Future Directions of Communities on the Web

Naohiro Matsumura[1,3] Yukio Ohsawa[2,3] Mitsuru Ishizuka[4]

[1] 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] PRESTO, Japan Science and Technology Corporation,
2-2-11 Tsutsujigaoka, Miyagino-ku, Sensai, Miyagi, 983-0852, Japan
[4] School of Information Science and Technology, University of Tokyo,
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan

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[1]. 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. The human genome project is getting into the limelight because we expect the conquest of fatal illnesses.

In many cases, 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[2, 4], each corresponding to a cluster of web pages sharing common interest. However, the communities are not independent but are related with each other in varying degrees. From this point of view, we are expecting the relations of communities might 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 link structure of 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 the process of ChanceFinder is described in Section 3. The experiments 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 is defined as a cluster of web pages which share common topics. However, there are many ways to detect the clusters.

For example, Broder et al.[2] reported on an algorithm of clustering web pages based on the contents. This approach can be applied not only to hypertext(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 judgment. Botafogo and Shneiderman[3] 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. On the other hand, Kumar et al.[4] 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. 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[5] and Brin and Page[6] 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. The highly ranked web pages tend to be the representative web pages of communities.

2.2 Discovery of Future Directions

In the broad sense, future directions refer to meaningful relations among communities in various scenes. Focusing on WWW, Matsumura et al.[8] discovered promising new topics on the web by visualizing new combinations of communities sharing common topics. Ohsawa et al.[11] 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 can be applied to show the future directions of textual data.

As for the human relations in communities, Kautz et al.[9] created REFERRAL WEB, a social network graph designed to find an expert who is both reliable and likely to respond to the user. Also, Foner et al.[10] 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, therefore, they show the future directions of individuals. Maarek et al.[12] embodied WebCutter which outputs a tailored map of the web according to the user-specified interests. The map is one of the suggestion of the future directions of the user because it shows 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 the overview of our idea, and then describe our approach in detail.

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 five phases as follows:

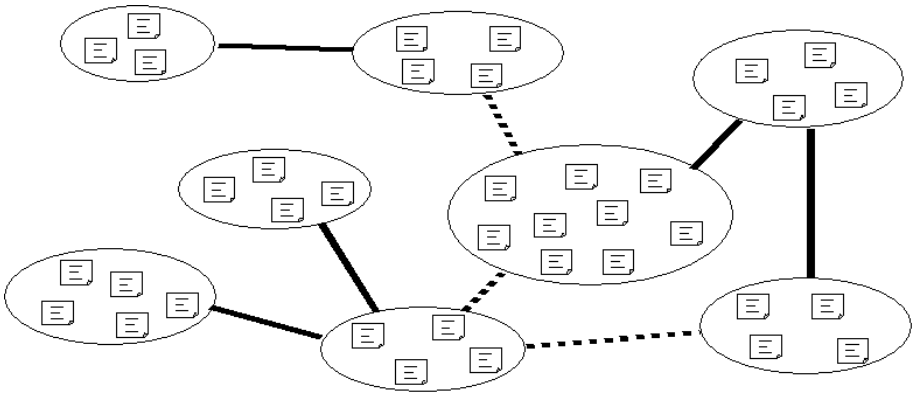

Fig. 1. An overview of the web. Each cluster corresponds to a community sharing common interest. Communities are often share the same interest with each other. Here, solid lines mean established relations and dotted lines show future directions.

Phase1. Collect web pages.
Phase2. Discover communities on the web.
Phase3. Discover established relations among the communities.
Phase4. Discover future directions among the communities.
Phase5. Visualize the future directions.

The accurate definition of a community on the web is an essential problem by itself. In Phase1, following Kumar's definition [4], we expediently define a simple bipartite graph as a community where a community consists of a much cited web page and its surrounding web pages. Next, we focus on the property of the web that communities are often somewhat related with each other because a web page often belongs to some communities. In our view, the relations 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[7]. 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 Fig. 1. Considering the fact that an established link arises only when a future direction grows, focusing on future links is useful for understanding where the changes happen.

3.2 The Detailed Process

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

Phase1. Preparations: First of all, let a user decide a target area/topic which s/he want to explore the future directions. Then, source web pages $\boldsymbol{D}$ are

collected by using Google [5] . Here, the first 500 web pages of Google's output for the query are downloaded.

Phase2. Discover Communities: For surveying the picture of communities by discovering the future directions among communities, we make use of only centered web pages in communities instead of all the web pages. The centered web page named as **core-page** is extracted as follows.

1. Count the frequency of links included in D.

2. Regard the top N_1 links $\boldsymbol{C}$ as the 'core-pages' of communities.

Phase3. Discover Established Relations: Measure the relations among core-pages by counting the number of co-citations, and regard strong relations as established links. The process is as follows.

1. For every pair of two core-pages in C, count the number of links included in both the core-pages.

2. Regard the top N_2 pairs as established links $\boldsymbol{L_1}$ (solid lines in Fig. 1).

Phase4. Discover Future Directions: Measure the relations among core-pages by counting the number of co-citations, and regard weak relations as future links. The process is as follows.

1. For every pair of two cores in C except for L_1, count the number of links included in both the cores.

2. Regard the top N_3 pairs as future links $\boldsymbol{L_2}$ (dotted lines in Fig. 1).

The movement of communities are shown by established and future relations. Therefore, future directions are expressed by the combination of these two kinds of relations.

Phase5. Visualization: Core-pages and its relations(C, L_1, and L_2) are visualized into 2-dimensional interface 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 ***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 Fig. 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

[5] Google is a search engine to which Brin and Page's algorithm[6] is applied. Google is available at http://www.google.com/.

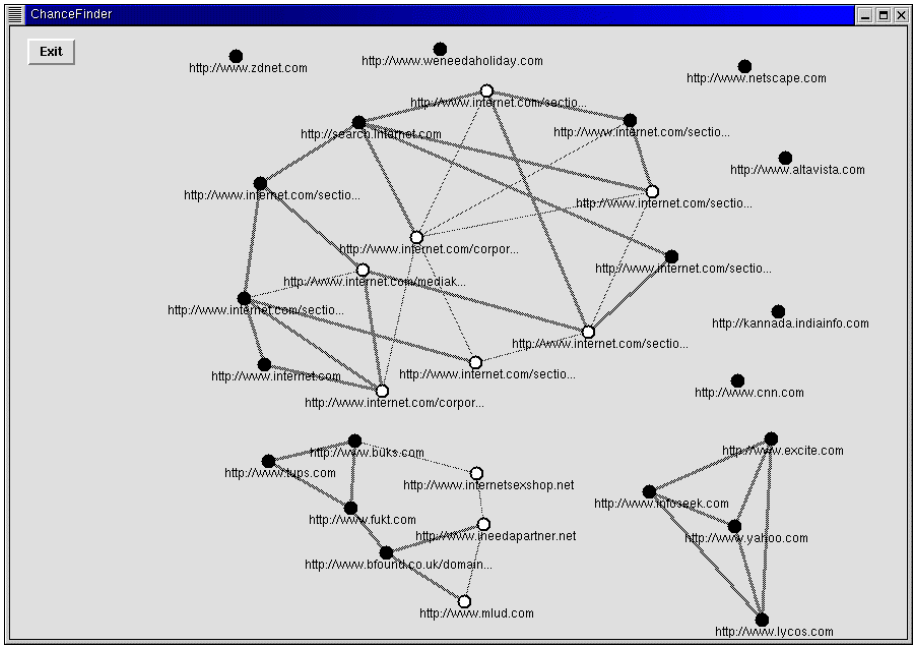

Fig. 2. An output of ChanceFinder for input query 'Portal Site'.

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 Fig.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[6] 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.

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

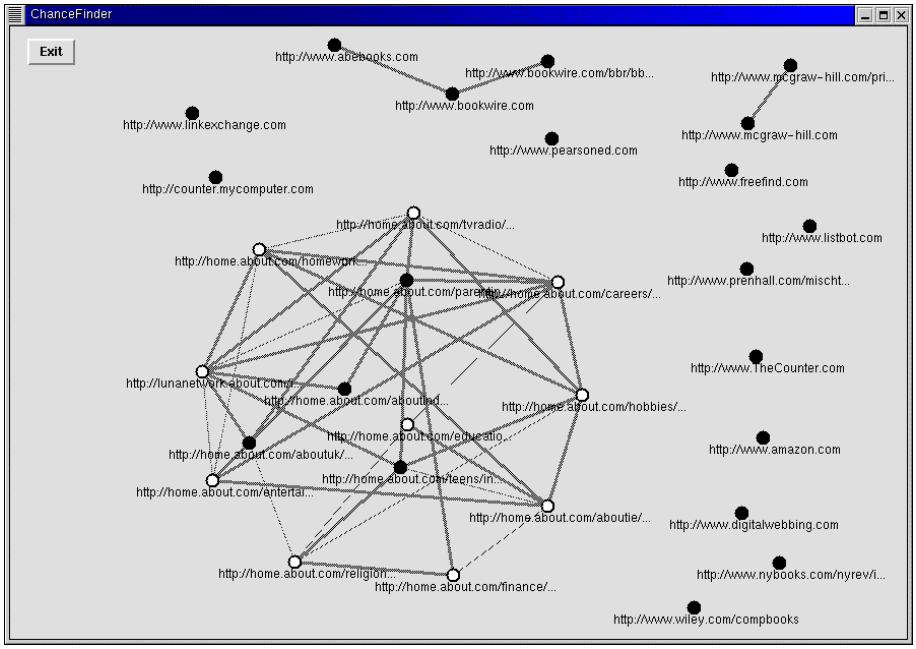

Fig. 3. An output for 'Book Site'.

4.2 Future Directions of Book Site

From the output of ChanceFinder for input query 'Book Site' shown in Fig.3, we can easily recognize one big cluster and two tiny clusters.

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 source 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 looks like tiny cluster at first sight, but these are the web page of McGraw-Hill company which is a time-honored publisher founded in 1909. Therefore, this cluster means a well established community.

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 leapfrog up Media Metrix chart of the Web's top sites' in December 1999, About.com is described as follows [7] :

> Excite@Home Corp., NBC Internet Inc. and About.com Inc. are on the rise, according to the latest traffic numbers from Internet measurement firm Media Metrix Inc.

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

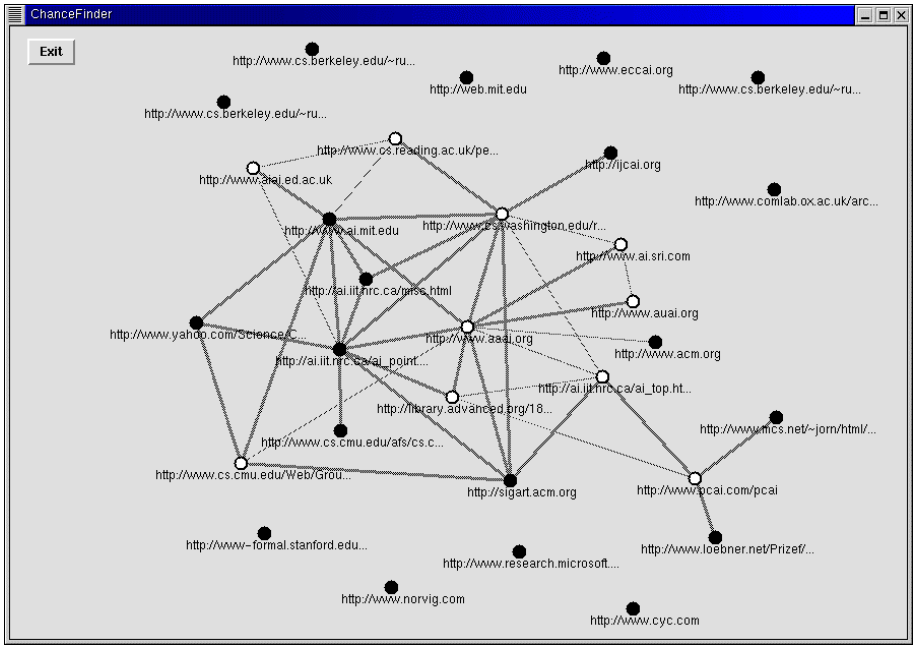

Fig. 4. An output for 'Artificial Intelligence'.

However, About.com seems to be a minor web site in the area of 'Book Site' yet (About.com does not appear in the rankings of 100hot.com). For these reasons, About.com is considered to be struggling to expand the influences, and this consideration can be read from Fig. 3.

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

4.3 Future Directions of Artificial Intelligence

The output for input query 'Artificial Intelligence' is shown in Fig. 4. Viewing Fig. 4, we can recognize only one big chunk of communities where 20 communities, 28 established links, and 11 future links are densely connected. Fig. 4 is essentially different from above two examples in the point that the cluster in Fig. 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 first insist on the importance of discovering new topics. Then, we describe the idea of discovering future directions of communities by chaining primitive communities to understand potential needs or demands. Through some experiments and their evaluations, we show that ChanceFinder certainly shows the relations of communities. However, we expect that whether the relations really become the future directions depends on the user's vision or imagination based on accurate information.

References

1. Gladwell, M.: THE TIPPING POINT: How Little Things Can Make a Big Difference. Little Brown & Company (2000)
2. Broder, A.Z., Glassman, S.C., Manasse, M.S.: Syntactic Clustering of the Web. Proc. World Wide Web Conference (1997)
3. Botafogo, R.A., Shneiderman, B.: Identifying Aggregates in Hypertext Structures. Proc. ACM Conference on Hypertext (1991) 63–74
4. Kumar, R., Raghavan, P., Rajagopalan, S., Tomkins, A.: Trawling the Web for Emerging Cyber-Communities. Proc. World Wide Web Conference (1999)
5. Kleinberg, J.M.: Authoritative Sources in a Hyperlinked Environment. Proc. ACM-SIAM Symposium on Discrete Algorithm (1998) 668–677
6. Brin, S., Page, L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine. Proc. World Wide Web Conference (1998)
7. White, H.D., McCain, K.W.: Bibliometrics. Annual Review of Information Science and Technology, Elsevier, **24** (1989) 119–186
8. Matsumura, N., Ohsawa, Y., Ishizuka, M.: Discovering Promising New Topics on the Web. Proc. Knowledge-Based Intelligent Engineering Systems & Allied Technologies (2000) 804–807
9. Kautz, H., Selman, B., Shah, M.: The Hidden Web. AI magazine **18** (1997) 27–36
10. Foner, L.N.: Yenta: A Multi-Agent, Referral-Based Matchmaking System. Proc. Autonomous Agents (1997) 301–307
11. Ohsawa, Y., Benson, N.E., Yachida, M.: KeyGraph: Automatic Indexing by Co-occurrence Graph Based on Building Construction Metaphor. Proc. Advances in Digital Libraries, (1998) 12–18
12. Maarek, Y.S., Jacovi, M., Shtalhaim, M., Ur, S., Zernik, D.: WebCutter: A System for Dynamic and Tailorable Site Mapping. Proc. World Wide Web Conference (1997)