Extracting Navigational Information from Link Structure in Blogoshere

Takashi Ishikawa[1] and Yuu Ishikawa[2]

1 Nippon Institute of Technology
4-1 Gakuendai, Miyashiro, Saitama, 345-8501 Japan
tisikawa@nit.ac.jp
2 Aoyama Gakuin University

Abstract

The paper describes a method to extract navigational information from link structure in Blogosphere such as Everything2. The purpose of the method is to support readers of weblogs in selecting hyperlinks from scattered anchor texts in the web pages by providing navigational information. The navigational information consists of hyperlinks that are organized with approximate subsumption relations between sets of hyperlinks in the web pages. A case study for Everything2 weblog exemplifies the effectiveness of the proposed method.

Introduction

Weblog is one of collaborative authoring tools to describe many kinds of contents in the web pages [1, 2]. A usual web page of weblogs is a list of entries or posts each of which is a hypertext including hyperlinks to connect to other web pages. The reader of a weblog selects an entry by the title or a hyperlink in other web page, reads the contents and proceeds to other entries in the same web page or other web pages by clicking an anchor text. The decision where to proceed is made by evaluating relevance of anchor texts to the interest of the reader in reading the entry. Therefore the reader has to read the whole contents of the entry to decide where to proceed. Contrastingly web pages in information sites such as Yahoo [3] have navigational information as category lists or local navigation menus so as to guide the readers in selecting anchor texts without reading the whole contents. The lack of navigational information in usual weblogs is a deficiency of media to describe a lot of information or knowledge in web pages.

The research hypothesis of this research is that adding navigational information to weblog improves the functionality of weblog as an information medium in retrieving information efficiently. In the hypothesis, navigational information means category lists or local navigation menus to guide the readers in selecting linked web pages. Looking navigational information is a kind of information medium. Issues to verify the hypothesis are that how the navigational information can be generated automatically from the contents and that how to add the navigational information to the web pages of weblog. An idea to approach the first issue is that the network structure of hyperlinks in web pages reflects the structure of information such as concept hierarchy so that the structural information must be extracted from the link structure [4, 5]. The second issue can be solved by implementing the method for extracting navigational information from link structure of weblogs

The aim of the research is to develop a computational method to extract navigational information from link structure of weblogs. The navigational information means a subset of anchor texts in a web page to reduce items to be selected by the readers for the sake of search efficiency. The items in the subset are preferable as being structured by the order of importance or relatedness and selected so as not to reduce web pages, which are reachable through the selected anchor texts. Therefore the problem to be solved is a kind of optimization problem to valance search efficiency and information amount.

Issues in browsing Blogosphere

Weblog (also referred to as blog) is defined as an online journal or diary that is frequently updated [6]. More specifically it is defined as a public web site where users post informal journals of their thoughts, comments, and philosophies, updated frequently and normally reflecting the views of the blog's creator [7]. In terms of format, a weblog is defined as a web page with small chunks of hypertext (known as ‘posts’), which are date-stamped, ordered reverse-chronologically, newest information at the top of the page, and posts consist of links and commentary [8]. From the feature of the last definition, one web page of a weblog consists of posts concerning different topics, while a web page in many information site treats one topic or theme. Therefore a weblog is a mixture of description for different topics, the feature makes it difficult to structure the information in the web page. This is an issue of weblog regarding inefficiency of information retrieval.

In the third definition described above, a weblog is characterized as posts consist of links and commentary. The links in a weblog are classified into hyperlinks and trackbacks [9]. The hyperlinks are written by authors of the weblog, while the trackbacks are made by readers of the weblog using automatic mechanism of weblog software. The trackback mechanism is an intrinsic feature of weblog

that combines automatically weblogs in mutual direction while usual hyperlinks combines weblog in one direction. Therefore a weblog has two types of links, one type is usual hyperlinks and another type is trackback links. A trackback link is merely a hyperlink to a web page of other weblog, but the linked web page has many posts in usual so that the relation between weblogs becomes many to many. The trackback mechanism of a weblog also makes it difficult to structure the information in the web page. This is another issue of weblog regarding inefficiency of information retrieval.

Link structure of Blogosphere

The word Blogosphere is used to describe the world or community of blogs and blogging [10]. Wikipedia defines it as follows [11]; Blogosphere (alternate: blogsphere) is the collective term encompassing all weblogs or blogs; blogs as a community; blogs as a social network. Weblogs are densely interconnected; bloggers read others' blogs, link to them, reference them in their own writing, and post comments on each others' blogs. Because of this, the interconnected blogs have grown their own culture. From these definitions, the essential feature of Blogosphere or weblogs is its densely interconnected structure. The structure is composed of hyperlinks as anchor texts in their posts and trackbacks. Therefore the network structure of Blogosphere is characterized with link structure of hyperlinks in weblogs.

On the basis of above mentioned results of previous researches, we adopt link analysis approach to verify the hypothesis that adding navigational information to weblog improves the functionality of weblog as an information medium in retrieving information efficiently. Our conjecture is that navigational information can be extracted from link structure of weblogs, which inherently have scale-free small world (SFSW) structure. The scale-free structure enables to reorganize posts in weblogs in nearly hierarchical structure and the small world structure enables to reduce items of related posts to select instead of reducing reachability to the related posts. In the following we describe first "writeup selection" to treat many to many relations in weblogs and then describe "link abstraction" to reduce items to select. Here writeup means a post in Everything2 [12] weblog, which is used in the experiment to demonstrate the effectiveness of the proposed method.

Writeup selection

To apply link analysis to web pages in weblogs, it needs to treat many to many relations between web pages that usually consists of multiple posts concerning different topics in one web page. The problem is to separate each posts in a web page and to determine which post is related to each posts. The problem is hard to solve in general, therefore we address a specific weblog as Everything2 to demonstrate the effectiveness of method.

On Everything2 a writeup is the text of an individual user's contribution to a node (web page) [13]. A user can contribute one writeup to any given node and only the author of a writeup can edit the writeup. There are two types of links in Everything2, one is a soft link that is specified with a title, and the other is hard link that is specified with a node id. A soft link refers to a web page (i.e., node) that consists of multiple writeups while a hard link refers to an individual writeup. A writeup may be displayed in a web page including the writeup or in a web page that consists of the writeup only. In Everything2 these two types of links are used without attention to the distinction of their functions. The situation induces complicated many to many relations between web pages.

In Everything2, when the node designated by a soft link with a node title does not exists, then some titles similar to the title in the soft link are displayed in a Findings page (Figure 1). This facility is convenient for human users, but requires an additional function of web crawlers to select the most relevant title from the titles in the Findings page. For the purpose we adopt least difference criterion that select a title with the least numbers of not coincident characters around contiguous coincident characters.

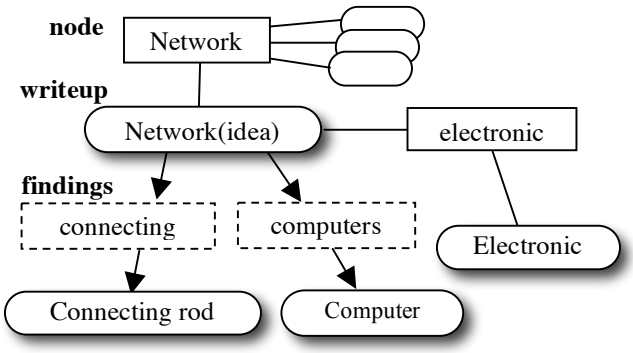

Figure 1. An example of writeup selection

Link abstraction

To make more efficient weblog browsing, it needs to reduce items to be selected in each web page without reducing number of web pages reachable from a current web page. If web pages constitute a complete graph by hyperlinks between them, then all of web pages are reachable with each other. But in usual situation some of web pages will not be reachable when a part of web pages are removed from. So the problem is to select a subset of web pages linked to a given web page so as to maximize the number of reachable web pages from the given web page for a given reduction ratio of the selected subset. If the ratio of reachable web pages to all linked web pages exceeds the reduction ratio, then abstraction of the network

structure is effective in the meaning that the loss of information reduced with suitable selection. So we call the selection process link abstraction.

A principle of link abstraction is that if a set of hyperlinks in a web page is a subset of hyperlinks in another web page, then the subsumed web page can be removed from a selected subset of web pages without loss of reachable web pages. But this situation is ideal, so we define subsumption ratio as follows.

Definition. A subsumption ratio SR(A, B) of a set A to a set B is a fraction that the number of common elements in A and B divided by the number of elements in A.

SR(A, B) ranges from 0 to 1. If A is subsumed by B then SR(A, B) is 1, and if A is partly subsumed by B then SR(A, B) is less than 1 and greater than 0. Subsumption ratio enables to treat approximately the hierarchical relations between web pages from the viewpoint of set theory.

Experiment

We have conducted an experiment to verify the effectiveness of the proposed method for extracting navigational information from link structure of weblogs. The details of the experiment are the following.

Objective

To validate the effectiveness of the proposed algorithms

Method

(1) Measures for writeup selection:
- **recall**: extracted writeups / linked writeups
- **precision**: correct writeups / extracted writeups
- **reduction rate**: selected writeups / linked writeups

(2) Measures for link abstraction:
- **reachable rate**: reachable writeips in navigation / reachable writeups
- **reduction rate**: selected writeups / linked writeups
- **grouping relevancy**: by subjective evaluation

Material

Everything2 [12]. The weblog makes it easy for potential authors to contribute, where the content for a node is entered in plain text, which Everything2 converts into HTML [2]. We used 15 sampled nodes of which titles are Blogosphere (place), Network (idea, thing), world wide web (idea, thing), weblog (idea, thing), link (thing), Internet (idea, thing), graph (idea), hypertext (thing, 2 writeups), blog (idea, 2 writeups).

Results

Followings are the averaged measurements for the samples n = 15.

(1) Measurement for writeup selection:
- **recall** = 230 / 327 = 0.7034
- **precision** = 200 / 230 = 0.8696
- **reduction rate** = 327 / 989 = 0.3306

(2) Measurement for link abstraction:
- **reachable rate** = 1510 / 2047 = 0.7381
- **reduction rate** = 154 / 230 = 0.6176
- **grouping relevancy** = GOOD

Conclusion

The paper has described a method to extract navigational information from link structure in Blogosphere such as Everything2. The navigational information consists of hyperlinks that are organized with approximate subsumption relations between sets of hyperlinks in the web pages. A case study for Everything2 weblog exemplifies the effectiveness of the proposed method that allows reader of weblogs to brows more efficiently. A future work is to implements the method in customized web browser for reading weblogs.

References

[1] Paquet, S. and Pearson, P. A Topic Sharing Infrastructure for Weblog Networks. In Proceedings of Conference on Communication Networks and Services Research, 2004.

[2] Emigh W, and Herring SC. Collaborative Authoring on the Web:. A Genre Analysis of Online Encyclopedias. In Proceedings of the 38th Annual Hawaii International Conference on System Sciences, pp. 99a-99a, 2005.

[3] Yahoo!. http://www.yahoo.com/

[4] Chakrabarti, S., Dom, B., Raghavan, P., Rajagopalan, S., Gibson, D., and Kleinberg, JM. Mining the Link Structure of the World Wide Web. IEEE Computer, 32(8):60-67, 1999.

[5] Chen, Z., Liu, S., Wenyin, L., Pu, G. and Ma, W. Building a Web Thesaurus from Web Link Structure. SIGIR, 2003.

[6] RSS Glossary. www.feedforall.com/rss-glossary.htm.

[7] WorldWideLearn. www.worldwidelearn.com/elearning-essentials/elearning-glossary.htm.

[8] Ó Baoill, A. Conceptualizing the weblog: understanding what it is in order to imagine what it can be. Interfacings.

[9] TrackBack from Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/TrackBack.

[10] Blogjargon. www.fzelders.nl/weblog/.

[11] Blogosphere from Wikipedia, the free encyclopedia. en.wikipedia.org/wiki/Blogosphere.

[12] Everything2. http://www.everything2.com/.

[13] Writeup. http://www.everything2.com/index.pl?node=writeup.