

## **Final Report**

**Title: Information monitoring system on World Wide Web**

**Contract Number: FA2386-09-1-4119**

**AFOSR/AOARD Reference Number: AOARD 094119**

**AFOSR/AOARD Program Manager: Hiroshi Motoda**

**Period of Performance: 10 June 2008 – 10 Dec 2009**

**Submission Date: 10 02 2010**

**PI:** Dr. Byeong Ho Kang/University of Tasmania

## Report Documentation Page

*Form Approved*  
*OMB No. 0704-0188*

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE <b>19 FEB 2010</b>	2. REPORT TYPE <b>Final</b>	3. DATES COVERED <b>10-06-2009 to 10-12-2009</b>			
4. TITLE AND SUBTITLE <b>Information monitoring system on World Wide Web</b>		5a. CONTRACT NUMBER <b>FA23860914119</b>			
		5b. GRANT NUMBER			
		5c. PROGRAM ELEMENT NUMBER			
6. AUTHOR(S) <b>Byeong Ho Kang</b>		5d. PROJECT NUMBER			
		5e. TASK NUMBER			
		5f. WORK UNIT NUMBER			
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) <b>University of Tasmania,GPO Box 252-100,Hobart TAS ,Australia,AU,7005</b>		8. PERFORMING ORGANIZATION REPORT NUMBER <b>N/A</b>			
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) <b>Asian Office of Aerospace Research &amp; Development, (AOARD), Unit 45002, APO, AP, 96338-5002</b>		10. SPONSOR/MONITOR'S ACRONYM(S) <b>AOARD</b>			
		11. SPONSOR/MONITOR'S REPORT NUMBER(S) <b>AOARD-094119</b>			
12. DISTRIBUTION/AVAILABILITY STATEMENT <b>Approved for public release; distribution unlimited</b>					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT <b>The overall objective of this effort is to explore the new usage of WebMon to study social behavior analysis and identify what need be added and the technical challenges to achieve the goal. The following two are the major tasks that need be undertaken: 1) enhance the monitoring capability of WebMon that is tailored to collect data useful for social network analysis, in particular public opinion formation, and 2) analyze the data collected and find important factors that affect public opinion formation. The enhancement consists of two components: smart multi-context filtering and smart scheduling. The former enables to handle various Web documents structures and collect information needed for detailed analysis of opinion formation, e.g. not only the content of the article but also the comment and the opinion about it. The latter enables to capture frequent update of the comments and the opinions. The heart of intelligent data collection is the use of incremental knowledge acquisition method MCRDR which can acquire and maintain rules to identify the different data structure.</b>					
15. SUBJECT TERMS <b>Internet, Meta Search Engine, World Wide Web, Knowledge Acquisition</b>					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT <b>Same as Report (SAR)</b>	18. NUMBER OF PAGES <b>13</b>	19a. NAME OF RESPONSIBLE PERSON
a. REPORT <b>unclassified</b>	b. ABSTRACT <b>unclassified</b>	c. THIS PAGE <b>unclassified</b>			

## **1 Objectives**

The WebMon, Web information monitoring system, was created to conduct web monitoring services for a few selected domains, such as Australian Government Web pages and health news web pages [7,8,9,11,13]. We have also conducted web publication analysis and search engine performance analysis using our web monitoring system. Information overload was one of our main problems in web monitoring systems in earlier studies, thus to overcome this problem, we developed a web document classification system. The MCRDR (Multiple Classification Ripple-Down Rules) knowledge acquisition methodology was employed to facilitate incremental knowledge encoding for classification knowledge. This project studied new technical challenges to resolve in order to extend the current Web monitoring system and make it applicable to collect data needed to investigate the influences of Internet based social networks to social issues and opinions.

## **2 Status of effort**

### **1) Background studies**

A recent field study found that 95% of people under 30 and 28% of people aged above 50 use the Internet in South Korea. 97% of Internet users in South Korea have purchased goods using the Internet. The same study showed that some other countries such as Australia show similar figures. However, the most significant figures are about social networking in South Korea. 58% of people have their own blog and 85% joined a certain form of social network. The average number of friends per person connected through this social network is about 75. The following figures show how many hours people in Korea are spending for various online activities during a week: Online news (8), Online chatting (11), Online game (7), online movie / music (9) Social network management (6), Blog management and visiting others (8), Internet Search (7) [2,3,4]

From this figure, we can notice that the number of hours on the Internet is significantly high and they spend much time (35 hours) on online chatting and games as well as Blog management. The main factor of this new life style in South Korea is bandwidth, not connectivity. This pattern does not appear in several other countries known as Internet developed countries with lower bandwidth.

In South Korea, because of overwhelming social services and connections via the Internet, it is difficult to get services without using the Internet, making the Internet a new must-have resource for households. This changes how people express their opinions to others. 71% of people who express their opinions about products use the Internet. 65% of these people use social medias such as blogs, community groups and messenger. They also give more credit to people's opinions that are found on the Internet when they have to make online shopping decisions. While they give more credit to the Internet community, which is 85%, they only give 65% credibility to articles of newspapers or magazines. [2,3,4]

## **2) Problems**

Although I explained a significant impact in South Korea, one of the leading Internet nations, influences of this type are still enduring for societies because the changes have only been made during the last several years. The outcomes and impacts of these changes are still progressing. Of course, the total impact of full scale Internet connectivity has yet to reach everybody in most other countries. Only a few countries are about to reach the full potential of super-fast Internet technologies and advanced internet applications.

The problem is that all countries are now joining this race towards super-fast Internet connections. Some people like Alvin Toffler warned of the change of society based on the new media like the Internet. The current impact among Internet leading countries showed that we are not well prepared for it. Comparing impacts between early Internet era (1970-1990) and present, we can only expect a few local micro-level impacts in early stage while we are now facing with problems caused by unexpected macro-level impacts.

The problem is that Internet developing countries as well as developed countries do not learn from previous mistakes and so they do not try and prevent it. Within 10 years of all countries joining the race [1], it is obvious that we should deal with impacts in super macro level. In order to keep this change in a "right and healthy" course whilst reducing damage to human societies, it is important to have a proper understanding for impacts and have appropriate tools to measure potential impacts qualitatively.

### 3) Conclusions

Develop a new research tool that collects data to identify new information patterns out of social media but also to provide analysis functions. Without proper visual tools based on statistical analysis methods, it is difficult to handle large set of data from Web environments such as social Webs. Investigations of the basic model for social Web analysis research and development of the system to convert collected Web data to this model are necessary.

- **Challenges to accomplish the recommendations**

The contents in social networks are different from traditional Web pages and there are two different contents generated by two different types of social network users, issue generators and conversation makers. Issue generators are people who bring the information about issues to their social networks. In fact, most journalists in newspapers or people who put up contents on Web sites are considered as issue generators. In addition, there are groups of people who comment on or express their opinion on certain issues. These are the conversation makers. Some issues raised by issue generators become public issues when the public shows their interests and express their opinions (public opinion) on these. The impact study of social networks finds specific patterns that make public issues and the rolling patterns of public opinions on these issues.

The problem of the original Web monitoring system is that it focuses on contents generated by issue generators and tries to find the best mechanism to deliver the information promptly. The pages monitored by the systems are normally header/front pages for information. For example, the front pages of CNN contains lists of links (Fig 1) to content pages (Fig 2 and 3). The real articles to be collected by the system are located in content pages, not front pages. In summary, after Web monitoring system identifies the location of updated information from the front page, the system follows sub layer to collect actual contents. In collecting contents from second layer, the system collects only the main content, not other extra information such as banners, menu, and advertisement (Fig 2). The problem is that comments added by conversation makers are not harvested.

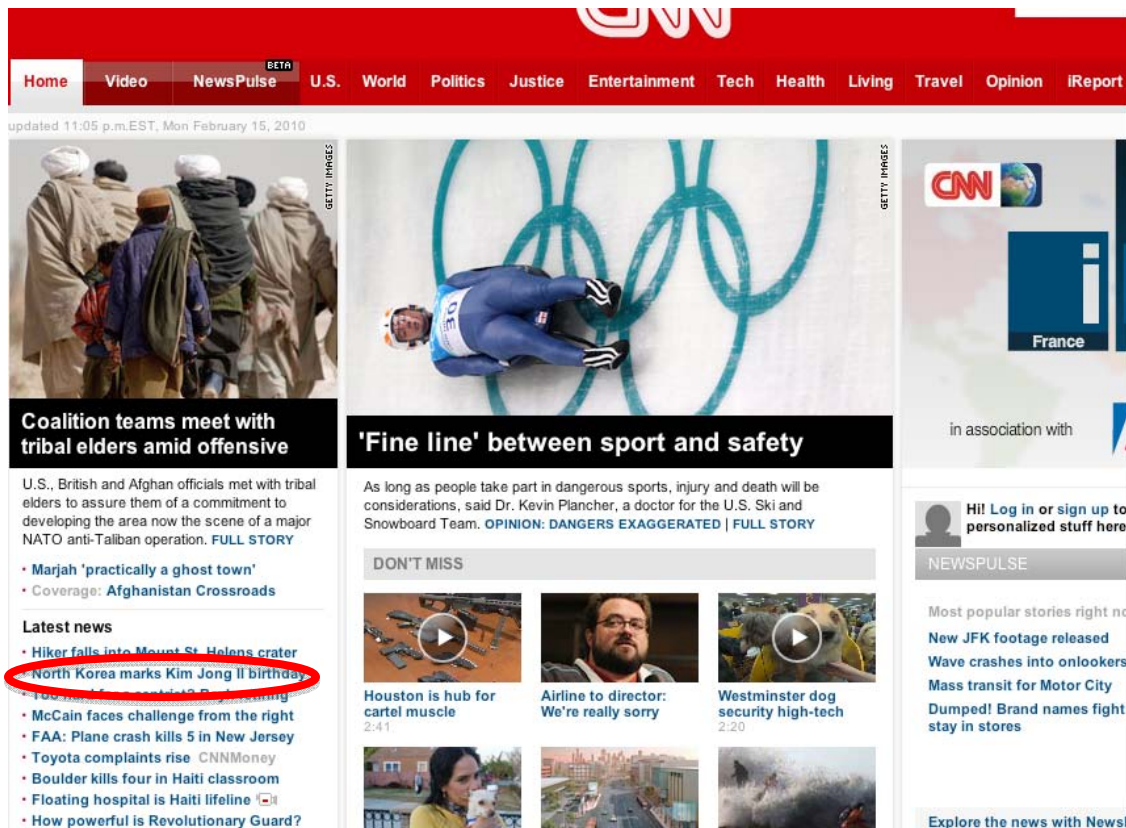



Fig 1. This is the front page of CNN Web sites. Pages are divided into several sections and Web monitoring system detects any updated links. Then the system follows the links to download full contents. For example the links in the red circle bring up the contents pages of this title (Fig 2 and Fig 3).

The contents added by the conversation makers are normally overlooked or ignored. For example, in the newspapers, the original articles are collected but comments sections are often ignored by the monitoring system. However, a recent study shows that these comments may have more impact on public opinions than original articles. Therefore, it is important to collect this information with relevant information such as dates of comments and number of replies. Note that these comments are part of one article and comments are also updated regularly. The more people have interests, the more people add comments. This requires another monitoring process for this page. Hence there should be an independent monitoring schedule for this page and section analysis technology including new style of automated information classification for comments sections. The classification here is whether the comment is in favour of or against to the

original article. New functions to study impact factors of these comments will be added to our system.

### North Korea marks Kim Jong Il's birthday

February 15, 2010 10:01 p.m. EST



This undated picture, released from North Korea's official Korean Central News Agency on December 11, 2009, shows Kim Jong-il inspecting the Kimchaek Iron and Steel Complex at Kimchaek city in North Hamgyong province.

**(CNN) --** North Korea celebrated the 68th birthday of Kim Jong Il on Tuesday with performances and festivals lauding the "Dear Leader," the country's state news agency reported.

Marked on the national calendar as a public holiday, "celebrations" included song and dance performances, and an ice sculpture festival, while senior party officials attended the screening of a documentary on Kim's "immortal exploits," North Korea's official KCNA news agency reported.

As a tradition on the occasion, on Saturday the regime gave children around the nation presents.

North Korea received rare animals as gifts for Kim from state and party leaders and other prominent figures of countries and overseas Koreans, KCNA reported.

**RELATED TOPICS**  
North Korea  
Kim Jong Il

Mix Facebook Twitter Share Email Save Print

Mix Facebook Twitter Share Email Save Print



ADVERTISEMENT

#### NewsPulse >>

Most popular stories right now

- New JFK footage released
- Wave crashes into onlookers
- Top Taliban commander captured
- High-tech dog security
- Dumped! Brand names fight to stay in stores

Explore the news with NewsPulse >



ADVERTISEMENT


FOLLOW THIS TOPIC

#### We recommend

- Vice Guide to North Korea
- 'No deal' in release of American from North Korea
- U.S. envoy: N. Korea uranium enrichment key issue

#### More World >>

- 18 killed in Belgian train crash, official says
- Jihadist Web sites warn al Qaeda leaders of possible exposure
- Trial starts in Argentina's 'robbery of the century'



Log in or sign up to comment

#### careerbuilder.com

- Part time Jobs
- Sales and Marketing Jobs
- Customer Service Jobs


#### Quick Job Search


Keywords City  
Job type State  
SEARCH more options >


Fig 2. The first part of the link from CNN front page (Fig 1). This includes the content section (marked) to be collected by the Web monitoring system.


North Korea marks Kim Jong Il's birthday – CNN.com 16/02/10 4:20 PM


soundoff (36 Comments)  
Show: Newest | Oldest | Most liked  
Showing 25 of 36 comments  
Sort by: Newest first | [Subscribe by email](#) | [Subscribe by RSS](#)


 **iransucks**  
iransucks  
Ha looks like a statue you would see at a wax museum in Disney World. There is a limitless amount of captions that could be written about this shot. simply priceless.  
5 minutes ago | Like | Report abuse


 **sgrayban**  
sgrayban  
die scumbag  
5 minutes ago | Like | Report abuse


 **Guest**  
Guest  
seriously, the most powerful man in N. Korea can't afford to see a dentist...well that's communism for y'all  
8 minutes ago | Like | Report abuse

 **Guest**  
Guest  
Classic Bond villain...  
13 minutes ago | Like (1) | Report abuse

 **Guest**  
Guest  
Maybe this fat worthless bum can spare some of his cake crumbs to the poor starving children of North Korea so they will not have to tear off tree barks, boil grass and bugs to make soup to stay alive. May he choke on his birthday cake so that his loyal subjects can like they did for his worthless b...more  
17 minutes ago | Like (2) | Report abuse

 **EKKadiddleho**  
EKKadiddleho  
I don't plan to wish birthday greetings to the world's most professional liar! This is the man who agreed to open the rail lines for trains to cross back and forth from China and Russia into South Korea, but has proven himself to be a liar. This is the man who accepts food from other nations to fe...more  
21 minutes ago | Like | Report abuse

 **Guest**  
Guest  
I o Jong un: Get out of P'yongyang now...while you can. I hat idiotic currency redenomination idea of yours didn't exactly make any believers in your ability to hold this fledgling regime together. You're in way over your head, pal. Trust nobody...especially "Dear Old Dad."  
34 minutes ago | Like (2) | Report abuse

 **paul9993**  
paul9993  
Can we please see those "immortal exploits"? Like the dam he built that caused massive flooding and destroyed most of the countries agriculture.....  
36 minutes ago | Like (1) | Report abuse

<http://www.cnn.com/2010/WORLD/asiapcf/02/15/north.korea.kim.birthday/index.html?hpt=T2> Page 2 of 5

Fig 3. The second part of linked page from CNN front page. This includes the comments section. These were considered as the junk information by the original Web monitoring system.



We also investigated the influence of social networks in the Internet leading countries as well as developing countries. We investigated how different information spreads in a society and how people influence each other on Internet based social networks. Are people really interactive or more obsessed by others while they make and change decisions? The challenge of this study was the method of extracting related information from collected data. For example, the system should be able to analyze information based on information such as location, authors, time, topics and quantity of data. We also identified important base factors for analysis and formed a new technology to extract these factors.

- **Recommendations**

Part 1: Web monitoring system.

We recommend to develop an advanced version of Web monitoring system for social Webs such as community Web sites including online news media. The main enhancements will be adding three components.

The smart multi-contexts filtering, smart scheduling system and sentimental context classification system are required. These functions are very important to collect and analyse contents from social network sites [17].

To study the pattern of comments such as the number of comments and impact of previous comments to the follow-on comments, it is necessary to distinguish comments sections from the main contents. Therefore, the system needs to be able to identify where the comments are located and divide it into individual contents. The implementation of multi-contexts filtering is expected to be a big challenge because Web page structures are very complex and non-uniform, making it difficult to identify all potential sections. The more dynamic the page is, the more complex the task becomes. This function will enable people to harvest valid information to analyse social behaviours and patterns with less garbage or noise in data set. The MCRDR method is recommended to maintain the rules to identify different sections on each page.

Then the smart scheduler is required for the system to estimate how often new comments are updated. The smart scheduler, event driven dynamic scheduler, was proposed in the

original Web Monitoring system. At each monitoring session the event detector finds anomaly of document classification for each category, not all the classification results. 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 of normal monitoring schedule. This original approach is not much useful in monitoring comments or opinion sections because the change of schedule is decided by the topics/issues, not opinions. In the new study, the patterns to make the prediction possible should be identified and we will integrate two different information, subject classification information and volume of opinions in the page in certain duration. By checking the weight of topics and size of people who put comments within a specific interval, the system may be able to predict the next schedule.

The third enhancement is going to be a classification system. In original Web monitoring system, there is a classification system using MCRDR method. However, this classification system focuses on subject like sports or politics. This is because the original Web monitoring system focuses on collecting issues (information) not opinions. Opinions normally have more sentimental expressions while original issues have factual information. It is commonly well known that the classification of sentimental information is more difficult. In this system, we have to investigate how MCRDR can classify the sentimental information in opinions.

Note that these smart functions are not only used for comment sections. Different styles of social network systems have different frameworks to exchange ideas or opinions of users. Therefore, the section filtering system should be smart enough to be adapted to these various styles of structures.

Part 2: Analysis of impact patterns in different social networks.

While the previous part is about the technology to develop the system to support the analysis of influences of various social networks, the recommendation of this part is to use the system to find to patterns that make online public opinion in Internet based social networks. There are different types of media from a single person media such as a blog to forum type social media ([http://en.wikipedia.org/wiki/List\\_of\\_social\\_software](http://en.wikipedia.org/wiki/List_of_social_software)). Their

impacts and acceptances are different in all societies. There are various different factors in these differences such as cultural backgrounds, ages and accessibilities. Several different media types should be considered during the study.

This study will focus on the impact social networks have on the forming of public issues and opinions. We are already aware that new media like the 'Internet' are upsetting the traditional media industries and officials such as TVs or Governments [20]. Because there have been many cases that they are no longer a monopoly in raising or making public issues. At present, big media companies nor governments have a strong grip on public issues in many countries. A big group of non-paid or non-professional journalists from networks named netizens are issue/opinion makers. They generate their own issues or express opinions on other issues. However, these issues are not yet qualified to become public issues or public opinions. Some of published issues independently from opinions of publishers are regarded as public issues if much more netizens reprint and spread them to other networks repeatedly. The public opinions require one more aspect. Discussions and opinions on such issues in each social network are necessary to be considered as public opinions. These discussions have different format in different social networks types. For example, comments on blogs are different from responses in forum style community. There is one commonly agreed hypothesis that netizens will be exposed to more variety of opinions and it will make people less biased by a specific direction. We propose the study that looks into this hypothesis in various circumstances. Different styles of social networks and different levels of network infra-structures should be observed.

It is not easy to find a qualitative research method using the Web monitoring system. This part of the study has to be divided into three stages, Sampling targets, defining target contents in each domain and interpretation of collected datasets.

Sampled targets/domains will be classified into three different perspectives and appropriate metrics will be used to define target domains for data collections. The first one will study the impact of accessibility and connectivity. The influence factors are not only the maturity of Internet infra-structure but also other factors such as age-groups and Internet skills. The second one is the cultural background. This has been always an important factor on human behaviour analysis. The last one is application type. The

social network is made based on a certain type of applications. Data from blog is different from Facebook style applications. Sampling domains will be derived from matrix of these perspectives.

Defining contents to be collected are related to each perspective described above and interpretations methods of data. We need to identify what should be collected. Also, we need to consider the factor that human behaviours are dependent on the situation-based. So, we need to collect the information with contextual information. For example, time stamps on each information and different type of links such as comments or responses between information are important information for analysis. Also we need to add some extra tags to each contents such as title, author and keywords.

The last stage should be interpretation of collected data. Of course, this interpretation should include statistical analysis including potentially data mining. The approach of quasi-foundationalists to interpret the collected data qualitatively can be considered. Hammersley [16], advocating this approach, defines his criteria for qualitative research to three terms, plausibility, credibility and relevance. The study should justify the claim by statistical analysis outcomes.

#### 4) References

1. Taylor, R. (2009). "Australia to build broadband network". Reuters. Retrieved 2009-04-07. Reuters. Canberra, Reuters.
2. (2009). Connect with the influencers. Seoul, A Hill & Knowlton Company. Survey Report
3. (2009). Connect. Seoul, A Hill & Knowlton Company. Survey Report
4. (2009). Trend of Online Opinion, Research Internation. Survey Report
5. Bingemann, M. (2009). "NBN boost for regional Australia." Retrieved 01.06.2009, 2009, from <http://www.australianit.news.com.au/story/0,25197,25717943-15306,00.html>.
6. Kim, Y. S., S.-K. Lee, et al. (2008). Coverage and Delay Forecast Modeling of Search Engine Services. Pacific Rim Knowledge Acquisition Workshop 2008, Hanoi, Vietnam, PRICAL.
7. Kim, Y. S., B. H. Kang, et al. (2008). "A Study on Monitoring Web Page Locating Heuristics." The 2008 International Conference on Information & Knowledge Engineering: 383 - 389.
8. Kim, Y. S. and B. H. Kang (2008). Search Query Generation with MCRDR Document Classification Knowledge. Knowledge Engineering: Practice and Patterns: 16th International Conference, EKAW 2008.
9. Kim, Y. S., B. H. Kang, et al. (2007). Search Engine Retrieval of Changing Information. The Sixteenth International World Wide Web Conference, Banff, Canada, W3C.
10. Kim, Y. S. and B. H. Kang (2007). "Tracking Government Web Sites for Information Integration." Information Research **12**(4).
11. Kim, Y. S. and B. H. Kang (2007). Coverage and Timeliness Analysis of Search Engines with Webpage Monitoring Results. The 8th International Conference on Web Information Systems Engineering, Nancy, France, Springer.
12. Kim, Y. S. and B. H. Kang (2007). Tracking Government Web Sites for Information

- Integration. the Sixth International Conference on Conceptions of Library and Information Science, Boras, Sweden.
13. Kang, B. H., Y. S. Kim, et al. (2007). Does Multi-User Document Classification Really Help Knowledge Management? The 20th Australian Joint Conference on Artificial Intelligence, Gold Coast, Australia, Springer, New York.
  14. "Indian Govt. plans for free broadband connectivity for all by 2009." Retrieved 2009, 2009, from <http://www.i4donline.net/news/news-details.asp?catid=10&newsid=8890>.
  15. Kang, B. H. and Y. S. Kim (2006). Noise Elimination from the Web Documents by Using URL paths and Information Redundancy. The 2006 International Conference on Information & Knowledge Engineering, Las Vegas, US, CSREA Press.
  16. Denzin, N. K. and Y. Lincoln, Eds. (2005). The Sage Handbook of Qualitative Research. New York, Sage Publications.
  17. Churcharoenkrung, N., Y. S. Kim, et al. (2005). Dynamic Web Content Filtering based on User's Knowledge. International Conference on Information Technology 2005, Las Vegas, USA, IEEE.
  18. Kim, Y. S., S. S. Park, et al. (2004). Adaptive Web Document Classification with MCRDR. International Conference on Information Technology: ITCC 2004, Las Vegas, Nevada, USA, IEEE.
  19. Kim, Y. S., S. S. Park, et al. (2004). Incremental Knowledge Management of Web Community Groups on Web Portals. Practical Aspects of Knowledge Management, Vienna, Austria, Springer-Verlag.
  20. Gillmor, D. (2004). "We the Media: The Rise of Citizen Journalists." National Civic Review **93**(3): 58-63.
  21. Park, S. S., Y. S. Kim, et al. (2003). Web Information Management System: Personalization and Generalization. IADIS International Conference: WWW/Internet 2003, Algarve, Portugal, IADIS Press.
  22. Wiggins, R. (2000). "AI Gore and the Creation of the Internet." First Monday **5**(10).
  23. Sapsford, R. and V. Jupp (1996). Data Collection and Analysis.
  24. Kang, B. H., P. Compton, et al. (1995). Multiple Classification Ripple Down Rules : Evaluation and Possibilities. The 9th Knowledge Acquisition for Knowledge Based Systems Workshop., Banff, Canada, SRDG Publications, Department of Computer Science, University of Calgary, Calgary, Canada.

### **3 Personnel Supported**

Prof. Tai Hoon Kim  
Hannam University, Korea

### **4 Publications**

A few publications will be submitted to Journals / conferences / workshops in 2010.

Target Conferences:

Pacific Knowledge Acquisition Workshop  
Pacific Rim International Conference on AI  
And a few other journals.

Titles of Papers (Proposed):

Incremental Webpage Segments Classification for Social Network Analysis.  
Classification of opinions in social media communication.

Note: Titles are subject to final contents and topics of papers.

**5 Interactions**

The technical challenges were presented at the Program Review of Dr. Terry Lyons on 27 Jan., 2010 in Arlington.

**6 News**

None

**7 Honors/Awards**

None

**8 Archival Documentation**

None

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

None