JOURNAL OF OBJECT TECHNOLOGY

Online at http://www.jot.fm. Published by ETH Zurich, Chair of Software Engineering ©**JOT, 2008**

Vol. 7, No. 4, May-June 2008

On Having Search Engines Deliver Hierarchies of Web Pages

Ok-Ran Jeong, Jiawei Han, Won Kim, Eunseok Lee

Abstract

There is large room for improving the usefulness of today's Internet search engines. There are various ways. One of the ways is, when given a sub-directory (subWeb) of a Web site in a certain domain, to automatically retrieve, or mine, corresponding sub-directories of other Web sites in a similar domain. In this article, we outline a methodology for doing this.

1 INTRODUCTION

There is no question that today's Internet search engines have become an indispensable tool for finding information on seemingly limitless subjects. However, there is large room for improving the usefulness of the search engines. [Kim et al. 2008] suggests a few major ways. In this article, we will explore one of them in some detail to encourage further research.

[Kim et al. 2008] points out that Web sites are hierarchically structured and the Web users need to navigate through them from the URLs of the roots of the Web sites, and that it would be helpful if the search engines can reduce the need for the users to navigate through the hierarchical structures of the Web pages. For example, if a student is interested in learning about the research and publication activities on "Web technology and u-commerce" of the computer science departments of several graduate schools in the US, he will need to visit the homepages of all the universities he is interested in, and navigate through them to learn about the professors in the "Web technology and u-commerce" or related research group, their publications, Ph.D. theses, etc. This is a rather tedious process. Since there is a reasonable degree of commonality in the hierarchical structures of the homepages in the same domain, it may often be helpful if the search engine can take as input a sub-directory (subWeb) of one homepage, or sub-directories of a set of homepages, in a certain domain, and by analyzing them (or "mining" them), return the hierarchical structures of other homepages in the same domain. For example, Figure 1 shows that the homepages of the universities in the US, being in the same domain, are similar in their hierarchical structures, and the labels on the Web pages are

Ok-Ran Jeong, Jiawei Han, Won Kim, and Eunsuk Lee: "On Having Search Engines Deliver Hierarchies of Web Pages", in *Journal of Object Technology*, vol. 7, no. 4, May-June 2008, pp. 33-41 http://www.jot.fm/issues/issue_2008_05/column4/

also similar. It may be useful, if the search engine can take as input the subWeb rooted in "the computer science department of the University of Illinois at Urbana-Champaign", and return corresponding subWebs of the computer science departments of several other universities. This may be applicable to other domains, such as e-markets, airlines, hotels, hospitals, newspapers, book publishers, etc.

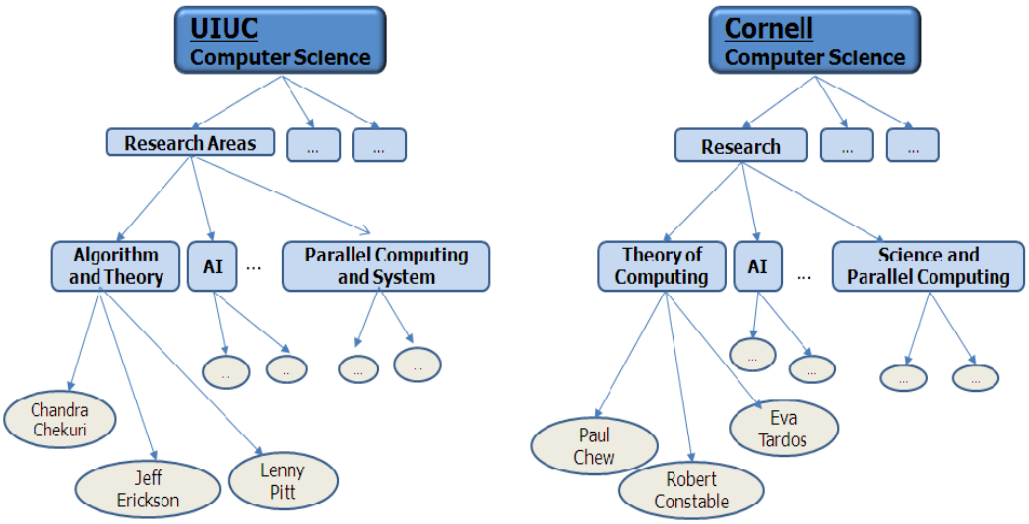
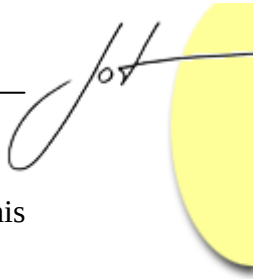


Figure 1: The structures of two subWebs in a similar domain

In this article, we propose a methodology for enabling search engines to take as input a subWeb in a certain domain, and retrieving, or mining corresponding subWebs in a similar domain. The methodology puts together four separate known techniques to solve an important unaddressed area of Web search. Relevant existing techniques include Web page classification algorithms [Brin and Page 1998, Chakrabarti 2002, Dumais and Chen 2002, Shen et al 2004, Shih and Karger 2004, Sun et al. 2000] and machine learning algorithms [Nigam et al. 2000, Raskutti et al 2002]. The ultimate objective of our methodology is to minimize the need for people to resort to tedious manual navigations of semantically related Web pages in similar domains.

2 THE METHODOLOGY

In this section, we describe our methodology, which takes as input only one subWeb in a particular domian, and analyzes it to mine, or predict, subWebs in a similar domain. It is based on analyzing the structure of a labeled training dataset (a subWeb) within a domain, and generating a prediction model (a modified Naïve Bayesian classifier) that can automatically categorize the structures of unlabeled data (other subWebs). The methodology includes four key elements: the creation of the schema of the input subWeb, the creation of a domain profile, the creation of modified Naïve Bayesian classifiers, and the training of the classifiers. Once the classifiers have been trained with the training



dataset, they are applied against the Web data of other subWebs. In the remainder of this section, we present these four steps in turn.

Creating the Schema

Web pages are structured in a hierarchy, and each Web page consists of content and links. There are three types of link; namely, incoming (inlink), outgoing (outlink), and co-citation (colink). An inlink is a gateway into a page (document); an outlink is a link through which one can exit the page, and colinks are connected with each other [Lim et al 1999]. The content provides the crucial classification data based on the importance and frequency of its elements.

The starting point of our methodology is the creation of the schema for the input subWeb. Since a subWeb is a hierarchy of Web pages, its schema includes the hierarchical structure information for the Web pages that comprise it, and the content and links for each of the Web pages. Figure 2 shows the structure of the "Research Areas" subWeb of the "University of Illinois at Urbana-Champaign computer science department" homepage. The "Research Areas" subWeb includes two different lists. One consists of links to its 10 sub-directories (i.e., child nodes): this includes "Algorithms and Theory," "Artificial Intelligence,",…"Database and Information Systems,",…"Scientific Computing," etc. Another consists of links to 13 directories that are not descendants of the subWeb, but that do appear on the subWeb's pages: this includes "Undergraduate," "Graduate," "Online Programs," "Research," "About Us," etc. The "Research" Web page is an ancestor of the "Research Areas" subWeb.

As a running example for illustrating our methodology, we will assume, for the remainder of this section, that, given the "Algorithms and Theory" subWeb of the UIUC CS department, we would like to receive as output, subWebs rooted in "Algorithms and Theory" or something comparable in other universities, such as Cornell University, Carnegie-Mellon University, the Massachusetts Institute of Technology, etc.

For the purpose of predicting other subWebs from the given subWeb, we need a dataset for training the classifier. Once the classifier has been trained, it can take other subWebs and classify their contents. The training dataset for the "Algorithms and Theory" subWeb includes all the terms that appear in all of the descendant Web pages rooted in the "Algorithms and Theory" Web page, and the outlinks from the "Algorithms and Theory" Web page. There is no need to consider any of the inlinks and colinks of the "Algorithms and Theory" Web page. This training dataset is called a labeled training dataset.

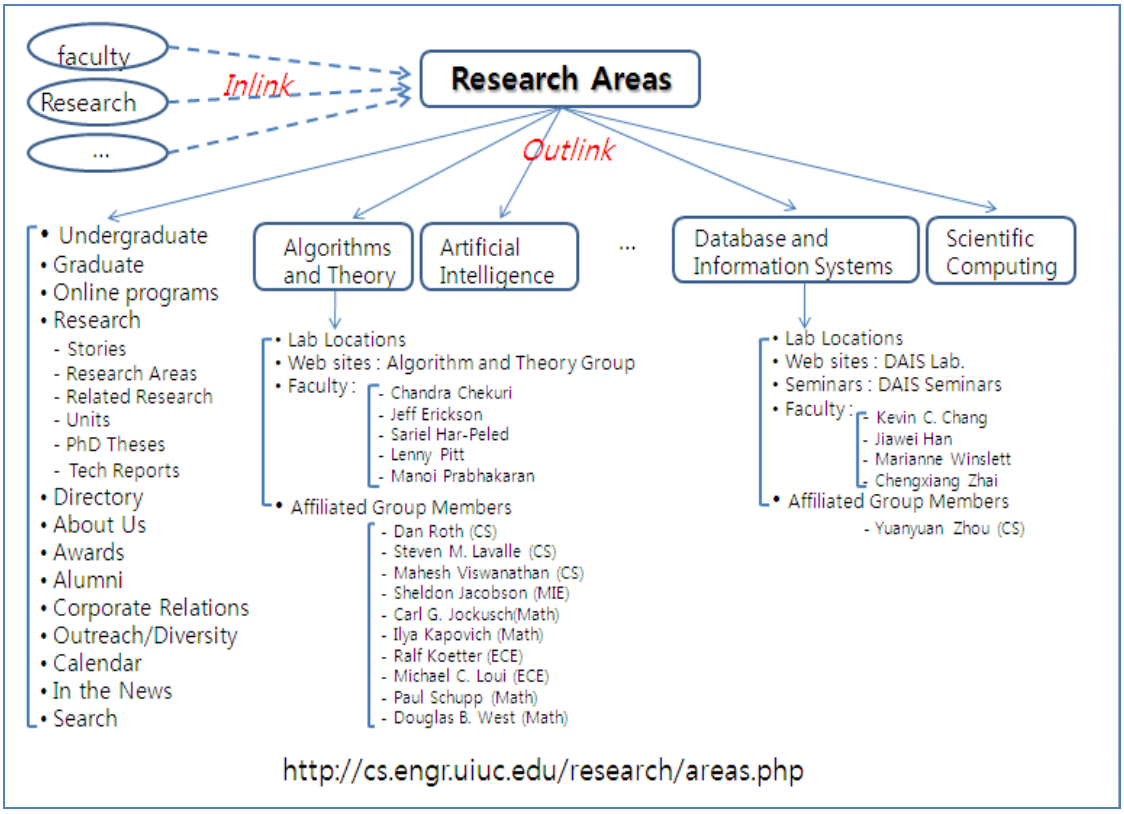


Figure 2. The Structure of the Research Areas subWeb of the UIUC CS Department homepage

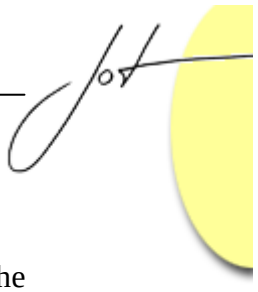
Creating the Domain Profile

Once the schema of a given subWeb has been created, the domain profile for the labeled training dataset needs to be created. The domain profile includes three attributes, namely page terms, link counts, and hierarchy terms. Each page term is basically a key term found on a Web page. Its frequency, that is, the term frequency, can be extracted from the page content using tf·idf. If a URL is extracted from a term, it can become an important keyword for use in subsequent classification.

Continuing with the example of Figure 2, the domain profile for the "Research Areas" subWeb consists of the following (in order not to clutter the Figure, not all of the details are shown):

1. tf (term frequency) of the page terms = { project: 60, Algorithm: 23, data mining: 15 …}
2. the number of inlinks: 11, the number of outlinks: 27, and the number of colinks: 3
3. hierarchy terms = {cs, uiuc, edu, research, areas}

The hierarchy terms can be used in the training of the classifier in the next step; when those terms appear in the page terms, the page terms are given additional weights for their semantic significance.



Creating Modified Naïve Bayesian Classifiers

A Naïve Bayesian classifier (NB) can be used to derive the classification rules from the labeled dataset. This classifier needs to be modified to account for the fact that Web data of a Web page has content (of the Web page) and links (i.e., other Web pages reached through outlinks from the Web page). We derive two NBs: one for the content data, and one for the link data.

Training the Modified Naïve Bayesian Classifiers

We use the modified NBs to perform two classification tasks: one on the labeled dataset and another on the unlabeled dataset. In other words, we apply a co-training algorithm [Blum and Mitchell 2000]. The modified NBs are first trained against the labeled training dataset. Continuing with our example of Figure 2, only the 10 sub-directories of the "Research Areas" page will be used to train the two modified NBs. Once the two classifiers have been trained, they will be applied against "unlabeled" data, that is, Web data of other subWebs (e.g., the "Research" subWeb of the Cornell University computer science department).

Some of the Web pages may be present in some of the "unlabeled" subWebs, but not in others; and some of key terms used may be different from subWeb to subWeb. For example, what is called "Algorithms and Theory" in the UIUC CS department homepage is called "Theory of Computing" in the Cornell University CS department homepage. The application of a "topic discovery" technique can address to some extent these types of heterogeneity in the subWeb. In particular, such a technique can group similar names into a common category, and assign names to "the unknown" categories, that is, the categories into which terms that are missing in other subWeb are thrown into.

The result of applying our methodology is a set of subWebs rooted in a Web page on "Algorithms and Theory" or comparable terms in the CS department homepages of Cornell University, Carnegie-Mellon University, and the Massachusetts Institute of Technology. It is shown in Figure 3. (Although the result includes the URLs of the Web pages of each subWeb, for simplicity, they are not shown in the figure. Further, the ancestors of each subWeb, namely, the CS department node and the "Research" node, are included in the figure only to make clear the context of the resulting subWebs.)

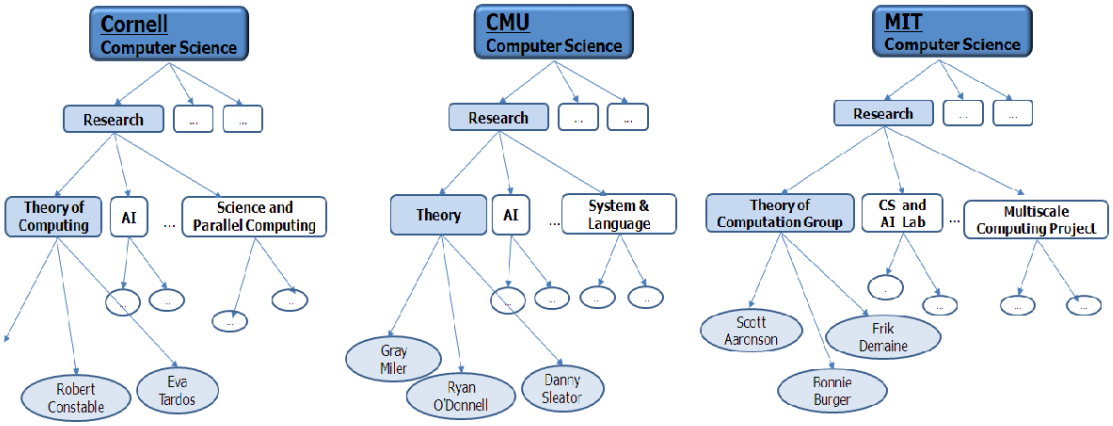


Figure 3. The subWeb Resulting from Applying the modified Naïve Bayesian classifers to three CS department homepages.

3 CONCLUDING REMARKS

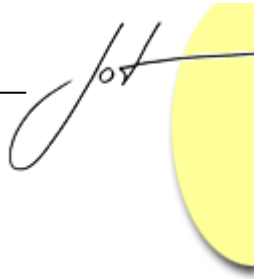
We outlined a methodology for mining or predicting hierarchical structures of subWebs in similar domains given the hierarchical structure of a subWeb. Some preliminary work to verify accuracy of the methodology has been encouraging. However, more extensive experiments need to be performed. Once verified, the methodology would help advance the current keyword-based search of the Web to the next level of sophistication, namely semantic-hierarchy-based search.

There are some technical issues to explore. Although we discussed training the classifiers using only one input subWeb, it may make sense to consider using a small set of subWebs as input in order to increase the accuracy of the classification rules. The heterogeneity in the structure and terms used in different subWebs in similar domains also needs to be explored further.

We have done preliminary work to verify the effectiveness of our methodology. We plan to report the results, along with the details of the methodology, in a forthcoming research paper.

ACKNOWLEDGMENTS

This research was supported by the MKE (Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Center) support program supervised by the IITA (Institute of Information Technology Advancement) (IITA-2008-(C1090-0801-0046))



REFERENCES

[Blum and Mitchell 2000] A. Blum and T. Mitchell: "Combining Labeled and Unlabeled Data with Co-Training," *Proceedings of the 11th Annual Conference on Computational Learning Theory(ICML)*, pp.327-334, 2000.

[Brin and Page 1998] Brin S. and Page L. The Anatomy of a Large-scale Hypertextual Web Search Engine. *In Proc. Of WWW7*, 107-117, Brisbane, Australia, April 1998.

[Buneman 1997] P. Buneman: "Tutorial: Semistructured Data," *Proceedings of ACM Symposium on Principles of Database Systems*, pp. 117-121, 1997.

[Chakrabarti 2002] S. Chakrabarti. *Mining the Web: Statistical Analysis of Hypertext and Semi-Structured Data*, Morgan Kaufmann, 2002.

[Dumais and Chen 2000] S. Dumais and H. Chen. Hierarchical classification of Web content. In Proc. *Int. 2000 ACM SIGIR Conf. Research and Development in Information Retrieval(SIGIR'00)*, pages 256-263, Athens, Greece, July 2000.

[Kim et al 2008] Won Kim, Ok-Ran Jeong, Hyungsuk Ji, and Sangwon Lee: "On Web Searches: Some Activities and Challenges", *Journal of Object Technology*, vol. 7, no. 2, March/April 2008 http://www.jot.fm/issues/issue_2008_03/column5/index.html.

[Lim et al 1999] J. M. Lim, H. J. Oh, S. H. Myaeng, and M. H. Lee: "Improving Efficiency with Document Category Information in Link-based Retrival," *Proc. Of the international Workshop on IRAL'99*, 1999.

[Nigam et al 2000] K. Nigam, A. K. Mccallum, S.Thrun, and T. Mitchell: "Text Classification from Labeled and Unlabeled Document using EM." *Machine Learning*, 39(2/3), pp.103-134, 2000.

[Raskutti et al 2002] B. Raskutti, H. Ferra and A. Kowalczyk: "Combining Clustering and Co-training to Enhance Text Classification Using Unlabeled data," *Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining(KDD)*, pp.620-625, 2002.

[Shen et al 2004] D. Shen, Z. Chen, Q. Yang, H. –J. Zeng, B.Zhang, Y. Lu, and W. –Y. Ma. Web-page classification through summarization. *In Proc. 2004 Int. ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR'04)*, pages 242-249, Sheffield, UK, July 2004.

[Shih and Karger 2004] L. K. Shih and D. R. Karger. Using urls and table layout for Web classification tasks. *In Proc. 13th Int. World Wide Web Conf. (WWW'04)*, pages 193-202, New York, NY, 2004.

[Sun et al 2002] A. Sun, E. P. Lim, W. K. Ng. Web classification using support vector machine. *In Proc. 4th Int. Workshop on Web information and data management (WIDM'02)*, pages 96-99, 2002

About the authors

Ok-Ran Jeong is a research professor with the School of Information and Communication Engineering at Sungkyunkwan University, Korea. She was a visiting scholar in the Department of Computer Science at the University of Illinois at Urbana-Champaign, and, before that, a post doctoral researcher in the Center for e-Business Technology in Seoul National University. She received a Ph.D. in computer science from Ewha Womans University. Her research interests include Web technology (Web architecture, Web mining, intelligent techniques) and u-commerce applications. She can be reached at orjeong@ece.skku.ac.kr.

Jiawei Han is a Professor in the Department of Computer Science at the University of Illinois. He has been working on research into data mining, database systems and data warehousing, with over 350 conference and journal publications. He has chaired or served in over 100 program committees of international conferences and workshops and also served or is serving on the editorial boards for Data Mining and Knowledge Discovery, IEEE Transactions on Knowledge and Data Engineering, Journal of Computer Science and Technology, and Journal of Intelligent Information Systems. He is currently the founding Editor-in-Chief of ACM Transactions on Knowledge Discovery from Data (TKDD). He has received three IBM Faculty Awards, the Outstanding Contribution Award at the 2002 International Conference on Data Mining, ACM Service Award (1999) and ACM SIGKDD Innovation Award (2004), and IEEE Computer Society Technical Achievement Award (2005). He is an ACM Fellow (2004). His book "Data Mining:Concepts and Techniques" (Morgan Kaufmann) has been used popularly as a textbook.

Won Kim is a Professor and Univeristy Fellow with the School of Information and Communication Engineering at Sungkyunkwan University, Suwon, S. Korea. He is Editor-in-Chief of ACM Transactions on Internet Technology (www.acm.org/toit). He is Global General Chair of the Human.Society@Internet International Conference. He is the recipient of the ACM 2001 Distinguished Services Award, and is an ACM Fellow. He can be reached at wonkim@skku.edu

Eunseok Lee is a Professor with the School of Information and communication Engineering at Sungkyunkwan University, Korea. He was an Assistant Professor at Tohoku Univ. in Japan. Before that, he was a Research Scientist in Information and Electronics Laboratory at Mitsubishi Electric Corporation. He received a Ph.D. in information engineering from Tohoku University, Japan. His research topics include methodologies in software engineering, autonomic computing, and Web-based agent technologies. He can be reached at eslee@ece.skku.ac.kr.