Discovering Promising New Topics on the WWW

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

†Dept. of Electronical Eng., School of Eng., Univ. of Tokyo.
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656 Japan
‡Graduate School of Systems Management, Univ. of Tsukuba.
3-29-1 Otsuka, Bunkyo-ku, Tokyo, 112-0012 Japan

Abstract

The phenomenon that a 'little' topic can make a big difference is called 'The Tipping Point'. As a strategy for marketing, this phenomenon could be a great economic chance (opportunity). For instance, it is a good strategy to promote items associated with a certain fashion. The Tipping Point refers to a situation where the topic matches potential needs of customers. In this paper, we analyze the mechanism of The Tipping Point in the context of the WWW, and present an algorithm named 'Expected Activation' for discovering promising new topics on the WWW.

1 Introduction

It often occurs that a new topic becomes suddenly popular. We understand 'topic' in the broad sense that covers fashion, an item, a song, food, and so on. The topic may seem insignificant at first sight, however, it may turn out to match potential needs of customers. The Tipping Point describes this kind of phenomenon where a 'little' (insignificant, marginal) topic makes a big difference [1]. For example, how does a novel written by an unknown author become a national bestseller? Why did the crime rate drop so dramatically in New York City in the mid-1990's? 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.

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 as a marketing strategy:

- Discovering chances could be an enormous advantage in our highly competitive world.
- Discovering chances may guarantee more safety, because we can avoid selling items that damage us severely.

The WWW is an attractive information source because of its sheer size and sensitivity to trends. In this paper, we present an algorithm called **Expected Activation** for discovering promising new topics on the WWW.

Discovering chances is different from predicting topics that will be popular in the future. The predicted topics will come into fashion to a certain extent. On the other hand, chances cannot come into fashion without one's action, e.g., promotion. That is, the active rôle of people is necessary for realizing the chances. The merit of discovering chances is that competitors are not aware of them. Previous data-mining methods in Knowledge Discovery in Databases (KDD) treat chances as noise for their rareness, so that promising new topics cannot be extracted.

A recent example of a chance is the use of cellular phones which spread exponentially all over the world. Because cellular phones were very rare before, pocket-bells were mainly used. Nobody predicted that cellular phones would become such a great sales hit. Considering the context of the appearance of cellular phones, there are essentially two factors which match people's potential needs. First, a pocket-bell was inconvenient, because people had to find a public phone when a pocket-bell rang. On the other hand, a cellular phone can be reached from anywhere. Second, a cellular phone is equipped with the functions of the Internet and E-mail services. Due to the synergy effects of these factors satisfying the user's potential needs, cellular phones began to get popular.

Needless to say, prediction is an important activity. In the case of determining where a new store should be built, the owner often predicts the estimated sales by analyzing various data, e.g., population or the ratio of young people. However, this form of prediction cannot find more (economically) relevant relations between customers and items (products). In another case, prediction is used for determining what items should be sold [2]. However, it is hard to find potentially needed new items, because new items are hardly detected from the sales in the first place. Although prediction might allow for simple estimations of potential needs, with a more detailed knowledge of potential needs (e.g., the relations between goods and customers), a shop manager could make more accurate strategic estimations. Of course, you might find out the potential needs of

[*] e-mail:matumura@miv.t.u-tokyo.ac.jp

customers if you conduct a survey of the needs of a large number of people. However, because of the short time span of (some) needs, this kind of survey would have to be done quite frequently, which is usually too costly. Therefore, the approach of discovering chances from data is much more beneficial for us.

The idea of discovering chances can be explained by an 'activation model' rather than a social epidemics referred to in The Tipping Point. The basic assumption of this model is that chances (e.g., promising new topics) are activated by human needs. This model is similar to KeyGraph [3], which was originally developed as a keyword extracting algorithm. KeyGraph makes use of the co-occurrence graph of terms in a document for extracting keywords. It assumes that a document consists of basic concepts from which assertions are activated. The difference between KeyGraph and the Expected Activation algorithm is that KeyGraph retrieves topics by micro-scale such as words, whereas our goal is to retrieve topics by macro-scale like web pages.

The rest of the paper is organized as follows. In Section 2, we describe how to discover web pages which contain potential customer needs by applying the activation model to the WWW. A more detailed description of the mechanism of Expected Activation is given in Section 3. In Section 4, we report on our experimental evaluation. Section 5 concludes the paper.

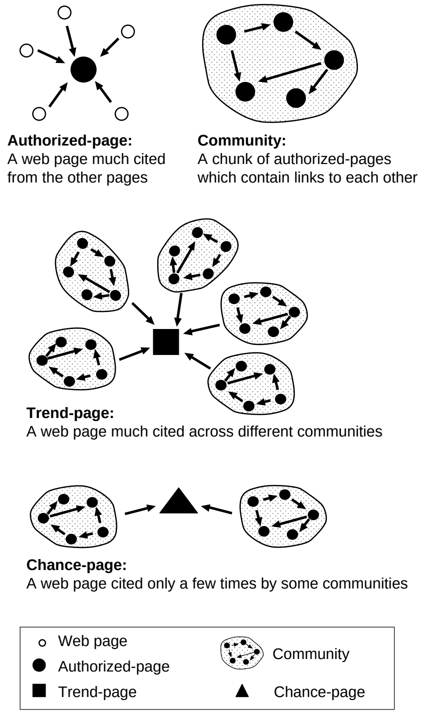
Authorized-page:
A web page much cited from the other pages
Community:
A chunk of authorized-pages which contain links to each other
Trend-page:
A web page much cited across different communities
Chance-page:
A web page cited only a few times by some communities
Web page
Authorized-page
Trend-page
Community
Chance-page

Figure 1: Illustration of authorized-pages, communities, trend-pages, and chance-pages.

2 How to Discover Chances?

In this section, we apply the idea described above to the WWW. More precisely, we aim to reveal how our mechanism detects chances (promising new topics) on the WWW.

First, we discuss what kind of web pages can be seen as containing actual needs. Recent search engines like Google [5] make use of the web's link structure (this technology is described in [4]). The value of importance (quality) of a web page is given as the number of citations from other pages (backlinks). The idea of using citations (backlinks) to evaluate the importance of web pages is based on social filtering [6]. Highly ranked web pages play a vital rôle in the WWW, therefore we call them **authorized-pages**. In the case where authorized-pages are linked with each other, we treat these pages as a **community**. A community is a big chunk of actual needs. In our view, chances are hidden between different communities.

Second, we mention the effects of activation of web pages originating from different communities. Consider a web surfer's next action starting from an authorized-page within a community, he/she will probably go to a page cited therein by tracing its link. A web page much cited (activated) across different communities will be visited repeatedly. However, such page is then in trend, so we call it a **trend-page**. On the other hand, a web page cited (activated) only a few times by some communities is not currently in trend, as some communities disregard it. Even though it might be a trivial web page, the page surely matches a potential need as some members of some community cite (activate) it. Accordingly, we call it a **chance-page**. Our eventual aim is to discover chance-pages that are supported by different communities. Our categorization of web pages for discovering promising new topics is graphically explained in Fig. 1.

The threshold of 'activation level' (number of times a certain web page is cited by different communities) could not be set to a certain value in advance because it depends on the scale of communities. In Expected Activation, (potential) chance-pages are extracted until the activation level reaches more than two by adding a community step by step, because a web page with activation level one is considered to have only too weak potential needs.

3 The Expected Activation Algorithm

Expected Activation is an algorithm for discovering chance-pages from the WWW by making use of the web's link structure. The algorithm is based on the activation model described in Section 1. The process of Expected Activation consists of five phases.

Phase 1: Input query and collect Web pages. Input queries composed of some keywords about the user's area of interest, and download the first 100 web pages for the query by using existing search engines. Our aim is to discover chance-pages supported by different communities, therefore we must input keywords from different areas.

Phase 2: Collect authorized-pages. Analyze the citations in web pages downloaded in **Phase1**, and collect much cited web pages as authorized-pages.

Phase 3: Form communities. Regard the authorized-pages that contain links to each other as a community.

Phase 4: Extract trend-pages. Regard much cited web pages from different communities as trend-pages.

Phase 5: Discover chance-pages. Discover web pages rarely cited from communities as chance-pages.

There is a variety of search engines which considerably differ in the way they rank retrieved pages. Since their internal ranking algorithm is mostly not open, we must treat the algorithm as a blackbox. In this paper, we experiment with Google[1] because its ranking algorithm is published. In the following, we give a brief overview of Google:

- Google searches web pages with the ‘PageRank’ criterion which selects web pages by importance and quality.
- PageRank is acquired by analyzing citations, backlinks, and tags of web pages recursively.

Though Expected Activation uses only links in the WWW, Google gives us much benefit because the idea underlying Google is similar to **Phase 2** in Section 3. Another benefit of using Google is that the web pages Google retrieves tend to be the most representative web pages of communities. Therefore, we do not have to form communities in **Phase 3** of Section 3. That is, we use Google to process **Phase 2** and **Phase3**.

4 Experimental Evaluations

We implemented a prototype of the Expected Activation system named **Chance Finder** on a Sun ENTERPRISE 450 workstation with perl5 and Perl/Tk.

It is difficult to evaluate the accuracy of chance-pages as a marketing strategy (see Section 1), because we cannot relate chance-pages to actual user needs easily. In this paper, we consider using questionnaire surveys for our experiment. The purpose of the questionnaire surveys is to identify the chance-pages’ effects to the test subjects. This process can be explained as follows:

1. Input queries composed of some keywords to both Google and Chance Finder.

2. Show Google’s search result to the test subjects, and let them report their thoughts about “Can you realize a chance?”.

3. Show the result of Chance Finder to the test subjects, and let them report their thoughts about “Can you realize a chance?”.

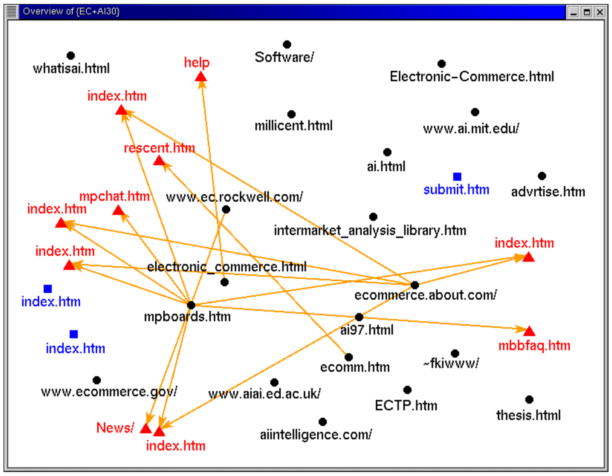

Figure 2: An output of Chance Finder for input query {”Electronic Commerce” OR ”Artificial Intelligence”}.

4.1 An Example of Chance Finder

In this subsection, we show an example experiment where chances are detected from the border area between EC (Electronic Commerce) and AI (Artificial Intelligence). The source web pages are collected by inputting queries ({“Electronic Commerce” OR “Artificial Intelligence”}) alternatively into Google, since Google does not support the logical “or” operator. Fig. 2 shows the result of Chance Finder for input query is {“Electronic Commerce”} OR {“Artificial Intelligence”}. In Fig. 2, the circle nodes, square nodes, and triangle nodes stand for authorized-pages, trend-pages, and chance-pages, respectively, and the arrows denote links from authorized-pages to chance-pages.

As can be seen from Fig. 2, we can recognize that two different authorized-pages play significant rôles for citing the same chance-pages which are not authorized in general. One of the authorized-pages is 1) ecommerce.about.com[2] that contains EC reports, statistics, technologies, and a variety of tips. The another authorized-page is 2) mpboards.htm[3] that covers almost all of the AI area. Those two authorized-pages strongly refer to three trend-pages: 3) index.htm[4], 4) index.htm[5], and 5) submit.htm[6]. 3) and 4) are representative index pages of the About.com domain which provide a variety of sites led by expert guides. Recently, About.com sites are rapidly growing. 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 [7]:

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

[1] http://www.google.com/

[2] http://www.ecommerce.about.com/

[3] http://www.ai.about.com/mpboards.htm

[4] http://home.about.com/index.htm

[5] http://a-zlist.about.com/index.htm

[6] http://www.perkinscoie.com/submit.htm

the latest traffic numbers from Internet measurement firm Media Metrix Inc.

5) is a homepage of PERKINS COIE LLP which is one of the largest law firms. Digital signature is a core technique for the security of EC, therefore the knowledge of the laws about digital signature is necessary. Therefore, 3), 4) and 5) are surely catch the present trend in the area of EC and AI. Moreover, the authorized-pages suggest five chance-pages (here, chance-pages are restricted to more than activation level two, because of 2.): 6) index.htm[7], 7) index.htm[8], 8) index.htm[9], 9) index.htm[10], and 10) index.htm[11]. Each of 6), 7), 8), 9), and 10) is created by the people who have a strong sense of religion, affiliates, travel, people, and internet, respectively. Considering these pages from a chance discovery point of view, we inspire the following ideas:

- AI might apply to improve the design of the web sites attractively.
- AI might apply to travel planning.
- AI might apply to consumer targeting for EC.

4.2 Evaluations on Questionnaire Surveys

We conducted questionnaire surveys to identify the chance-pages' effects to the three test subjects, whose reports can be summarized as follows:

Google

- I had to read all the web pages to see the picture of the results.
- It was hard to find chances because the contents of the results should have been compiled.

Chance Finder

- I easily found the border areas between the communities and could focus on a few profitable web pages.
- The arrows often gave us triggers to convert contents into chances.
- Chance Finder is an unprecedented system, because it focuses not only on authorized web pages but also on non-authorized web pages.

To summarize the reports, we observed that chances are recognized only after presented consciously and explicitly, since we could not compile the contents of many web pages well. From a chance discovery point of view, Chance Finder's results are more beneficial than Google's outputs, because Chance Finder cuts the extra information and shows the relations among authorized-pages and chance-pages.

[7] http://www.about.com/religion/index.htm

[8] http://affiliates.about.com/index.htm

[9] http://home.about.com/people/index.htm

[10] http://home.about.com/travel/index.htm

[11] http://home.about.com/internet/index.htm

5 Conclusion

In this paper, we describe the idea of discovering promising new topics on the WWW, and propose the Expected Activation algorithm to automatically retrieve pages that contain material about new topics. From the point of view of chance discovery, we can show that the results of Chance Finder are more profitable than Google's outputs. Experiments have been performed based on questionnaire surveys.

Nowadays, the WWW is one of a most important and vast information sources, and will be more so in the future. In this context, the web is a good starting point. Discovering chance-pages could be a major economic factor in our highly competitive world.

As future work, we consider acquiring new customers by promoting the chance-pages to the community, and plan to apply Chance Finder to a consumer targeting technique of online advertisement[8].

References

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

[2] Paul Resnick and H.R. Varian, "Recommender system", *Introduction to special section of Communications of the ACM(ACM'98), pp.56–58, 1998.*

[3] Yukio Ohsawa, Nels E. Benson and Masahiko Yachida, "KeyGraph: Automatic Indexing by Co-occurrence Graph based on Building Construction Metaphor," *Proceedings of the Advanced Digital Library Conference (IEEE ADL'98), pp.12-18, 1998.*

[4] Soumen Chakrabarti, Byron E. Dorn, S. Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, Andrew Tomkins, David Gibson, and Jon Kleinberg, "Mining the Web's Link Structure", *IEEE Computer, pp.60–67, 1999.*

[5] Sergey Brin and Lawrence Page, "The anatomy of a Large-Scale Hypertextual Web Search Engine", *7th World Wide Web conference(WWW7), 1998.*

[6] Upendra Shardanand and Pattie Maes, "Social information filtering: algorithms for automating word of mouth", *Proceedings of the ACM Conference on Human Factors in Computing Systems(CHI'95), pp.210–217, 1995.*

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

[8] Marc Langheinrish, Atsuyoshi Nakamura, Naoki Abe, and Yoshiyuki Koseki, "Unintrusive Customization Techniques for Web Advertising", *8th International World Wide Web Conference(WWW8), 1999.*