

Using Markov Models for Web Site Link Prediction

Jianhan Zhu, Jun Hong, John G. Hughes

School of Information and Software Engineering, University of Ulster at Jordanstown

Newtownabbey, Co. Antrim BT37 0QB, United Kingdom

{jh.zhu, j.hong, jg.hughes}@ulster.ac.uk

ABSTRACT

Markov models have been extensively used to model Web users' navigation behaviors on Web sites. The link structure of a Web site can be seen as a citation network. By applying bibliographic co-citation and coupling analysis to a Markov model constructed from a Web log file on a Web site, we propose a clustering algorithm called CitationCluster to cluster conceptually related pages. The clustering results are used to construct a conceptual hierarchy of the Web site. Markov model based link prediction is integrated with the hierarchy to assist users' navigation on the Web site.

Categories and Subject Descriptors

H.5.4 [Information Interfaces and Presentation (e.g., HCI)]: Hypertext/Hypermedia – navigation.

General Terms

Algorithms, Human Factors.

Keywords

Markov models, hierarchy, link prediction

1. INTRODUCTION

By viewing each Web user's navigation process on a Web site as a Markov chain, we can build a Markov model of the Web site using past users' traversals on the hyperlinks as their accumulated navigation behaviors. We construct a link graph, in which Web pages are the nodes, hyperlinks between pages are the links between the nodes, and the numbers of traversals on the hyperlinks by past users are the weights on the links from a Web log file on a Web site [8]. We assume that most users have followed the links to the pages that they are interested in. By viewing the weights on the links as past users' implicit feedback of their preferences in following the hyperlinks in each page [8], we can use the link graph to calculate a probability transition matrix containing one-step transition probabilities between the states in the Markov model. Bibliographic analysis is a quantitative method for the study of citations between scientific literatures. Almind and Ingwersen [1] contend that the World

Wide Web (WWW) is a citation network and bibliographic analysis can be applied to WWW. Hyperlinks between Web pages are conceptually similar to citation links. One page refers to another as in a bibliography, one paper refers to another. Based on the similarity between the Web link structure and a collection of cited and citing scientific papers, we try to extend co-citation and coupling analysis to the Markov model. We define co-citation similarity of two Web pages as a distance based measure of the one-step transition probabilities on their in-links, and coupling similarity of the two pages as a distance based measure of the one-step transition probabilities on their out-links. We propose a hierarchical clustering algorithm called CitationCluster to cluster Web pages to form conceptual clusters based on their co-citation and coupling similarities. Based on the citation relationships between the clusters and un-clustered individual pages, we can construct a conceptual hierarchy of the Web site, which has a hierarchical organization of information.

2. RELATED WORK

In [7], we presented PageRate algorithm to give search results ratings based on past users' accumulated navigation behaviors on Web sites, and PageClustering algorithm to cluster Web pages with similar in-links to form conceptual categories to integrate with search results. In [6], we used Markov models to find conceptual clusters of Web pages based on link similarities between Web pages. In [8], we used a transition matrix compression algorithm to compress the Markov model of a Web site to an optimal size for efficient link prediction on the Web site. Sarukkai [4] used Markov model based method for link prediction on Web sites. Kleinberg [2] proposed HITS algorithm to find authorities and hubs based on the Web link structure. Thimbleby [5] used Markov models to study Web site usability.

3. RESEARCH GOALS

We have two major goals in our research. First, is to use a Markov model built from the Web link structure of a Web site for clustering conceptually related Web pages. Second, is to use the clustering results for constructing a conceptual hierarchy of the Web site to integrate with link prediction for user navigation on the Web site.

4. RESEARCH APPROACH

4.1 CitationCluster Algorithm

Based on the similarity of hyperlinks between Web pages and citations between scientific literature, we try to extend co-citation and coupling analysis to the Web link structure of a Web site. We define co-citation and coupling similarities of two Web pages as distance based similarities of the transition probabilities on their

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

HT'02, June 11-15, 2002, College Park, Maryland, USA.

Copyright 2002 ACM 1-58113-477-0/02/0006...\$5.00.

in-links and out-links, respectively. Co-citation and coupling similarities are used for measuring the conceptual relationships between Web pages. We propose CitationCluster, a three-stage agglomerative hierarchical clustering algorithm, to find three kinds of conceptual clusters, namely, navigation, category, and reference clusters. A navigation cluster is defined as a group of Web pages hierarchically clustered together based on both their co-citation and coupling similarities. A category cluster is defined as a group of Web pages hierarchically clustered together based on their co-citation similarity. A reference cluster is defined as a group of Web pages hierarchically clustered together based on their coupling similarity. We calculate a similarity matrix consisting of both co-citation and coupling similarities between every pair of pages from the Markov model. Firstly, we use both co-citation and coupling similarities to cluster Web pages for navigation clusters. Secondly, we use co-citation similarity to cluster Web pages for category clusters. Thirdly, we use coupling similarity to cluster Web pages for reference clusters. Each cluster is given a title through conceptual learning [3] to conclude contents of all the pages in the cluster. There are also un-clustered pages after clustering.

4.2 Conceptual Hierarchy Construction

We use the navigation, category, reference clusters, and un-clustered pages as building blocks to construct a conceptual hierarchy of the Web site for navigation. We create semantic virtual links (as opposed to physical hyperlinks) between the building blocks based on the semantic relationships between them reflected in the transition probabilities between them in the Markov model. For a hierarchical Web site, the conceptual hierarchy has the homepage as the root, which links to a set of clusters and pages on a general concept level. Each of them links to a set of clusters and pages on a less general concept level, and so on to a set of clusters and pages on the most specific concept level. The conceptual hierarchy is visualized in a prototype called ONE (Online Navigation Explorer) [6] to assist user navigation on the Web site.

4.3 Link Prediction on Conceptual Hierarchy

Given the user's current page and a set of visited pages and clusters as history, we can use the Markov model to calculate the probabilities of visiting other clusters and pages in the conceptual hierarchy for link prediction. In ONE, we visualize the user history by expanding clusters containing visited pages, highlighting current page, and using icons to indicate their order of being visited. We highlight pages and clusters with the highest probabilities of being visited in the future, expand higher level clusters linking to pages and clusters with the highest probabilities, and use icons to show different levels of possibilities.

4.4 Approach Evaluation

We built a Markov model using a Web log file from University of Ulster Web site. CitationCluster was applied to the Markov model. The clustering results were used to construct a conceptual hierarchy of the Web site, which is visualized in ONE and integrated with link prediction to assist user navigation. Our

group members have used ONE to navigate the university Web site. Compared with link prediction presented in [4,8], the conceptual hierarchy has given them a clearer notion of their current locations on the Web site, and the semantic relationships among visited and recommended pages/clusters. We observed that this notion had contributed to improved efficiency and accuracy in finding Web pages that they are interested in during their visits to the Web site assisted by ONE.

5. CONCLUSIONS AND FUTURE WORK

A hierarchical clustering algorithm called CitationCluster is proposed to cluster conceptually related Web pages on a Web site based on co-citation and coupling similarities between Web pages defined on the transition probabilities on their in-links and out-links, respectively. The clustering results are then used to construct a conceptual hierarchy of the Web site. Finally, Markov model based link prediction is integrated with the hierarchy to assist user navigation in a prototype called ONE.

Link prediction in ONE needs to be evaluated by a larger user group. We plan to select a group of users including students, staff in University of Ulster, and people from outside the university to use ONE. Their interaction with ONE will be logged for analysis. We plan to use Web log files from some commercial Web site to build a Markov model for link prediction and evaluate the results on different user groups.

6. ACKNOWLEDGEMENTS

We would like to thank Mark Bernstein and Jayne Klenner for their valuable comments to previous versions of this paper.

7. REFERENCES

1. Almind, T. C. and Ingwersen, P., (1997). Informetric Analysis on the World Wide Web: Methodological Approaches to "Webometrics". *Journal of Documentation* 53, no. 4: 404-426.
2. Kleinberg, J. M., (1999). Authoritative sources in a hyperlinked environment. *Journal of ACM*, 46:604-632.
3. Perkowski, M. and Etzioni, O., (1999). Adaptive web sites: conceptual cluster mining. In *Proceedings of IJCAI 1999*.
4. Sarukkai, R. R., (2000). Link prediction and path analysis using Markov chains, WWW9, Amsterdam.
5. Thimbleby, H., Cairns, P., and Jones, M., (2001). Usability Analysis with Markov Models. *ACM Transactions on Computer-Human Interaction*, Vol. 8, No. 2, pp. 99-132.
6. Zhu, J., (2001). Using Markov Chains for Structural Link Prediction in Adaptive Web Sites. In *Proc. of User Modeling 2001*, pp. 298-300.
7. Zhu, J., Hong, J., and Hughes, J., (2001). PageRate: Counting Web Users' Votes. In *Proc. of ACM Hypertext'01*, pp. 131-132.
8. Zhu, J., Hong, J., and Hughes, J., (2002). Using Markov Chains for Link Prediction in Adaptive Web Sites. In *Proc. of Software 2002: Computing in an Imperfect World*, Springer-Verlag LNCS 2311, pp. 60-73.