Title: Myths on bi-direction communication of Web 2.0 based social networks: Is social network truly interactive?

Contract Number: FA2386-10-1-4027

AFOSR/AOARD Reference Number: AOARD-10-4027

AFOSR/AOARD Program Manager: Hiroshi Motoda

Period of Performance: 26 03 2010 – 28 02 2011

Submission Date: 10 03 2011

PI:	Associate Professor Byeong Ho Kang/University of Tasmania
CoPI:	Professor Tae Hoon Kim/Han Nam University (South Korea)
Researcher:	Mr. Jae-koo Song/Han Nam University (South Korea)

Report Documentation Page	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE **20 MAR 2012**	2. REPORT TYPE **Final**	3. DATES COVERED **01-03-2010 to 01-03-2012**

4. TITLE AND SUBTITLE **Myths on bi-direction communication of Web 2.0 based social networks: Is social network truly interactive**	5a. CONTRACT NUMBER **FA23861014027**
	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,Australia,NA,NA**	8. PERFORMING ORGANIZATION REPORT NUMBER **N/A**
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) **AOARD, UNIT 45002, Australia, APO, AP, 96338-5002**	10. SPONSOR/MONITOR'S ACRONYM(S) **AOARD**
	11. SPONSOR/MONITOR'S REPORT NUMBER(S) **AOARD-104027**

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

A new Social Interaction Monitoring System was developed based on the monitoring system that was developed in the previous project in which it focused on new contents created by the generators (i.e. online news, personal blog, and government agencies and departments). The SIMS focuses on monitoring social interactions by the followers, as well as new contents. It consists of smart crawler, section filter, dynamic scheduler, event detector , social interaction analyzer and MCRDR classifier. Korean Twitter service was monitored from October 2010 to February 2011 and user activities were analyzed. The results show very limited proportion of trend words are common among three services: Google Trend, Google News and Twitter, and Twitter is more closely related to Google News rather than Google Trends. Social network moves faster than conventional media as they directly reflect public opinions.

15. SUBJECT TERMS

Social Network, Internet, World Wide Web, Data Mining

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**31**	

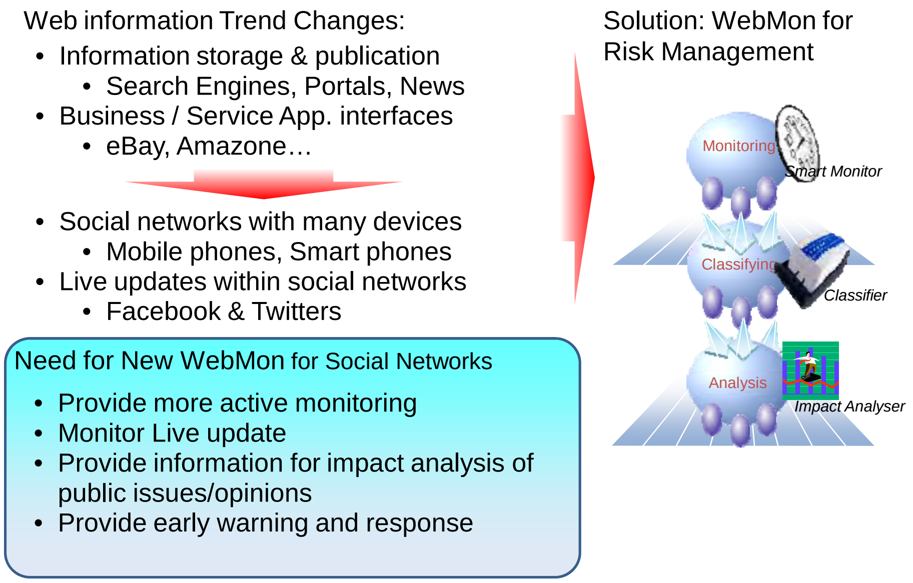
1 Objectives

This project focuses on the development of the system that monitors *social interactions* on the Internet and analysis of the collected *social interactions data*. *Social interactions* are "the acts, actions, or practices of two or more people mutually oriented towards each other, that is, any behaviour that tries to affect or take account of each other's subjective experiences or intentions." [1]. Previously major social interactions took place in the off-line. After the Internet provided various mediums for social behaviours, more and more social interactions are happening on the on-line. Especially recent uptake of the social network sites (SNSs), such as Facebook (http://www.facebook.com/) and Twitter (http://twitter.com/), accelerated the growth of online social behaviours. Even traditional online media such as online news web sites introduce functions that support social interactions (i.e., comment and sharing functions).

Understanding social interactions on the Internet and analyzing the impact of the social interactions to the society have tremendous importance. Existing social interaction analysis for the social network websites usually focuses on the static analysis, and an analysis is conducted after collecting data set (i.e., [2], [3-4]). However, the social interactions dynamically happened on the Internet and it is desirable to develop methods that monitor and analyze the social interactions in real-time. Our concept of monitoring social interaction is illustrated in Figure 1 (next page).

In this project, there are two significant objectives as follows:

- Development of an internet application that can harvest related data to find out useful information for social interactions and information flow patterns, and
- Investigation of the soundness and usefulness of the proposed application for analysing social behaviours in different domains.



Web information Trend Changes:

- Information storage & publication
  - Search Engines, Portals, News
- Business / Service App. interfaces
  - eBay, Amazone…

- Social networks with many devices
  - Mobile phones, Smart phones
- Live updates within social networks
  - Facebook & Twitters

Need for New WebMon for Social Networks

- Provide more active monitoring
- Monitor Live update
- Provide information for impact analysis of public issues/opinions
- Provide early warning and response

Solution: WebMon for Risk Management

Figure 1 Social Interaction Monitoring

2 Status of Efforts

1) System Development

We have studied web-monitoring systems for several years [5-13]. Previous monitoring system focused on new contents on new contents created by the generators (i.e. online news, personal blog, and government agencies and departments). Our Social Interaction Monitoring System (**SIMS**) focuses on monitoring social interactions by the followers, as well as new contents. Once **SIMS** detects a new content, it continually tracks the social interactions of followers and analyzes implications of the social interactions.

The system architecture of **SIMS** is illustrated in Figure 2, which has the following modules that support social interactions monitoring:

- The **Smart crawler** collects social interactions, as well as new contents from the target web sites. The behaviour of the smart crawler is controlled by the **section filter** and the **dynamic scheduler**;
- The **Section filter** identifies the specific section of a web page to extract information, such as social behaviours (i.e. comments or tweets) and contents;
- The **Dynamic Scheduler** changes revisit times of the target web sites, based on the events, which are detected by the **event detector**. The main aim of the dynamic scheduler

is to the cost caused by the revisiting activity of the smart scheduler, but collect new information (new contents and new social interactions) in a timely and politely;

- The **Event Detector** detects special events on the target web sites from the social interaction results obtained from the social interaction analyser, and classification results obtained from the MCRDR classifier;
- The **Social Interaction Analyzer** conducts statistic and semantic analysis of the collected social interaction data. The analysis results are directly reported to the users, and also used to detect special events of followers and to schedule revisiting of the smart crawler dynamically.
- The **MCRDR classifier** incrementally acquires classification knowledge and automatically classifies the collected contents. Classification results are used to detect events of the generators.

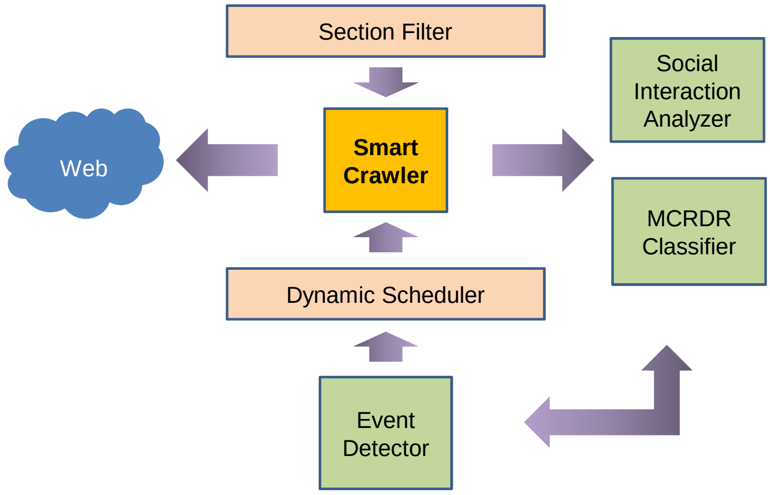
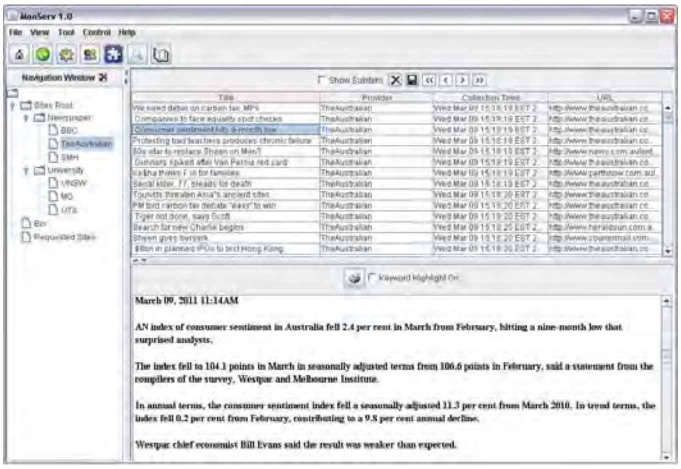


Figure 2 System Architecture of SIMS

SIMS has been developed with Java program language and PostgreSQL database. **SIMS** is based on the previous web monitoring system (**WMS**), which collects new contents on the target web sites and classifies the collected web pages by applying MCRDR (Multiple Classification Ripple-Down Rules). **SIMS** aims to extend **WMS** to monitor social interactions. Figure 3 illustrates the interface of the **Smart crawler**, currently the system supports static crawling, but its scheduling will be dynamically changed by the **Dynamic Scheduler** and the **Event Detector**. These two modules are under development now.



Title	Provider	Collection Time	URL
We need debate on carbon tax: MPs	TheAustralian	Wed Mar 09 15:18:19 EST 2...	http://www.theaustralian.co...
Companies to face equality spot checks	TheAustralian	Wed Mar 09 15:18:19 EST 2...	http://www.theaustralian.co...
Consumer sentiment hits 9-month low	TheAustralian	Wed Mar 09 15:18:19 EST 2...	http://www.theaustralian.co...
Protecting bad teachers produces chronic failure	TheAustralian	Wed Mar 09 15:18:19 EST 2...	http://www.theaustralian.co...
60s star to replace Sheen on Men?	TheAustralian	Wed Mar 09 15:18:19 EST 2...	http://www.news.com.au/ent...
Gunners spiked after Van Persie red card	TheAustralian	Wed Mar 09 15:18:19 EST 2...	http://www.theaustralian.co...
Ka$ha thinks F is for families	TheAustralian	Wed Mar 09 15:18:19 EST 2...	http://www.perthnow.com.au/...
Serial killer, 77, pleads for death	TheAustralian	Wed Mar 09 15:18:19 EST 2...	http://www.theaustralian.co...
Tourists threaten Asia's ancient sites	TheAustralian	Wed Mar 09 15:18:20 EST 2...	http://www.theaustralian.co...
PM told carbon tax debate "easy" to win	TheAustralian	Wed Mar 09 15:18:20 EST 2...	http://www.theaustralian.co...
Tiger not done, says Scott	TheAustralian	Wed Mar 09 15:18:20 EST 2...	http://www.theaustralian.co...
Search for new Charlie begins	TheAustralian	Wed Mar 09 15:18:20 EST 2...	http://www.heraldsun.com.a...
Sheen goes berserk	TheAustralian	Wed Mar 09 15:18:20 EST 2...	http://www.couriermail.com...
$1bn in planned IPOs to test Hong Kong	TheAustralian	Wed Mar 09 15:18:20 EST 2...	http://www.theaustralian.co...

March 09, 2011 11:14AM

AN index of consumer sentiment in Australia fell 2.4 per cent in March from February, hitting a nine-month low that surprised analysts.

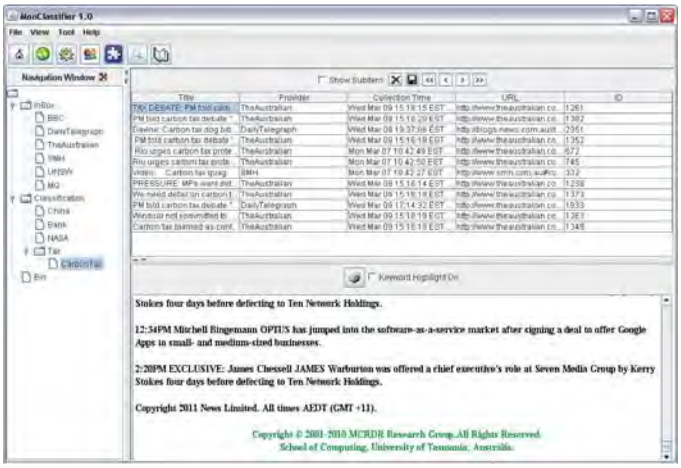
The index fell to 104.1 points in March in seasonally adjusted terms from 106.6 points in February, said a statement from the compilers of the survey, Westpac and Melbourne Institute.

In annual terms, the consumer sentiment index fell a seasonally adjusted 11.3 per cent from March 2010. In trend terms, the index fell 0.2 per cent from February, contributing to a 9.8 per cent annual decline.

Westpac chief economist Bill Evans said the result was weaker than expected.

Figure 3 Smart Crawler of SIMS

Figure 4 illustrates the MCRDR document classifier of SIMS, which is similar to the classifier that was developed for the previous project. Whereas the previous system only uses the terms in a document, the new system uses not only terms in a document, but also more advanced information for document classification (i.e. word location in the document and co-occurrence of terms).



Title	Provider	Collection Time	URL	ID
TAX DEBATE: PM told carb...	TheAustralian	Wed Mar 09 15:18:15 EST ...	http://www.theaustralian.co...	1261
PM told carbon tax debate "...	TheAustralian	Wed Mar 09 15:18:20 EST	http://www.theaustralian.co...	1382
Devine: Carbon tax dog bill...	DailyTelegraph	Wed Mar 09 19:37:16 EST	http://blogs.news.com.au/t...	2351
PM told carbon tax debate "...	TheAustralian	Wed Mar 09 15:16:19 EST	http://www.theaustralian.co...	1352
Rio urges carbon tax prote...	TheAustralian	Mon Mar 07 10:42:49 EST	http://www.theaustralian.co...	672
Rio urges carbon tax prote...	TheAustralian	Mon Mar 07 10:42:50 EST	http://www.theaustralian.co...	745
Video: Carbon tax spag	SMH	Mon Mar 07 10:42:27 EST	http://www.smh.com.au/...	332
PRESSURE: MPs want deb...	TheAustralian	Wed Mar 09 15:16:14 EST ...	http://www.theaustralian.co...	1238
We need debate on carbon t...	TheAustralian	Wed Mar 09 15:18:18 EST	http://www.theaustralian.co...	1373
PM told carbon tax debate "...	DailyTelegraph	Wed Mar 09 17:14:32 EST ...	http://www.theaustralian.co...	1833
Windsor not committed to ...	TheAustralian	Wed Mar 09 15:18:19 EST	http://www.theaustralian.co...	1361
Carbon tax blamed as cont...	TheAustralian	Wed Mar 09 15:18:18 EST ...	http://www.theaustralian.co...	1348

Stokes four days before defecting to Ten Network Holdings.

12:34PM Mitchell Bingemann OPTUS has jumped into the software-as-a-service market after signing a deal to offer Google Apps to small- and medium-sized businesses.

2:20PM EXCLUSIVE: James Chessell JAMES Warburton was offered a chief executive's role at Seven Media Group by Kerry Stokes four days before defecting to Ten Network Holdings.

Copyright 2011 News Limited. All times AEDT (GMT +11).

Copyright © 2001-2010 MCRDR Research Group.All Rights Reserved.
School of Computing, University of Tasmania, Australia.

Figure 4 MCRDR Classifier of SIMS

We also developed a system to report social interaction monitoring results as illustrated in Figure 5 to **Figure 9**. The system displays various analysis results, such as topic and issue tracking, brand volume and sentiment analysis results, social map, and social influencers.

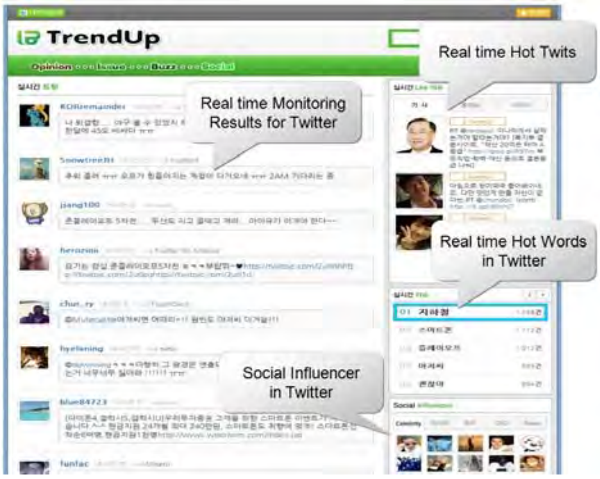

TrendUp
Opinion Issue Buzz Social
Real time Hot Twits
Real time Monitoring Results for Twitter
Real time Hot Words in Twitter
Social Influencer in Twitter

Figure 5. Topic & Issue Tracking Interface

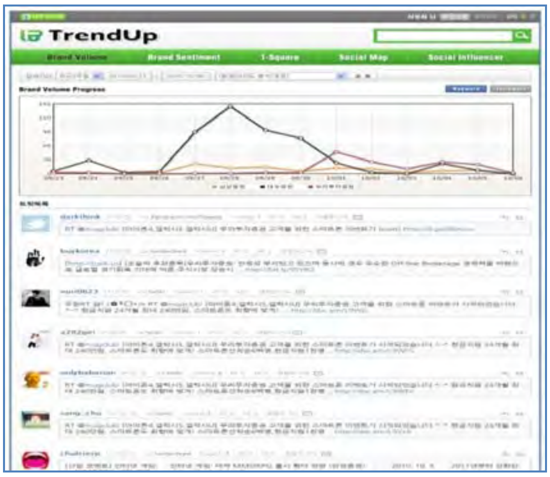


Figure 6. Brand Volume Interface

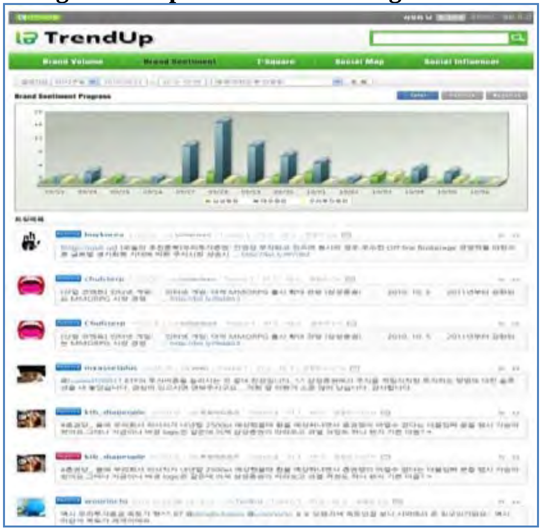


Figure 7. Brand Sentiment Analysis Interface

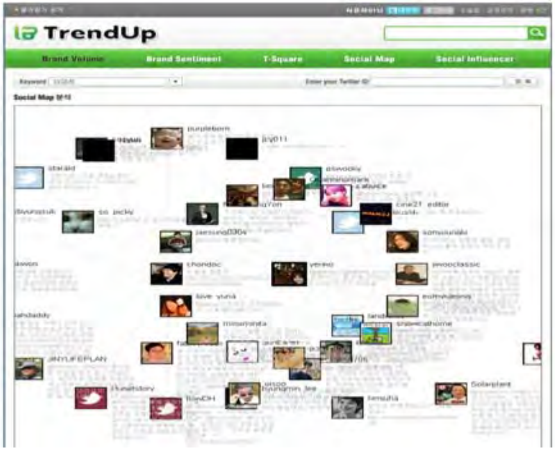


Figure 8. Social Map Interface

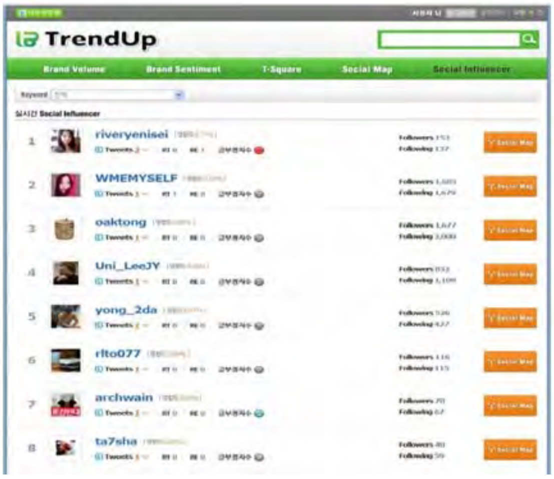


Figure 9. Influencer Analysis Interface

2) Social Interaction Analysis

Twitter Data Analysis

We monitored Korean Twitter service from October 2010 to February 2011 to analyze user activities. The results give overviews on social interactions on a popular social network site. As each twitter account has different characteristics based on relationships (i.e., business account and popular account), different crawling schedules were set for different accounts as summarized in **Table 1**. About 1.9 million accounts were monitored. Each account was monitored with different schedule. The focused group accounts, such as companies and popular accounts (about 1 % of all accounts), are monitored every one or two minutes. The active accounts (tweets ≥ 5000) are monitored every hour and the moderate accounts (100 ≤ tweets < 5000) are monitored every 12 hours and other accounts are monitored once a day. Note that the number of accounts, which should be monitored, is about 80% of all accounts.

Table 1 Crawling Schedule

Target	Monitoring Interval	Number of account	Ratio
Focused accounts	1~ 2 minutes	19,730	1.0%
tweets ≥ 5000	1 hours	69,196	3.5%
100 ≤ tweets < 5000	12 hours	139,824	7.2%
11 ≤ tweets < 99	1 day	931,432	47.7%
1 ≤ tweets < 10	1 day	633,192	32.4%
Inactive account	1 day	158,685	8.1%
Total		1,952,059	100.0%

Monitoring results are summarized in **Table 2**. A total of 185.8 million tweets were collected for five months, which means on average about 1.2 million tweets are collected every day. There are fluctuations in the number of collected tweets (about 30~40 million per month) because of real world affairs.

Table 2 Monitoring Results

	2010			2011		Total	Average
	10	11	12	1	2		
Monthly Avg.	31,778,140	37,940,009	42,846,564	41,743,512	31,480,702	185,788,927	37,157,785
Daily Avg.	1,025,101	1,264,667	1,382,147	1,346,565	1,124,311	6,142,791	1,228,558

Weekly trends of monitoring results illustrated in Figure 10 clearly show that social interactions in Twitter are related to the social events. For example, when North Korea attacked Yeon-Pyeong Island in Week 8, it causes the increase in the number of tweets.

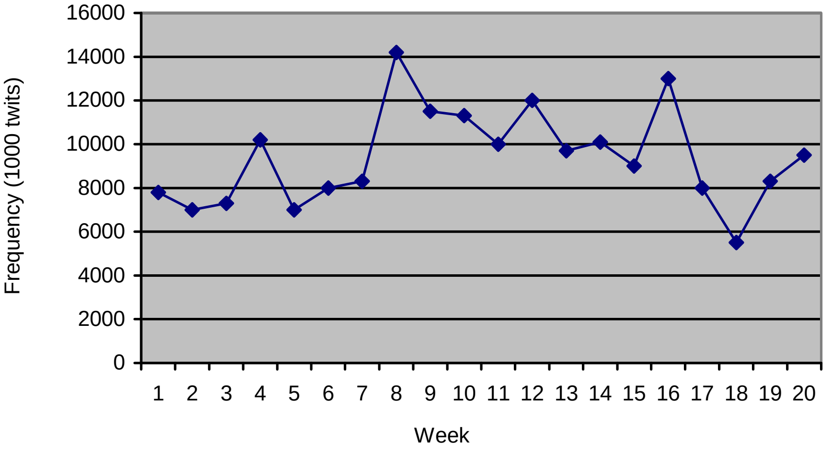


Figure 10. Weekly Monitoring Trends

There are significant changes in the number of tweets in a day as illustrated in Figure 11. Korean Twitter users normally use Twitter from 8:00 am to 10:00 pm. Most frequent time is from 10:00 pm to 1:00 am. Weekly trends and daily trends support we need to have an intelligent scheduler that changes the monitoring schedules, according to real world events.

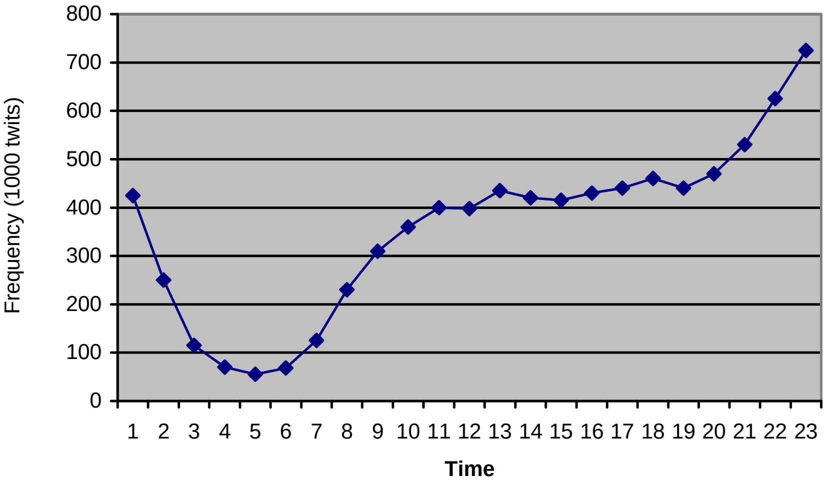


Figure 11. Hourly Monitoring Trends

3) Trend Correlation Analysis of Social Network, Search Engine, Online News

People use the web to share or disseminate information. News companies publish their articles to the public and individuals post their private stories on their blogs and share their interests using social network sites. On the other hand, people try to find information using search engines. As these activities are conducted in social contexts, they convey social trends. Manifests and disappearances of trends are different for each service. This paper uses a news service (Google News), a search engine service (Google Trend), and a social network service (Twitter) to analyse their relationships.

Data Collection Method: Trend Words Collection

We collected trend words from three commercial services, Google Hot Trends, Twitter, and Google News. First, we collected trends words from Google Trends, which shows how often a particular search-term is entered relative to the total search-volume. Google Trends, http://www.google.com/trends, an additional service of Google, provides the top 10 popular search terms of the past hour in the United States. For each of the search-terms, it provides a 24-hour search-volume graph, as well as related blog, news and web search results. Search keywords can be obtained by using Atom web feed at an hourly basis. Google Trends has been used for detecting disease outbreaks [14-18], suicide risk [19-20], software engineering trends[21].

Second, we used social network sites to collect trend words. Social networks do not publish information like online news, however messages usually reflect social interests. Therefore, social networks have been used to detect social trends, such as disease tracking [22], earthquake [23], emergency situation [24], public reaction to disease [25], and sentiment [26]. For this research, we used Twitter, which is one of most influential social network sites [27]. We do not directly extract trend words from the message; instead we used top 10 trend words provided by Twitter.

The third trend words were collected from Google News. Google News is a news aggregation service by Google Inc, which collects most up-to-date news articles from around the world. Google News also provides the trends service, which is called ‘Google News Top Stories’. The service provides the top 10 popular stories. We therefore collected top 10 stories from the “Top Stories” of Google News [28].

We collected three datasets of trend words from these three services for about one and half month (22 December, 2011 ~ 7 February, 2012) by one hour interval. Table 3 summarise trend word collections. A total of 33,715 words collected, but a total of distinct words are only 4,535 since many words continually appeared during the collection period. Note that the total number of distinct words differs from sum of the number of distinct word by each service (4,960). While Twitter provided more unique words compared to those of Google Trends and Google News, Google Trend and Google News provided similar number of words.

Table 3 Trend Word Collections

Provider	# of words	#of words per day	# of distinct words
Google Trends	11,240	468	1,033
Google News	11,218	467	1,040
Twitter	11,257	469	2,887
Total	**33,715**	**1,405**	**4,535**

Data Analysis

We conducted the following analysis with the collected data sets: First, we analysed how words are distributed over the collection time period. Some words appear frequently, but others appear only very few times. It is expected that three services have different trends reflecting each service characteristic. Second, we analyzed how much proportions of the trend words are duplicated by services. Lastly, we conducted time-based analysis for the common trend words. This analysis is interesting since the results show which service reflects trends more rapidly.

Observation 1: Trend Words Distributions

Trend words differently appeared by each service. Some trend words appear very frequently as summarized in Table 4, where top 10 trend words are presented. Google Trends and Google News have many person names (i.e. Susan G Komen, Rick Santorum, and Mitt Romney) in the lists, while Twitter has seasonal or event words, such as “HAPPY NEW YEAR, CHRISTMAS, and SUPER BOWL”. This difference comes from the differences in trend words generation methods of each service. Surprisingly there is no duplication between these high manifested trend words.

Table 4 Top 10 Trend Words during Collection Period

Ranking	Google Trends	Counts	Google News	Counts	Twitter	Counts
1	GIRL SCOUT COOKIES	133	MITT ROMNEY	1070	MPH	112
2	REPUBLICAN DEBATE	89	VLADIMIR PUTIN	288	HAPPY NEW YEAR	78
3	SUSAN G KOMEN	84	SYRIA	267	NYE	59
4	RICK SANTORUM	83	TIM TEBOW	266	CHRISTMAS	55
5	ROSE BOWL	71	RON PAUL	229	KOBE	54
6	PRIME RIB RECIPE	68	PEYTON MANNING	193	2012	41
7	MARY TYLER MOORE	65	JOE PATERNO	183	SUPER BOWL	40
8	COLBERT SUPER PAC	62	TIGER WOODS	143	HAPPY FOUNDERS DAY	39
9	SAUL ALINSKY	61	NEW YORK KNICKS	141	THE DEVIL INSIDE	38
10	GOP DEBATE	60	RAFAEL NADAL	130	NEW ORLEANS	38

Figure 12 summaries trend words distributions by each service. As small numbers of trend words show very high frequency, we analysed word distributions between 1 to 20 manifests. The results show that Twitter (57% of all manifests) has absolutely high proportion of one time manifested trend words compared to other services (30% for Google News and 15% for Google Trends). Within 24 manifests, all services have most of manifests. 99%, 91%, and 89% manifests of Twitter, Google News, and Google Trends appeared within 24 manifests. As we collected trend words every hour, if we assume that the trend words appear consecutively, the manifest frequency implies how long the trend words retained in the trend word list. For example, 57%, 30% and 15% of Twitter, Google News, and Google Trends stay for one hour and 99%, 91%, and 89% of them disappears within one day. Therefore, we can conclude the trend words change very rapidly.

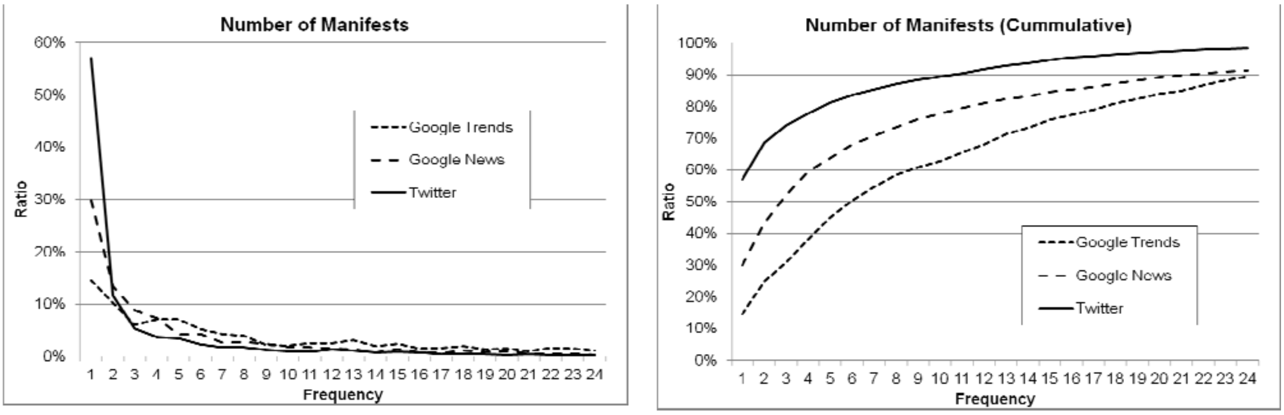


Figure 102 Number of Word Manifests

Observation 2: Common Trend Words

Table 5 summarises common trend words among three services. All three services have few common trend words (60). We expected that Google Trends has more common trend words with Google News compared to Twitter, but the results show it has more common trend words with Twitter. While Google Trends and Twitter have 130 and 139 common trend words with Google News respectively, Google Trends has 216 common trend words with Twitter. Most of them are person names and a few words are related to specific events (i.e. EARTHQUAKE, IOWA CAUCUS).

Table 5 Common Trend Words

Intersect	# of words
Google Trends – Google News	130
Google News – Twitter	139
Twitter - Google Trends	216
Google Trends – Google News– Twitter	60

Observation 3: Timelines

Table 6 summarises the number of first manifests and delays between all services. Google Trends and Twitter gave trend words earlier than Google News. Out of 60 common trends words of three services, Google Trends, Google News and Twitter, gave 31, 6, and 27 first notifications of the trend words. Note that there are duplications between services in the first notifications. Table 6 also illustrates how many days delayed after a service provided the common trend words. If Google Trends provided a common trend word first, Google News and Twitter provides the same trend word after 9 and 10.5 days later respectively. If Google News provided a common trend word first, the same trend-word appeared on Google Trends and Twitter after 2.0 and 1.50 days. Interestingly, if Twitter provided a trend word, Google Trends and Google News provided it on the same day. Note that these delays are based on median, not average. If average is used for aggregation, the delays increase significantly (see Table 6). For example, if Twitter provided the first notification, the same trend-word appeared on Google Trends and Google News after 5.02 and 4.67 days. This implies that if Twitter provided the first notifications, most same trend words were also provided rapidly, but a few of them are significantly delayed in Google Trends and Google News.

Table 6 The Number of First Manifests and Delay after First Manifest of All Services

		Google Trends	Google News	Twitter
# of first manifests		31	6	27
Google Trends	Avg	0.00	11.57	11.53
	Median	0.00	9.00	10.50
Google News	Avg	5.67	0.00	5.67
	Median	2.00	0.00	1.50
Twitter	Avg	5.02	4.67	0.00
	Median	0.00	0.00	0.00

Table 7 summarizes the number of first manifests and delays between Google Trends and Google News. Google Trends picks the issue earlier than Google News. While Google Trends provided 87 trend words first, Google News provided 40 trend words first. Note that each number includes the case that both services provide same collection time (3 trend words). For the common trend words, two services provided them within very short time. Based on median, if Google Trends provided a trend word, Google News provided it 1.00 days later. On the contrary, if Google News provided a trend word, Google Trends provide it 2.0 later.

Table 7 Delays between Google Trends and Google News

		Google Trends	Google News
# of first manifests		87	40
Google Trends	Avg	0.00	7.60
	Median	0.00	1.00
Google News	Avg	5.25	0.00
	Median	2.00	0.00

Table 8 summarizes the number of first manifests and delays between Google Trends and Twitter. Twitter is slightly earlier than Google Trends. While Google Trends provided 62 first notifications, Twitter provided 70 first. The gap of notification time between two services is greater than those between Google Trends and Google News. Based on median, if Google Trends provided a trend word, Twitter provided it 13.00 days later. On the contrary if Twitter provided a trend word, Google Trends provide it 6.00 later.

Table 8 Delays between Google Trends and Twitter

		Google Trends	Twitter
# of first manifests		62	70
Google Trends	Avg	0.00	14.66
	Median	0.00	13.00
Twitter	Avg	8.43	0.00
	Median	6.00	0.00

Table 9 summarizes the number of first manifests and delays between Google News and Twitter. Twitter is far better in picking up the common trend words than Google News. While Google News provided 45 first notifications, Twitter provided 165 first. The gap of notification time between two services is very short. Based on median, if Google News provided a trend word, Twitter provided it 2.00 days later. On the contrary, if Twitter provided a trend word, Google News provided it on the same day. Therefore, these results imply Google News and Twitter are more closely related to each other rather than Google Trends, so Twitter can be used as a predictor of trends in Google News.

Table 9 Delays between Google News and Twitter

		Google News	Twitter
# of first manifests		45	165
Google News	Avg	0.00	6.89
	Median	0.00	2.00
Twitter	Avg	2.61	0.00
	Median	0.00	0.00

Summary

This research first analyzed trend words of three popular web services – search engine, online news, and social network site. Our research shows very limited proportion of trend words are common among three services (8.6% of 4,535 distinct trend words). The results show that Twitter is more closely related to Google News rather than Google Trends. Between Twitter and Google News, Twitter picked trend word more rapidly and they usually appear in Google News on the same day. This result means that social network moves faster than conventional media as they directly reflect people’s attentions.

4) Conclusions

The project has progressed well and many interesting outcomes are found. It is going well as they are presented above. As we planned to develop the tool to analysis the social issue evolvement, the SIMS can collect social issue data. We have demonstrated some examples of data collection and analysis in this report. More detailed analysis will be conducted for the collected data. We try to submit 2-3 research papers to the referred international conferences that are derived from this project.

3 Personnel Supported

Jae-Koo Song, PhD Student, Hannam University (Korea)
Soyeon Han, PhD Student, University of Tasmania

4 Publications

Two publications planned. The papers are attached.

5 Interactions

Intermediate results have been discussed with Dr. Hiroshi Motoda when he visited PI.

6 New

7 Honors/Awards

8 Archival Documentation

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

Preliminary SIMS will be submitted with this report.
http://www.cis.utas.edu.au/iweb2

References

[1] Rummel, R.J.: *Understanding Conflict and War* Vol. 2. Beverly Hills, California: Sage Publications(1976).

[2] Lewis, K., J. Kaufman, M. Gonzalez, A. Wimmer, and N. Christakis: *Tastes, Ties, and Time: A New Social Network Dataset Using Facebook.Com.* Social Networks. **30**(4), pp. 330-342 (2008).

[3] Kwak, H., C. Lee, H. Park, and S. Moon: *What Is Twitter, a Social Network or a News Media?*, in *Proceedings of the 19th international conference on World wide web*. 2010, ACM: Raleigh, North Carolina, USA. pp. 591-600.

[4] Weng, J., E.-P. Lim, J. Jiang, and Q. He: *Twitterrank: Finding Topic-Sensitive Influential Twitterers*, in *Proceedings of the third ACM international conference on Web search and data mining*. 2010, ACM: New York, New York, USA. pp. 261-270.

[5] 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), pp. 383-389. Monte Carlo Resort, Las Vegas, Nevada, USA, (2008).

[6] 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, Acitrezza, Catania, Italy, (2008).

[7] Kim, Y.S. and B.H. Kang. *Coverage and Timeliness Analysis of Search Engines with Webpage Monitoring Results*. In: The 8th International Conference on Web Information Systems Engineering (WISE 2007), pp. 361-372. Nancy, France, (2007).

[8] Kim, Y., B.H. Kang, P. Compton, and H. Motoda. *Search Engine Retrieval of Changing Information*. In: 16th International World Wide Web Conference (WWW2007), Banff, Canada, (2007).

[9] Kim, Y. and B.H. Kang: *Tracking Government Websites for Information Integration.* Information Research. **12**(4) (2007).

[10] Park, S.S., Y.S. Kim, and B.H. Kang. *Web Document Classification: Managing Context Change*. In: IADIS International Conference WWW/Internet 2004, pp. 143-151. Madrid, Spain, (2004).

[11] Kim, Y.S., S.S. Park, B.H. Kang, and Y.J. Choi. *Incremental Knowledge Management of Web Community Groups on Web Portals*. In: 5th International Conference on Practical Aspects of Knowledge Management, pp. 198-207. Vienna, Austria, (2004).

[12] 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'04), pp. 476 (2004).

[13] Park, S.S., S.K. Kim, and B.H. Kang. *Web Information Management System: Personalization and Generalization*. In: the IADIS International Conference WWW/Internet 2003, pp. 523-530. Algarve, Portugal, (2003).

[14] Carneiro, H.A. and E. Mylonakis: *Google Trends: A Web-Based Tool for Real-Time Surveillance of Disease Outbreaks.* Clinical Infectious Diseases. **49**(10), pp. 1557-1564 (2009).

[15] Breyer, B.N., S. Sen, D.S. Aaronson, M.L. Stoller, B.A. Erickson, and M.L. Eisenberg: *Use of Google Insights for Search to Track Seasonal and Geographic Kidney Stone Incidence in the United States.* Urology. **78**(2), pp. 267-271 (2011).

[16] Malik, M.T., A. Gumel, L.H. Thompson, T. Strome, and S.M. Mahmud: *"Google Flu Trends" And Emergency Department Triage Data Predicted the 2009 Pandemic H1n1 Waves in Manitoba.* Canadian Journal of Public Health-Revue Canadienne De Sante Publique. **102**(4), pp. 294-297 (2011).

[17] Seifter, A., A. Schwarzwalder, K. Geis, and J. Aucott *The Utility Of "Google Trends" For Epidemiological Research: Lyme Disease as an Example.* Geospatial Health. **4**(2), pp. 135-137 (2010).

[18] Ginsberg, J., M.H. Mohebbi, R.S. Patel, L. Brammer, M.S. Smolinski, and L. Brilliant: *Detecting Influenza Epidemics Using Search Engine Query Data.* Nature. **457**(7232), pp. 1012-1014 (2009).

[19] McCarthy, M.J.: *Internet Monitoring of Suicide Risk in the Population.* Journal of Affective Disorders. **122**(3), pp. 277-279 (2010).

[20] Page, A., S.S. Chang, and D. Gunnell: *Surveillance of Australian Suicidal Behaviour Using the Internet?* Australian and New Zealand Journal of Psychiatry. **45**(12), pp. 1020-1022 (2011).

[21] Rech, J.: *Discovering Trends in Software Engineering with Google Trend.* SIGSOFT Softw. Eng. Notes. **32**(2), pp. 1-2 (2007).

[22] Christakis, N.A. and J.H. Fowler: *Social Network Sensors for Early Detection of Contagious Outbreaks.* Plos One. **5**(9) (2010).

[23] Sakaki, T., M. Okazaki, and Y. Matsuo: *Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors*, in *Proceedings of the 19th international conference on World wide web*. 2010, ACM: Raleigh, North Carolina, USA. pp. 851-860.

[24] Watanabe, K., M. Ochi, M. Okabe, and R. Onai: *Jasmine: A Real-Time Local-Event Detection System Based on Geolocation Information Propagated to Microblogs*, in *Proceedings of the 20th ACM international conference on Information and knowledge management*. 2011, ACM: Glasgow, Scotland, UK. pp. 2541-2544.

[25] Szomszor, M., P. Kostkova, and C.S. Louis: *Twitter Informatics: Tracking and Understanding Public Reaction During the 2009 Swine Flu Pandemic*, in *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01*. 2011, IEEE Computer Society. pp. 320-323.

[26] Thelwall, M., K. Buckley, and G. Paltoglou: *Sentiment in Twitter Events.* J. Am. Soc. Inf. Sci. Technol. **62**(2), pp. 406-418 (2011).

[27] Xu, Z., L. Ru, L. Xiang, and Q. Yang: *Discovering User Interest on Twitter with a Modified Author-Topic Model*, in *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01*. 2011, IEEE Computer Society. pp. 422-429.

[28] Liu, J., P. Dolan, E. R, #248, and n. Pedersen: *Personalized News Recommendation Based on Click Behavior*, in *Proceedings of the 15th international conference on Intelligent user interfaces*. 2010, ACM: Hong Kong, China. pp. 31-40.

Amalgamating Web Trends (DRAFT)

Soyeon Caren Han[1], Byeong Ho Kang[1], and Yang Sok Kim[2]

[1] School of Computing and Information System,
Tasmania 7005, Australia
{ Soyeon.Han, Byeong.Kang }@utas.edu.au

Abstract. Social networking services have received a great amount of attention so that the discussion of certain issues is becoming more dynamic. Many large Internet based companies provide a new service that displays the list of the trending social issues. Nowadays people publish and/or share information on the web via various technologies. Web application services such as online news, search engines, and social networks, track the issues, and provide them at real-time. This paper first reports how they are related each other.

Keywords: Scoial issues, Social networks, Google Trends, Trending topic

1 Introduction

Social network services are People use the web to share or disseminate information. News companies publish their articles to the public and individuals post their private stories on their blogs and share their interests using social network sites. On the other hand, people try to find information using search engines. As these activities are conducted in social contexts, they convey social trends. Manifests and disappearances of trends are different for each service. This paper uses a news service (Google News), a search engine service (Google Trend), and a social network service (Twitter) to analyze their relationships.

2 Related Work

dddd1. Social networking service
2.

3 Method

3.1 Trending keyword collection

We collected trend words from three commercial services, Google Hot Trends, Twitter, and Google News. First, we collected trends words from Google Trends, which shows how often a particular search-term is entered relative to the total search-volume. Google Hot Trends http://www.google.com/trends/hottrends, an additional service of Google Trends, provides the top 10 hot words of the past hour in the United States. For each of the search-terms, it provides a 24-hour search-volume graph as well as blog, news and web search results. Search keywords can be obtained using Atom web feed at an hourly basis. Google Trends has been used to detect disease outbreaks [1-5], suicide risk [6-7], software engineering trends[8].

Second, we used social network sites to collect trend words. Social networks do not publish information like online news, but messages on them usually reflect social interests. Therefore, social networks have been used to detect social trends, such as disease tracking [9], earthquake[10], emergency situation[11], public reaction to disease[12], and sentiment[13]. For this research, we used Twitter it is one of most influential social network sites [14]. We do not directly extract trend words from the message; instead we used top 10 trend words provided by Twitter.

The third trend words were collected from Google News. Google News is a news aggregator service by Google Inc, which collects most up-to-date information from thousands of publications. Unlike Google Trends and Twitter, Google News does not provide trend words. We therefore collected top ten nouns from the articles in the "Top Stories" of Google News, since most trend words of Google Trends and Twitter are noun. Google News places Top Stories reflecting user behaviours [15].

We collected three datasets of trend words from these three services for about one and half month (22 December, 2011 ~ 7 February, 2012) by one-hour interval. Table 1 summarise trend word collections. A total of 33,715 words collected, but a total of distinct words are only 4,535 since many words continually appeared during the collection period. Note that the total number of distinct words differs from sum of the number of distinct word by each service (4,960). While Twitter provided more unique words compared to those of Google Trends and Google News, Google News and Google News provided similar number of words.

Table 1 Trend Word Collections

Provider	# of words	#of words per day	# of distinct words
Google Trends	11,240	468	1,033
Google News	11,218	467	1,040
Twitter	11,257	469	2,887
Total	**33,715**	**1,405**	**4,535**

3.2 Data Analysis

We conducted the following analysis with the collected data sets: First, we analysed how words are distributed over the collection time period. Some words are more frequently manifested and others only appear only very few times. It is expected the three services have different trends reflecting each service characteristics. Second, we analysed how much proportions of the trend words are duplicated by services. Lastly, we conducted time based analysis for the common trend words. This analysis is interesting since the results show which service is more rapidly reflect the trends compared to others.

4 EXPERIEMENT AND DISCUSSION

4.1 Trend Words Distributions

Trend words are differently appeared by each service. Some trend words appear very frequently as summarized in Table 2, where top 10 trend words are presented. Google Trends and Google News have many person names (i.e. Susan G Komen, Rick Santorum, and Mitt Romney) in the lists, while Twitter has seasonal or event words such as “HAPPY NEW YEAR, CHRISTMAS, and SUPER BOWL”. This difference comes from the differences in tend words generation methods of each service. Surprisingly there is no duplication between these high manifested trend words.

Table 2 Top 10 Trend Words during Collection Period

	Google Trends		Google News		Twitter	
1	GIRL SCOUT COOKIES	133	MITT ROMNEY	1070	MPH	112
2	REPUBLICAN DEBATE	89	VLADIMIR PUTIN	288	HAPPY NEW YEAR	78
3	SUSAN G KOMEN	84	SYRIA	267	NYE	59
4	RICK SANTORUM	83	TIM TEBOW	266	CHRISTMAS	55
5	ROSE BOWL	71	RON PAUL	229	KOBE	54
6	PRIME RIB RECIPE	68	PEYTON MANNING	193	2012	41
7	MARY TYLER MOORE	65	JOE PATERNO	183	SUPER BOWL	40
8	COLBERT SUPER PAC	62	TIGER WOODS	143	HAPPY FOUNDERS DAY	39
9	SAUL ALINSKY	61	NEW YORK KNICKS	141	THE DEVIL INSIDE	38
10	GOP DEBATE	60	RAFAEL NADAL	130	NEW ORLEANS	38

Figure 1 summaries trend words distributions by each service. As small number of trend words show very high frequency, we analyzed word distributions between 1 to 20 manifests. The results show that Twitter (57% of all manifests) has absolutely high proportion of one time manifested trend words compared to other services (30% for Google News and 15% for Google Trends). Within 24 manifests, all services have most of manifests - Twitter 99%, 91%, and 89% manifests of Twitter, Google News, and Google Trends appeared within 24 manifests. As we collected trend words every hour, if we assume that the trend words appear consequently, the manifest frequency implies how long the trend words retained in the trend word list. For example, 57%, 30% and 15% of Twitter, Google News, and Google Trends stay one hour and 99%, 91%, and 89% of them disappears within one day. Therefore, we can conclude the trend words change very rapidly.

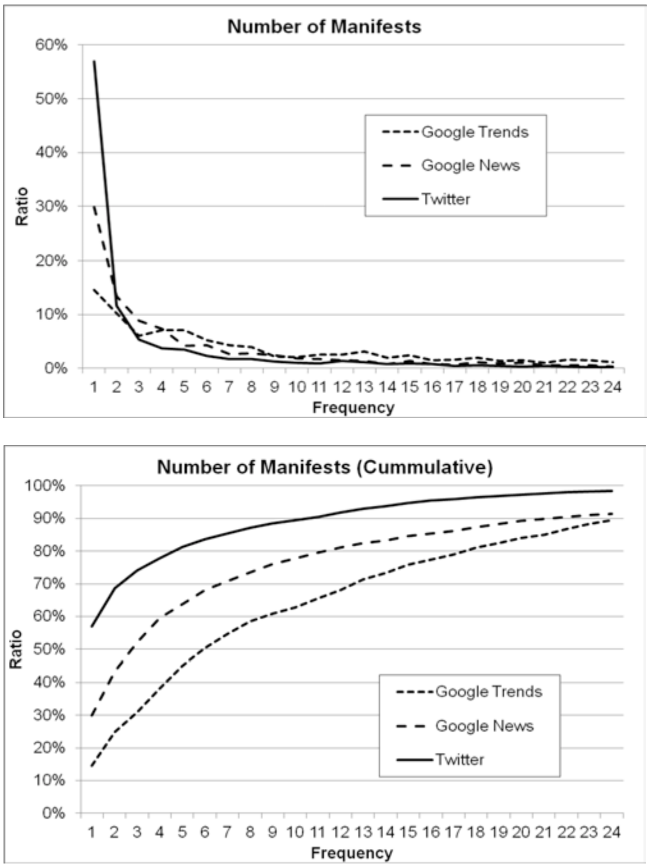


Fig. 7. The accuracy of four sub-categories

4.2 Common Trend Words

Table 3 summarizes common trend words between three services. All three services have few common trend words (60). Google Trends was expected to have more common trend words with Google News compared to Twitter, but our results show it has more common trend words with Twitter. While Google Trends and Twitter have 130 and 139 common trend words with Google News respective, Google Trends has 216 common trend words with Twitter. Most of them are person names and a few words related to specific events are included in the common words (i.e. EARTHQUAKE, IOWA CAUCUS).

Table 3 Common Trend Words

Intersect	# of words
GT-GN	130
GN-TW	139
TW-GT	216
GT-GN-TW	60

4.3 Timelines

Table 4 summarizes the number of first manifests and delays between all services. Google Trends and Twitter gave trend words earlier than Google News. Out of 60 common trends words of three services, Google Trends, Google News and Twitter gave 31, 6, and 27 first notifications of the trend words. Note that there are duplications between services in the first notifications. Figure 2 also illustrates how many days delayed after a service provided the common trend words. If Google Trends provided a common trend word first, Google News and Twitter provides the same tend word after 9 and 10.5 days later respectively. Similarly, if Google News provided a common trend word first, the same tend word was provided after 2.0 and 1.50 days by Google Trends and Twitter. Interestingly, if Twitter provided a trend word, Google Trends and Google News provided on the same day. Note that these delays based on median not average. If average is used for aggregation, the delays increase significantly (see Table 4). For example, if Twitter provided the first notification, the same tends word was provided on average after 5.02 and 4.67 days later by Google Trends and Google News. This implies that when Twitter provided the first notifications, most same trend words were also provided rapidly, but few of them are significantly delayed because of Google Trends and Google News.

Table 4 The Number of First Manifests and Delay after First Manifest of All Services

		Google Trends	Google News	Twitter
# of first manifests		31	6	27
Google Trends	Avg	0.00	11.57	11.53
	Median	0.00	9.00	10.50
Google News	Avg	5.67	0.00	5.67
	Median	2.00	0.00	1.50
Twitter	Avg	5.02	4.67	0.00
	Median	0.00	0.00	0.00

Table 5 summarises the number of first manifests and delays between Google Trends and Google News. Google Trends pick the hot issue earlier than Google News. While Google Trends provided 87 trend words first, Google News provided 40 trend words first. Note that each number includes the case that both services provide same collection time (3 trend words). For the common trend words, two services provided within very short time. Based on median, if Google Trends provided a trend word, Google News provided it 1.00 days later. Inversely if Google News provided a trend word, Google Trends provide it 2.0 later.

Table 5 Delays between Google Trends and Google News

		Google Trends	Google News
# of first manifests		87	40
Google Trends	Avg	0.00	7.60
	Median	0.00	1.00
Google News	Avg	5.25	0.00
	Median	2.00	0.00

Table 6 summarises the number of first manifests and delays between Google Trends and Twitter. Twitter is slightly more early notification compared to Google Trends. While Google Trends provided 62 first notifications, Google News provide 70 first. Note that each number includes the case that both services provide same collection time (7 trend words). Notification gaps between two services are greater than those between Google Trends and Google News. Based on median, if Google Trends provided a trend word, Twitter provided it 13.00 days later. Inversely if Twitter provided a trend word, Google Trends provide it 6.00 later.

Table 6 Delays between Google Trends and Twitter

		Google Trends	Twitter
# of first manifests		62	70
Google Trends	Avg	0.00	14.66
	Median	0.00	13.00
Twitter	Avg	8.43	0.00
	Median	6.00	0.00

Table 7 summarises the number of first manifests and delays between Google News and Twitter. Twitter is far better in picking up the common trend words between Google News and Twitter. While Google Trends provided 45 first notifications, Google News provides 165 first. Note that each number includes the case that both services provide same collection time (6 trend words). Notification gaps between two services are very short. Based on median, if Google News provided a trend word, Twitter provided it 2.00 days later. Inversely if Twitter provided a trend word, Google News provided it on the same day. Therefore, these results imply Google News and Twitter are more closely related to each other compared to Google Trends and Twitter can be used as a predictor of trends in Google News.

Table 7 Delays between Google Trends and Twitter

		Google News	Twitter
# of first manifests		45	165
Google News	Avg	0.00	6.89
	Median	0.00	2.00
Twitter	Avg	2.61	0.00
	Median	0.00	0.00

5 Conclusion

This research first analysed trend words of three popular web services – search engine, online news, and social network site. Our research show very limited proportion of trend words is common between three services (8.6% of 4,535 distinct trend words). The results show Twitter is more closely related to Google News compared to Google Trends. Between Twitter and Google News, Twitter picked trend word more rapidly and they usually appear in Google News on the same day. This result means that social network moves faster than conventional media as they directly reflect people's attentions.

6 Acknowledgements

This project was supported by AOARD, Japan.

References

1. Carneiro, H.A. and E. Mylonakis, *Google trends: A web-based tool for real-time surveillance of disease outbreaks.* Clinical Infectious Diseases, 2009. 49(10): p. 1557-1564.
2. Breyer, B.N., et al., *Use of Google Insights for Search to Track Seasonal and Geographic Kidney Stone Incidence in the United States.* Urology, 2011. 78(2): p. 267-271.
3. Malik, M.T., et al., *"Google Flu Trends" and Emergency Department Triage Data Predicted the 2009 Pandemic H1N1 Waves in Manitoba.* Canadian Journal of Public Health-Revue Canadienne De Sante Publique, 2011. 102(4): p. 294-297.

4. Seifter, A., et al., *The utility of "Google Trends" for epidemiological research: Lyme disease as an example.* Geospatial Health, 2010. 4(2): p. 135-137.
5. Ginsberg, J., et al., *Detecting influenza epidemics using search engine query data.* Nature, 2009. 457(7232): p. 1012-1014.
6. McCarthy, M.J., *Internet monitoring of suicide risk in the population.* Journal of Affective Disorders, 2010. 122(3): p. 277-279.
7. Page, A., S.S. Chang, and D. Gunnell, *Surveillance of Australian suicidal behaviour using the Internet?* Australian and New Zealand Journal of Psychiatry, 2011. 45(12): p. 1020-1022.
8. Rech, J., *Discovering trends in software engineering with google trend.* SIGSOFT Softw. Eng. Notes, 2007. 32(2): p. 1-2.
9. Christakis, N.A. and J.H. Fowler, *Social Network Sensors for Early Detection of Contagious Outbreaks.* Plos One, 2010. 5(9).
10. Sakaki, T., M. Okazaki, and Y. Matsuo, *Earthquake shakes Twitter users: real-time event detection by social sensors*, in *Proceedings of the 19th international conference on World wide web*. 2010, ACM: Raleigh, North Carolina, USA. p. 851-860.
11. Watanabe, K., et al., *Jasmine: a real-time local-event detection system based on geolocation information propagated to microblogs*, in *Proceedings of the 20th ACM international conference on Information and knowledge management*. 2011, ACM: Glasgow, Scotland, UK. p. 2541-2544.
12. Szomszor, M., P. Kostkova, and C.S. Louis, *Twitter Informatics: Tracking and Understanding Public Reaction during the 2009 Swine Flu Pandemic*, in *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01*. 2011, IEEE Computer Society. p. 320-323.
13. Thelwall, M., K. Buckley, and G. Paltoglou, *Sentiment in Twitter events.* J. Am. Soc. Inf. Sci. Technol., 2011. 62(2): p. 406-418.
14. Xu, Z., et al., *Discovering User Interest on Twitter with a Modified Author-Topic Model*, in *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01*. 2011, IEEE Computer Society. p. 422-429.
15. Liu, J., et al., *Personalized news recommendation based on click behavior*, in *Proceedings of the 15th international conference on Intelligent user interfaces*. 2010, ACM: Hong Kong, China. p. 31-40.

Analysing the trends of social issue (DRAFT)

Soyeon Caren Han
School of Computing and Information Systems
University of Tasmania
Sandy bay, Tasmania 7005, Australia
Soyeon.Han@utas.edu.au

Byeong Ho Kang
School of Computing and Information Systems
University of Tasmania
Sandy bay, Tasmania 7005, Australia
Byeong.Kang@utas.edu.au

***Abstract*— Social networking services have received a lot of attention recently so that the discussion of certain issues is becoming more dynamic. Many websites provide a new service that displays the list of the trending social issues. Responding to those issues is very crucial because their impact can be significant to organizations or individuals. In this paper, we focus on proposing the method that identifies the personalized relevance of social issues to targets, such as individuals or organizations. To achieve this aim, we first collected trending issues from Google Trends, micro-blog, and Internet news. Then, we obtained the well-structured document management system as a target domain that contains all activities and document regarding target objects. We applied the Term Frequency Inverse Document Frequency (TFIDF) to obtain the personalized relevance weight of the social issue to a target. Our experiments prove that we can identify the meaningful relevance of social issues to a target such as an individual user or an organization.**

Keywords* Goole Trends, Social Issues, Social Networks, Twitter, Related Keyword, Trending Topic

I. INTRODUCTION

Social networking services (SNS) have received a great deal of attention recently [15]. These services enable the users to communicate with others in a new way and reflect the users' real-life interests [16]. SNSs do not only change the way that people communicate but also increase the speed of sharing information. There are two reasons for the latter. Firstly, unlike other online communication services, SNSs provide push-based information. For example, while the e-mail is like a letter that a person places in somebody else's mailbox, so that it can be opened when the user wants to, SNS can be likened to the user tapping another person's shoulder and forcefully placing a message in the latter's hand. Secondly, SNS messages are broadcast to all the people linked to the sender while e-mails are sent only to the addresses specified by the sender. As the speed of the communication flow has been increased by SNS, a large amount of information exists on the Web; and because humans are social beings and are thus intensely interested in what others are doing, there are those who want to see what information people are looking for.

Many web companies have not passed up this opportunity. They develop the new service that displays the list of top trending social issues. For example, Google and Twitter provide the new service of showing the list of trending topics in Google Trends and Twitter Trending Topics, respectively. While Google Trends displays the list of top 10 fastest-rising search terms based on hourly data from Google Search, Twitter Trending topic provides the list of top 10 most discussed topics based on tweets in Twitter [10]. The flow of Twitter is influenced by the "big mouths" like celebrities or special groups. Google Trends, however, is based on the search results so that it is affected by general users. According to Rech, among the existing trending services, Google Trends provides a highly reliable list of social issues. Google Trends is a good indicator of the evolution of world interests in certain topics of search [13]. Therefore, if you define the social issues as "the events that many people are interested in", the keywords published by Google Trends should be considered as representing people's interests.

Responding to those social issues is very crucial for both individual users and organizations, since certain trending topics may have a significant impact on them. If they know the relevance of a certain trending issue to themselves, they will be able to see and exploit the opportunities and threats that such a trending topic may present for them. Unfortunately, there is no service that identifies the relevance of those trending issues to people.

In this paper, we describe our research on developing a service that identifies the personalized relevance of social issues to targets, such as individuals or organizations. First, in order to obtain the social issues, we collected social issue keywords from Google Trends for a period of 195 days, approximately over 5 months. Each keyword from Google Trends represents a certain social issue. However, there is great ambiguity in defining the exact meaning of a certain social issue by using one term from Google Trends. To reduce the ambiguity, we decided to extract several related keywords. As the top 10 social issues keywords from Google Trends are real-time information so that the related keywords should be collected from the services which provide real-time information, such as micro-blog or Internet news. In this paper, Twitter and Google News were chosen as the related keywords extractor. In order to calculate the relevance weight of social issues to a target, we applied TFIDF. TFIDF is common in calculating relevance weight. We will show the effectiveness of our method by conducting several types of experiments.

The paper is structured as follows: Section 2 presents the related work, followed by the methodology of this proposed system in Section 3. In Section 4, we describe the evaluations conducted and discuss the results. Finally, we conclude this paper in Section 5.

II. RELATED WORKS

In various aspects, social networking services have been researched including their characteristics [2, 16] and the reason why people are enthusiastic about them. In this regard, there are many works that analyze the behavior of SNSs: Putnam described it as a social capital maintainer. Boyd and Ellison investigated the difference between SNSs and other communication services, such as email or messenger [2]. There are some researchers who have analyzed different types of SNSs, such as Facebook [7], YouTube [12], and Twitter [10, 15]. Having the SNSs drawn enormous interest in a short time span, trending topic are becoming more dynamic. Because of this, tracking trends recently become the important issue in every field [13]. Many websites did not miss this opportunity and, consequently now provide the service that displays trending topics [16]. The method of trends tracking can be classified as three main sections: search-based [10, 16], social networking-based [10, 15] and news-based tracking trends [11]. Even if they use the same tracking method, the result would be different.

Unfortunately, it is hard to identify the exact meaning of a trending topic by using only a keyword from trends tracking services. It is necessary to utilize query expansion. Query expansion is widely-used in the field of information retrieval. The process of query expansion generally in-cludes four steps: resources selection, seed query construc-tion, search results review, and query reformulation [4]. Most researchers perform query expansion based on either local or global analysis [1].

To develop the personalized system with a certain target object, such as individual users or organizations, they always need to provide the digitalized domain. Fortunately, most activities of both individuals and organizations are now saved in assortment of digital information [9]. Most users utilize the information management system that enables them to manage their knowledge in a well-structured and categorized. Moreover, those systems offer centralized storage, which covers almost all activities of a target object, such as email [3], blog [6] or knowledge management system (KMS) [8].

To identify the relevance of a query to a certain document, string comparison and matching methods are briefly reviewed. A method of string matching that enables the system to make decisions using the actual content flow. This method applied in many pattern-matching and Web search areas [5]. There are several kinds of methods that are widely-used, such as the edit-distance method [14], Jaro-Winkler distance [5], Jaccard distance [18] and TFIDF distance [17].

III. PERSONALISED RELEVANCE IDENTIFICATION

In this paper, we present our research on proposing the method that identifies the personalized relevance of trends to target objects, such as individuals or organizations. To provide this personalized relevance identification, the methodology employed in this research can be divided into four phases, as follows: (1) trending social issue keyword collection; (2) related keywords extraction; (3) personalized/adapted domain identification; and (4) relevance identification.

A. *Social issue collection*

The first phase involves how to collect the trending social issues that show what people are currently most interested in. Fortunately, many websites provide the services that display the trending social issues. For example, Google, Yahoo, and Twitter provide the trends service that shows the list of trending topics in Google Trends, Yahoo Buzz, and Twitter Trending Topic, respectively.

In this paper, Google Trends has been chosen as the trending topic collector. Google Trends displays the list of the top 10 fastest-rising search terms based on hourly data from Google Search. The search-terms indicate what topics people are interested in and looking for. It is evident that Google Search is currently the most popular search engine. Because of this, Rech indicated that Google Trends most effectively provides the most searched terms and phrases [13]. Thus, Google Trends has been chosen as the trending topic collector in this study so that more accurate results would be obtained.

B. *Related Keyword extraction*

Even though the top 10 trending social issue keywords per hour were collected, ambiguity occurs when the exact meaning of a trend topic is obtained by using each trend from Google Trends. For instance, assuming that “Apple” is one of the fastest-rising search terms in Google Trends, most people may think that the keyword “Apple” is an American multinational corporation that sells computer materials. The keyword “Apple,” however, may be related to the fruit or orchard thereof. Therefore, it is necessary to expand a trending keyword by extracting several related keywords. As Google Trends displays the list of fastest-rising search terms, which are considered as real-time social issue keywords, the related keywords must be extracted from services that publish real-time publishing, such as micro-blog and Internet news [10]. If related keywords are extracted from general documents published at any time, semantically related keywords will be extracted, not keywords that are related to the trending social issue.

In this paper, Twitter and Google News were chosen as the micro-blog and Internet news service, respectively. To extract the appropriate related keywords from those services, articles related to a Google Trends keyword were first searched. As it is necessary to extract documents related to an hourly-trending social issue keyword, we extract only articles that people uploads in an hour. After collecting the articles, we applied Term Frequency (TF) method to find the most relevant nouns on a Google Trends keyword. TF weight will be defined by dividing the occurrence count of a certain term by the total number of words in the given document [17]. Then, term weights are sorted in descending order. The higher the term weight, the more the keyword is related. The best number of related keywords will be analyzed in the evaluation session.

C. *Personalized Domain*

After finishing the trending topic collection, it is necessary to obtain the digitalized document management system that contains all activities and information regarding target objects, such as individual users or organization. The document management system should be well-structured and categorized. The typical examples of a digitalized document management system are email, blog, and Knowledge Management System (KMS). Most document management systems are categorized the sections by genre. The way to categorize the document is a personal decision so that it might be subjective. However, the relevance will be viewed by people who classified that way so that it is not an issue.

In this paper, as KMS in a certain organization might contain private information, we create the virtual personalized domain by collecting the several kinds of food blogs. Since each individual's blog is concentrated on only few topics, it might not show the relevance into various trends. Therefore, we used the combination of food blogs as a target domain. It was categorized by the names of each continent and country. The combination of several kinds of food blogs is classified by continent and country folder that is defined by the International Cartographic Association (ICA). All food blogs are collected by the Google Search, with the form of such search terms as 'Nation_food_blog'. For example, to find the blogs for the 'japan' folder, we searched by using 'Japanese food blog'. We collected only the blogs that are shown in the first page of Google Search.

In the target domain, there are 4 continent categories (e.g. Asia), 14 area categories (e.g. East Asia) and 26 countries categories. We crawled 22933 documents.

D. *Relevance identification*

The goal of this paper is to identify the relevance of the collected trending topics to a target object. In this paper, trending topics are identified by using Google Trends, Twitter, and Google news. The target domain comprises the combination of different countries' food blogs. In order to identify the relevance of social issues to the target domain, we applied the Term Frequency Inverse Document Frequency method that is usually used by search engines that need to rank a document's relevance given a query. The process of identifying the relevance is as follows.

The set of trending keywords includes one Google Trends keyword and several related keywords from Twitter and Google news. What we want to obtain is the relevance weight of each document to each set of trending keyword. First, the TF should be applied. The system removes the documents that do not contain all five trending keywords. After that, the system counts each number of terms in a document and totals them. However, if the trending social issue keywords contain common words, such as 'cook', it will emphasize these words. To filter out the common terms, Inverse document frequency (IDF) is utilized. The IDF weight can be measured by dividing the total number of documents by the number of documents that contain the trending keywords. Therefore, the higher TFIDF weight is calculated by both a higher TF (in one document) and a lower DF of the term in the whole target domain [17].

After calculating the relevance weight, we decided how to visualize the relevance of trends to the target domain. Considering different characteristics of each target, the way that visualizes the relevance weight is separated by three different types as follows. The first type of relevance visualization is document-based relevance visualization. This is useful for a user who does not have a large number of documents or a complicated structure. The second type is category/folder-based relevance visualization. Most organizations have a complicated structure so that it is hard to identify all documents for them. Therefore, it might be essential for them to understand the highly-related category. The third combines both document-based and category/folder-based relevance visualization.

E. *Summary*

In this paper, the relevance value (RV) is defined as:

Gn is one of the top 10 search terms from Google Trends.

$$RV_n = \sum_{D=1}^{k} \left(TFIDF(TF(G_n, R_{m+i}), T_D) \right)$$

n is a number from 1 to 10. Then, the system searched related documents by using Google Trends key-word (Gn). Rm and Ri represent the related documents from microblog, internet news. To find the highest related keywords, TF was conducted. TD is a digitalized domain that contains all information of a target object. D is the number of the documents, from 1 to the maximum number, k. To identify the relevance of the set of trending keywords to a target domain, we totaled all documents' TFIDF weight.

IV. EVALUATION AND DISCUSSION

Evaluations of the proposed system were carried out in order to examine the success of the method. With this in mind, we collected data for evaluating the proposed meth-od. First, to extract trending social issues, we crawled Google Trends keywords for a period of 195 days, approx-imately over 5 months. As described in the introduction section, we obtain 17559 unique topics. Secondly, in order to reduce the ambiguity of the social issue, we extracted several related keywords from Twitter and Google News hourly. The target domain is the combination of different countries' food blogs, which are collected from Google search. In the target domain, there are 4 continent catego-ries (e.g. Asia), 14 area categories (e.g. East Asia) and 26 countries categories. We crawled 22933 documents. We collected only the blogs that are shown in the first page of Google Search. Each data-set contains one Google Trends keyword, several related keywords, date, and relevance weight. We calculated not only each target's relevance weight, but also each document's and each category's weight.

In the first part of evaluation section, we explain the reason why we extracted several related keywords. To do this experiment, we extracted 10 related keywords for each Google Trends Keyword, and calculated their relevance weights. Figure 1 displays the relevance weights for the number of related keywords. First, when we did not extract any related keyword, most relevance weights are almost 0, which can be seen the blue line in the bottom. If the relevance weights are almost 0, it might be very hard to distinguish which social issue is highly related to a target. On the other hand, you can clearly see the difference when we extracted at least one related keyword. This result proves the importance of the related keywords extraction.

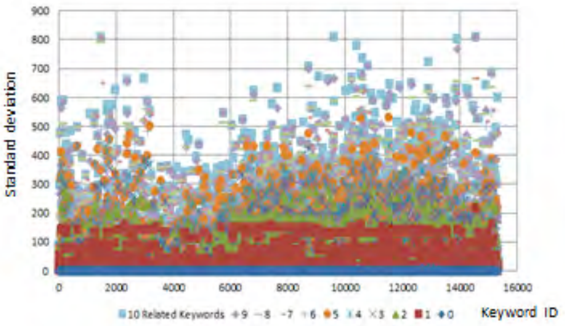


Figure 1. Relevance weight (using TFIDF) for the number of related keywords

Even though the Figure 1 represents the importance of the related keyword extraction, it is not easy to see how the relevance weights are changed. In this section, we provide the Figure 2 which displays the standard deviation of relevance weights for the number of related keywords. In Figure 2, the x-axis represents the number of related keyword. As can be seen in the graph, the more we extract the related keywords, the higher the standard deviation is obtained. This result indicates that the more related keywords are extracted, the clearer distinction of the document is derived.

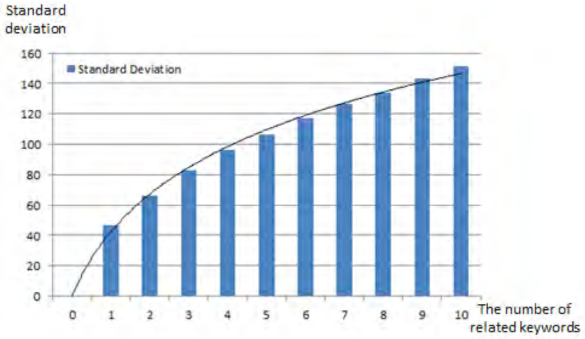


Figure 2. Standard deviation for TFIDF

Next, we consider the appropriate number of related keywords. In Figure 2, you can see the gap between each stand deviation is dwindling. This result might show the proper number of related keywords to identify the personalized relevance of social issues to a target object. There are two reasons why we would like to obtain the most appropriate number of related keywords. First, we need to consider the time consumption. We collected related articles from Twitter and Internet news hourly; Tweets are almost 90 and news articles are almost 10. It depends on the number of articles that people uploads in an hour. Extracting over 10 related keywords may not be consumed a lot of time, but it does consume a great amount of time to calculate the personalized relevance of Google keyword and over 10 related keywords to a target. Secondly, the number of the related articles is limited. If we extract over 10 related keywords, some keywords might be not really related to that social issue. In other words, some keywords may be just very general words that are no relevant with a Google Trends social issue keyword. Therefore, for these two reasons, it is necessary to get the suitable number of the related keywords. With this in mind, we present the Figure 3.

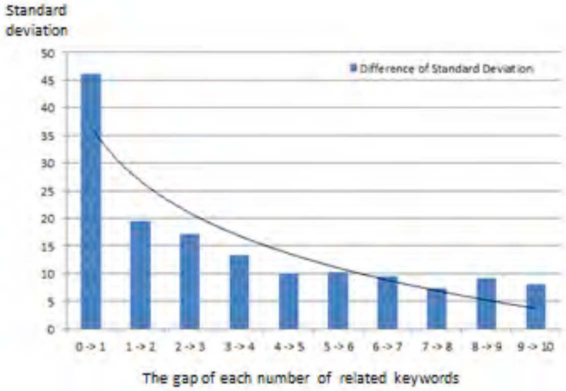


Figure 3. The difference between each standard deviation for the number of the related keywords

The Figure 3 indicates the difference between each standard deviation for the number of the related keywords. The number in x-axis represents the number of related keyword. For example, the '0->1' indicates that the difference in standard deviation between '1 Google keyword + 0 related keyword' and '1 Google keyword + 1 related keyword'.

As can be seen in the graph, at first section, 0 to 1, the difference is the highest in this graph. Then, the rate of those three sections, '1 to 2', '2 to 3', and '3 to 4', follow that of '0 to 1' section. From the '4 to 5' section, the differences become similar or less. Therefore, it seems that it is appropriate to extract 5 related keywords hourly. It is obvious that 5 related keywords are suitable so that we will conduct the user study of the relevance weight accuracy for the number of related keywords in the future.

As mentioned before, the reason why we extracted the related keywords is to reduce the ambiguity of a Google Trends social issue keyword and improve the ability to identify the relevance of a social issue to a target.

In this section, we examined whether the related keywords that we extracted are useful to display the exact meaning of social issue and identify the relevance of a social issue to a target, such as individual or organization. To do so, we present a qualitative comparison among three types of related keywords; 5 related keywords by using TF, related searches from Google Trends, and related keywords from WordNet. To show the various keywords, we choose 20 Google Trends social issue keywords, which are top10/bottom10 in relevance weight rankings

	Keyword	Related Keywords					TFIDF	Related Searches provided by Google					TFIDF	WordNet			TFIDF
Top 10	air jordans	sniaggac	yahoosports	frenzy	cause	shoppers	679.4217	air jordan 11 retro concords	air jordan	jordan shoes	new air jordans	air jordan 11 concord	0	N/A			0
	x factor winner	nxvh	winnerthe	myspace	finale	winner	676.2475	x factor winner 2011	x factor finale	melanie amaro x factor	x factor 2011	xfactor	0	N/A			0
	work it	time	things	day	one	relatablequote	637.5518	bosom buddies	last man standing	new tv shows 2012	revenge	bachelor	1.1925	N/A			0
	friday the 13th	day	today	year	one	jason	636.1167	friday the 13	fredag den 13	freddy krueger	friday 13th	friday the 13th quotes	0	N/A			0
	truffles	world	food	minutes	chocolate	truffle	634.9386	truffle	fedora	perugia	stem cell	60 minutes	21.4043	fungus	vegetable	candy	26.1856
	truffles	food	world	chocolate	posts	foie	614.8527	fedora	perugia	stem cell	60 minutes	empathy	2.5708	fungus	vegetable	candy	26.1856
	phish	blog	york	post	year	city	594.3573	francesca woodman	sugarland	uk basketball	nikon	girl with the dragon tattoo	2.8419	N/A			0
	friday the 13th	today	year	day	people	lt	588.8346	friday 13th	friday the 13	viernes 13	13th friday	fredag den 13	0	N/A			0
	restaurant week	time	lunch	food	site	diego	588.4931	mutual funds	nyc.gov	nyc restaurant week	flowers of war	mlk quotes	0.0000	N/A			0
	taylor lautner	people	blog	news	year	day	581.1989	bolo	hallo pizza	hayley williams	la sirenetta	lily collins	0.4350	N/A			0
Bottom 10	mega upload	megaupload	denovo	anonymouswiki	gomegaupload	gomegaupload	0.1557	megauplaupload	anonymous	department of justice	fbi	icefilms	0.0594	N/A			0
	coachella	wid	rolln	snoopdogg	coachizzle	ticket	0.1532	coachella 2012	coachella 2012 lineup	coachella line up	coachella lineup	tomorrowland	0.1404	N/A			0.1404
	doj	fbi	riaa	anonymousirc	mpaa	warner	0.0995	department of justice	universal music	justice.gov	universal	anonymous	0	executive department			0
	honey badger	beatthesaints	badgerrt	musburger	brent	gt	0.0984	alabama crimson tide	alabama football schedule	alabama national championships	honeybadger	jarrett lee	0	musteline mammal			0
	ifl	owens	iflifl	terrell	iflhttp	nfl	0.0934	ifc	indoor football league	itl	terrell owens	gary carter	0.3828	N/A			0
	the weeknd echoes of silence	silencert	svdxoqzn	mrmiketubbz	np	chills	0.0802	the weeknd	clams casino	michael jackson	alan hansen	xfactor	0	N/A			0
	nfl playoff schedule	broncos	tebow	timtebow	wtcommunities	help	0.0672	h and r block	raven	bcs	dr mercola	peyton manning	0.5864	N/A			0
	ben roethlisberger	offseason	steelers	deadspin	broncos	tebow	0.0672	ben roethlisberger wife	eli manning	peyton manning	andy dalton	big ben	0	N/A			0
	doj	megaupload	fbi	umg	websitesattack	technica	0.0594	universal music	department of justice	mpaa	universal	justice.gov	0	executive department			0
	emily maynard	bachelorette	enews	womack	keystone	brad	0.0053	brad womack	girl scout cookies	the bachelor	hysteria	joe magrane	0.5128	N/A			0

Table 1. Top10/ Bottom10 of relevance weight based on related keywords, related searches and WordNet

The first section in the table covers the related keywords by using TF and their relevance weight. The related keywords seem quite understandable and help users to figure out the exact meaning of the each related keywords. The related searches from Google Trends website are also quite understandable and enables user to get the idea of that social issue. However, Word-net is able to extract only few related keywords. This is because most Google Trends keywords are related to celebrities name or event. Wordnet provides semantically related words/terms so that the wordnet system cannot find much related keywords.

Let's move to the next step. As you can see in the table, the relevance weights of those two groups, the related searches from Google and the related terms Wordnet, are almost 0. However, with the related keyword by TF, they display the relevance very clearly. Compare to other two groups, it derives much better and recognizable results. For the future work, we will conduct the user study to figure out which related keywords group is useful.

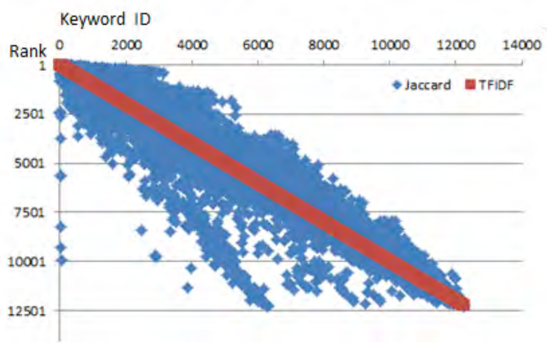


Figure 4. Trend of TFIDF and Jaccard

In this paper, we used TFIDF as a primary approach to calculate the relevance weight of social issues to a target. TFIDF is good approach to calculate relevance weight and it is usually used for ranking relevance weight in most search engines. However, it has never used in this area before. Therefore, we conducted the experiment to prove the efficiency of TFIDF by comparing another relevance weight approach, Jaccard.

To see a similarity of trend between TFIDF and Jac-card weight, we ranked each keyword based on each ap-plied relevance value. Then we ascended the ranks of the social issue keywords that are applied TFIDF method and matched with the rank of same keywords that are applied Jaccard. In general, similar trends are observed in both of two methods, TFIDF and Jaccard. Therefore, we can obtain the similar relevance weight regardless of relevance weight approach. For the future work, it might be good to propose new relevance weighting approach that will suitable to this project.

V. Conclusion

As described in this paper, we present our research on developing a system to identify the personalized rele-vance of trends to target objects, such as individuals or organizations. The outcome of the initial tests proved that we have achieved the three primary goals: (1) collecting the trending social issues, (2) identifying a target domain, and (3) demonstrating the relevance of the trending topic to a target domain. First, we collected social issues from Google Trends, Twitter and Internet news. The target domain for this paper is the combination of different countries' food blogs. We constructed the virtual target domain that is well-structured and categorized so that the system can identify the relevance weight of each document and category. Finally, we applied TFIDF method to obtain the personalized relevance of social issues to a target, such as an individual or an organization.

We conducted several types of experiments. Firstly, we proved that it is necessary to extract the related keywords and show the appropriate number of related keyword in this paper. We analyzed and compared the extracted related keyword by using TF with other related keywords groups. The advantage of our related keywords is proved. We also analyzed the comparison between TFIDF and Jaccard to prove the efficiency of TFIDF. As mentioned in evaluation part, for the future work, we will conduct further analysis and evaluation, including user study.

ACKNOWLEDGMENT

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression, “One of us (R.B.G.) thanks . . .” Instead, try “R.B.G. thanks”. Put applicable sponsor acknowledgments here; DO NOT place them on the first page of your paper or as a footnote.

REFERENCES

[1] 1. Aly, AA 2008, ‘USING A QUERY EXPANSION TECHNIQUE TO IMPROVE DOCUMENT RETRIEVAL’, International Journal "Information Technologies and Knowledge", vol. 2, pp. 343-348.

[2] 2. Boyd, DM & Ellison, NB 2007, ‘Social Network Sites: Definition, History, and Scholarship’, Journal of Computer-Mediated Communication, vol. 13, no.1, pp. 210-230.

[3] 3. Brutlag, JD & Meek, C 2000, ‘Challenges of the Email Domain for Text Classification’, Microsoft Research, One Microsoft Way, Redmond, WA, USA.

[4] 4. Chum, O, Mikulik, A, Perdoch, M & Matas, J 2011, ‘Total recall II: Query expansion revisited’, Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pp.889-896.

[5] 5. Cohen, WW, Ravikumar, PD & Fienberg, SE 2003, ‘A Comparison of String Distance Metrics for Name-Matching Tasks’, IIWeb, pp. 73-78.

[6] 6. Fitzpatrick, K 2007, ‘The Pleasure of the Blog: The Early Novel, the Serial, and the Narrative Archive’, POMONA FACULTY PUBLICATIONS AND RESEARCH.

[7] 7. Joinson, AN 2008, ‘Looking at, looking up or keeping up with people?: motives and use of facebook’, CHI '08 Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems, ACM New York, NY, USA, pp. 1027-1036.

[8] 8. Juang, YS, Lin SS & Kao, HP 2008, ‘A knowledge management system for series-parallel availability optimization and design’, Expert Systems with Applications, pp. 181-193.

[9] 9. Kolbitsch, J & Maurer H 2006, ‘The Transformation of the Web: How Emerging Communities Shape the Information we Consume’, Journal of Universal Computer Science, vol. 12, no. 2, pp. 187-213.

[10] 10. Kwak, H, Lee, C, Park, H & Moon, S 2010, ‘What is Twitter, a social network or a news media?’, WWW '10 Proceedings of the 19th international conference on World wide web, ACM New York, NY, USA, pp. 591-600.

[11] 11. Liu, J, Dolan, P & Pedersen, ER 2010, ‘Personalized news recommendation based on click behavior’, IUI '10 Proceedings of the 15th international conference on Intelligent user interfaces, ACM New York, NY, USA, pp.31-40.

[12] 12. Paolillo, JC 2008, ‘Structure and Network in the YouTube Core’, Proceedings of the 41st Annual Ha-waii International Conference on System Sciences (HICSS 2008), pp. 156.

[13] 13. Rech, J 2007, ‘Discovering trends in software engineering with google trend’, ACM SIGSOFT Software Engineering Notes, ACM New York, NY, USA, vol. 32, no. 2.

[14] 14. Ristad, ES & Yianilos, PN 1998, ‘Learning string-edit distance’, IEEE Trans Pattern Annual mach. Intell., pp. 522-532.

[15] 15. Sakaki, T, Okazaki, M & Matsuo, Y 2010, ' Earthquake shakes Twitter users: real-time event detection by social sensors', WWW '10 Proceedings of the 19th international conference on World wide web, ACM New York, NY, USA, pp. 851 – 860.

[16] 16. Tirado, JM, Higuero, D, Isaila, F & Carretero, J 2011, ‘Analyzing the impact of events in an online music community’, SNS '11 Proceedings of the 4th Work-shop on Social Network Systems, ACM New York, NY, USA, no.6.

[17] 17. Wu, HC, Luk, RWP, Wong, KF & Kwok, KL 2008, ‘Interpreting TF-IDF term weights as making relevance decisions’, ACM Transactions on Information Systems (TOIS), ACM New York, NY, USA, vol. 26, no. 3.

[18] 18. W. Cohen, P. Ravikumar, and S. Fienberg 2003. A comparison of string metrics for matching names and records. In Proceedings of the workshop on Data Cleaning and Object Consolidation at the International Conference on Knowledge Discovery and Data Mining.