A PRELIMINARY STUDY ON THE EXTRACTION OF SOCIO-TOPICAL WEB KEYWORDS

KULWADEE SOMBOONVIWAT

Graduate School of Information Science and Technology, University of Tokyo,
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan
E-mail: kulwadee@tkl.iis.u-tokyo.ac.jp

SHINJI SUZUKI

Institute of Industrial Science, University of Tokyo,
4-6-1 Komaba, Meguro-ku, Tokyo, 153-8505, Japan
E-mail: suzuki@tkl.iis.u-tokyo.ac.jp

MASARU KITSUREGAWA

Institute of Industrial Science, University of Tokyo,
4-6-1 Komaba, Meguro-ku, Tokyo, 153-8505, Japan
E-mail: kitsure@tkl.iis.u-tokyo.ac.jp

In recent years, the Web has become a popular medium for disseminating information, news, ideas, and opinions of the modern society. Due to this phenomenon, the Web information is reflecting current events and trends that are happening in the real world which, in turn, has attracted a lot of interest in using the Web as a sociological research tool for detecting the emerging topics, and social trends. To facilitate such kind of sociological research, in this paper, we study the characteristics of socio-topical web keywords sampled from a series of Thai web snapshots. The socio-topical web keyword, extracted from the content of some web pages, is a keyword relating to some topics of interest in a real-world society. The study was conducted as follows. First, the socio-topical keywords were sampled from the inverted index of each Thai web snapshot. Then, for each sampled keyword, we observe the pattern of changes of the number of documents containing the keyword, and the inverse document frequency (IDF) scores. Finally, we try to find the relationships between the observed patterns of changes and their corresponding real-world events in the Thai society.

1. Introduction

The Web has been increasingly gaining popularity as a tool for disseminating information, news, ideas, and opinions of the modern society. Due to this phenomenon, the Web information is reflecting current events and trends that are happening in the real world which, in turn, has attracted a lot of interest in using the Web as a tool for sociological, marketing, and survey research.

For example, brand popularity and penetration may be observed from the Web using a simple technique such as plotting the number of occurrences of topical keywords (e.g. the brand names) in web pages over time.

In this paper, we will study the characteristics of *socio-topical web keywords*. The socio-topical web keyword, extracted from the content of some web pages, is a keyword relating to some topics of interest in a real-world society. The characteristics found in this study should provide some insights into the development of techniques and tools for detecting the emerging topics, and social trends from the Web.

We study the characteristics of socio-topical web keywords sampled from a series of Thai web snapshots. The Thai web snapshots were created by periodically crawling a subset of the Thai Web (we use the word "Thai Web" to represent a set of all web pages written in the Thai language that are accessible on the Web via some URLs). The series of Thai web snapshots consists of three dataset crawled on October 2006, December 2006, and January 2007.

The study was conducted as follows. First, the socio-topical keywords were sampled from the inverted index of each Thai web snapshot. Then, for each sampled keyword, we observe the pattern of changes of the number of documents containing the keyword, and the inverse document frequency (IDF) scores. Finally, we try to find the relationships between the observed patterns of changes and their corresponding real-world events in the Thai society by considering some examples of the socio-topical keywords extracted from the datasets.

The rest of the paper is organized as follows. In Section 2, we review some related works on the study of web evolution. Section 3 explains related basic concepts. Next, Section 4 describes our experimental environment and reports the results of our study. Finally, Section 5 concludes the paper.

2. Related Works

Much research has been done on detecting the emerging trends on the Web based on link analysis ([3, 4, 5,]). In [3], Amitay *et al.* proposed a time-stamped links based trend detection method. The time stamp of each link is determined from the Last-Modified time of the corresponding web pages. For a given keyword, they collected top pages of search results and pages that pointed to the top pages. Then, they plotted a histogram of time-stamped links for these set of pages to check the trend of the given keyword.

In September 2003, the Internet Archive [1] started providing a text search engine service (Recall [2]) over its 11 billion archived web pages. The Recall

search engine offered time-based search and provided a graph showing changes of the frequency of the search query keyword over time along with the list of relevant results. The emerging and fading of the search query keyword may be determined from that graph. This time-based search service, which was once provided by the Internet Archive, represents an example of works on content-based trend detection from Web data.

3. Background

3.1. *Language Identification of Web Pages*

TextCat [7] is an implementation of an n-gram based text classification proposed by Cavnar and Trenkle [8]. In short, the concept of n-gram based text classification technique is to calculate a profile of an input document with an unknown category. Then, based on a comparison metric (such as an out-of-place metric) the profile of the input document will be compared with the profiles of all known categories which were created earlier using a number of documents of which the categories are known. The output of the classification is the name of the categories with the closest matches.

The n-gram based text classification technique can be easily applied to the task of language identification of web pages. In this context, any category can be viewed as corresponding to a language; a category's profile is used to represent a profile of a language. And, the classification output is the name of the languages with the closest matches.

In this research, we will use TextCat to identify the languages of crawled web pages. The TextCat library [7] comes with several built-in language profiles. Because our main concern here is to precisely identify the Thai web pages, the TextCat library has to be configured so that the precision of Thai web pages identification is maximized. This may be done, for example, by eliminating from the TextCat library some profiles of languages that are rarely found in the input data to reduce false positive classification of Thai web pages. In our experiment, we have eliminated some minority language profiles such as Welsh, Sanskrit, and Drents from the TextCat library.

Our language identification method can be described as follows.

(1) Remove all html tags from an input web page.

(2) A web page frequently includes English navigation menu, disclaimer, copyright texts, *etc.* at its top and bottom parts. These kinds of textual content can deteriorate the precision of the language classifier. To prevent the effects of such unrelated textual content on the classification

results, a portion of the first and the last parts of the resulting text from the previous step will be removed.

(3) Submit the remaining text from (2) to TextCat.

(4) If the output string of TextCat is UNKNOWN, then try to infer the language from the charset specified in html META tag. The input web page with html charset equals to "tis-620" or "windows-874" will be classified as a Thai web page.

(5) If the output string of TextCat contains [thai], the input web page will be classified as a Thai web page. Otherwise, the input web page will be classified as a non-Thai web page.

3.2. *Inverse Document Frequency (IDF score)*

An inverse document frequency (IDF score) is a measure of the general importance of the term. The IDF score is widely used in the TF*IDF weighting scheme to scale down the effects of terms that occur in many documents irrespective of the content. The intuition behind the IDF is that a term which occurs in many documents should be given less weight than one which occurs in few documents because it has less discrimination power.

The IDF score of a term t may be calculated by:

$$\text{IDF\_SCORE}\ (t) = 1 + \log\ (D/D_{+}) \qquad (1)$$

where, D : is the total number of documents in the collection, and

D_{+} : is the number of documents containing term t.

4. Experiments and Results

4.1. *Datasets*

In this section, we will describe a procedure for constructing the datasets used in our experiment. Firstly, a series of Thai web snapshots was first created by periodically crawling of Thai web pages. In order to selectively collect Thai web pages from the World Wide Web, we have applied a language specific web crawling strategy as proposed in [6] and [11]. In all of the crawls, the start URLs used are: http://webindex.sanook.com/ and http://directory.truehits.net/ (both URLs are the homepages of Thai web directories). The crawls were conducted consecutively in October 2006, December 2006, and January 2007. Then, for each Thai web snapshot, only the Thai web pages were selected and added into the dataset. The number of Thai web pages in each dataset is as shown in Table 1.

Table 1. Number of Thai web pages and number of distinct keywords in each dataset.

Dataset	Crawl Period	Number of Thai web pages	Number of distinct keywords
Thai200610	October 2006	210,889	1,611,055
Thai200612	December 2006	108,415	1,046,739
Thai200701	January 2007	280,429	2,074,516

Note that, a Thai web page is a web page that is likely to be written in Thai language. The language identification of web pages was done using the method explained earlier in Section 3.1.

4.2. *Experimental Setup*

In order to study the characteristics of the socio-topical web keywords in the datasets, it is necessary to tokenize textual content of all web pages in the datasets and create inverted indexes which store statistics about the tokens found in the documents. In the following subsections, we will discuss about the tokenization of Thai web pages and the creation of the inverted indexes.

4.2.1. *Tokenization of Thai web pages*

Tokenization refers to the process of converting textual content of the documents into a stream of words which will later be used as the index terms in an inverted index [13, 12].

In Thai written text, there is no word boundary. Thai words are implicitly recognized based on judgment of individual reader. There are many works on word segmentation problem for Thai language *e.g.* [14, 15 and 16]. In this paper, the tokenization of textual content of Thai web pages was done using the ThaiAnalyzer [10], which is a software package providing Java APIs for tokenizing Thai text.

4.2.2. *Creation of the Inverted indexes*

An inverted index [17] is a data structure that maps terms to the documents that contain them. The inverted index is widely used to allow full text search on a collection of documents. In our experiment, we will use the inverted indexes of the Thai datasets to extract socio-topical web keywords.

The Apache Lucene API [9] is an open-source text search engine library written entirely in Java. In Lucene, the inverted index maps terms to documents containing it. In order to make term-based search more efficient, The Lucene's

index also stores statistics about terms, such as term frequency and the number of documents containing the terms.

We have created inverted index for each dataset using Apache Lucene API. The number of distinct keywords (*i.e.* terms) in the resulting inverted index for each dataset is as shown in Table 1.

4.3. *Results: Characteristics of Some Socio-topical Web Keywords*

As the first step toward the extraction of important topical keywords from a series of web snapshots, in this paper, we will study the characteristics of such keywords from our Thai web snapshots. We have selected some topical keywords to study their characteristics (see Table 2). In the selection of these keywords, we have tried to choose the keywords that are related to events in the real world according to the crawl time of the web snapshot. For example, the first keyword "ปฏิวัติ" (coup) has been selected as an interesting keyword because during the end of September 2006, there was a coup in Thailand.

4.3.1. *The number of documents containing the socio-topical keywords*

Let us now consider the number of documents containing the keywords (D_+). From Table 2, it can be seen that the D_+ value is increasing when the corresponding real-world event is approaching or when the corresponding topic is gaining more interest from the society. For example, in the case of socio-topical keyword "New Year", it can be seen that its D_+ value is increasing sharply when the "New Year" is approaching (*i.e.* from 273 in December 2006 to 1772 in January 2007).

4.3.2. *IDF scores of the socio-topical keywords*

The IDF score of each keyword in the inverted indexes was calculated using Equation (1). The range of the IDF scores for the keywords in our dataset is from 1.0 to 12.0. Table 2 shows the IDF scores of some selected socio-topical web keywords in our Thai dataset.

According to our observation, the IDF scores of the keywords that are related to the real-world events are mostly in the middle part (*i.e.* 3.5 to 6.5) of the IDF scores ranking. From Table 2, the IDF scores of all keywords do not significantly vary over time.

Table 2. IDF scores and D_+ of some socio-topical web keywords.

Keyword **(meaning)**	**IDF scores (Number of documents containing keywords; D_+)**		
	Thai2006210	**Thai200612**	**Thai200701**
ขึ้นปีใหม่ (New Year)	**8.01** (190)	**6.98** (273)	**6.06** (1772)
ปฏิวัติ (Coup)	**5.96** (1476)	**6.29** (545)	**6.21** (1517)
ทักษิณ (Name of a former Prime Minister of Thailand)	**4.58** (5705)	**4.78** (2448)	**5.22** (4091)

The observations from the study of the characteristics of the socio-topical keywords in Section 4.3.1 and 4.3.2 are as follows.

- The evolution of the number of documents containing the selected socio-topical web keywords (D_+) correctly reflects the trends of corresponding real-world events.
- The evolution of the IDF scores of the selected socio-topical web keyword does not significantly reflect changes in the real-world.
- However, based on the observation that the socio-topical web keywords typically have IDF scores within the range of 3.5 to 6.5, it may be possible to use the IDF scores for automatically discriminating or determining the socio-topical keywords out of a large number of keywords in the inverted indexes of the datasets.

5. Conclusions and Future Works

The popularity of the Web as a medium for disseminating information, news, ideas, and opinions of the modern society has attracted a lot of interest in using it as a tool for observing and understanding emerging social trends, ideas, and opinions. To facilitate this kind of sociological research, we have conducted a preliminary study on the characteristics of socio-topical web keywords sampled from a series of Thai web snapshots.

Two attributes of each sampled socio-topical web keywords were investigated: (1) the number of documents containing the keywords, and (2) the inverse document frequency (IDF) scores of the keywords. According to our experimental results, we found that the evolution of the number of documents

containing the socio-topical web keywords correctly reflects the trends of corresponding real-world events. We also found that a socio-topical web keyword typically has an IDF score within the range of 3.5 to 6.5.

As for the future works, we will evaluate and improve the quality of the inverted indexes produced by the ThaiAnalyzer package [10]. We also planned to crawl a larger number of Thai web pages more frequently.

References

1. Wayback Machine, *The Internet Archive.* http://www.archive.org/.
2. A. Patterson, *CobWeb Search.*http://ia00406.archive.org/cobwebsearch.ppt.
3. E. Amitay, D. Carmel, M. Herscovici, R. Lempel and A. Soffer, *Trend Detection through Temporal Link Analysis.* J. Am. Soc. Inf. Sci. Technol, **55(14)**, 1270-1281 (2004).
4. R. Kumar, J. Novak, P. Raghavan and A. Tomkins, *On the Bursty Evolution of Blogspace.* Proceedings of the 12th International Conference on World Wide Web, 568-576 (2003).
5. D. Gruhl, R. Guha, D. Liben-Nowell and A. Tomkins, *Information Diffusion through Blogspace.* Proceedings of the 13th International Conference on World Wide Web, 491-501 (2004).
6. K. Somboonviwat, T. Tamura and M. Kitsuregawa, *Finding Thai Web Pages in Foreign Web Spaces.* ICDE Workshops 2006, 135 (2006).
7. WiseGuys Internet B.V., *libTextCat – lightweight text categorization.* http://software.wise-guys.nl/libtextcat/.
8. William B. Cavnar and John M. Trenkle, *N-gram based text categorization.* Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, 161-175 (1994).
9. The Apache Software Foundation, *Apache Lucene.* http://lucene.apache.org/java/docs/index.html.
10. National Electronics and Computer Technology Center (NECTEC) Thailand, *ThaiAnalyzer Package.* http://sansarn.com/look/download.html.
11. K. Somboonviwat, T. Tamura and M. Kitsuregawa, *Simulation Study of Language Specific Web Crawling.* ICDE Workshops 2005, 1254 (2005).
12. Soumen Chakrabarti, *Mining the Web: Discovering Knowledge from Hypertext Data.* Morgan Kaufmann Publishers (2003).
13. R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval.* Addison Wesley Longman Limited (1999).
14. V. Sornlertlamvanich, T. Potipiti, C. Wutiwiwatchai, P. Mittrapiyanuruk, *The State of the Art in Thai Language Processing*. Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, 1-2 (2000).

15. S. Meknavin, P. Charoenpornsawat and B. Kijsirikul, *Featured Based Thai Word Segmentation.* Proceedings of Natural Language Processing Pacific Rim Symposium, 41-46 (1997).
16. V. Sornlertlamvanich, T. Potipiti and T. Charoenporn, *Automatic corpus-based Thai word extraction with the c4.5 learning algorithm.* Proceedings of the 18th conference on Computational linguistics, **2**, 802-807 (2000).
17. J. Zobel and A. Moffat, *Inverted files for text search engines.* ACM Computing Surveys, **38(2)**, Article No. 6, (2006).
18. Sparck Jones, K. *A statistical interpretation of term specificity and its application in retrieval*. Journal of Documentation, **28**, 11–21 (1972).