Discovery of Emerging Topics by Co-citation Graph on the Web

Naohiro Matsumura[1,3] Yukio Ohsawa[2,3] Mitsuru Ishizuka[1]
[1] *Graduate School of Engineering, University of Tokyo,*
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656 Japan
[2] *Graduate School of Systems Management, University of Tsukuba,*
3-29-1 Otsuka, Bunkyo-ku, Tokyo, 112-0012 Japan
[3] *TOREST, Japan Science and Technology Corporation*
2-2-11 Zakurooka, Miyagino-ku, Sendai, Miyagi, 983-0852 Japan

Abstract. Discovering new topics covering profitable items and ideas (e.g., mobile phone, global warming, human genome project, etc) is important and interesting. However, since we cannot completely encode the world surrounding us, it's difficult to detect such topics and their mechanisms in advance. In order to support the detection, we show a method for revealing the structure of WWW by using the KeyGraph algorithm. Empirical results are reported.

1 Introduction

In our daily lives, a new topic sometimes become suddenly popular. The topic might seem insignificant at first sight, however, it turns out to match potential needs of us. *The Tipping Point*[1] describes this kind of phenomenon where a 'little' thing can make a big difference in the future. However, we cannot detect new topics and their mechanisms in advance since we cannot completely decode the world surrounding us. Detection of a *Tipping Point*, in face of this obstacle, could be a big chance for our various activities because competitors are not aware of such new topics. We here interpret 'topics' in the broad sense that cover new items, problems, ideas, and so on (e.g., mobile phone, global warming, human genome project, etc).

These topics were born when new collaborations of existing interests satisfy our potential needs or demands. Although the hidden factors might be 'submerged' in human mind, we believe that a few signs can be mined from a database on human behaviors reflecting human mind. For this purpose, the web is an attractive information source for its size and sensitivity to trends. The web consists of an abundance of communities[2, 3], each corresponding to a cluster of web pages sharing common interest. Since a community means a chunk of shared interest, it is considered that a web page supported(or linked) from some communities satisfies their interests, and shows the movement direction of the widen human world. From this point of view, we are expecting the structure of WWW might be a key to understand the real world.

In this paper, we show a method for revealing the structure of WWW by using KeyGraph algorithm[8] to inspect that WWW reflects the real world, and that the revealed structure of WWW supports our detection of new significant topics.

2 KeyGraph Algorithm

KeyGraph[8] is originally an algorithm for extracting assertions based on co-occurrence graph of terms from textual data. The strategy of KeyGraph comes from considering that a document is constructed like a building for expressing new ideas based on traditional concepts. The processes of KeyGraph are composed by four phases.

0)Document preparation: Prior to processing a document D, *stop words*[5] which have little meaning are discarded from D, words in D are stemmed[6], and phrases in D are identified[7]. Hereafter, a *term* means a word or a phrase in processed D.

1)Extracting foundations: Graph G for document D is made of nodes representing terms, and links representing the *co-occurrence* (term-pairs which frequently occur in same sentences throughout D). Nodes and links in G are defined as follow:

- **Nodes** Nodes in G represent high-frequency terms in D because terms might appear frequently for expressing typical basic concept in the domain. High frequency terms(HF) are the set of terms above the 30th highest frequency.
- **Links** Nodes in HF are linked if the association between the corresponding terms is strong. The association of terms w_i and w_j in D are defined as

$$assoc(w_i, w_j) = \sum_{s \in D} \min(|w_i|_s, |w_j|_s), \tag{1}$$

where $|x|_s$ denotes the count of x in sentence s. Pairs of high-frequency terms in HF are sorted by *assoc* and the pair above the (*number of nodes in* G) - 1 th tightest association are represented in G by links between nodes.

2)Extracting columns: The probability of term w to appear if all the foundations in G are considered by the author is defined as $key(w)$, and the $key(w)$ is defined by

$$key(w) = 1 - \prod_{g \subset G} \left(1 - \frac{\sum_{s \in D} |w|_s \, |g - w|_s}{\sum_{s \in D} \sum_{w \in s} |w|_s \, |g - w|_s}\right). \tag{2}$$

Sorting terms in D by *keys* produces a list of terms ranked by their association with cluster, and the 12 top *key* terms are taken for *high key terms.*

3)Extracting roofs: The strength of column between a *high key term* w_i and a high frequency term $w_j \subset HF$ is expressed as

$$column(w_i, w_j) = \sum_{s \subset D} \min(|w_i|_s, |w_j|_s). \tag{3}$$

Columns touching w_i are sorted by $column(w_i, w_j)$, for each *high key term* w_i. Columns with the highest *column* values connecting term w_i to two or more clusters are selected to create new links in G. Finally, nodes in G are sorted by the sum of *column* of touching columns. Terms represented by nodes of higher values of these sums than a certain threshold are extracted as the keywords for document D.

3 Our Approach

By focusing on the analogy between a document and other textual data, KeyGraph can be applied to a variety of topics. For example, KeyGraph has been adoped to

- find areas with the highest risks of near-future earthquakes from data of observed past earthquakes[9],
- get timely files from visualized structure of our working history [10],
- construct planning to guide concept understanding in WWW[11],
- make tools for shifting human context into disasters[12],
- discover potential motivations and fountains of chances [13].

In a document D, high-frequency terms are used for expressing typical basic concept, and term-pairs which frequently occur in the same sentences mean strong association throughout D (see Sect. 2).

In this paper, we extend the use of KeyGraph to another kind of data, i.e., Web-page set (corresponding to D, document in Sect. 2) including Web-pages (each corresponding to a sentence in Sect. 2) having URL-links, each corresponding to words in Sect. 2. That is, high-frequency links (which are the URLs pointing to other web pages) in a collection W of web pages show popular web pages, and link-pairs which frequently occur in the same web pages show strong relations in W[4]. Our fundamental hypothesis here is that the occurrence of a document and a collection of web pages have common causal structures, and our strategy for applying KeyGraph is based on this analogy.

Let us be more formal. A web page(which URL is u) is translated to a sentence as

$$u \; u_1 \; u_2 \; u_3 \; \cdots \; u_i \cdots u_n. \tag{4}$$

Where $u_i (i = 1, 2, 3, \ldots, n)$ are the URLs contained in the web page. A document is formed by combining sentence, shown in eq. (4), for each web page of a collection. By this translation, we can obtain the document reflecting the link structure of WWW.

4 An Example of Experiment

In this section, we report on our experiment where we applied KeyGraph to two sets of collections C_A and C_B, each of which contains 500 popular web pages obtained by Google for the input query 'human genome', to follow the changes of the communities with time. The difference between the collections is the date: C_A is obtained on November 26, 2000, and C_B is on March 11, 2001.

After C_A and C_B were translated into two documents like Sect. 3, for each document KeyGraph output URLs as *roof*(asserted) keywords. The URLs for C_A and C_B are shown in Table 1. Comparing the output URLs for C_A and C_B in Table 1, we can recognize the movement among them. For example,

```
http://www.ncbi.nlm.nih.gov
http://www.nhgri.nih.gov
http://www.sanger.ac.ul
```

Table 1: Output URL lists of KeyGraph for collections C_A and C_B.

Output URLs for C_A	*Affiliation*
`www.ncbi.nlm.nih.gov`	National Center for Biotechnology Information
`gdbwww.gdb.org`	The Genome Database
`www.ornl.gov`	Oak Ridge National Laboratory
`www.nhgri.nih.gov`	The National Human Genome Research Institute
`www.gene.ucl.ac.uk`	The Galton Laboratory
`www.ebi.ac.uk`	European Bioinformatics Institute
`www.gdb.org`	The Genome Database
`lpg.nci.nih.gov`	CGAP Genetic Annotation Initiative
`www.sanger.ac.uk`	The Sanger Centre
`www.genetics.utah.edu`	Human Genetics Department in University of Utah

Output URLs for C_B	*Affiliation*
`www.ncbi.nlm.nih.gov`	National Center for Biotechnology Information
`www.nhgri.nih.gov`	The National Human Genome Research Institute
`www.ornl.gov`	Oak Ridge National Laboratory
`www.cnn.com`	CNN.com
`genome.wustl.edu`	Genome Sequence Center in Washington University
`onhealth.webmd.com`	OnHealth Network Company
`www.gdb.org`	The Genome Database
`www.sanger.ac.uk`	The Sanger Centre
`www.tigr.org`	The Institute for Genomic Research
`www.celera.com`	Celera.com

appear in both the outputs. These are the most authorized research institutes in the area of human genomics. On the other hand,

`http://www.tigr.org`
`http://www.celera.com`

appear only in URLs for C_B. These are newly growing research organizations in the area. This movement reflects events/situations of the real world in the topic of human genome, and means that KeyGraph could detect the major changes in the society. However, we cannot see why these changes occurred, from these tables.

In the field of human genome, revolutionary events were occured in 2000 and 2001. Celera Genomics and the U.S. Human Genome Project has been entered into a keen competition for leadership. Both two rivals independently accomplished the analysis of most of the human genome at almost the same time, and their papers were appeared in *Science*[15] and *Nature*[14] respectively in February 2001.

Considering these real events/situations, the emerging URLs(`http://www.tigr.org`, `http://www.celera.com`) are considered to reflect the real society.

5 Conclusions

In this paper, we introduced a method for aiding human awareness on significant novel, i.e., emerging topics. Here, the algorithm of KeyGraph is extended to be a method for the analysis and visualization of cocitations between Web pages. Communities, each having members (Web pages, their authors/readers) with common interests are obtained as graph-based clusters, and an emerging topic is detected as a Web page relevant to multiple communities. Experiments, an example of which is presented in this paper, show that the aimed effect of our method is realized.

References

[1] Malcolm Gladwell: THE TIPPING POINT: How Little Things Can Make a Big Difference. Little Brown & Company, 2000.

[2] Andrei. Z. Broder, Steven. C. Glassman, and Mark. S. Manasse: Syntactic Clustering of the Web. Proceedings of the 6th World Wide Web Conference, 1997.

[3] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins: Trawling the web for emerging cyber-communities. Proceedings of the 8th World Wide Web Conference, 1999.

[4] H. D. White and K. W. McCain: Bibliometrics. Annual Review of Information Science and Technology, volume 24, pages 119 – 186, Elsevier, 1989.

[5] G. Salton and M. J. McGill: Introduction to Modern Information Retrieval, McGraw-Hill, 1983.

[6] M. F. Porter: An Algorithm for Suffix Stripping, Automated Library and Informations Systems, Vol. 14, No. 3, pp. 130 – 137, 1980.

[7] J. Cohen: Highlights: Language- and Document- Automatic Indexing Terms for Abstracting, Journal of Amerimcan Society for Information Science, 46, pp. 162 – 174, 1995.

[8] Yukio Ohsawa, Nels E. Benson and Masahiko Yachida: KeyGraph: Automatic Indexing by Co-occurrence Graph Based on Building Construction Metaphor. Proceedings of the Advances in Digital Libraries Conference, pages 12 – 18, 1998.

[9] Yukio Ohsawa, Masahiko Yachida: Discover Risky Active Faults by Indexingan Earthquake Sequence. Proceedings of the International Conference on Discovery Science, 1999.

[10] Yukio Ohsawa: Get Timely Files from Visualized Structure of Your Working History, Proceedings of the 3rd International Conference on Knowledge-Based Intelligent Engineering Systems & Allied Technologies, 1999.

[11] Seiji Yamada, Yukio Osawa: Navigation Planning to Guide Concept Understanding in the World Wide Web, Proceedings of Autonomous Agents, 2000.

[12] Yumiko Nara and Yukio Ohsawa: Tools for Shifting Human Context into Disasters, Chance Discovery and Management session, Proceedings of the 4th International Conference on Knowledge-Based Intelligent Engineering Systems & Allied Technologies, 2000.

[13] Yukio Ohsawa, Hisashi Fukuda: Potential Motivations and Fountains of Chances, Chance Discovery from Data session, Proc. International Conference on Industrial Electronics, Control and Instrumentation, 2000.

[14] International Human Genome Sequencing Consortium: Initial sequencing and analysis of the human genome, Nature 409, pp. 860 – 921, 2001

[15] J. Craig Venter, et al.: The Sequence of the Human Genome, Science 291: pp. 1304-1351, 2001.