Future Directions of Communities on the Web

Naohiro Matsumura † **Yukio Ohsawa** ‡ **Mitsuru Ishizuka** †

†Graduate School of Engineering, University of Tokyo
‡Graduate School of Systems Management, University of Tsukuba

Abstract

Discovering new topics which cover new items, problems, and ideas (e.g., mobile phone, global warming, human genome project, etc) is truly profitable, important, and interesting for us. For instance, 1. Companies producing 'mobile phones' have made large profits by the great sales, 2. The awareness of 'global warming' has improved the environment of the earth by regulating exhaust emissions, 3. Fatal illnesses might be conquered by the human genome project. However, since we cannot completely decode the world surrounding us, we cannot know the topics and their mechanisms in advance. Considering this situation, these phenomena could be a big chance for our activities. In this paper, we describe our approach for discovering the future directions of communities on the web to detect chances.

1 Introduction

Often, a new topic suddenly becomes popular although it seems insignificant at first sight. The Tipping Point describes this kind of phenomenon where a 'little' thing can make a big difference[Gladwell, 2000]. We are deeply confused by changes that happen suddenly. However, since we cannot completely decode the world surrounding us, we cannot know the chances and their mechanisms in advance. Considering this situation, the Tipping Point could be a big chance for our activities. We understand 'topics' in the broad sense that cover new items, problems, ideas, and so on. Below, we show you some recent examples of new topics:

Mobile Phone: Considering the context of 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 environmental preservation communities, and consequently brought about hybrid automobiles which have minimal exhaust emissions for preserving the environment of the earth.

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 topics satisfy our potential needs or demands. Although the hidden factors might only be 'submerged' in the human mind, we believe that a few signs can be mined from a database reflecting human's thought. For this purpose, the web is an attractive information source for its sheer size and sensitivity to trends. The web consists of an abundance of communities[Broder *et al.* , 1997; Kumar *et al.* , 1999] which show the smallest unit of potential needs or demands. However, the communities are not independent but are related with each other in varying degrees. In our view, the relations of communities show the future directions of communities, and suggest the potential needs or demands.

In this paper, we describe our approach for discovering the future directions of communities by exploring the web. We have implemented a prototype system named *ChanceFinder* that visualizes the future directions of communities and ranks promising web pages and links. Empirically, ChanceFinder showed some interesting directions for some topics.

The rest of this paper is organized as follows. In Section 2, we introduce related researches, and in Section 3, the process of ChanceFinder is described. The evaluations are discussed in Section 4, and finally we conclude this paper in Section 5.

2 Related Researches

Our research consists of two parts: The discovery of communities, and the discovery of relations among these communities. In this section, we introduce researches related to these two processes.

2.1 Discovery of Communities

A community on the Web corresponds to a cluster of web pages which share common topics. Broder[Broder *et al.* , 1997] reported on an algorithm of clustering web pages based on the contents. This approach can be applied 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 meaningful.

In contrast to the content-based approach, links in web pages can be reliable information because they reflect human judgement. Botafogo and Shneiderman[Botafogo and Shneiderman, 1991] proposed an idea for abstraction called *aggregate* based on graph theory. Their algorithm removes 'indics'(nodes with high number of out-links) and 'references'(nodes with high number of in-links) iteratively to clear the graph. However, removed nodes often become very important elements to understand the web.

Kumar[Kumar *et al.* , 1999] defined a community on the web as a dense directed bipartite subgraph, and discovered over 100,000 communities. However, the scale of subgraphs depends on its parameters, and the number of communities depends on the scale. This implies the difficulty in detecting communities from the web since the communities are often somewhat related with each other. We think the relations show the future directions of these communities.

As another use of links, Kleinberg[Kleinberg, 1998] and Brin and Page[Brin and Page, 1998] used the link structures for ranking web pages. Their main idea was based on mutually reinforcing that 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.

2.2 Discovery of Future Directions

In the broad sense, future directions refer to some meaningful relations among communities. Matsumura[Matsumura *et al.* , 2000] discovered promising new topics on the web by finding new combinations of communities sharing common topics. The combinations did not show general future directions but partial future directions of communities on the web.

Ohsawa et al.[Ohsawa *et al.*, 1999] proposed KeyGraph, which is an algorithm for extracting assertions based on co-occurrence graph of terms from textual data. KeyGraph visualizes the relations between assertions and foundations to help us understand potential needs or demands. Accordingly, KeyGraph shows the future directions of textual data.

Kautz[Kautz *et al.*, 1997] created REFERRAL WEB, a social network graph designed to find an expert who is both reliable and likely to respond to the user. And also, Leonard[Leonard, 1997] described a matchmaker system named Yenta for finding people with similar interests and introduce them to each other. Because both systems reveal the potential relations between individuals, they show the future directions of individuals.

Maarek[Maarek *et al.*, 1997] embodied WebCutter which outputs a tailored map of the web according to the user-specified interests. The map might suggest the future directions of the user by showing essentially related web pages.

3 Future Directions of Communities

For the discovery of new topics on the web, we aim to discover the future directions of communities and to understand the potential needs or demands. In this section, we first represent our idea, and then describe our approach in detail.

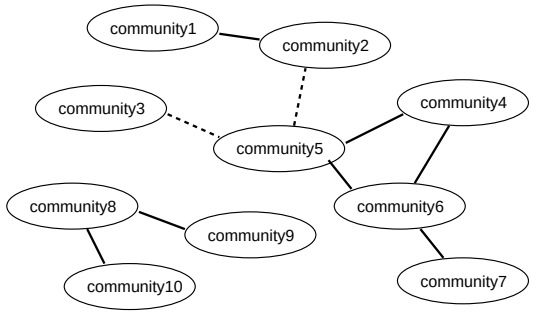

Figure 1: An overview of the web. Communities are often related with each other. Here, solid lines mean established relations and dotted lines show future directions.

3.1 How to Discover the Future Directions?

Our approach for discovering the future directions is based on link analysis because links can be more reliable information than terms (see 2.1). The outline of our process consists of three phases as follows:

Phase1. Detection of communities on the web.

Phase2. Finding the relations among these communities.

Phase3. Discovery of the future directions of these communities.

The accurate definition of a community on the web is an essential problem by itself. In Phase1, following Kumar's definition [Kumar *et al.* , 1999], we expediently define a simple bipartite graph as a community where a community consists of a much cited web page(core) and its surrounding web pages. Next, we focus on the property of the web that communities are often somewhat related with each other as a web page often belongs to many communities. In our view, the relations found may include established(well-known) relations as well as the future directions of these communities. The degree of relation among two communities can be measured by the number of web pages included in both the communities. This idea is based on the co-citation concept originated in the bibliometrics[White and McCain, 1989]. In this way, we regard strong relations as established relations in Phase2, and weak relations as the future directions in Phase3. Our idea is graphically shown in Figure 1. An established link arises only when a future direction grows, we only focus on future links for understanding where the changes happen.

3.2 The Detailed Process

Here, we describe our approach sketched in 3.1 in detail.

Phase1: Discover Communities In preparation, we collect source web pages $\boldsymbol{D}$ by downloading the first 500 web pages of Google's [1] output for the query a user enters.

Then, we count the frequency of links included in D, and regard the top N_1 links $\boldsymbol{C}$ as the 'cores' of communities.

[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/.

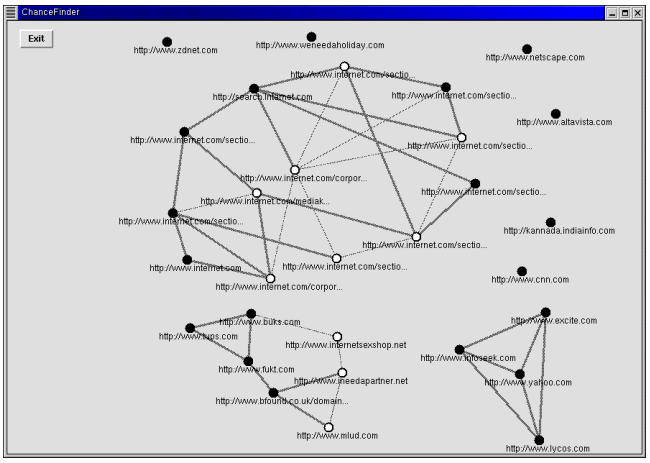

Figure 2: An output of ChanceFinder for input query 'Portal Site'. ChanceFinder shows three clusters.

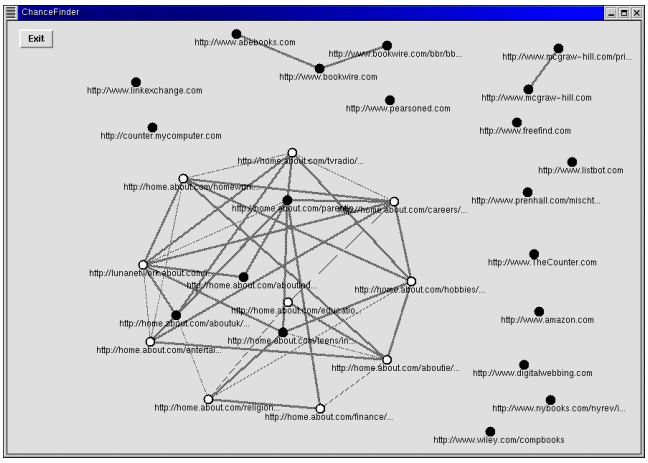

Figure 3: An output for 'Book Site'.

Phase2: Discover Established Relations For every pair of two cores in C, count the number of links included in both the cores, and regard the top (N_2) pairs as established links $\boldsymbol{L_1}$ (solid lines in figure 1).

Phase3: Discover Future Directions For every pair of two cores in C except for L_1, count the number of links included in both the cores, and regard the top N_3 pairs as future links $\boldsymbol{L_2}$ (dotted lines in figure 1).

We visualize the cores and its relations(C, L_1, and L_2) to piece out the connections of communities and to understand the potential needs or demands.

4 Experiments and Discussions

We have implemented a prototype system named $\boldsymbol{ChanceFinder}$ on a Sun Enterprise450 with perl5 and Perl/Tk. ChanceFinder visualizes future directions. In this section, we show three experiments of ChanceFinder with $N_1 = 30$, $N_2 = 29$, and $N_3 = 10$, and discuss them (These experiments were done on 17th of January in 2001).

4.1 Future Directions of Portal Sites

The output of ChanceFinder for input query 'Portal Site' is shown in figure 2. Each node stands for a community, and especially each white node represents a core with many future links. Strong relations of communities are expressed by thick lines(established links), and promising future direction of communities are shown by thin lines(future links). The URL below each node shows the core of each community. Considering the near future, future links might change into established links or disappear. In either event, we should focus on only future links to predict the future. That is to say, the output shows the present and future map of communities.

We can perceive three clusters in figure 2. The lower right-hand cluster is constructed by 4 major portal sites: 'Yahoo!', 'Infoseek', 'Excite', and 'Lycos'. The cluster is considered to be matured since every node links to each other by established links, and this assumption actually matches well accepted norms.

All the communities in the lower left-hand cluster are strongly related to 'Bfound.co.uk' which is a company conducting web design, internet solutions, and e-commerce. This cluster seems to be a community in early development.

The upper middle cluster consists of web pages belonging to 'internet.com' communities. According to the 100hot.com[2] which is the Web's leading ranked directory where the rankings are based on the Internet habits of more than 100,000 Web surfers each month, internet.com got 77th in the same date as the experiment. This means that 'internet.com' is not a major portal site at present. However, we can see that the cluster is in energetic development because the cluster is composed of 13 communities, 17 established links, and 8 future links.

4.2 Future Directions of Book Site

We can recognize a big cluster and two tiny clusters from the output for input query 'Book Site' shown in figure 3.

The upper-middle cluster is composed of two 'bookwire.com' sites and one 'abebooks.com' site. The former is the book industry's most comprehensive and thorough online information source, and the latter is a the world's largest sourse of out-print books. That is, this cluster shows information sources of books.

The upper-right cluster includes two communities of 'mcgraw-hill.com' sites. These sites are the web page of McGraw-Hill company which is a time-honored publisher founded in 1909.

The largest cluster comprises 14 about.com communities. The cluster seems to be already connected densely since it has 25 established links, and 11 future links. In fact, according to the survey on 'Portals leapfrong up Media Metrix chart of the Web's top sites' in December 1999, About.com is described as follows [3]:

> Excite@Home Corp., NBC Internet Inc. and About.com Inc. are on the rise, according to the lat-

[2] http://www.100hot.com

[3] http://www.zdnet.com/zdnn/stories/news/0,4586,2424687,00.html

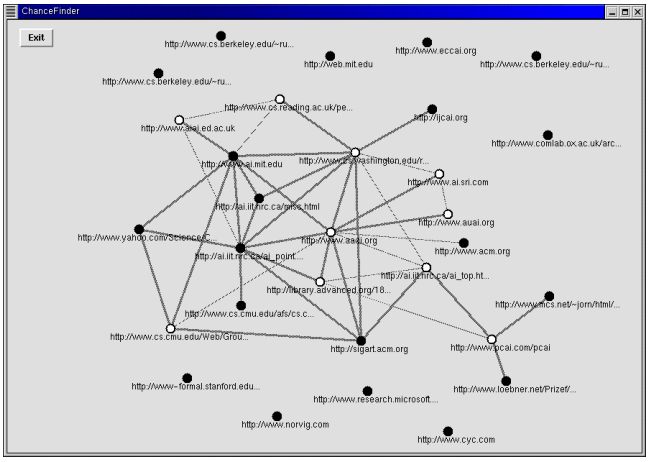

Figure 4: An output for 'Artificial Intelligence'.

est traffic numbers from Internet measurement firm Media Metrix Inc.

However, 'about.com' is not yet a major site in the area of 'Book Site' ('about.com' does not appear in the rankings of 100hot.com). That is, 'about.com' is considered to be struggling to expand the influences, and this consideration can be read from figure 3.

Interestingly, 'amazon.com', the most famous and giant book site exists alone in the middle-right in figure 3. This implies that almost all the communities rival each other, and figure 3 clearly shows this situation.

4.3 Future Directions of Artificial Intelligence

The output for input query 'Artificial Intelligence' is shown in figure 4. Viewing figure 4, we can recognize only one big chunk of communities where 20 communities, 28 established links, and 11 future links are densely connected. Figure 4 is essentially different from above two examples in the point that the cluster in figure 4 consists of different communities. This may show the maturity of the area of 'Artificial Intelligence'. If this assumption is true, we must seek a new area which collaborates with 'Artificial Intelligence' to create future directions.

5 Conclutions

In this paper, we describe the idea of discovering future directions by chaining primitive communities to understand the trend or potential needs or demands.

ChanceFinder certainly shows the relations of communities. However, we expect that whether the relations become the future directions depends on the user's vision or imagination based on accurate information.

References

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

[Broder *et al.* , 1997] Andrei. Z. Broder, Steven. C. Glassman, and Mark. S. Manasse. Syntactic Clustering of the Web. *In Proceedings of the Sixth World Wide Web Conference*, 1997.

[Botafogo and Shneiderman, 1991] Rodrigo A. Botafogo and Ben Shneiderman. Identifying Aggregates in Hypertext Structures. *In Proceedings of the Third ACM Conference on Hypertext*, pages 63-74, 1991.

[Kumar *et al.* , 1999] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins. Trawling the web for emerging cyber-communities. *In Proceedings of the Eighth World Wide Web Conference*, 1999.

[Kleinberg, 1998] Jon M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. *In Proceedings of the Ninth ACM-SIAM Symposium on Discrete Algorithm*, 1998.

[Brin and Page, 1998] Sergey Brin and Lawrence Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *In Proceedings of Seventh World Wide Web Conference*, 1998.

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

[Matsumura *et al.* , 2000] Naohiro Matsumura, Yukio Ohsawa, and Mitsuru Ishizuka. Discovering Promising New Topics on the Web. *In Proceedings of the Fourth International Conference on Knowledge-Based Intelligent Engineering Systems & Allied Technologies*, pages 804 – 807, 2000.

[Kautz *et al.*, 1997] Henry Kautz, Bart Selman, and Mehul Shah. The Hidden Web. *AI magazine*, Vol. 18, No. 2, pages 27 – 36, 1997.

[Leonard, 1997] Leonard N. Foner. Yenta: A Multi-Agent, Referral-Based Matchmaking System. *In Proceedings of the First International Conference on Autonomous Agents*, pages 301 – 307, 1997.

[Ohsawa *et al.*, 1999] Yukio Ohsawa, Nels E. Benson and Masahiko Yachida. KeyGraph: Automatic Indexing by Co-occurrence Graph Based on Building Construction Metaphor. *In Proceedings of the Advances in Digital Libraries Conference*, pages 12 – 18, 1998.

[Maarek *et al.*, 1997] Yoelle S. Maarek, Michal Jacovi, Menachem Shtalhaim, Sigalit Ur, and Dror Zernik. WebCutter: A System for Dynamic and Tailorable Site Mapping. *In Proceedings of the Sixth World Wide Web Conference*, 1997.