisit time ($T_{revisit}$). The revisiting time (Trevisit) is affected by the publication frequency of the source pages and the user's need for information. If the publication frequency is high, Trevisit should decrease otherwise increase. If the user wants to get information as soon as possible, Trevisit should decrease otherwise increase. As we had no prior information about the publication patterns of the selected domains, we employed a single fixed scheduling strategy. We set the revisit time ($T_{revisit}$) as 2 hours for all the Web information source pages.

When the system receives an HTTP response message from the server, the system extracts URLs and their link texts from the HTTP response body (H_c) and compares them with those of the last HTTP response body (H_p).

H_c and H_p are defined as follows

Hc = $\{(U_i,T_i)\}$, where H_c is a set of URLs and link texts of the current HTTP response message body. (U_i,T_i) is the ith pair of hyperlink and link text of a hyperlink.

Hp = $\{(U_j,T_j)\}$, where Hp is a set of URLs and link texts of the last HTTP response message body. (U_j,T_j) is the jth pair of URL and link text of a hyperlink.

Newly updated information is $I_n = H_c - H_p$

For more clarity, we would like to indicate the following things. Firstly, according to this definition, if the URL or link text is changed, it is regarded as new information. However, this approach cannot detect new information when the URL and the link text are not changed, but the linked content has changed. Although this limitation may cause under-detecting of new information, we employed this approach to save cost. Secondly, we excluded URLs that contain session id in the URL path. When a URL contains session id, it is detected as new information whenever the monitoring system requests the Web pages, because session id is created on the fly and therefore the URL is continually changed. Thirdly, we eliminated some URLs from the new URLs by registering filtering URLs. For

example, advertising URLs are collected as new information, but our URL filter (H_f) eliminates these URLs before they are recorded in the database. Lastly, we excluded URLs that were already recorded in the database. Some URLs may have been harvested already before the last session (H_e). For example, a URL is collected during the first monitoring session and disappears in the second session, and reappears in the third session. In this case the URL is regarded as new information at the third session according to the above definition. Therefore, the system checks if a new URL already exists in the database, and it is only recorded in the database if it is not already listed there. New information is redefined as follows:

$$I_n = H_c - H_p - H_f - H_e$$

We did not use data that was collected from the first session, because they included the navigational information as well as old information that was published before monitoring was started.

Workflows and benefits

Government Web information integration workflow with a Web monitoring system is summarized in Figure 1. The system automatically identifies new information on the selected federal, state and local government Web pages, and they are reported to the SLT cataloguers via the Webbased change reporting system, which reports new information in real time. The cataloguers need to review the records in the tracking reporting service and identify which records need follow up and adding to the Service Tasmania Online and Tasmania Online Websites. Public users can access new information from these two Websites.

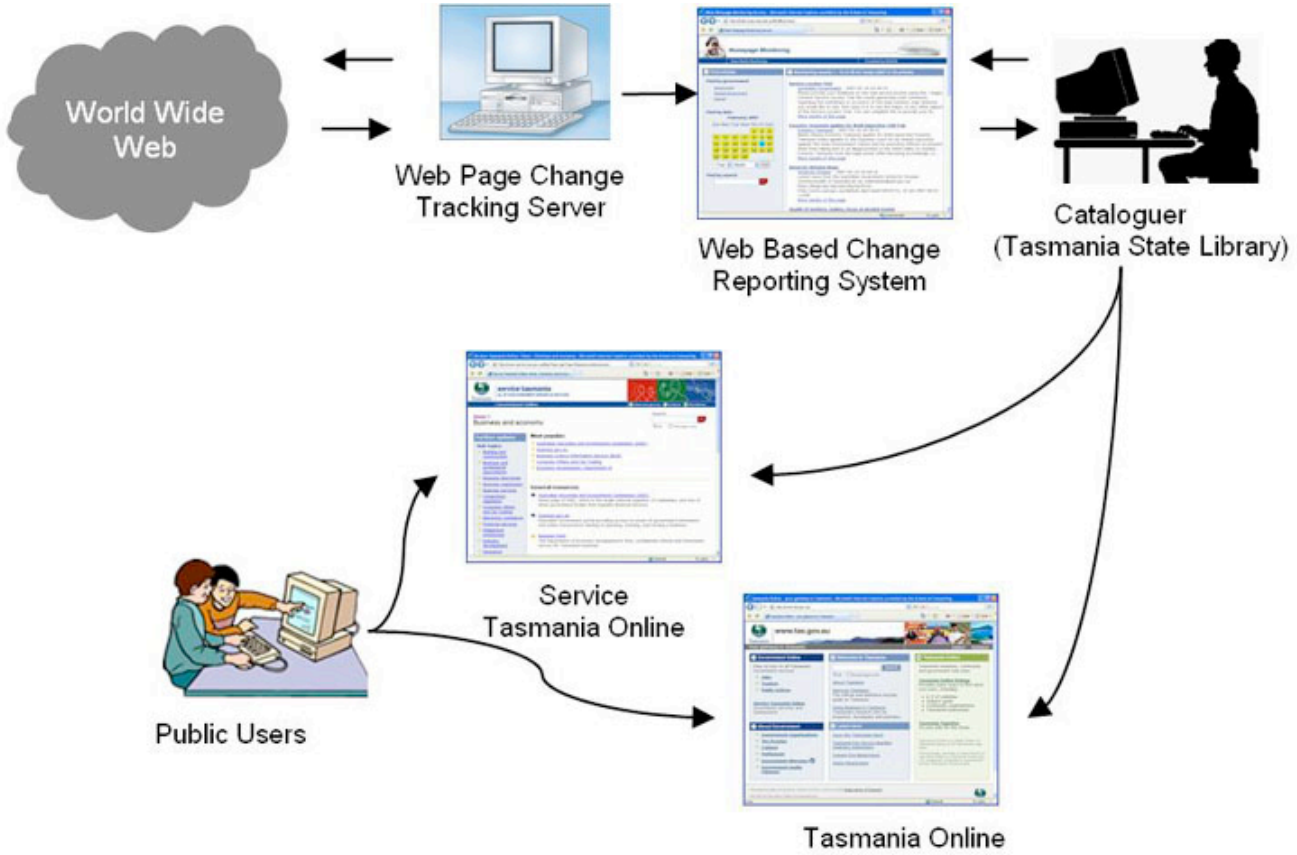


Figure 1: Government Web information Integration Workflow

The SLT reported the following benefits of using our system and applying the new workflow.

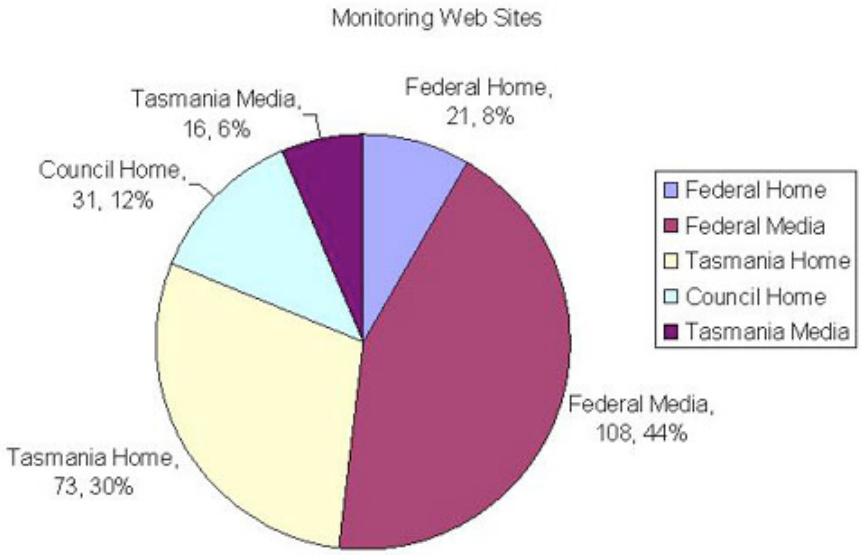
- Savings in staff time.
- Elimination of a significant component of a repetitive and routine task.
- Increased staff satisfaction as they no longer navigate to a site only to find there is no new content.
- A more comprehensive and timely scan for new and changed content is now possible. Certain key Websites were previously checked daily (e. g., http://www.premier.tas.gov.au, http://www.dhhs.tas.gov.au, and http://www.australia.gov.au). Other major state departments were checked weekly, with other agencies - especially commonwealth and local government agencies checked on a monthly or irregular basis. Much new content was not identified or was dated by the time it was identified.
- State wide coverage is improved significantly with daily monitoring of council sites.
- Inexperienced staff can monitor sites as no prior knowledge of the site content is required. They are assessing only new resources on the change reporting system. The previous manual process meant that it was difficult to assess new content unless staffs were very familiar with the content of agency Web sites.
- The appearance of many new Web pages from one agency on the change reporting system can alert cataloguers to a major change in that

agency's Website. The cataloguers will then undertake a review of all URLS for that agency listed on Service Tasmania Online. A major agency recently changed its Website and cataloguers were alerted on the day of the change. In the past they may have waited for three days until the resources appeared in a broken URL report. If redirects had been put in place, a number of months until the change was identified in a regular quality assurance check.

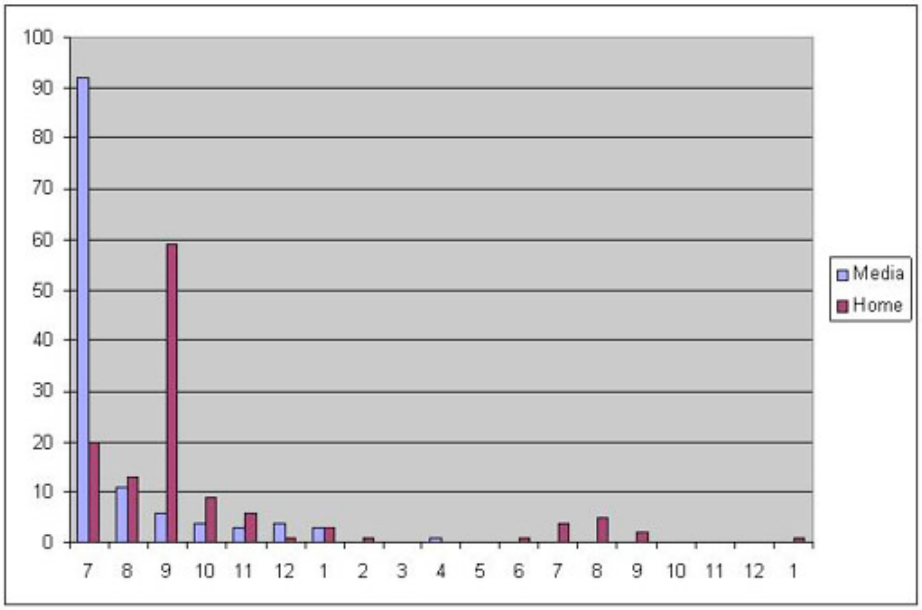
- New Web pages mentioned on the change reporting system may lead the cataloguers to related content that they decide to index. For example, they did not index a media release on grants to Greening Australia (http://www.deh.gov.au/minister/env/2005/mr01 nov2005.html), but investigation of this media release led them to identify a site for the Tasmanian page for Greening Australia (http://www.greeningautralia.org.au/GA/TAS/). This site was then indexed on Tasmania Online.

Results

We have been monitoring 249 federal, state and local government Websites for the Web information integration since July 2005. Monitoring Websites consist of federal government homepages (21 sites, 8%), federal government media release pages (108 sites, 44%), Tasmanian government homepages (73 sites, 30%), Tasmanian government media release pages (16 sites, 6%), and Tasmanian local government homepages (31, 12%) as illustrated in Figure 2 (a). At first we started with 112 Websites, but we gradually added additional Websites as requested by the SLT (Figure 2 (b)).



(a) Monitoring sites by domains



(b) Monitoring site registration

Figure 2: Monitoring sites

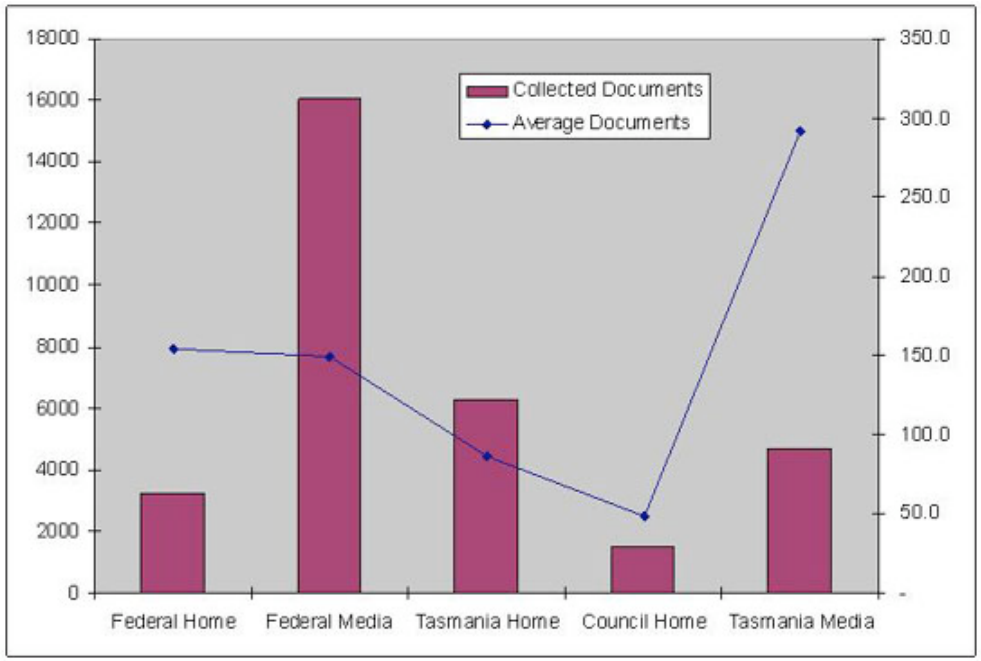
Overall results

Figure 3 summarizes overall monitoring results from July 2005 to January 2007. As seen in Figure 3 (a), the most prolific domain is federal media release pages (16,075 documents, 51%). This is followed by Tasmanian government homepages (6,288 documents, 20%), Tasmanian government media release pages (4,673 documents, 15%), federal government homepages (3,243 documents, 10%), and local council homepages (1,493 documents, 5%).

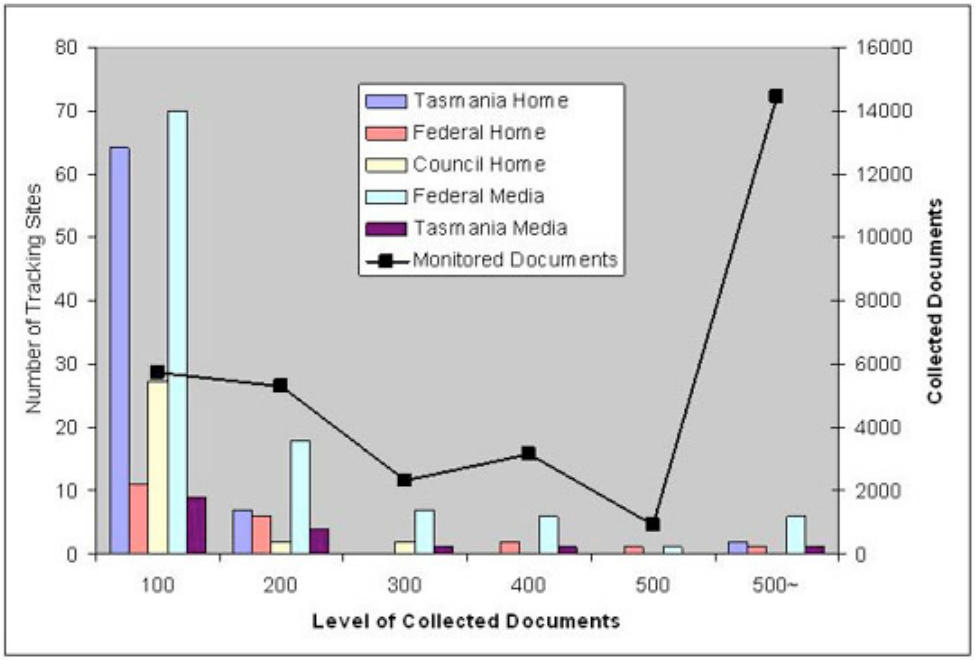
Figure 3 (b) illustrates distribution of Web pages by monitoring result frequencies. The horizontal axis represents the level of collected documents of each Web page, the left vertical axis shows the number of monitoring sites of each domain in each publication level, and the right vertical axis represents the total number of collected documents in each level. Though the federal media release pages and Tasmanian government homepages are more prolific than Tasmanian government homepages (see Figure 3(a)), they have more monitoring results in the 100 publication level compared to that of the Tasmanian government media release pages. For this reason, the Tasmanian government media release pages, the third most prolific domain, shows the highest number in average documents per Web pages.

Figure 3(b) illustrates another fact. Even though the number of monitoring sites in the level of over 500 publication (500~) is smaller

than that of other sites, they contribute large parts of the overall monitoring documents. That is, only a small number of Web pages contribute large parts of the total collected documents. This result means that we need to reconsider the monitoring scheduling strategy, because it implies that some new information checking sessions of the monitoring system is unnecessary. The scheduling time should be effectively changed according to the publication characteristics of the monitoring Websites to minimize the monitoring costs.



(a) Monitoring Results by Domains

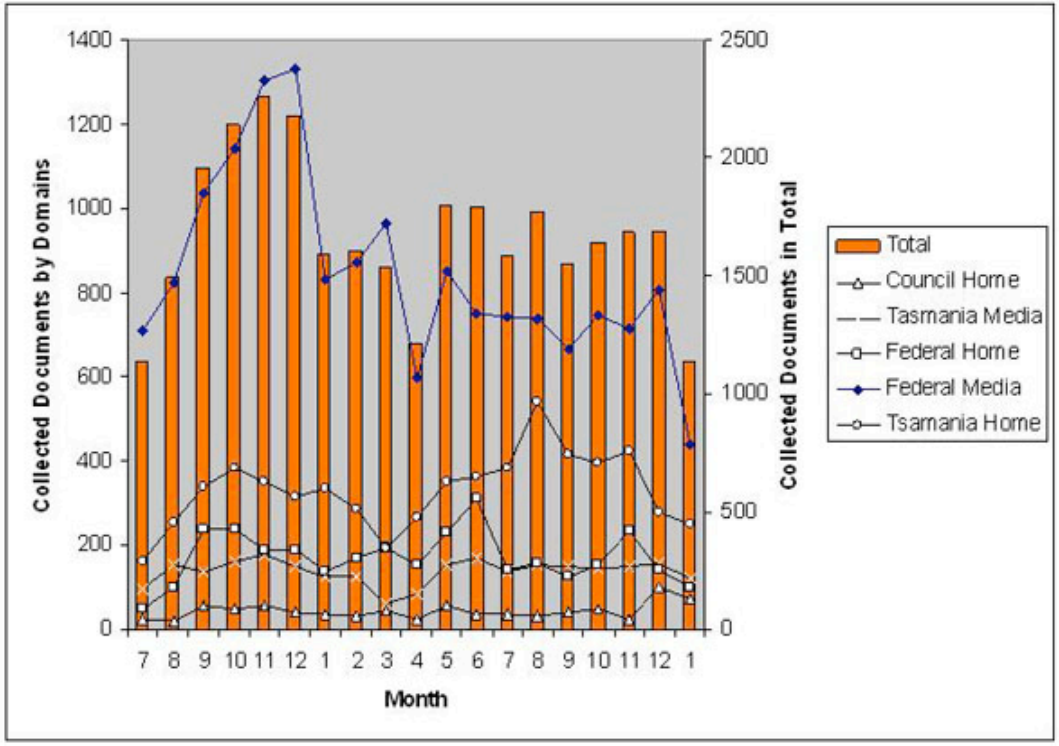


(b) Distribution of Monitoring Results

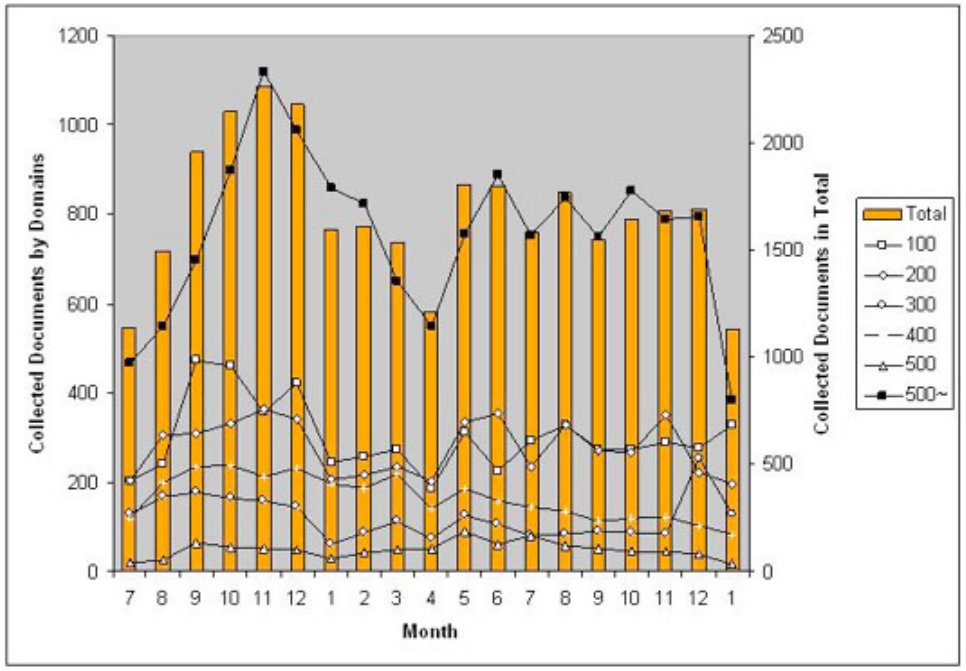
Figure 3. Monitoring Results

Monthly trends

Figure 4 (a) illustrates monthly monitoring results by domains over the operation period. We need to consider the trends from September, 2005 since we added most Websites from September, 2005. First of all, we can see that the total of collected documents from May to December is greater than those from January to April. However, though there are significant drops in January, 2007, we can not conclude that this trend is stable, because we only reviewed about 1.5 years. This trend requires further investigation, although given the nature of government activity during the summer holiday period it is probable this trend will continue. The trend of the total monitoring results is similar to those of each domain. Especially the trend of the federal media releases represents high similarity. Figure 4 (b) illustrates monthly monitoring results by the collected documents frequency level. The result shows that the overall trend is closely related to that of the 500 ~ publication level.



(a) Monthly Trends

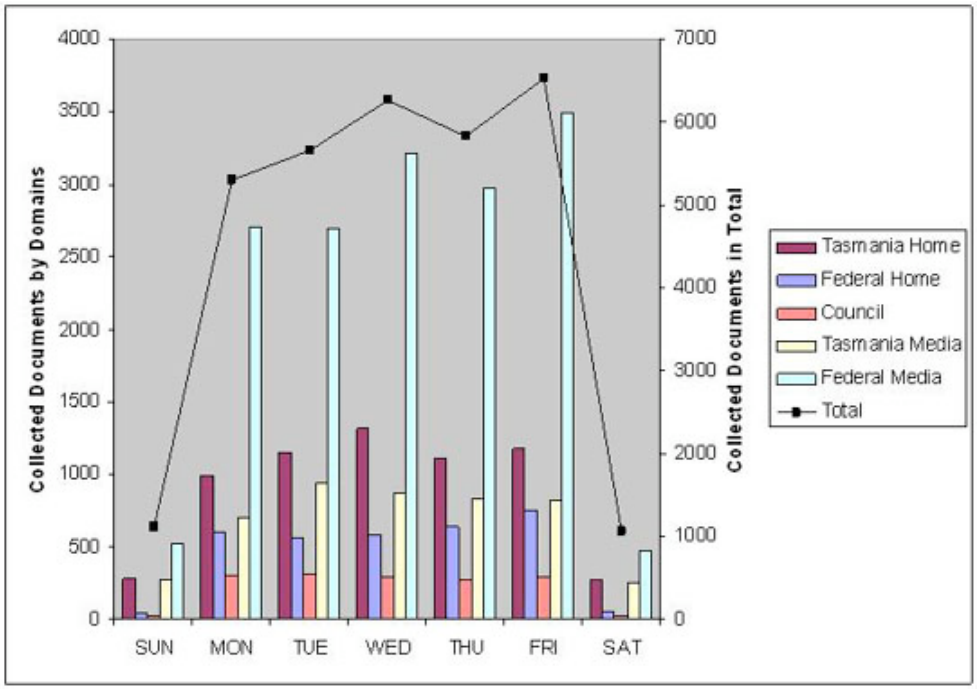


(b) Monthly Trends By Publication Level

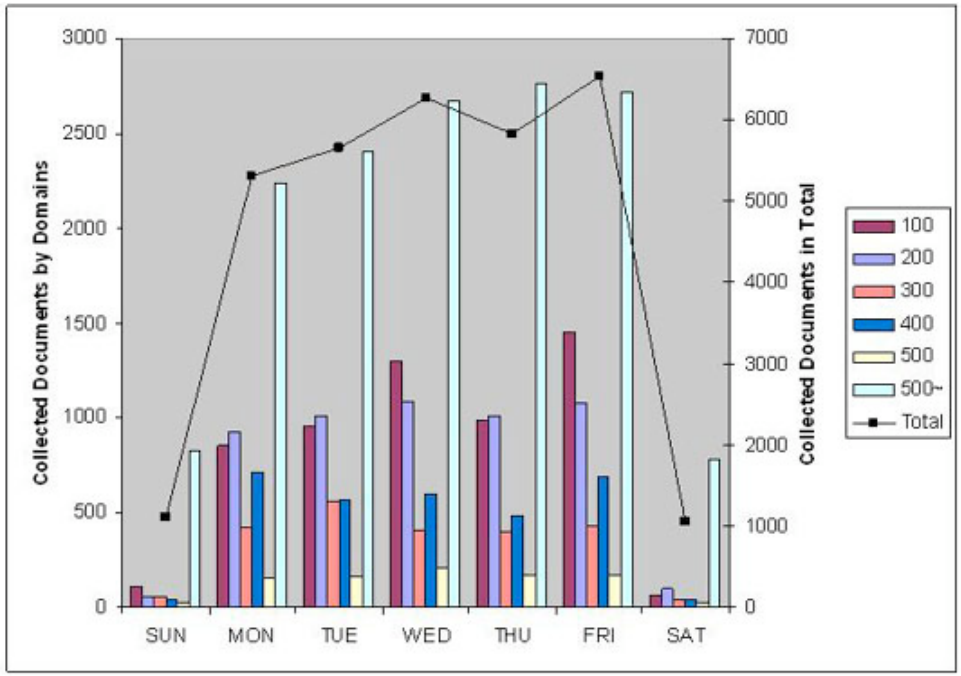
Figure 4: Monthly Monitoring Results

Weekly trends

Figure 5. illustrates the weekly trend of the collected documents. Figure 5 (a) shows weekly trends by domains and Figure 5 (b) shows weekly trends by publication level. Not surprisingly, more documents are collected during working days, Monday to Friday. There are no significant differences among domains and different publication levels except the documents from the 500~ publication level indicates that they published more new information on Saturdays and Sundays than other level Websites. This result implies that we wasted computing resources to process unnecessary monitoring sessions in the current fixed monitoring scheduling strategy and we may use weekly variations to create a more dynamic scheduling strategy.



(a) Weekly trends

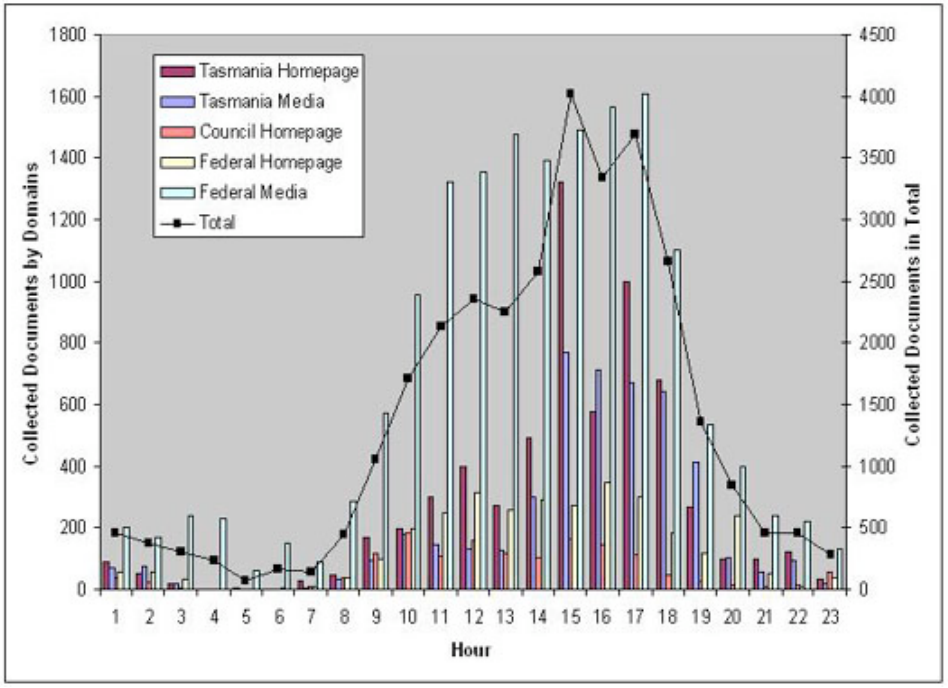


(b) Weekly trends by publication level

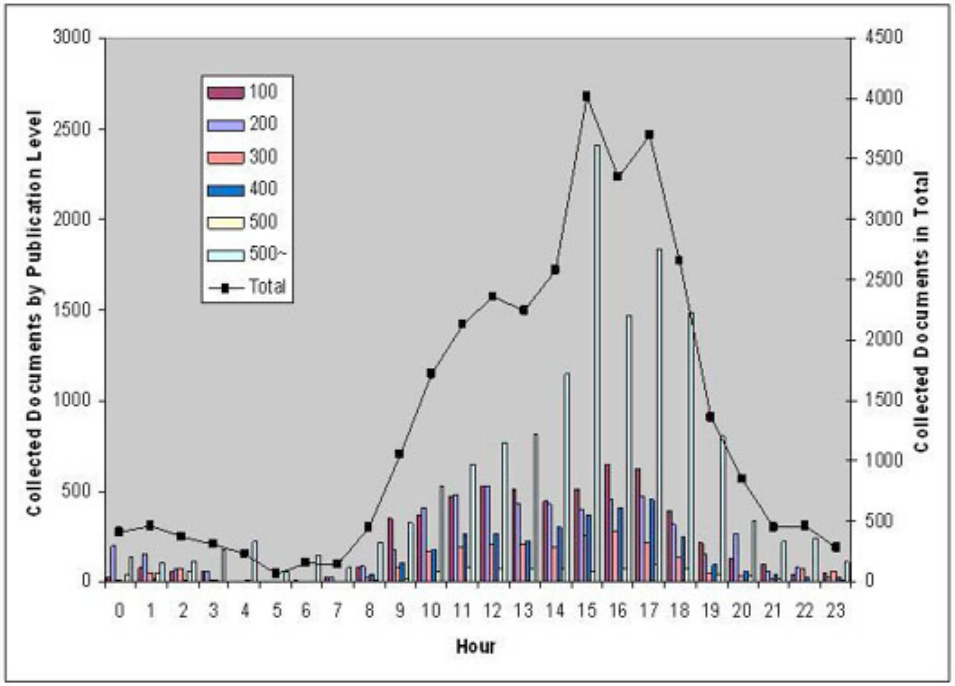
Figure 5. Weekly monitoring results

Daily trends

Figure 6 illustrates the daily trend of the collected documents. Figure 6 (a) compares daily trend by domains and Figure 6 (b) by publication levels. The results show that most documents were collected around working hours, from 9 am to 8 pm. There is no significant difference among domains and publication levels. Daily trend results also propose a further research on the scheduling strategy of our monitoring system because we wasted some resources for unnecessary monitoring sessions in a day and we may minimize monitoring cost by employing an appropriate scheduling strategy.



(a) Daily Trends



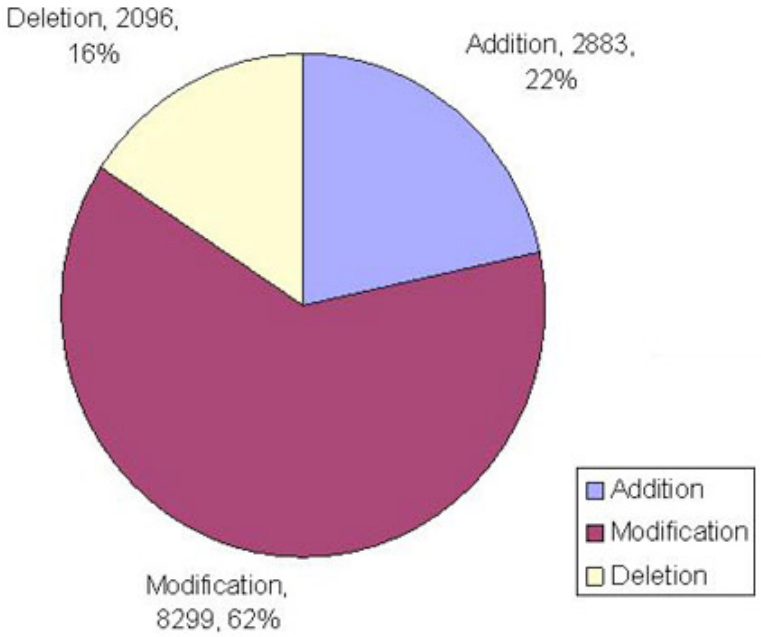
(b) Daily Trends By Publication Level

Figure 6: Daily monitoring results

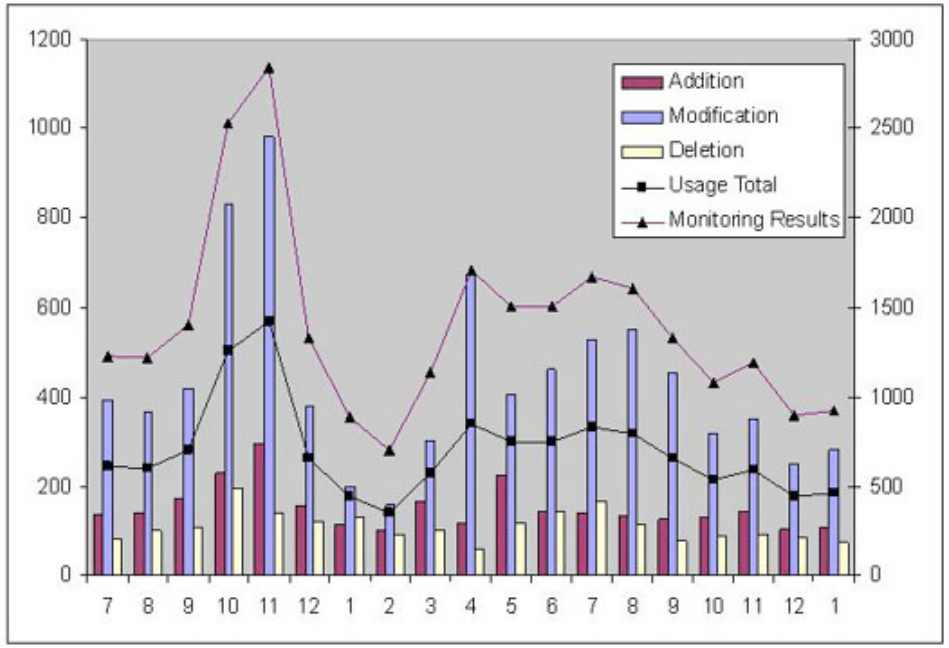
Usage results

Figure 7 illustrates overall usage results and monthly trend, which is are reported by the SLT. According to their report, they used other sources such as newspapers and internal reports on Web pages changes to change Web information on the Service Tasmania Online and Tasmania Government Online. Therefore, the report on usage trends not exactly match with the total usages of monitoring results. However, we can draw trends of monitored document usage result under this assumption. Figure 7 (a) shows the overall usage statistics. Modification of the current Web information (62%) is the largest part of usage, which is followed by 'Addition' (22%) and 'Deletion* (16%). Figure 7(b) shows usage trend sin July, 2005. The result shows that overall usage trends very closely related to the monitoring results trends.

However, the SLT cataloguers reported that the number of documents from the monitoring system indexed directly on Tasmania Online and Service Tasmania Online is very small compared to the total number of monitored pages. This is not necessarily a problem for the SLT cataloguers as they wish to be alerted to new and changed content even though this content may not always be appropriate to be added to the Tasmania Online and Service Tasmania Online Websites. The cataloguers assess the monitored pages and make a professional judgement as to whether resources need to be added or updated on Tasmania Online and Service Tasmania Online. The SLT cataloguers continually review the URL seed list to remove redundant or unhelpful URLS. Some technical refinements may require investigation to improve the performance of the monitoring service. Possible refinements include the removal of duplicated resources and the elimination of noise with certain categories of updated information being excluded from monitoring. For example updated weather reports from monitored pages could be excluded.



(a) Overall results



(b) Usage trends

Figure 7: Monitoring service usage results

Conclusions and further work

Government Web information integration is an important issue in e-Government implementation. In this paper, we proposed a monitoring system based Web information integration method and reported 1.5 years operation results. This approach is very useful because it supports Web information integration without requiring any changes in the current system. Although our system successfully supported the Web information integration for the SLT, the following significant challenging issues were raised during the project. These issues will be the basis for further research. Firstly, we need to extend our scheduling system to minimize monitoring costs. We found that there were yearly, monthly, weekly, and daily variation patterns in the past data set. These patterns can be used to establish improved scheduling strategies. Secondly, we need to provide recommendation system that helps the SLT cataloguers. The current system only display the newly updated information from the target Websites and the cataloguers manually add, delete, or modify Web information on the Service Tasmania Online and the Tasmania Government Online. The further system should provide further recommendations such as addition, deletion, and modification with the monitored results.

Acknowledgements

This work was supported by the Asian Office of Aerospace Research and Development (AOARD) (AOARD-06-4006)

References

- Drumm, C. (2006). *Integrating eGovernment Services using Semantic WebTechnologies*. Paper presented at the The Semantic Webmeets eGovernment (2006 AAAI Spring Symposium), Calif., USA.
- Gugliotta, A., Cabral, L., Domingue, J., Roberto, V., Rowlatt, M., & Davies, R. (2005). *A Semantic WebService-based Architecture for the Interoperability of E-government Services*. Paper presented at the ICWE 2005 - 5th International Conference on WebEngineering, Sydney, Australia.
- Liu, L., Pu, C., & Tang, W. (2000, Nov. 7-10, 2000). *WebCQ: Detecting and Delivering Information Changes on the Web*. Paper presented at the International Conference on Information and Knowledge Management (CIKM), Washington D.C.
- McIlraith, S. A., Son, T. C., & Zeng, H. (2001). Semantic Webservices. *Intelligent Systems, IEEE*, **16**(2), 46- 53.
- Pandey, S., Dhamdhere, K., & Olston, C. (2004). *WIC: A General-Purpose Algorithm for Monitoring Web information Sources*. Paper presented at the 30th VLDB Conference, Toronto, Canada.
- Pardo, T. A. (2000). Realizing the Promise of Digital Government: It's More than Building a WebSite. *Information Impact Magazine*, 2000
- Powers, S. (2005). *What Are Syndication Feeds.* Cambridge: O'Reilly.
- Tan, B., Foo, S., & Hui, S. C. (2002). Web information monitoring for competitive intelligence. *Cybernetics and Systems*, **33**(3), 225-251.
- UN. (2004). *Global E-Government Readiness Report 2004: Towards Access for Opportunity*. (No. UNPAN/2004/11)New York:

UNPAN.

- Wagner, C., Cheung, K. S. K., Ip, R. K. F., & Bottcher, S. (2006). Building Semantic Webs for e-government with Wiki technology. *Electronic Government, an International Journal*, **3**(1), 36 - 55.
- West, D. M. (2004). *Global E-Government*. Providence, RI: Brown University.
- Wimmer, M. A. (2001). *European Development towards Online One-stop Government: The "eGOV" Project*. Paper presented at the ICEC2001 Conference, Vienna.

How to cite this paper

Kim, Y. S., Kang, B.H. (2007). "Tracking Government WebSites for Information Integration" *Information Research*, 12(4) paper colis09. [Available at http://InformationR.net/ir/12-4/colis/colis09.html]

Contents | Author index | Subject index | Search | Home

Search Query Generation with MCRDR Document Classification Knowledge*

Yang Sok Kim and Byeong Ho Kang

School of Computing, University of Tasmania, Sandy Bay,
7005 Tasmania, Australia
{yangsokk,bhkang}@utas.edu.au

Abstract. The MCRDR (Multiple Classification Ripple-Down Rules) Classifier was developed to classify documents incrementally. A knowledge base of MCRDR-Classifier consists of two types of rules (refining and stopping rules), categories into which documents are classified, and cornerstone cases used for creating new rules. As document classification knowledge reflects user's preference for documents, it can be used to generate search queries to retrieve relevant web pages from public search engines. This research aims to propose various query generation methods using MCRDR knowledge base and evaluates them to choose the best one. For this purpose, search queries were generated from ten users' knowledge bases using the proposed query generation methods and then they were submitted to MSN web search service to retrieve search results. Search results were evaluated with discriminative power (how search results are distinctive?) and domain similarity (how search results are similar to the user's interest?) criteria to select the best query generation methods.

Keywords: MCRDR, Knowledge Reuse, Search Engines, Search Query Generation.

1 Introduction

Web monitoring systems have been proposed as a complementary WIFT (Web Information Finding Technologies) to the crawling systems. Web monitoring system collects new information by revisiting registered web pages and by comparing objects in newly retrieved web page with those of its old version web page [1]. Therefore, appropriate monitoring web pages should be registered before web monitoring operation. They can be added manually, but the process is potentially tiresome and annoying. Therefore, it is useful to register new monitoring web pages automatically. This paper proposes a relevant web page finding method used for this purpose. This paper consists of the following contents: Section 2 explains MCRDR classification system. Section 3 summarizes four types of query generation methods reusing the MCRDR classification knowledge base. Section 4 describes our evaluation methodology of query generation methods. The experimental results are summarized in Section 5. Section 6 concludes this paper.

* This work was supported by the Asian Office of Aerospace Research and Development (AOARD).

A. Gangemi and J. Euzenat (Eds.): EKAW 2008, LNAI 5268, pp. 292–301, 2008.

2 MCRDR Document Classification System

A document classification system, which employed MCRDR knowledge acquisition method [2], has been developed by a research at the University of Tasmania [1, 3]. It was used to classify web documents that were collected by WebMon, a web information monitoring system, incrementally.

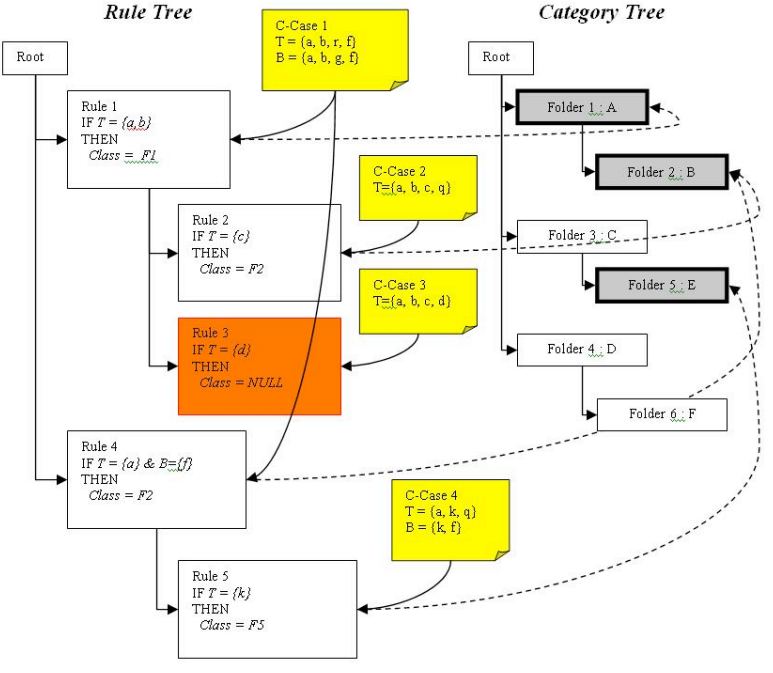


Fig. 1. The knowledge base of MCRDR document classification consists of rule tree, category tree, and cornerstone cases

It is essential to understand its knowledge base structure for formulating search query using it. Fig. 1 illustrates an MCRDR knowledge base structure, where relationship between rule, category, and cornerstone cases is explained. A common folder structure can be used to represent a category tree, because it can be easily maintained by domain experts for managing a conceptual domain model by using simple folder manipulation. The right tree of Fig. 1 represents a category tree, where the category is named Folder i, where there is a hierarchical relationship among categories. The user's heuristic classification knowledge is maintained by an n-ary rule tree. The left tree of Fig. 1 represents a rule tree, where each node is a rule and has a parent-child relationship. A child node refines its parent node rule – each child rule is added as an exception of the parent rule. For example, Rule 2 is an exception rule of Rule 1. One special exception rule is the stopping rule, which has no indicating category in the conclusion part. Rule 3 is an example of a stopping rule. More detailed explanations of the MCRDR-Classifier is explained in [3].

3 Query Generation with MCRDR Knowledge Base

Rule Condition-Based Search Query. Rule condition-based queries were constructed using rule conditions selected by the user as representative words of a cornerstone case. Table 1 summarizes search queries based on rule conditions used in the MCRDR knowledge base exemplified in Fig. 1. The following principles were considered when search queries were formulated:

- As the MCRDR classification rules have a hierarchical relationship, condition words of distinct rule paths were considered ;
- The leaf rule nodes of the rule tree were the only ones considered for rule condition based query formulation. Because of this principle, rule paths 1 and 5 were excluded from the query generation (Fig. 1) ;
- If the leaf node is a stopping rule, the rule path was not considered, because if the search engine retrieved documents to satisfy the rules of this rule path, these would not be relevant documents. Rule path 4 was discarded for this reason ;
- Even though the types of condition words, title or body, have significant meaning in classification, they were not considered because there is no method for submitting a query with those kinds of options in commercial search engines ;
- Search queries were formulated by adding each rule's condition words in top-down and bottom-up order; and
- Stop word elimination and stemming were not conducted because search engines have their own internal processes to achieve this.

Table 1. Rule-Condition-Based Search Query

No.	Rule Path	Rule Type	Rule Condition	Used	Formulation Method	Query Type	Query
1	Rule 1	Normal	T={a, b}	No			
2	Rule1–Rule 2	Normal	T={a, b}&T={c}	Yes	Top-down	Type 1	{a b c}
3	Rule 1–Rule 2	Normal	T={a, b}& T={c}	Yes	Bottom-down	Type 2	{c a b}
4	Rule 1–Rule3	Stopping	T={a, b}& T={d}	No			
5	Rule 4	Normal	T={a}&B={f}	No			
6	Rule 4–Rule 5	Normal	T={a}&B = {f}&T={k}	Yes	Top-down	Type 1	{a k f}
7	Rule 4–Rule 5	Normal	T={a}&B = {f}&T={k}	Yes	Bottom-down	Type 2	{k a f}

Category-Based Search Query. Category based-search queries were formulated utilizing the category tree used by the users. As illustrated in Fig. 1, the category tree is a hierarchical structure and each category (folder) has a specific name such as "Tasmanian Government" or "IT Policy". Each category is used as the conclusion of a rule. Table 2 summarizes the category based search queries extracted from the MCRDR knowledge base exemplified in Fig. 1. The following principles were considered when the search queries are generated from the category tree:

- The leaf nodes of the category tree were only considered for query formulation. Due to this principle, Category paths 1, 4 and 7 were excluded (Fig. 1) ;

- Only category paths that had been used for classification were employed for search query formulation. Thus Category path 8 was excluded for this reason (Fig. 1);
- Search queries were formulated by adding each category name in a top-down and bottom-up order; and
- Stop word elimination and stemming were not conducted as search engines have their own internal processes.

Table 2. Category-Based Search Query

No.	Category Path	Used	Folder Name	Used for Query	Formulation Method	Query Type	Query
1	Folder1	Yes	A	No			
2	Folder1–Folder2	Yes	A & B	Yes	Top-Down	Type 3	{A B}
3	Folder1–Folder2	Yes	A & B	Yes	Bottom-Up	Type 4	{B A}
4	Folder3	No	C	No			
5	Folder3–Folder5	Yes	C&E	Yes	Top-Down	Type 3	{C E}
6	Folder3–Folder 5	Yes	C&E	Yes	Bottom-Up	Type 4	{E C}
7	Folder 4	No	D	No			
8	Folder4–Folder6	No	D	No			

Table 3. Combined Search Query

No.	Rule Path	Rule Condition		Category		Query Type	Query
		Formation	Query	Formation	Query		
1	Rule1–Rule2	Top-down	{a b c}	Top-down	{A B}	Type 5	{A B a b c}
2	Rule1–Rule2	Top-down	{a b c}	Top-down	{A B}	Type 6	{a b c A B}
3	Rule1–Rule2	Top-down	{a b c}	Bottom-down	{B A}	Type 7	{B A a b c}
4	Rule1–Rule2	Top-down	{a b c}	Bottom-down	{B A}	Type 8	{a b c B A}
5	Rule1–Rule2	Bottom-down	{c a b}	Top-down	{A B}	Type 9	{A B c a b}
6	Rule1–Rule2	Bottom-down	{c a b}	Top-down	{A B}	Type 10	{c a b A B}
7	Rule1–Rule2	Bottom-down	{c a b}	Bottom-down	{B A}	Type 11	{B A c a b}
8	Rule1–Rule2	Bottom-down	{c a b}	Bottom-down	{B A}	Type 12	{c a b B A}
9	Rule4–Rule5	Top-down	{a k f}	Top-down	{C E}	Type 5	{C E a k f}
10	Rule4–Rule5	Top-down	{a k f}	Top-down	{C E}	Type 6	{a k f C E}
11	Rule4–Rule5	Top-down	{a k f}	Bottom-down	{E C}	Type 7	{E C a k f}
12	Rule4–Rule5	Top-down	{a k f}	Bottom-down	{E C}	Type 8	{a k f E C }
13	Rule4–Rule5	Bottom-down	{k a f}	Top-down	{C E}	Type 9	{k a f C E}
14	Rule4–Rule5	Bottom-down	{k a f}	Top-down	{C E}	Type 10	{C E k a f}
15	Rule4–Rule5	Bottom-down	{k a f}	Bottom-down	{E C}	Type 11	{k a f E C}
16	Rule4–Rule5	Bottom-down	{k a f}	Bottom-down	{E C}	Type 12	{E C k a f}

Combined Search Query. The rule condition-based search query and the category-based search query can be combined to generate new types of search queries. There are two ways of combining these query sets. On the one hand, the combined query can be generated by joining the conclusion category path of a rule with its own rule condition query. This outcome is derived from the fact that each rule has conditions and a relevant conclusion category. On the other hand, the combined queries may be constructed by joining a category path with multiple rules that indicate the category as a conclusion. This option can be considered because the MCRDR classifier allows multiple rules indicating the same category. However, it was discarded because as the

number of rules that indicate the same category increases, the term count of these combined queries increases and the query becomes too specific. In total eight types of search query sets can be constructed from the first option. Table 3 summarizes these search query sets created by Rule1-Rule 2 and Rule4-Rule 5 paths.

Case-based Search Query. Case based-search queries were generated from each rule's cornerstone case document, which was saved when each rule was created. A normalized term frequency was used to choose search query terms from each cornerstone case, and was measured by the following formula:

$$n(i) = 0.5 + 0.5\frac{tf(i)}{\max tf},$$

where *tf(i)* is a term frequency of term *i* in a cornerstone case, and *maxtf* is a maximum term frequency in a cornerstone case. This formula normalizes term frequencies so they range from 0.5 to 1.0. The top 10%, 20%, and 30% of terms in a cornerstone case were chosen as search query terms. However, the search queries formulated with the top 10% threshold may include too few terms (e.g., one or two) when the distribution of term frequency is too skewed. For this reason, instead of the highest term frequency, the second highest term frequency was also considered, which allows at least two words can be obtained from a case. The top 10% and 20% of terms in each cornerstone document were chosen as search queries, but the 30% threshold was discarded because too many words were chosen. Table 4 summarizes five types of the cornerstone case-based search queries.

Table 4. Case-based Search Query

No.	Term Frequency Normalization	Threshold	Query Type
1	Maximum frequency based normalization	Top 10%	Type 13
2	Maximum frequency based normalization	Top 20%	Type 14
3	Maximum frequency based normalization	Top 30%	Type 15
4	Second maximum frequency based normalization	Top 10%	Type 16
5	Second maximum frequency based normalization	Top 20%	Type 17

4 Evaluation Methodology

Fig. 2 illustrates our approach, where a web search service and a MCRDR (Multiple Classification Ripple-Down Rules) document classification knowledge base are reused to retrieve search results relevant to user interest. At first, search queries are generated by the query generator from the user's knowledge base and they are submitted to the search engines such as Google, Yahoo, and MSN to obtain relevant web pages. The query evaluator check search results using each user's MCRDR-Classifier. Evaluation results were analyzed by two criteria – the discriminative power and the domain similarity. Firstly, discriminative power was used to examine how distinctive the search results were when a set of search queries was employed to retrieve search results. Discriminative power (*D*) is defined by the following formula:

$$D = \frac{S_d}{S_t},$$

where S_d is the number of distinct search results and S_t is the number of total retrieved search results. The higher the value of D, the more search query sets are deemed to contain distinctive search results. Secondly, domain similarity (S) was measured to check how relevant the search results to the user's interests using each user's classification knowledge. Domain similarity (S) is measured by the following formula:

$$S = \frac{S_f}{S_t},$$

where S_f is the number of search results classified by knowledge bases that used for generating search queries and S_t is the number of total retrieved search results.

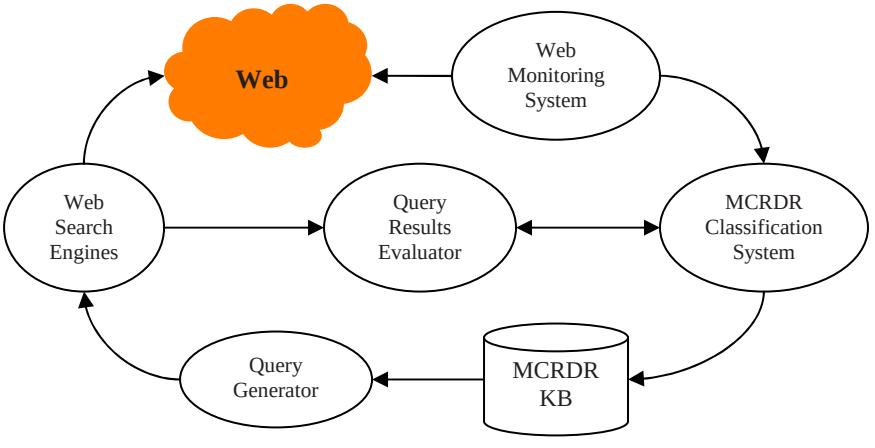
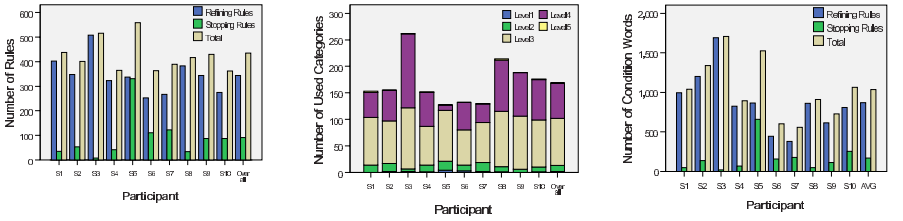


Fig. 2. A web search service and a MCRDR document classification knowledge base are reused to retrieve search results relevant to user interest

The participants were ten Masters or Honours course students from the School of Computing at the University of Tasmania. For this experiment, WebMon, a web monitoring system, collected newly uploaded web documents from 249 Australian and Tasmanian Government homepages and media release pages. Each participant created new rules to classify unclassified web pages or to reclassify misclassified web pages. The hierarchical category tree, which has five levels and 769 categories, was given to the participants at the beginning of the experiment. The students incrementally constructed their classification knowledge base, even though the category tree was given to them. In the trial phase, MCRDR-Classifier was provided to the participants for two weeks before the main classification phase to give them some experience with it. After resetting the trial period knowledge base, classification of the monitored web pages was undertaken for eight weeks from 21 August, 2006, to 9 October, 2006. A total of 2,567 web pages were collected during the experiment period. On average, 46 documents were presented to the participants each day. Each participant's classification results are illustrated in Fig. 3.



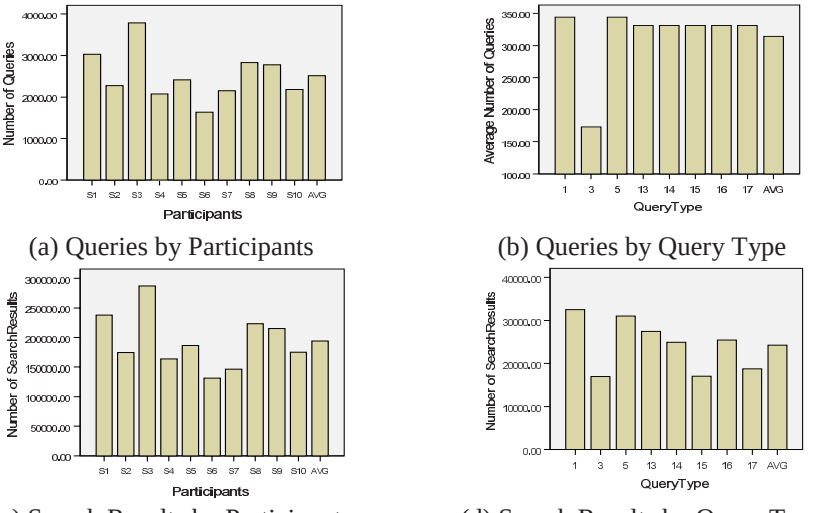
(a) Rule Creation (b) Category Usage (c) Number of Condition

Fig. 3. Classification knowledge bases created by each participant are significantly different

5 Experimental Results

5.1 Query Generation and Search Results

Among 17 types of search query generation methods, the rule-condition-based search query (Type 1), the category-based search query (Type 3), the combined search query (Type 5), the maximum-term-frequency-based case search query (Type 13 ~ 15), and the second-maximum-term-frequency-based case search query (Type 16 ~ 17) were only used for evaluation, because of large volumes of search queries.



(a) Queries by Participants (b) Queries by Query Type

(c) Search Results by Participants (d) Search Results by Query Type

Fig. 4. Queries and search results are significantly varies among participants

The number of each participant's queries significantly varies among participants and it were affected by the number of distinct leaf rule paths, the number of used category paths, and the number of cornerstone cases respectively (Fig. 4 (a)). The average numbers of rule-condition-based queries and the combined queries (344) and the case-based

queries (331) by query types were similar, but the number of category-based queries (173) was significantly less than other types of search query (Fig. 4.(b)). In total, 1,941,352 search results were obtained from the MSN Live search engine. Average number of search results by participants is 194,135 web pages and it varies from 131,165 (S6) to 286,927 (S3) (Fig. 4 (c)). Each participant's average number of search results by query type is 24,267 and it varies from 16,950(Type 3) to 32,531 (Type 1) (Fig. 4 (d)).

5.2 Evaluation Results

Discriminative power by query type is illustrated in Table 5, where search results overlaps within each search type for each participant were considered in order to calculate the discriminative power. The average discriminative power by search query type ranges from 0.79 (Type 3 and Type 5) to 0.97 (Type 13) and the average discriminative power by participant ranges from 0.88 (S 5) to 0.92 (S 2). Individual differences in discriminative power between queries of the same query type are most significant in the rule-condition-based query (Type 1) and the combined query (Type 5). In contrast, the category-based query (Type 3) and the case-based query (Type 13 ~ 17) display low differences in the discriminative power. If individual differences are large, it means that the characteristics of the search query are significantly different among participants. That is, the query characteristics of the rule-condition-based query (Type 1) and the combined query (Type 5) are quite diverse among participants. In addition, it is noticeable that the rule-condition-based query (Type 1) presents similar discriminative power to the case-based query (Type 13 ~ Type 17), even though the former contains a lower query word count and length compared to that of the latter. This means that similar distinctive search results can be obtained by using both the rule-condition-based query and the case-based query.

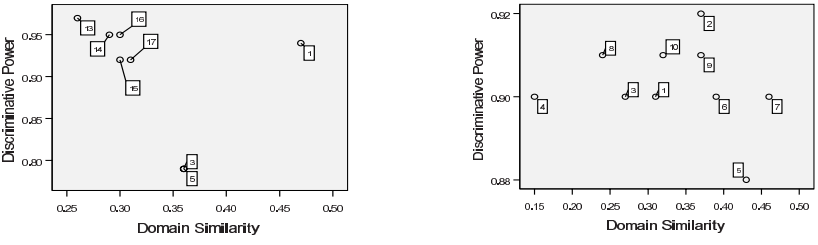
Table 5. Discriminative Power Results

	Type 1	Type 3	Type 5	Type 13	Type 14	Type 15	Type 16	Type 17	AVG	ST DEV
S1	0.97	0.78	0.81	0.97	0.95	0.91	0.94	0.91	0.90	0.07
S2	0.93	0.80	0.85	0.98	0.97	0.95	0.97	0.94	0.92	0.06
S3	0.97	0.78	0.81	0.96	0.94	0.90	0.93	0.89	0.90	0.07
S4	0.94	0.80	0.81	0.97	0.94	0.92	0.93	0.91	0.90	0.06
S5	0.86	0.80	0.70	0.97	0.94	0.92	0.95	0.91	0.88	0.09
S6	0.89	0.80	0.75	0.97	0.96	0.93	0.96	0.93	0.90	0.08
S7	0.96	0.80	0.75	0.97	0.95	0.91	0.94	0.92	0.90	0.08
S8	0.98	0.77	0.82	0.97	0.96	0.92	0.95	0.92	0.91	0.07
S9	0.97	0.80	0.80	0.97	0.96	0.92	0.95	0.92	0.91	0.07
S10	0.95	0.79	0.82	0.97	0.97	0.93	0.97	0.93	0.91	0.07
AVG	0.94	0.79	0.79	0.97	0.95	0.92	0.95	0.92	0.90	0.07
ST DEV	0.037	0.010	0.044	0.005	0.010	0.014	0.013	0.013	0.011	

Domain similarity results are summarized in Table 6. On average, the rule-condition-based query (Type 1) exhibits the highest domain similarity (0.47), followed by the category-based query (Type 3, 0.36), the combined query (Type 5,

Table 6. Domain Similarity Results

	Type 1	Type 3	Type 5	Type 13	Type 14	Type 15	Type 16	Type 17	AVG	ST DEV
S1	0.48	0.32	0.36	0.23	0.27	0.26	0.28	0.29	0.31	0.08
S2	0.51	0.43	0.41	0.28	0.31	0.34	0.32	0.34	0.37	0.08
S3	0.36	0.26	0.28	0.19	0.25	0.27	0.26	0.28	0.27	0.05
S4	0.48	0.20	0.10	0.08	0.08	0.08	0.07	0.08	0.15	0.14
S5	0.46	0.47	0.47	0.41	0.42	0.41	0.42	0.42	0.43	0.03
S6	0.55	0.36	0.42	0.32	0.35	0.35	0.40	0.38	0.39	0.07
S7	0.61	0.50	0.53	0.38	0.41	0.40	0.41	0.41	0.46	0.08
S8	0.31	0.28	0.28	0.19	0.21	0.23	0.21	0.24	0.24	0.04
S9	0.54	0.37	0.42	0.28	0.32	0.34	0.32	0.34	0.37	0.08
S10	0.42	0.38	0.36	0.25	0.27	0.30	0.28	0.31	0.32	0.06
AVG	0.47	0.36	0.36	0.26	0.29	0.30	0.30	0.31	0.33	0.07
ST DEV	0.09	0.10	0.12	0.10	0.10	0.10	0.11	0.10	0.10	



(a) Query Type (b) Participants

Fig. 5. Best query generation method can be selected by combining domain similarity and discriminative power

0.36), and the case-based-query (Type 13 ~ 17). The levels of domain similarity in the cased-based query (Type 13 ~ 17) slightly increase as the threshold increases, that is, as the number of query terms increases. This means domain similarity is not significantly affected by the threshold degrees of the case-based query generation methods. Domain similarity varies significantly between participants within the same query types. For example, the lowest domain similarity for the rule condition based query is 0.31 (S8) and the highest 0.61 (S7). Unlike discriminative power, domain similarity is significantly different among participants, ranging from 0.15 (S 4) to 0.46 (S 7).

5.3 Query Generation Method Selection

The best query can be chosen by considering domain similarity and discriminative power as illustrated in Fig. 5. Fig. 5 (a) illustrates discriminative power and domain similarity of each search query type as data points. The average domain similarity and discriminative power of query types are 0.33 and 0.90 respectively. The rule-condition-based query (Type 1) displayed the best query generation because both discriminative power and domain similarity are higher than their average. The

category-based query (Type 3) and the combined query (Type 5) were slightly better than average, but their discriminative powers were far less than average. On the contrary, the case-based queries (Type 13 ~ Type 17) were better discriminative power than average, but their domain similarities were lower than average. Fig. 5 (b) illustrates each participant's discriminative power and domain similarity. The participants' average domain similarity and discriminative power are 0.33 and 0.90 respectively. Participants S2, S6, S7, and S9 have more efficient search queries, because they exhibited discriminative power and domain similarity above average. However, participants S5 exhibited its discriminative power lower than average and participants S1, S4, S8, and S10 exhibited their domain similarity lower than average.

6 Conclusions

This paper proposed a new personalized search query generation methods reusing MCRDR document classification knowledge. A total 17 types of search query generation methods were proposed using the rule tree, category tree, and cornerstone cases. They were evaluated according to discriminative power and domain similarity. The experiment results show that the rule-condition-based query (Type 1) and the various case-based queries (Type 13 ~ 17) presented higher discriminative power compared to the category based query (Type 3) and the combined query (Type 5). Each participant's discriminative power was similar even though the number of his/her search results was significantly different. The rule-condition-based query (Type 1) exhibited the highest domain similarity. Unlike the discriminative power, the domain similarity differed significantly among participants. Finally, when discriminative power and domain similarity were considered simultaneously, the rule- condition-based query generation method was shown to be the most effective query method.

Acknowledgments. This work was supported by the Asian Office of Aerospace Research and Development (AOARD)

References

[1] Park, S.S., Kim, S.K., Kang, B.H.: Web Information Management System: Personalization and Generalization. In: The IADIS International Conference WWW/Internet 2003, Algarve, Portugal (2003)

[2] Kang, B., Compton, P., Preston, P.: Multiple Classification Ripple Down Rules: Evaluation and Possibilities. In: 9th AAAI-Sponsored Banff Knowledge Acquisition for Knowledge-Based Systems Workshop, Banff, Canada. University of Calgary, Menlo Park (1995)

[3] Park, S.S., Kim, Y.S., Kang, B.H.: Web Document Classification: Managing Context Change. In: IADIS International Conference WWW/Internet 2004, Madrid, Spain (2004)

A Study on Monitoring Web Page Locating Heuristics

Yang Sok Kim and Byeong Ho Kang

#School of Computing and Information Systems, University of Tasmania
Private Bag 100 Hobart TAS 7001, Australia
{yangsokk, bhkang}@utas.edu.au

Abstract— **Monitoring web page locating is a key process in the web monitoring systems or in the automated client pull systems. Previous systems usually register these web pages manually. For an automated monitoring web page registration system, it is essential to know human heuristics. In this research, we summarised human heuristics based on 12 participants' reports on monitoring web page locating strategies.**

I. Introduction

Various Web information finding technologies (WIFT) have been suggested since the beginning of the Web. Nowadays web crawling is regarded as the most convincible one. However, it is not perfect solutions for all information needs. Therefore, several complementary WIFTs of web crawling have been proposed by researchers. Web monitoring is one of them and its goal is to provide more complete coverage and timeliness than web crawling in a specific domain. The web monitoring system requires monitoring web pages to obtain new web pages from them. These web pages may be maintained manually, but it is useful to manage them dynamically according to the users' preference. This research focuses on finding heuristics used for locating monitoring web pages relevant to the users' interests. These heuristics include the following sub processes:

(1) **Find monitoring web pages relevant to the users' interest**: Web search services can be used to find relevant web pages by submitting appropriate search queries. Document classification knowledge reflecting user's interests may be used to formulate search queries. In the previous research, various query generation methods derived from the MCRDR (Multiple Classification Ripple–Down Rules) classification knowledge were examined to decide which methods were appropriate for finding relevant web pages using web search engines. The research results showed that the 'rule condition based query' is the best method from the discriminative power and the domain similarity criteria.

(2) **Find candidate monitoring web pages**: Once relevant search results were obtained by using search quires, it is required to find candidate web monitoring pages from them. In this research, search results were classified into three types – monitoring web page, document web page, and ad hoc web page. These web page types are defined as follows:

- Monitoring web pages: Monitoring web pages contain links to document web pages and the main purpose of these pages is to provide list of linked web pages. They are usually main pages of a web site (e.g., index.html, index.php, index.jsp), but sometimes they may be non-main pages. ;
- Document web pages: Document pages are linked from a monitoring web page and contain contents to be read by the users and their main purpose is to provide information. ; and
- Ad-hoc web pages: Some web pages may not be clearly classified into the monitoring web pages or the document pages and these web pages were classified into "ad hoc pages" in this research.

If a search result is a monitoring web page, this page becomes a candidate monitoring web page without any further process. However, if a search result is either a document web page or an ad-hoc web page, it is required to find its parent (monitoring) web.

(3) **Recommend monitoring web pages**: As all candidate monitoring web pages are not appropriate for monitoring, it is required to decide whether they are to be monitored or not.

This paper focuses on finding human heuristics for these three sub processes. This paper contains the following contents: Section 2 provides study background and related study. Section 3 explains our experiment methodology and Section 4 summarizes our experiment results. Conclusions and further study directions will be given in Section 5.

II. Background and Related Study

A. Web Page Type Classification

Web page type classification differs from web document classification. Whereas the latter aims to place documents into the specific topics, the former tries to classify web pages into one of specific page types. Web document classification usually based on similarity between documents' contents or their hyperlink structures. There are lots of previous research as summarized in [1, 2, 3].

Web page type classification gained little attention from the researchers compared to web document classification. Matsuda and Fukushima [4] identified web page type classification problem, and tried to solve it by human description of structural characteristics of a type. Glover et al. [5] manually classified a large amount of documents to collect negative training data into "personal homepage" and "call-for-paper" using SVM (Support Vector Machines) with fixed set

of features. Elsas and Efron [6] proposed a HTML tag based metrics for web page type classification. They used table tag ratio, anchor text ratio, and text per table data tag to develop thresholds that classify web pages into (data) table, index/table of content, and content pages.

Web page type classification usually depends on the particular tasks. For example, Glover et al. stated as follows in [5] "A personal homepage is a difficult concept to define objectively. The definition we used is a page made by an individual (or family) in an individual role, with the intent of being the entry point for information about the person (or family)." (p.6). Therefore, web page classification method should take into account specific task of this research. As this study focuses on finding candidate web monitoring pages, web page type classification needs to be conducted according to this purpose and the classification approach should be developed in relation to this task.

B. Finding Monitoring Web Pages

Each page type's ratio within search results varies according to search query. For example, two sets of search query results collected from Google with two search queries – "search engine" and "search engine coverage" – contain different search results. Whereas search results of "search engine" consist of main pages except one result from Wikipedia, search results of "search engine coverage" do not contain any main page and they consist of three types of web pages.

If a search result is not a web monitoring page, it is required to find its parent page that links the search result page. One might think the backlinks of a current page as a parent page, because as summarized in Fig. 1, commercial search engines provide search options for finding backlinks of a specific web page.

- Google Web Search: Search for «link:website address», example: link:http://www.daff.gov.au/
- Yahoo Site Explorer: Search for your website address, example: link:http://www.daff.gov.au/
- Windows Live Search: Search for «+link:website address», example: + link:http://www.daff.gov.au/

Fig. 1 Commercial search engines provides search options for backlink retrieving.

However, there are some limitations in using backlinks as a method for finding parent pages. Firstly, the parent page does not mean backlink pages. The latter may contain the parent page, but they also include lots of other web pages as backlink pages. Secondly, search engines do not provide all backlinks. For example, as illustrated in Fig 2. (a), MSN search provides www2005.org/cdrom/docs/p1190.pdf as search results when the "Focused Crawling" search word was submitted, but MSN search does not provide backlinks for this URL (see Fig 2. (b)). In fact, the provision of backlinks totally depends on search engine companies' decision [7]. For this reason, this research study the heuristics of finding the parent page from the given search result page.

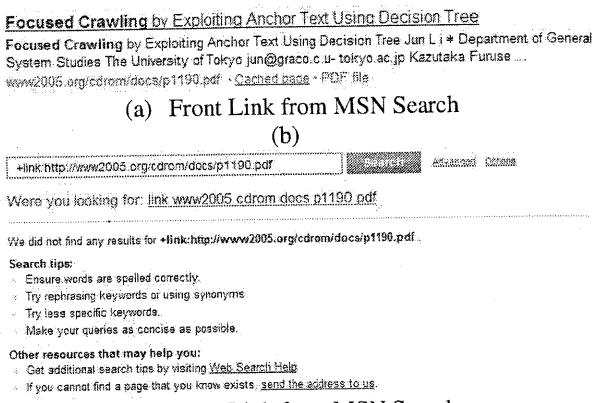


Fig. 2 Commercial search engines provides search options for backlink retrieving.

C. Relevant Web Page Recommendation

In the commercial area, relevant web page recommendation was introduced when Netscape released a 'What's Related?' feature in version 4.06 of the Netscape Communicator browser. Details about the approach used to identify related pages in their algorithm are obscure. However, according to the What's Related FAQ page indicated that the algorithm uses connectivity information, usage information, and content analysis of the pages to determine relationships [8]. Nowadays similar service is provided by Google named as 'Similar Pages'. Google explains its service as follows:

When you click on the "Similar Pages" link for a search result, Google automatically scouts the web for pages that are related to this result. ...If you are interested in researching a particular field, Similar Pages can help you find a large number of resources very quickly, without having to worry about selecting the right keywords. (http://www.google.com/help/features.html#related)

In the research area, there have been significant researches on finding relevant web pages based on hyperlink analysis. Kleinberg [9] applied the HITS algorithm to find relevant web pages. In his theory, each page is assigned "hub" value and an "authority" value. The authority and hub weights are updated based on the following equations:

$$\text{Authority (p)} = \sum_{(q \to p)} Hub(q) \quad (1)$$

$$\text{Hub (p)} = \sum_{(q \leftarrow p)} Authority(q) \quad (2)$$

Equation (1) implies that if a page is pointed by many good hubs, its authority weight should increase (i.e., it is the sum of current hub weights of all of the pages pointing to it).

Equation (2) implies that if a page is pointing to many good authorities, its hub weight should increase (i.e., it is the sum of the current authority weights of all of the pages it pints to). Pages with high authority scores are expected to have relevant content while with high hub scores are expected to contain links to relevant content. Kleinberg suggested that the HITS algorithm could be used for finding related pages as well, and provided anecdotal evidence that it might work well.

Dean and Henzinger [8] extended the HITS algorithm to exploit not only links but also their order on a page. They construct page source in different way for their relevant page finding algorithm, Companion. The Companion algorithm takes as input a starting URL u and consists of four steps:

- Step1: Build a vicinity graph for u;
- Step 2: Contract duplicates and near-duplicates in this graph;
- Step 3: Compute edge weights based on host to host connections;
- Step 4: Compute a hub score and an authority score for each node in the graph and return the top ranked authority nodes.

Hou and Zhang[10] proposed a more direct algorithm named LLI (Latent Linkage Information) for finding relevant pages. Firstly, it builds one reference page set (Pu) by selecting a group of its parent (inlink) pages and the other reference page set (Cu) by a group of its child (outlink) pages. Then, it builds one candidate page set (BPS) by adding a group of child pages from each of the Pu pages. Similarly, it builds the other candidate set (FPS) by adding a group of parent pages from each of the Cu pages. Both candidate sets BPS and FPS are presumably rich in relevant pages, and their pages are ranked and returned by the algorithm. The neighborhood page construction is finalized by merging some of the reference pages and their outlink sets. For page ranking in LLI, two matrices (Pu-BPS, Cu-FPS) are constructed to represent the hyperlink relations among the pages. In both Pu-BPS and Cu-FPS matrices, each column represents a reference page and each row represents a candidate page. The binary matrix entries indicate if there are page links between the corresponding reference and candidate pages. LLI then applies the singular value decomposition (SVD) on the matrices to reveal deeper relationships among the pages in the rank-reduced SVD spaces.

However, these link analysis based approaches may not be applied for this research. Firstly, the link analysis based approaches only focus on finding authority pages as relevant pages from their own page graphs, rather than on directly finding relevant pages from page similarities. Therefore, if the page graph is not constructed properly, it is hard to find relevant pages [11]. For this reason, some researchers tried to integrate content and link structure to find relevant web pages [12]. Secondly, the context that the link analysis based approaches applied significantly differs from this research. Whereas above research focuses on the "relevant pages", this research focuses on the web pages that produce "relevant pages". Lastly, the purpose of finding relevant pages is significantly different between whereas the link analysis based approaches and this research. Whereas the link analysis based approach focused on the generalized relevant pages, this research focuses on the personalized relevant pages reflecting the user's interests.

III. Research Methodology

A. Data Set Collection

This research was designed to elicit human heuristics that may be used for locating monitoring web pages using search engines. Data collection process is summarised as follows: A total of ten Master and Honours students at the University of Tasmania took part for experiment data collection, our web monitoring system, WebMon, collected newly uploaded web documents from Australian Government homepages and media release pages and forwarded them to the participants. Each participant created rules to classify unclassified documents or to reclassify misclassified documents using MCRDR classifier. Hierarchical category tree, which has three levels and 769 categories, was given to the participants. Classification of the monitored documents was undertaken for eight weeks from 21 August, 2006, to 9 October, 2006. Detailed classification results were reported [13].

A total 3,439 of the rule condition based queries were generated by using the above MCRDR classification knowledge base and they were submitted to MSN search API to obtain top 100 search results. A total 249,230 distinct web pages were collected. A total 200 of the search results were randomly sampled from these search results for this experiment.

B. Participants and Procedures

The following monitoring scenario was given to the participants for this experiment.

We wish to operate an Australian Government Information Portal, which provides Web information from both government agencies and non-government organisations. Therefore, you need to decide the monitoring Webpages based on their contents, not only their domain name type such as "gov.au" or "org". You need to consider classification procedure, because the collected documents from the registered Webpages will be classified by users. The classification structure is same to that of the current MCRDR classifier.

Total 40 Masters and Honours students at the University of Tasmania participated in this experiment. For this experiment, participants were divided into 12 groups and each group was required to write a report on strategies used for locating web monitoring pages. The topics include (1) the web page classification strategies, (2) candidate web monitoring page finding strategies, and (3) web monitoring recommendation strategies. Each topic should be written in one page in A4 size with 12 font size. Their reports were analysed to extract human heuristic for locating monitoring web pages.

IV. Results

A. Web Page Classification

The web page classification strategies that were proposed by participants are summarized in TABLE I, where the strategies are ordered by total frequency. Each participant suggested one to six strategies. 'Links vs. content ratio' and 'file name' were the most frequently stated strategy.

TABLE I
Web Page Classification Strategy

Classification Strategy	Total
Links vs. content ratio	9
File name	9
Web page layout & design	5
Specific links	4
Special words in the URL	4
URL depth	3
Link characteristic	3
Tags (<p>, tags)	3
Meta tag	2
Link position	2
Use graph structure	1
Total	45

Participants reported the following criteria as web page classification strategies:

- **Links vs. content ratio**: Usually the number of links is larger in the monitoring pages than the document pages. One could analyse the text to link ratio in a web page, using some statistical method to find a point at which a certain ratio means a web page is a monitoring page and another is a document page. This should be done on a website-to-website basis, as this ratio would be context specific. ;
- **File name**: The types of web pages can be identified by examining the postfixes of the URLs. Whereas most URL addresses of monitoring pages have an extension such as 'index.asp', 'index.html', 'index.php' etc, if the postfixes of their URLs are '.doc' or '.pdf', it commonly can be considered as a document web page. Sometimes, the web pages that have an extension such as 'index.html', 'index.php', and 'index,jsp' could be a home page or sub-home page. ;
- **Web page layout and design**: Some participants reported the layout and design of the web pages among the monitoring pages, the document pages, and ad hoc pages. For example, a participant writes, "The length of candiate monitoring web pages normally is less then two pages with 1.5 line spaces.; ii) There are various type font, style, size and colour, for example, underline and bold are often used in the monitoring web pages. In contrast, document web pages comprise full sentences, and each line exceeds 15 characters with 1 line space. There are consistent type font, style, size and colour. Most monitoring web pages are divided into multiple blocks. While document web pages usually are divided into two or three columns including left, centre and right. Text paragraph are generally aligned in the middle areas, navigation and links are aligned on left and right areas" and another participant writes, "Images are small, like icons or photos to accent the pages, rather than be part of the content themselves."
- **Specific Links**: Participants reported that the specific links can be used for web page classification.
- **Special words in the URL**: Some participants reported that if there are specific words in the URL, the web pages may be classified into one of three web page categories.
- **Others**: Some participants indicate URL depth, specific tags (e.g., <p>,
, and <meta>), link characteristics, and link positions as web classification criteria.

B. Candidate Monitoring Web Page Finding

If a document web page is relevant to the user's interests, it is required to locate its parent page. In general, the candidate monitoring webpage is the parent page of the document page, such as the index page or home page. Candidate monitoring Webpage could be different from the home page or index page. The candidate monitoring web page finding strategies are summarized in TABLE II, where the strategies are ordered by total frequency. Each participant suggested one to five strategies. Sub-URL, Navigation links in the current page, and Breadcrumbs are most frequently stated in the participants' report.

TABLE II
Candidate Monitoring Web Page Finding Strategy

Monitoring Page Locating Strategies	Total
Sub URLs	12
Navigation in the current page	8
Breadcrumbs	8
Specific links in the current page	4
Navigation in the homepage	3
Sitemap	1
Search	1
RSS link	1
All links in the current page	1
Total	39

The participants reported the following strategies for finding the candidate monitoring web page:

- **Sub-URLs**: The most efficient and well known approach is to get the sub-URL of candidate monitoring web pages from the given URL. For example, a document webpage URL is "http://www.qrd.org/qrd/usa/legal/lgln/1995/06.95", the candidate monitoring web page "http://www.qrd.org/qrd/usa/legal/lgln/1995" can be get by deleting last part of the URL, "/06.95". However, not all candidates monitoring web page can be gained by Sub-URL strategy. For example, a web page

http://www.operationit.com/OperationIT/Articles/B2B3rd Wave.html can't find its candidate monitoring web page by Sub-URL. It may depend on web administrator making website address regulation.

- **Breadcrumbs**: Using breadcrumb is another effective way. Breadcrumb clearly lists the routine of the current web page and provides links to every web page in the routine. Generally speaking, the last second hierarchy is the current document web page's candidate monitoring web page. For example, a document page shows a breadcrumb "Home>Board of Trustees>By-laws", which tells user the current document page is "By-laws", "Board of Trustees" item provide the hyperlink to the document page's candidate monitoring page. However, some bread crumbs do not provide hyperlink functions, which mean the breadcrumb is meaningless for finding candidate monitoring webpage at this situation. Mostly, the breadcrumb can show the relationship between links, and viewers can find the candidate monitoring webpage via breadcrumb. Unlike navigation, breadcrumb is always dynamic. Its purpose is to support a way to keep track of their location within programs or documents.
- **Links in the current page**: Navigation links always provides the useful or important links to the important sections of the website. Navigation links always lists the useful or important index page links. When sub-URL can not link to candidate monitoring webpage, the sub-URL is also useful. The document page URL http://www.mothersmovement.org/features/ddavis_interview.htm can not find the candidate monitoring webpage via "Sub-URL". However, its sub-URL shows a key word "features". The key word "features" can be found in the navigation to link a main index page. Actually, this main index page is also the candidate monitoring webpage what we want to find.
- **Navigation links in the home page**: Navigation bar is not only a good tool to guide visitor to read information, but also we can find the parent page or candidate monitoring pages from the links which provided by homepage's navigation bar. Normally, this kind of menu usually appears on the top of the webpage or the left column, which provides many links organized by tree structure. This method is working efficiently on the webpage, such as http://www.npgcable.net/cable_tv_serv.html.
- **Specific links**: Many document pages support links of other documents which is from the same candidate monitoring webpage. They also provide a "more" link at the bottom to approach the candidate monitoring webpage. Another specific link is "previous" link in document page can come back to candidate monitoring pages. While some "previous" link just go to other document pages.
- **Search**: Users can use the search box of the website to find the candidate monitoring webpage. According to the content or title of the document, they can input the key words and get the search result pages to find the candidate monitoring webpage.
- **Sitemap**: Relevant candidate monitoring web pages can be found from site map by locating its document webpage. However, not all website provide site map, or site maps just provide links to a majority of web pages, so it is possible that some document web page can not be found in the site map.
- **RSS links**: RSS link can be used as a candidate monitoring web page. For example, web page of http://news.bbc.co.uk/2/hi/entertainment/7031366.stm contains a link referring to a RSS feed, which would provide a very efficient way to monitor the website.

C. Monitoring Recommendation

It is required to select monitoring web pages from the candidate monitoring web pages, because not all the candidate monitoring page has value for web monitoring. The monitoring web page recommendation criteria are summarized in TABLE III, where the criteria are ordered by total frequency.

TABLE III
Monitoring Recommendation Criteria

Monitoring Recommendation Criteria	Total
Update regularity	9
Relevant to the user's interest	8
Number of links on the page	7
Credibility and popularity	5
Accessibility and usability	4
Total	27

Each participant suggested one to five strategies. Sub-URL, Navigation links in the current page, and Breadcrumbs are most frequently stated in the participants' report. The participants reported the followings as web monitoring page recommendation strategies:

- **Relevancy to the topic of interests**: Many participants indicated that if the candidate monitoring pages or their child pages are relevant to the topic of interests, they should be monitored. The first and top most of all rules is to look at the query words which are used to generate this result. Relevancy can be checked by using in the meta tag, <meta name = "keywords">(e.g., <meta name = "keywords" content = "pharmacy, pharmacies, pharmacy practice">). If the content of this tag always update, that candidate monitoring web pages should be draw attentions by other users. Another approach may be to classify the page using MCRDR classifier.
- **Update regularity**: Many participants indicate that update frequency is an important factor in deciding the value of parent page. The value of candidate monitoring web page for web monitoring depends on whether the candidate monitoring web page attracts users to access. Generally, the newest information is more attractive to user and valuable to be monitored. Therefore, the candidate monitoring web pages should be updated regularly that web monitoring system can gain real time

information using monitoring web pages list. There are a number of ways for checking the frequency a page is being updated. The first of these is to simply check the 'last updated' date often found on a webpage. Other methods to check the rate of which a page is updated include checking various attributes of the page itself. The page can be crawled, checking for updates to attributes such as text, links, and images. Depending on the type of web page a person is looking for, a combination of these could be used in determining whether or not a page is worth monitoring. To effectively use this as a metric for determining whether or not to monitor a page, the page would have to be continuously monitored using a sliding window to adapt to trends within the website.

- **Number of links**: The number of links found on the page can be used to determine a page's usefulness. If there are a number of outgoing links found, this can indicate that the website has a wealth of information that could be monitored and retrieved. To ascertain the number of links of a page, the HTML code can be examined, locating each instance of link tags (<a href>). This however has inherent problems in that all links on a page may not lead to relevant documents (for example a link could simply be an advertisement on the page). Another problem associated with using links as a metric is that the links could simply be links to areas on the same page, though this could be detected by the algorithm. Therefore while using links as a metric does have its inherent problems, a page that has a high ratio of links to content is more likely to be a good page to be monitored.
- **Accessibility and usability**: If the parent page of a given document page expired by some reasons, it is definitely no worth to monitor. No one would like to waste time on the useless web pages. If the parent pages require privileges to access, they are also not worth monitoring, such as the forums. One more type of parent pages not worth monitoring is that the given document page's parent page is garbage page, like the advertisement web page. The document page could be a description of an image advertisement.
- **Similarity between monitoring page and linked page**: One other way in which the parent page's usefulness can be determined is by checking the similarity in keywords between it and documents linked from it. For example if a key word found in the originally located document is matched with words found on the parent page, it could be assumed that the website will be useful for that subject area. One downfall to this approach is that if the main page supplies a summary of linked documents, it will always match key words, therefore it would be advantageous to match blocks of text, and if they are identical do not consider these, rather only consider similar blocks of text.
- **Publication credibility and web page popularity**: Official publication shows more authoritative and reliable information, while individual publication may be unreliable. Thus, when candidate monitoring web pages provide formal publication, it is better to be monitored. It may possible to set up an online register web page, and ask government to cooperate to make a policy, that force every government website need to register on the register webpage.

V. Conclusions and Further Study

This paper researched human heuristics for locating monitoring web pages. The main finding as follows:

- Users most frequently use link-content ratio and file name to classify web pages into monitoring web page, document page, and ad-hoc page
- The 'sub-URL' strategy is used most frequently to find candidate monitoring web pages and followed by navigation in the current page and navigation menu in the home page. ; and
- Update regularity, relevancy to the user's interest, and the number of links on the page were suggested as the most significant strategies.

This paper focuses on identifying human heuristics for locating monitoring web page locating approach. However, we need to conduct the following study. Firstly, we need to conduct user study how really user classify web pages, because our current study provide only strategies. Secondly, we need to check how classification knowledge can be used to recommend web pages for monitoring. Lastly, an automated system will be implemented on the basis of this and the following study.

Acknowledgement

This paper was supported by AOARD (Asian Office of Aerospace Research and Development).

References

[1] Sebastiani, F., *Machine learning in automated text categorization.* ACM Computing Surveys, 2002. **34**(1): p. 1-47.

[2] Sebastiani, F., *Text categorization*, in *The Encyclopedia of Database Technologies and Applications*, L.C. Rivero, J.H. Doorn, and V.E. Ferraggine, Editors. 2005, Idea Group Publishing: Hershey, US.

[3] Sebastiani, F., *Text categorization*, in *Text Mining and its Applications*, A. Zanasi, Editor. 2004, WIT Press, Southampton, UK. p. pp. 109--129.

[4] Matsuda, K. and T. Fukushima. *Task-oriented world wide web retrieval by document type classification.* in *the eighth international conference on Information and knowledge management.* 1999. Kansas City, Missouri, United States: ACM New York, NY, USA.

[5] Glover, E.J., G.W. Flake, S. Lawrence, W.P. Birmingham, A. Kruger, C.L. Giles, and D.M. Pennock. *Improving Category Specific Web Search by Learning Query Modifications.* in *Symposium on Applications and the Internet, SAINT 2001.* 2001. San Diego, California: IEEE Computer Society.

[6] Elsas, J. and M. Efron. *HTML tag based metrics for use in web page type classification.* in *American Society for Information Science and Technology Annual Meeting.* 2004. Providence, Rhode Island, USA.

[7] Wilson, R.F., *Google's Index Shows Only a Few Backlinks.* 2006.

[8] Dean, J. and M.R. Henzinger. *Finding related pages in the World Wide Web.* in *Eighth International Conference on World Wide Web.* 1999. Toronto, Canada: Elsevier North-Holland, Inc. New York, NY, USA.

[9] Kleinberg, J.M., *Authoritative sources in a hyperlinked environment.* Journal of the ACM, 1999. **46**(5): p. 604-632.

[10] Hou, J. and Y. Zhang, *Effectively Finding Relevant Web Pages from Linkage Information.* IEEE Transactions on Knowledge and Data Engineering, 2003. **15**(4): p. 940-951.
[11] Yan, H. and Richards. *Enhanced searching algorithms for relevant web pages using hyperlink graphs.* in *43rd ACM Southeast Regional Conference*. 2005. Kennesaw, Georgia: ACM New York, NY, USA.
[12] Davis, D. and E. Jiang. *Exploring Content and Linkage Structures for Searching Relevant Web Pages.* in *Third International Conference on Advanced Data Mining and Applications(ADMA 2007)*. 2007. Harbin, China: Springer Berlin / Heidelberg.
[13] Kang, B.H., Y.S. Kim, and Y.J. Choi. *Does Multi-user Document Classification Really Help Knowledge Management?* in *Twentieth Australian Joint Conference on Artificial Intelligence AI 2007)*. 2007. Queensland, Australia: Springer-Verlag Berlin Heidelberg.