Discovering Emerging Topics from WWW

Naohiro Matsumura*1,3 Yutaka Matsuo*1,3
Yukio Ohsawa*1,2 Mitsuru Ishizuka*3

*1 PRESTO, Japan Science and Technology Corporation,
2-2-11 Tsutsujigaoka, Miyagino-ku, Sendai, Miyagi, 983-0852 Japan

*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 Graduate School of Engineering, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656 Japan
{matumura, matsuo, ishizuka}@miv.t.u-tokyo.ac.jp

Discovering emerging topics from WWW has been attracting attention of business professionals, especially marketing researchers. For this purpose, WWW can be a valuable source of information because it reflects the dynamics of human society. In this paper we aim at revealing the structure of WWW by using *KeyGraph,* a visualization method of hidden structure behind data, for understanding emerging topics.

Introduction

We experience that new topics suddenly become popular. Such a topic, which might seem insignificant at first, can turns out to match our potential needs. *The Tipping Point* (Gladwell, 2000) 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 calls these phenomena *social epidemics*, i.e., new topics sometimes behave just like outbreaks of infectious disease (Gladwell, 2000). However, we cannot detect the social epidemics (new topics) and their mechanisms in advance because the real world surrounding us is too complex to decode. Detecting a *Tipping Point*, in face of this obstacle, could be a big chance for one's activity, of which competitors are not aware. We interpret 'topics' in the broad sense that cover ideas, behavior, messages, products and so on. Let us introduce some recent examples of new significant topics:

The Mobile Phone: For the appearance of mobile phones, essentially two factors were present. First, mobile phones conquered the inconvenience of pagers whose user had to find a public phone when a pager rang. Second, mobile phones came to be equipped with the functions of the Internet and E-mail services. Due to the synergy effects of these factors satisfying users' needs, mobile phones began to get popular.

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

As we can easily realize from above descriptions, these topics are born when new collaborations of existing interests satisfy our potential needs or demands. Although the hidden factors might be 'submerged' in the human mind, we believe that a few signs can be mined from a database on human behavior reflecting the human mind. For this purpose, the web is an attractive source of information because of its size and sensitivity to trends. The web consists of an abundance of communities (Kumar, 1999), each corresponding to a cluster of web pages sharing common interests. Since a community means a chunk of shared interest, a web page supported (or linked) by multiple communities is considered to satisfy their interests, and shows the movement direction of the wider human world, considering the synergy effects mentioned above. From this point of view, we are expecting the structure of WWW to be a key to understand the real world. In this paper, we aim at revealing the structure of WWW by using the *KeyGraph* algorithm (Ohsawa, 1999a), and then inspect the revealed structure of WWW supports our detection of new significant topics.

The rest of this paper is organized as follows. We first overview the structure of web communities from related research, and clarify the difference between two types of relations, i.e., *direct relation* and *co-citation*. Then, we describe our approach employing *KeyGraph* for understanding the real world, and report experiment evaluations.

The Structure of Web Community

WWW is a good source of information to detect the movement of human society, because it reflects the movement of the real world very quickly. On top of that, WWW is a part of the human social network (Adamic, 2001). The creation of a hyperlink by the author of a web page is an implicit type of 'endorsement' of the page being pointed to. By mining social interests contained in the set of such endorsements, we can obtain a better understanding of the movement of human society. In the following, we overview related researches on web communities by the link structure.

The Discovery of Web Communities

The web harbors a large number of communities -- groups of content creators -- each sharing a common interest that manifests itself as a set of web pages. Although some communities have explicitly defined common interests (newsgroup, resource collections in portals, etc.), other are implicit (Kumar, 1999). Kumar et al. defined a community on the web as a dense *directed bipartite sub-graph*, one whose nodes can be partitioned into two sets A and B such that every link in the sub-graph is directed from a node in A to a node in B. They actually discovered over 100,000 communities from the entire web (Kumar, 1999).

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 (Kleinberg, 1999) and Brin and Page (Brin and Page, 1998) used link structures for ranking web pages. Their main idea was based on mutual reinforcing, i.e., the more a web page is referred to, the more authoritative the web page becomes. The more authoritative a web page becomes, the higher the web page ranks. Thus, highly ranked web pages tend to be the representative web pages of communities.

The Discovery of Related Web Pages

Chakrabarti et al. have suggested using co-citation and other forms of connectivity to identify related web pages (Chakrabarti, 1998). Simply put, if page *A* points to both pages *B* and *C*, then *B* and *C* might be related. Terveen et al. used the connectivity structure of a web page to find related web pages (Terveen, 1999). Dean et al. also found related pages only by the connectivity information where the input to the search process is the URL of a page (Dean, 1999). Netscape browser equipped a 'What's Related?' button that lists related pages to help us understand where to go next when we are surfing the web or drilling for information (Netscape Communication Company). Ohsawa et al. tried to discover web pages that absorb attentions of people from multiple communities (Ohsawa, 2001). Topics in such pages can be triggers for personal or social progress of interests, beyond the bounds of existing communities. Kautz et al. made REFERRAL WEB, a social network graph designed to find an expert both reliable and likely to respond to the user (Kautz, 1997). Finally, Matsuo et al. defined a intuitive distance of web pages based on the link structure (Matsuo, 2001).

Direct Relation and Co-citation

The web can be viewed as a graph, where nodes represent web pages and links represent the relation between web pages. Two major relations of web pages are at hand:

- Direct relation: a node represents a web page and a link represents a hyperlink between two web pages;

- Co-citation: a node represents a web page and a link represents the relation of co-citation between web pages.

We consider a community as a set of web pages aiming at similar interests. Web pages in the same community do not frequently refer to one another. They may, for one reason, be in a competitive relation. In an extreme case, they are not aware of each others' presence because they keep secrets from each other. For example, a laboratory in the University of Tokyo (http://www.miv.t.u-tokyo.ac.jp), JST (http://www.jst.go.jp/EN/) and the University of Tsukuba (http://www.tsukuba.ac.jp) do not link one another (although these are our affiliations), however our homepages have some hyperlinks to theirs. In this sense, they are co-cited and have a relation. In the following, we clarify the difference between direct relation and co-citation.

The overview of a given area is obtained as follows: We first use query terms to collect a set of pages from the Google search engine [1]. We get a list of authoritative web pages related to a given query. Then, we download the content of each web page and extract hyperlinks. Finally, we make a graph using one of the following strategies:

Algorithm 1: Obtain nodes representing the top authoritative web pages and links representing the direct relation between them.

Algorithm 2: Obtain nodes representing the top authoritative web pages and links representing the relation of co-citation.

Algorithm 3: Obtain nodes representing the most frequently cited web pages and links representing the relation of co-citation.

In Algorithm 1, a link has a direction in a natural sense; if page *A* points to page *B*, we make a link from page *A* to *B*. In Algorithm 2 and 3, assuming page

[1] Google is a search engine to which Brin and Page's algorithm (Brin and Page, 1998) is applied. Google is available at http://www.google.com/.

$C \in C_{A,B}$ has hyperlinks to both page A and B, we calculate the strength of co-citation of A and B as:

$$rel(A,B) = \sum_{C \in C_{A,B}} \frac{1}{OutDegree(C)^2},$$

where $C_{A,B}$ represents a set of authoritative pages which points to both A and B, and $OutDegree(C)$ represents the number of hyperlinks in page C. This index is based on the random surfer model (Brin and Page, 1998), where a random surfer keeps clicking on successive links selected at random, and the probabilistic retrieval model (Rolleke, 1995). A link by co-citation has no direction.

An example for a query 'abortion -- pro life' for Algorithm 1, 2 and 3 is respectively shown in Figure 1, 2 and 3. For each figure, the number of nodes is 32 and the number of links is 31. In Figure 1, we can see well-linked web pages in the center of the figure. These web pages include organizations (.org domain site) such as 'National Right to Life Committee,' 'Priests For Life' and 'Republican National Coalition for Life.' On the other hand, in Figure 2, we see a less centralized structure. The web pages in the center are still organizations, but the left-hand-side web pages are companies' web pages (.com domain site) and right-hand-side web pages are AOL and Amazon pages.

In Figure 3, we see more clearly the clusters of organizations, companies, and free web sites such as AOL and Geocities. In this case, although the nodes are not necessarily authoritative web pages, we can find some interesting web pages. For example, one web page in the middle of the figure, http://www.afterabortion.org, provides an important source of information on the aftereffects of the abortion, but this site is currently ranked very low (below 300) on the list by Google. This web page is well cited and often co-occurs with http://www.abortionfacts.com and http://www.prolifeinfo.org.

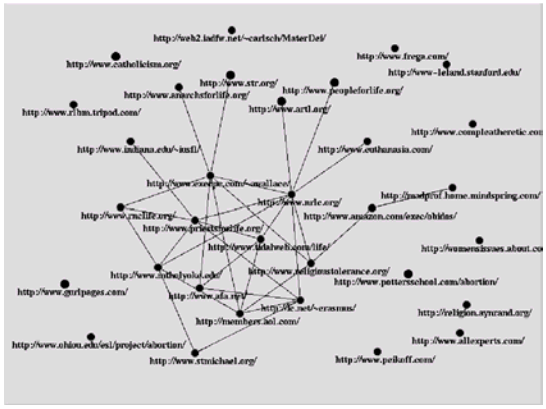


Figure 1: The structure of web pages related to 'abortion -- pro life' by Algorithm 1

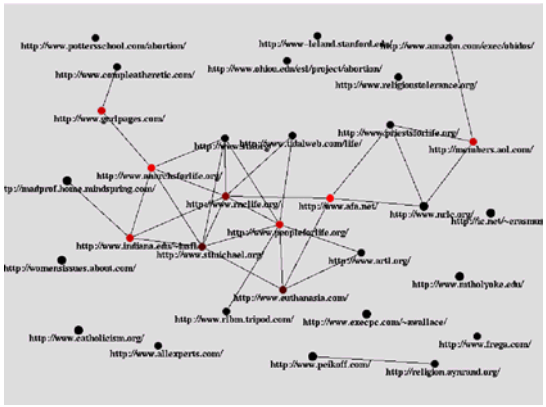


Figure 2: The structure of web pages related to 'abortion -- pro life' by Algorithm 2

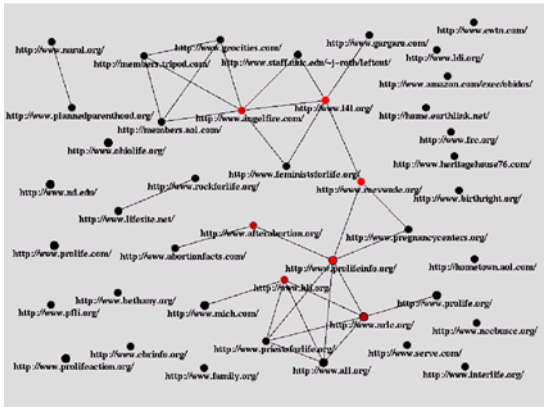


Figure 3: The structure of web pages related to 'abortion -- pro life' by Algorithm 3

We show another example for query term 'web mining'. Web mining is a relatively new topic and the authoritative web pages relevant to this topic are not tightly connected, as shown in Figure 4. However, we can see the relation more clearly for Algorithm 2 and 3 as Figure 5 and Figure 6. In other words, co-citation can detect more subtle relations between web pages than direct relations can. If query terms are even newer and more rare, e.g., 'Ichiro' in major league[2], there is no way of finding the communities by direct links. In fact, if we make a graph by Algorithm 1 for the query, none of two nodes are linked. By using the co-citation information, we can find communities even before participants realize that they have formed their own community.

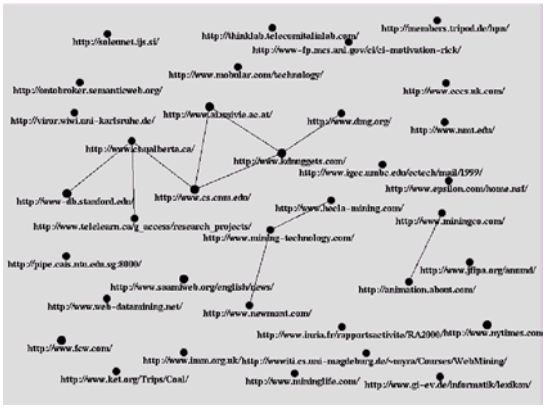
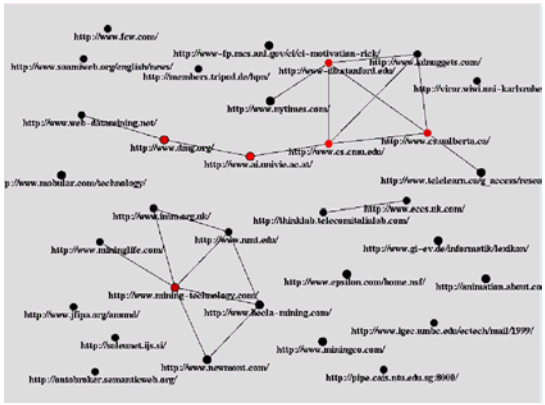


Figure 4: The structure of web pages related to 'web mining' by Algorithm 1



[2] Ichiro is the first Japanese fielder in Major League Baseball and won the leading hitter title and MVP in the season 2001.

Figure 5: The structure of web pages related to 'web mining' by Algorithm 2

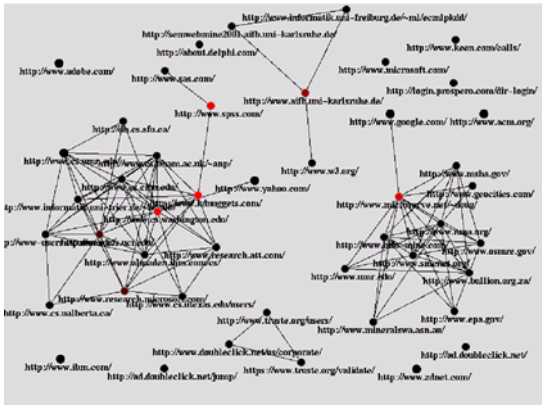


Figure 6: The structure of web pages related to 'web mining' by Algorithm 3

In summary, the link structure of web pages is a good source of information, however, if we look at direct links, we cannot find emerging topics. Instead, we should focus on the co-citation information for detecting emerging topics. In the following, we focus not only on communities but also on web pages implying a big change in the real world.

.

Discovering New Topics on the Web

We aim at understanding the movement of human society through the structure of WWW that is composed of communities and their relations. In this section, we first introduce the *KeyGraph* algorithm (Ohsawa, 1999a), and then describe our approach.

The Algorithm of KeyGraph

KeyGraph (Ohsawa, 1999a), originally an algorithm for extracting terms (words or phrases), expresses assertions based on the 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:

A building has *foundations* (statements for preparing basic concepts), walls, doors and windows (ornamentation). But the *roofs* (main 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 process of *KeyGraph* consists of four phases:

0) Document preparation: Prior to processing a document *D*, *stop words* (Salton, 1983) that have little meaning are discarded from *D*, words in *D* are stemmed (Porter, 1980), and phrases in *D* are identified (Cohen, 1995). 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 their *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 (black nodes in Figure 7). 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* is defined as

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

where $|w|_s$ denotes the count of *w* in sentence *s*. Pairs of high-frequency terms in *HF* are sorted by *assoc* and the pairs above the (*number of nodes in G*) * 1 th tightest association are represented in *G* by links between nodes (solid lines in Figure 7). Then, each cluster -- called a foundation -- is obtained as a set of nodes and links forming a connected graph (gray parts in Figure 7).

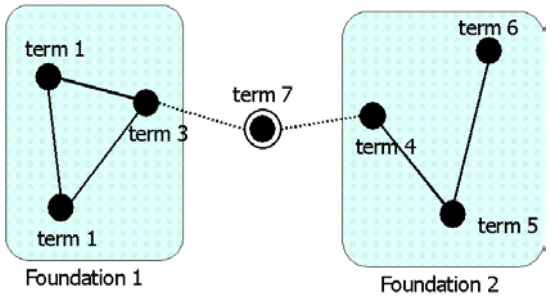


Figure 7: An overview of *KeyGraph*

2) Extracting columns: The probability of term *w* to appear near clusters is defined as *key(w),* and the *key(w)* is defined by

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

where *g* represents each cluster in *G*. Sorting terms in *D* by *key* produces a list of terms ranked by their association with clusters, 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

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

Columns touching w_i are sorted by $column(w_i, w_j)$ for each *high key term*

w_j. Columns with the highest *column* values are selected to create new links in *G*. We depict such links representing columns by dotted lines (see Figure 7). Then, each term w_i is connected by these attached columns to terms in two or more clusters. Finally, nodes in *G* are sorted by the sum of *column* values of its touching columns. Terms represented by nodes of higher values of these sums than a certain threshold are extracted as the keywords for document *D*, as depicted by node of term7 in Figure 7.

Our Approach

By focusing on the analogy between a document and other textual data (data formed by readable letters), *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 (Ohsawa, 1999b),
- get timely files from visualized structure of one's working history (Ohsawa, 1999c),
- computer plan to guide concept understanding in WWW (Yamada, 2000),
- make tools for shifting human context into disasters (Nara, 2000),
- discover potential motivations and fountains of chances (Ohsawa, 2000).

In a document *D*, high-frequency terms are used for expressing typical basic concept, and term-pairs that frequently occur in the same sentences mean strong association throughout *D*. In this paper, we extend the use of *KeyGraph* to another kind of data, i.e. a web page set (corresponding to *D*, a document) including web pages (each corresponding to a sentence) having URL-links, each corresponding to a word. 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, i.e., the co-citation, in *W*. Our fundamental hypothesis here is that the co-occurrence of terms in a document and the co-citation of web pages are common in that both carry the underlying important shared context. Our strategy for applying *KeyGraph* is based on this analogy.

More formally, a web page (which URL is u_0) is translated to a 'sentence' as:

$$u_0 \;\; u_1 \;\; u_2 \;\; u_3 \;\; \ldots \;\; u_i \;\; \ldots \;\; u_n , \qquad (1)$$

Where u_i (i=1, 2, 3, …, n) are the URLs contained in the web page. Combining the (virtual) sentences, shown in eq. (1), for each web page of a collection, forms a document. By this translation, we can obtain a 'document' reflecting the link structure of WWW.

For example, Figure 8 depicts the result of *KeyGraph*, where foundations (solid lines and their touching black nodes), columns (dotted lines), and roofs (double circles) are obtained. Some infrequent nodes as depicted by page8 can be obtained as a roof. In this case, a foundation corresponds to an established web community because each node shows a well-cited web page and each link shows a strong tie among such nodes in a foundation. Our aim is to detect significant emerging topics from web pages (i.e., node of page8 in Figure 8) relevant to multiple established communities based on the assumption that Figure 8 reflects the structure of the real world.

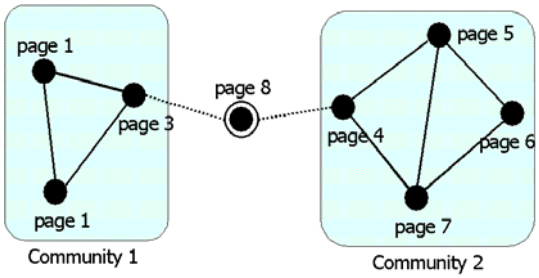


Figure 8: An overview of *KeyGraph* for web page set

We expect that a graphical output of *KeyGraph* helps understand potential interests and the underlying relation between them, and leads us to the understanding of the structure of the interests of people in the real human society. This is a realization of looking at weak-ties between strongly tied communities (Granovetter, 1973).

Experimental Examples and Discussions

We report our experiments where we applied *KeyGraph* to two sets of collections C_A and C_B, each containing 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, for each document *KeyGraph* outputs URLs as *roof* (asserted) keywords. The roofs for C_A and C_B are shown in Table 1 and Table 2, and the graphical outputs are in Figure 9 and in Figure 10 respectively. In the figures, the single-circle and double-circle nodes show *foundation* and *roof* pages respectively, and links among nodes show *columns.*

Comparing Table 1 with Table 2, we can recognize the movement among them. For example, the roofs: NHGRI (http://www.nhgri.nih.gov), NCBI (http://www.ncbi.nlm.nih.gov) and Sanger (http://www.sanger.ac.ul) appear in both tables. NHGRI (National Human Genome Research Institute) is one of 24 institutes, centers, or divisions that make up the National Institute of Health (NIH), the federal government's primary agency for the support of biomedical research. NCBI (National Center for Biotechnology Information) is also one of the institutes of NIH. Sanger (The Sanger Institute) is a research center that provides a major focus in the UK for mapping and sequencing the human genome, and genomes of other organisms. These are the most contributed institutes for the Human Genome Project, 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.). We can also understand this fact from Figure 9 because these institutes are densely connected to each other.

On the other hand, the roofs GSC (http://genome.wustl.edu), TIGR (http://www.tigr.org), Celera (http://www.celera.com) and CNN (http://www.cnn.com) appear only in Table 2. GSC (The Genome Sequencing Center) is a leading contributor to the Human Genome Project and TIGR (The Institute for Genomic Research) is one of the original centers conducting large-scale human genome sequencing. Note that TIGR already appeared in Figure 9 as a node of foundation.

Here, let us focus on Celera (The Celera Genomics), an ambitious venture corporation sequencing the human genome from September 1999. As you can see from Figure 9 and Figure 10, the situation around Celera changes dramatically from November 2000, to March 2001, i.e. Celera began to be supported by the cluster of the Human Genome Project and CNN, among the world's leaders in online news and information delivery. Looking back on the

real events and situations, we can understand the leap of Celera.

In the field of human genome, revolutionary events occurred in 2000 and 2001. The Human Genome Project team and Celera announced the completion of the draft sequence of the human genome in June 2000, and the subsequent articles were published in *Nature* (International, 2001) and *Science* (Venter, 2001) in February 2001. Both are the most important milestones for the human genome analysis. In fact, J. Craig Venter, president and chief scientific officer of Celera and Francis S. Collins, director of the Human Genome Project, were celebrated by U.S. President Bill Clinton and British Prime Minister Tony Blair for the progress of the human genome analysis at the White House on June 26 2000.

Considering these real events and situations, the changes of the structures shown by Figure 9 and Figure 10 (e.g., Celera grew to be widely supported) are considered to reflect the real society.

Conclusions

In this article, we introduced a method that can help understand significant and novel -- i.e. emerging -- topics. Here, the algorithm of *KeyGraph* is extended to be a method for the analysis and visualization of co-citations between web pages. Communities, each having members (web pages, their authors and readers) with common interests are obtained as graph-based clusters, and an emerging topic is detected as a web page relevant to multiple communities, corresponding to weak ties between strongly tied communities. Experiments, an example of which is presented in this paper, show that the aimed effect of our method is realized.

The co-occurrence of links in WWW often suffers from problems specific to WWW (Bharat, 1998). In the future research, we plan to improve *KeyGraph* algorithm to fit the link structure of WWW by considering the contents of web pages.

The emergence of significant sites on WWW as we seek to attract attention to various domains where social evolution forms a key issue, because affairs in the real communities of man are reflected in the topics in the virtual communities on WWW. This is true for the global communities among nations, as well as for local communities among humans. The CyPRG project (La Porte et al., 2001) for example, seeks to understand the factors underlying the diffusion of WWW into national, state and local governments worldwide. In deepening their methods surveying each government in terms of organizational openness and internal effectiveness, it can be important to consider inter-government interactions global and local levels via real and virtual ties. An ultimate application of our methods in this article might be to understand the chances for governments and citizens, i.e. for discussing and deciding how we should deal with essential factors underlying emergent social events.

.

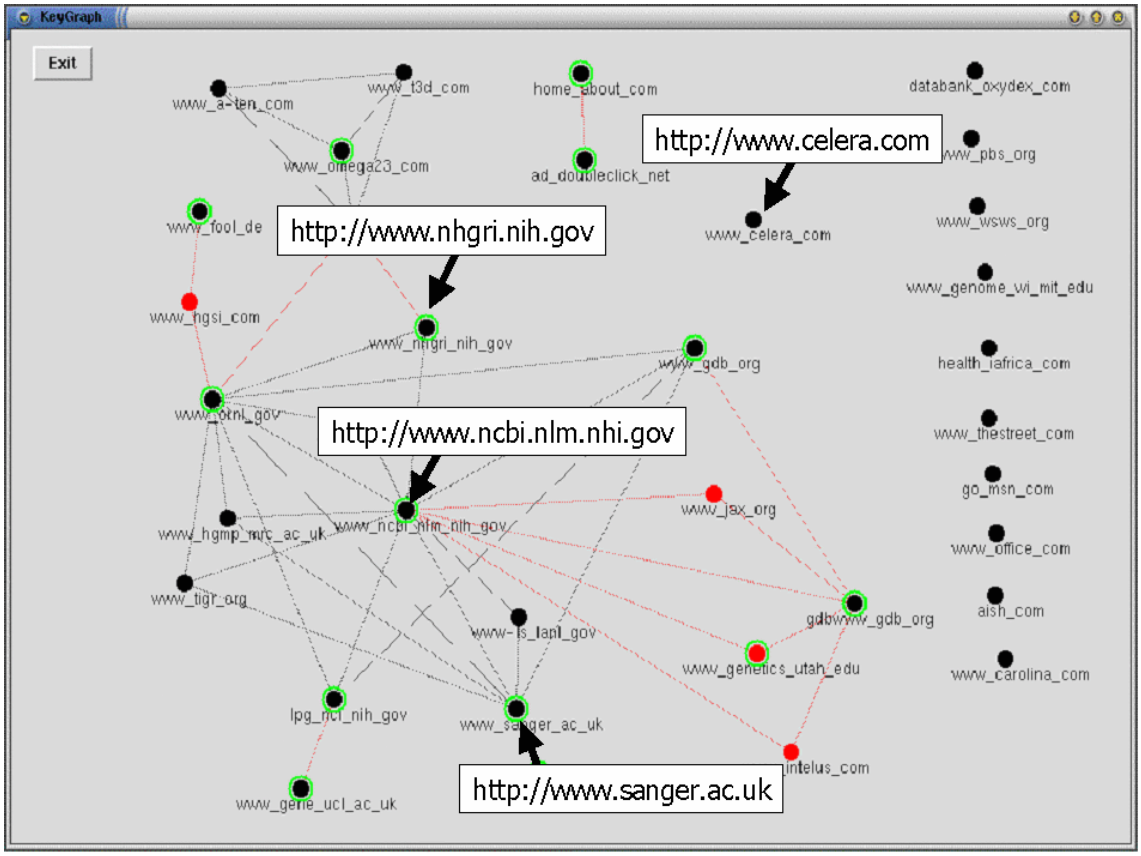


Figure 9: The graphical output of *KeyGraph* for the input query 'human genome' (November 26, 2000). We can recognize a large cluster, which is composed of the institutes that contributed most to the Human Genome Project, such as NCBI (http://www.ncbi.nlm.nih.gov), NHGRI (http://www.nhgri.nih.gov), Sanger (http://sanger.ac.ul), etc. Note that Celera (http://www.celera.com) is isolated from the big cluster

Table 1: The textual output of *KeyGraph* for a collection of web pages on 'human genome' (November 26, 2000)

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

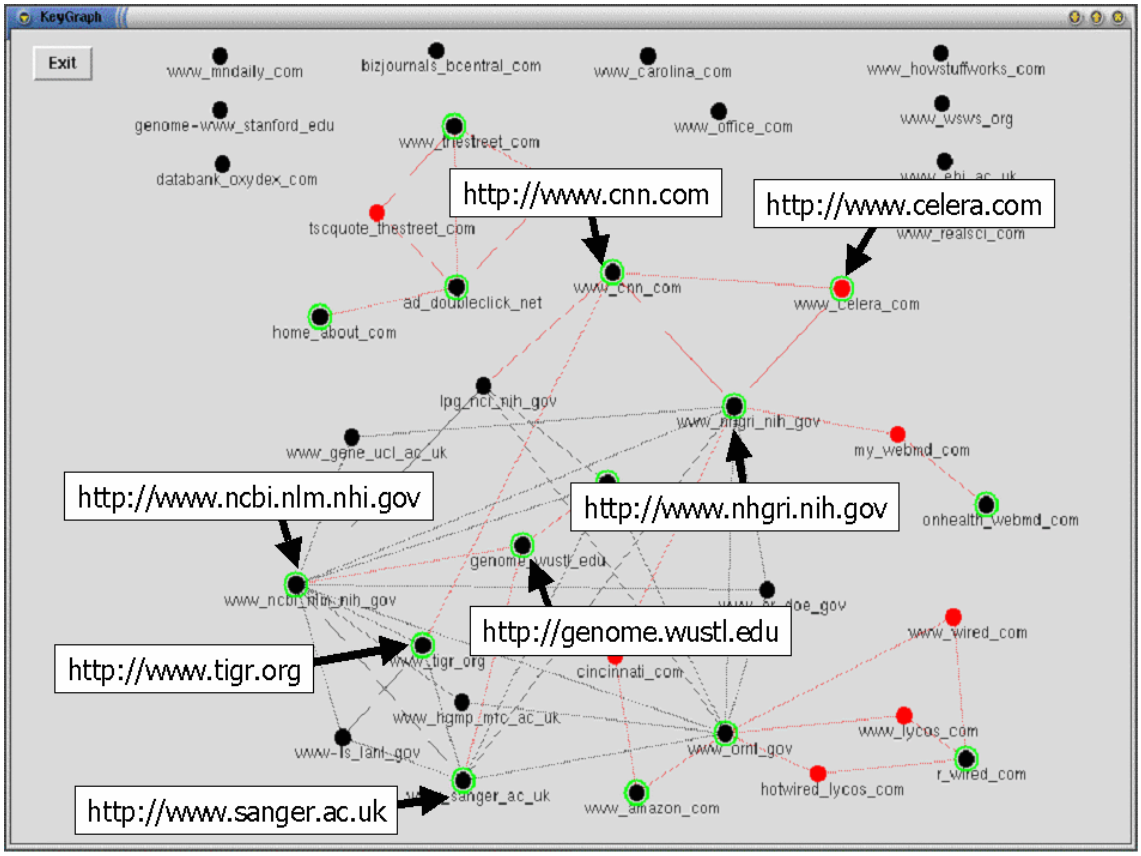


Figure 10: The graphical output of *KeyGraph* for the input query 'human genome' (March 11, 2001). The major web pages of the large cluster in Figure 10 are almost the same as the cluster in Figure 9, i.e., the Human Genome Project cluster. However, Celera (http://www.celera.com) began to be supported by multiple clusters, i.e., the Human Genome Project cluster and CNN (http://www.cnn.com), the mass media cluster

Table 2: The textual output of *KeyGraph* for a collection of web pages on 'human genome' (March 11, 2001)

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

References

Adamic, L.A. and Adar, E. (2001), 'Friends and Neighbors on the Web', to appear, (http://www.hpl.hp.com/shl/people/eytan/fnn.pdf)

Bharat, K. and Henzinger, M.R. (1998), '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.

Brin, S. and Page, L. (1998), 'The Anatomy of a Large-Scale Hypertextual Web Search Engine', *Proceedings of the 7th World Wide Web Conference.*

Chakrabarti, S., Dom, B. and Indyk, P. (1998), 'Enhanced hypertext categorization using hyperlinks', *Proceedings of ACM SIGMOD International Conference on Management of Data*, pp. 307--318.

Cohen, J. (1995), 'Highlights: Language- and Document- Automatic Indexing Terms for Abstracting', *Journal of American Society for Information Science*, Vol. 46, pp. 162--174.

Dean, J. and Henzinger, M. R. (1999), 'Finding Related Pages in the World Wide Web', *Proceedings of the 8th World Wide Web Conference*.

Gladwell, M. (2000), 'THE TIPPING POINT: How Little Things Can Make a Big Difference', *Little Brown & Company.*

Granovetter, M. (1973), 'Strength of Weak Ties', *American Journal of Sociology*, 8, pp. 1360--1380.

International Human Genome Sequencing Consortium (2001), 'Initial sequencing and analysis of the human genome', *Nature*, 409, pp. 860--921.

Kautz, H., Selman, B. and Shah M. (1997), 'The Hidden Web', *AI magazine*, Vol. 18, No. 2, pp. 27--36.

Kleinberg, J.M. (1999), 'Authoritative Sources in a Hyperlinked Environment'. *Journal of the ACM*, Vol. 46, No. 5, pp. 604--632.

Kumar, R., Raghavan, P., Rajagopalan, S. and Tomkins, A. (1999), 'Trawling the web for emerging cyber-communities', *Proceedings of the 8th World Wide Web Conference.*

La Porte, T.M., Demchak, C.C. and Friis, C. (2001), 'Webbing Governance: Global Trends across National Level Public Agencies', *Communications of the ACM*, Vol. 44, No. 1, pp. 63--67.

Matsuo, Y., Ohsawa, Y. and Ishizuka, M. (2001), 'Average-clicks: A new measure of distance on the World Wide Web', *Proceedings of the 1st Web Intelligence*, pp. 106--114.

Nara, Y. Ohsawa, Y. (2000), 'Tools for Shifting Human Context into Disasters', *Proceedings of the 4th Knowledge-Based Intelligent Engineering Systems & Allied Technologies*.

Netscape Communication Corporation, 'What's Related' web page, (http://home.netscape.com/escapes/related/.)

Ohsawa, Y., Benson, N.E. and Yachida, M. (1998), 'KeyGraph: Automatic Indexing by Co-occurrence Graph Based on Building Construction Metaphor', *Proceedings of Advances in Digital Libraries Conference*, pp. 12--18.

Ohsawa, Y. and Yachida, M. (1999), 'Discover Risky Active Faults by Indexing an Earthquake Sequence', *Proceedings of Discovery Science*, pp. 208--219.

Ohsawa, Y. (1999), 'Get Timely Files from Visualized Structure of Your Working

History', *Proceedings of the 3rd Knowledge-Based Intelligent Engineering Systems & Allied Technologies.*

Ohsawa, Y. and Fukuda, H. (2000), 'Potential Motivations and Fountains of Chances', *Proceedings of Industrial Electronics, Control and Instrumentation.*

Ohsawa, Y. Matsumura, N. and Ishizuka M. (2001), 'Discovering Topics to Enhance Communities' Creation from Links to the Future', *Proceedings of the 10th World Wide Web Conference.*

Porter, M. F. (1980), 'An Algorithm for Suffix Stripping', *Automated Library and Information Systems*, Vol. 14, No. 3, pp. 130--137.

Rolleke, T. and Blomer, M. (1997), 'Probabilistic Logical Information Retrieval for Content, Hypertext, and Database Querying', *Hypertext - Information Retrieval - Multimedia* 1997, pp.147--160.

Salton, G. and McGill, M.J. (1983), 'Introduction to Modern Information Retrieval', *McGraw-Hill.*

Terveen, L.G., Hill, W.C. and Amento, B. (1999), 'Constructing, Organizing, and Visualizing Collections of Topically Related Web Resources', *ACM Transactions on Computer-Human Interaction* 6, 1, pp. 67--94.

Venter, J. C., et al. (2001), 'The Sequence of the Human Genome', *Science* 291: pp. 1304--1351.

Yamada, S. and Osawa, Y. (2000), 'Navigation Planning to Guide Concept Understanding in the World Wide Web', *Proceedings of Autonomous Agents*, pp. 114--115.