Discovery of Emerging Topics between Communities on WWW

Naohiro Matsumura[1,3], Yukio Ohsawa[2,3], and Mitsuru Ishizuka[1]

[1] Graduate School of Engineering, University of Tokyo,
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113–8656 Japan
{matumura, ishizuka}@miv.t.u-tokyo.ac.jp
[2] Graduate School of Systems Management, University of Tsukuba,
3-29-1 Otsuka, Bunkyo-ku, Tokyo, 112–0012 Japan
osawa@gssm.otsuka.tsukuba.ac.jp
[3] TOREST, Japan Science and Technology Corporation,
2-2-11 Tsutsujigaoka, Miyagino-ku, Sendai, Miyagi, 983–0852 Japan

Abstract. In the real world, 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. For example, how does a novel written by an unknown author become a bestseller? Why did the crime-rate drop so dramatically in New York City? Malcolm Gladwell answers to these questions as follows [1]:

> ... ideas and behavior and message and products sometimes behave just like outbreaks of infectious disease. They are social epidemics. The Tipping Point is an examination of the social epidemics that surround us.

The infectious disease usually spreads through the virus. Whereas, we cannot detect the social epidemics(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. Here we introduce some recent examples of new significant topics:

Mobile Phone: Considering the appearance of mobile phones, there were essentially two factors. First, mobile phones conquered the inconvenience of beepers that people had to find a public phone when a beeper rang. Second, mobile phones were equipped with the functions of the Internet and E-mail services. Due to the synergy effects of these factors satisfying our needs, mobile phones began to get popular.

Global Warming: The awareness of global warming realized the collaboration of automobile and ecological preservation communities, and consequently brought about hybrid automobiles which have minimal exhaust emissions for preserving the earth ecology.

Human Genome Project: Many researchers in the field of artificial intelligence, biology, and medical science are collaborating on the human genome project to analyze the human genome and to reveal its effects. As we expect the conquest of fatal illnesses, the human genome project is in the limelight.

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, considering the synergy effects mentioned above. 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 [13] to inspect that WWW reflects the real world, and that the revealed structure of WWW supports our detection of new significant topics.

The rest of this paper is organized as follows. In Section 2, we describe our approach for understanding the real world through WWW, and an experiment is shown in Section 3. The results are discussed on in Section 4, and finally we conclude this paper in Section 5.

2 Understanding Human Society on WWW structure

We try to understand the movement of the human society through the structure of WWW which is composed of communities and their relations. In this section, we first introduce previous studies for the discovery of communities, and the discovery of their relations. Then, we introduce KeyGraph algorithm [13] and our approach for applying KeyGraph to WWW.

2.1 The Discovery of Communities

Broder et al.[2] reported on an algorithm of clustering web pages based on the similarities of contents. The merit of this approach is to be able to apply not only

to hyper-text(e.g., web pages) but also plain-text. However, indexing web pages accurately is difficult because the contents of web pages are not always concentrated on certain themes. In contrast to the content-based approach, links in web pages can be reliable information because they reflect human judgement[5]. Kumar et al.[3] defined a community on the web as a dense directed bipartite subgraph, and actually discovered over 100,000 communities. His idea was innovative because he was the first who formulated a community mathematically in our knowledge. The bipartite graph, however, comes to include pages of different interests, if it is expanded to a wide area at the Web. As another use of links, Kleinberg [4] and Brin and Page [5] used the link structures for ranking web pages. Their main idea was based on mutual reinforcing, i.e., the more a web page is referred, the more authoritative the web page becomes, and the more authoritative a web page becomes, the higher the web page ranks. The highly ranked web pages tend to be the representative web pages of communities. This method is useful for finding reliable pages, but is not suitable for our aim because we prefer premature significant pages to authorized ones. Compared with these methods, we aim at communities each having a shared interest.

2.2 The Discovery of Relations

Matsumura et al.[7] tried to find new combinations of different communities sharing common topics to discover promising new topics on the web. His idea was based on the co-citation concept originated in the bibliometrics [6]. However, the community was different from our aim in this paper in the point that he regarded each of the web pages obtained by Google [1] as a community. On the other hand, Kautz et al.[8] made REFERRAL WEB, a social network graph designed to find an expert who is both reliable and likely to respond to the user. Also, Leonard [9] described a matchmaker system named Yenta for finding people with similar interests and introduce them to each other. Both systems reveal the potential relations between individuals. Our aim is also to discover potential and interesting relations among the communities on WWW.

From Subsec. 2.1 and 2.2, we need a new method for discovering such latent relations among communities, each having an interest shared by Web pages in the community. For this purpose, the next section introduce KeyGraph.

2.3 KeyGraph Algorithm

KeyGraph [13] 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 as follows:

> This building has *foundations* (statements for preparing basic concepts), walls, doors and windows(ornamentation). But, after all, the *roofs*(main

[1] Google is a search engine to which Brin and Page's algorithm [5] is applied. Google is available at `http://www.google.com/`.

ideas in the document), without which the building's inhabitants cannot be protected against rains or sunshine, are the most important. These roofs are supported by *columns*. Simply put, KeyGraph finds the roofs.

The processes of KeyGraph are composed of four phases.

0)Document preparation: Prior to processing a document D, *stop words* [10] which have little meaning are discarded from D, words in D are stemmed [11], and phrases in D are identified [12]. 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 are the set of terms above the 30th highest frequency(we denote this set by HF).
- **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 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.

2.4 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 adopted to

- find areas with the highest risks of near-future earthquakes from data of observed past earthquakes [14],
- get timely files from visualized structure of our working history [15],
- construct planning to guide concept understanding in WWW [16],
- make tools for shifting human context into disasters [17],
- discover potential motivations and fountains of chances [18].

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 Subsect. 2.3).

In this paper, we extend the use of KeyGraph to another kind of data, i.e., Web-page set (corresponding to D, document in Subsec. 2.3) including Web-pages (each corresponding to a sentence in Subsec. 2.3) having URL-links, each corresponding to words in Subsec. 2.3. 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 [6]. 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.

In order to understand the real world through the document, we have to piece out the situation between asserted keywords(*roofs*) and the basic concepts(*foundations*). In the metaphor of KeyGraph, the context structure expressed by links(*columns*) connecting assertions with basic concepts. We expect that a graphical output of KeyGraph helps us in understanding potential interests and the underlying relation between them, and loads us to the understanding of the structure of the interests of people in the real human society.

3 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 as described in Subsect. 2.4, for each document KeyGraph output URLs as *roof*(asserted) keywords. The URLs for C_A and C_B are shown in Table 1 and Table 2, and the graphical outputs are in Fig. 1 and in Fig. 2 respectively. Comparing Table 1 with Table 2, 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`

appear in both the Tables. 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 Table 2. 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.

Next, let us pay attention to the Fig. 1 and Fig. 2 to piece out the movement. In the figures, the single-circle and double-circle nodes show *fundation* and *roof* pages respectively, and links among nodes show *columns*. Comparing both the figures, we can imagine two situations as follows.

- `http://www.ncbi.nlm.nih.gov`,`http://www.nhgri.nih.gov`,`http://www.tigr.org`, `http://www.sanger.ac.ul`, etc. are densely connected to each other. That is, these web pages are considered to be well established web pages in the topic of human genome.
- The situation around Celera Genomics(`http://www.celera.com`) changes dramatically from Fig. 1 to Fig. 2. Therefore, it can be assumed that something big event might have occured between November 26, 2000 and March 11, 2001. From this, we can clearly understand how much(and whose) acceptance Celera won from various established communities.

These are natural interpretations of the URLs and figures, with imagination based on common sense. In the next section, we discuss these interpretations by looking back the real events/situations.

4 Discussions

In the field of human genome, revolutionary events were occured in 2000 and 2001. In the White House on June 26, 2000, J. Craig Venter, president and

Table 1. An output list of KeyGraph for a collection of web pages on 'human genome'(Nov. 26, 2000).

URL	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.amazon.com	Amazom.com
www.gene.ucl.ac.uk	The Galton Laboratory
www.ebi.ac.uk	European Bioinformatics Institute
ad.doubleclick.net	DoubleClick Inc.
home.about.com	About.com
www.omega23.com	Omega23.com
www.fool.de	The Motley Fool
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

Table 2. An output list of KeyGraph for a collection of web pages on 'human genome'(Mar. 11, 2001).

URL	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
ad.doubleclick.net	DoubleClick Inc.
www.thestreet.com	TheStreet.com
r.wired.com	Not Found
www.amazon.com	Amazom.com
home.about.com	About.com
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

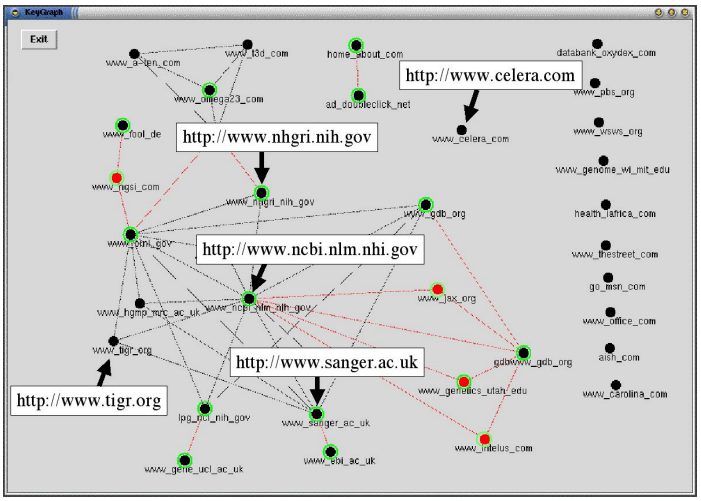

Fig. 1. A graphical output of KeyGraph for the input query 'human genome'(November 26, 2000). We can recognize a big cluster, which are composed of `http://www.ncbi.nlm.nih.gov`, `http://www.nhgri.nih.gov`, `http://www.tigr.org`, `http://sanger.ac.ul, etc`. Note that `http://www.celera.com` is isolated from the big cluster, although the company had a worldwide fame.

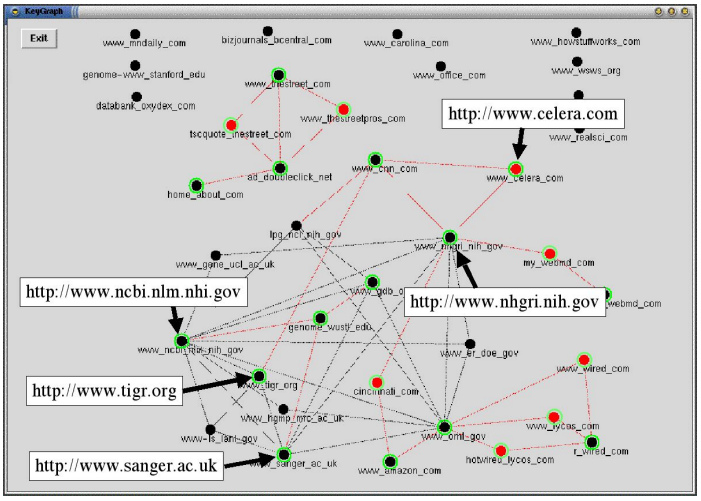

Fig. 2. A graphical output of KeyGraph for the same query(March 11, 2001). The major web pages of the clusters is almost the same as above cluster in Fig. 1. However, `http://www.celera.com` began to be supported by the clusters.

chief scientific officer of Celera Genomics corporation(`http://www.celera.com`), Francis S. Collins, Director of the National Human Genome Research Institute (sponsored by the U.S. National Institutes of Health) (`http://nhgri.nih.gov`) were celebrated by U.S. President Bill Clinton and British Prime Minister Tony Blair for the achievements of the human genome analysis. Celera Genomics and the U.S. Human Genome Project has been entered into a keen competition for leadership. Celera Genomics is an ambitious venture corporation, which began to sequence the human genome on September 8, 1999 by using the whole genome shotgun technique. On the other hand, the U.S. Human Genome Project is an international scientific effort to map and sequence the 3 billion genetic codes, involving more than 1000 scientists from five countries (China, France, Japan, the U.K., and the U.S.A.). 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* [20] and *Nature* [19] respectively in February 2001. With the complete sequencing of the human genome, pharmaceutical companies will create new medicine for patients of all ages. Since Celera Genomics goes release the sequences of the human genome, Fig. 2 might show the sign of the genesis of post-genome era.

Considering these real events/situations, the changes of the structures shown by Fig. 1 and Fig. 2 (e.g., Celera Genomics grew to be widely supported) 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.

The co-occcurrence of links in WWW often suffers from problems specific to WWW[21]. In the future work, we plan to improve KeyGraph algorithm to fit the link structure of WWW.

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. Jon M. Kleinberg: Authoritative Sources in a Hyperlinked Environment. Journal of the ACM, Vol. 46, No. 5, pp. 604–632, 1999.
5. Sergey Brin and Lawrence Page: The Anatomy of a Large-Scale Hypertextual Web Search Engine. Proceedings of 7th World Wide Web Conference, 1998.
6. H. D. White and K. W. McCain: Bibliometrics. Annual Review of Information Science and Technology, Vol. 24, pp. 119–186, Elsevier, 1989.
7. Naohiro Matsumura, Yukio Ohsawa, and Mitsuru Ishizuka: Discovering Promising New Topics on the Web. Proceedings of the 4th International Conference on Knowledge-Based Intelligent Engineering Systems & Allied Technologies, pp. 804–807, 2000.
8. Henry Kautz, Bart Selman, and Mehul Shah: The Hidden Web. AI magazine, Vol. 18, No. 2, pp. 27–36, 1997.
9. Leonard N. Foner: Yenta: A Multi-Agent, Referral-Based Matchmaking System. Proceedings of the 1st International Conference on Autonomous Agents, pp. 301–307, 1997.
10. G. Salton and M. J. McGill: Introduction to Modern Information Retrieval, McGraw-Hill, 1983.
11. M. F. Porter: An Algorithm for Suffix Stripping, Automated Library and Informations Systems, Vol. 14, No. 3, pp. 130–137, 1980.
12. J. Cohen: Highlights: Language- and Document- Automatic Indexing Terms for Abstracting, Journal of Amerimcan Society for Information Science, Vol. 46, pp. 162–174, 1995.
13. 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, pp. 12–18, 1998.
14. Yukio Ohsawa, Masahiko Yachida: Discover Risky Active Faults by Indexingan Earthquake Sequence. Proceedings of the International Conference on Discovery Science, pp. 208–219, 1999.
15. 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.
16. Seiji Yamada, Yukio Osawa: Navigation Planning to Guide Concept Understanding in the World Wide Web, Proceedings of Autonomous Agents, pp. 114–115, 2000.
17. 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.
18. 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.
19. International Human Genome Sequencing Consortium: Initial sequencing and analysis of the human genome, Nature 409, pp. 860–921, 2001
20. J. Craig Venter, et al.: The Sequence of the Human Genome, Science 291: pp. 1304–1351, 2001.
21. Krishna Bharat, Monika R. Henzinger: Improved Algorithms for Topic Distillation in a Hyperlinked Environment, Proceedings of the 21st ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 104–111, 1998.