Automatic Creation of Links: An Approach Based on Decision Tree

Peng Li[1] and Seiji Yamada[2]

[1] CISS, IGSSE, Tokyo Institute of Technology
4259 Nagatuta, Midori-ku, Yokohama, Japan
liorlee@nii.ac.jp
[2] National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan
seiji@nii.ac.jp

Abstract. With the dramatic development of web technologies, tremendous amount of information become available to users. The great advantages of the web are the ease with which information can be published and made available to a wide audience, and the ability to organize and connect different resources in a graph-based structure using hyperlinks. However, most of these links are created manually and the page that the link represents must be known to the author of the link. In this paper, we propose a decision-tree-based approach to solve this problem. We set up a system that gathers information about the candidate pages, evaluates them and creates links to them automatically.

1 Introduction

The World Wide Web contains over 2 billion pages and this number is growing at a breakneck speed. Links between these pages are generated and removed very constantly. Over the past years, a number of approaches such as Adaptive Web sites [1] , FWEB [2] , ARC [3] and some link- or topology-based approaches[4][5][6] have been developed to improve link structure.

In most of the cases, *official sites* (organizations, products, people, services, facilities, events, etc) possess links to the other sites that deal with the same topics. These links are created manually. This can be a disadvantage if information in a particular field is incomplete and expanding rapidly over time and where a page author cannot be expected to know which pages are the most appropriate to link to and when they become available.

Our aim is to create these links automatically. We focus on official sites because these sites are usually so popular that other pages which deal with the same topics are willing to be linked from them. In this research, our targets are the pages that have sent link requests to the specified official sites. We call these pages *link request pages*. When receiving a request, our system will gather information about this page such as its access amount, maintenance frequency, page similarity, etc. We set up a decision tree to handle these information and determine whether the link request page should be linked or not. Links are created, modified and deleted automatically with the changes of link request pages.

L. Wang and Y. Jin (Eds.): FSKD 2005, LNAI 3614, pp. 1200 – 1203, 2005.

2 Automatic Link Creation System Based on Decision Tree

2.1 System Overview

The purpose of our system is to create and modify links on the links page automatically. Broken links are removed and new links are added. Moreover, with the growing popularity and richness of the contents, the position of a link will rise and vice versa.

2.2 Decision Tree

A decision tree is a tree-shaped structure that represents a set of decisions. These decisions generate rules for the classification of a dataset. In a decision tree, each "branch node" (feature) represents a choice between a number of alternatives, and each "leaf node" (class) represents a class or decision. OC1 (Oblique Classifier 1) [7] is a decision tree induction system designed for applications where the instances have numeric (continuous) feature values. OC1 builds decision trees that contain linear combinations of one or more attributes at each internal node; these trees then partition the space of examples with both oblique and axis-parallel hyperplanes. OC1 has been used for classification of data representing diverse problem domains, including astronomy, DNA sequence analysis and others. In this research, we use OC1 to construct decision trees because the values of our features and classes are numeric.

2.3 Evaluation Features

We use 4 types of features in this system, access amount, maintenance frequency, page similarity and contents richness. These features are chosen because they are very important characteristics for the evaluation of a page. On the other hand, these information are extractable with current web technologies. We plan to add a few more features to raise the accuracy of our system in the evaluation experiment.

2.4 System Construction

Fig.1 shows the system construction. The details of several modules are described below.

Training Data. Training data are used to construct a decision tree. They are information of a set of valued pages. Each page has its class value given by the administrator and 4 attributes values gathered by web robots.

Test Data. Classification is performed according to test data. They are information of a set of unvalued pages. Each page only contents 4 attributes values.

Decision tree. A decision tree is constructed by OC1 through training data. When unvalued pages are given, the decision tree classifies these pages to give each page a class value.

Check Engine. When check engine receives requests from link request pages, it sends web robots to gather 4 types of attributes from these pages. Usually, this information is sent to the decision tree as test data. But if the administrator evaluates some pages, they are given as training data to construct a decision tree. The check engine will gather information from link request pages regularly, even if no new request is detected.

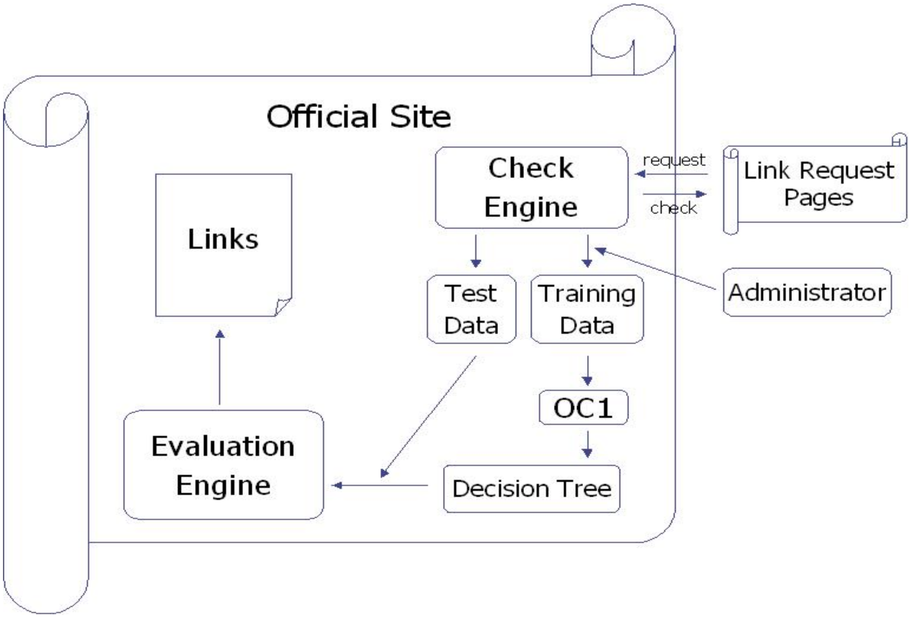

Fig. 1. System Construction

Evaluation Engine. Evaluation engine sorts the link request pages by their class values and creates Links.

2.5 Links Creation Flow

(1) Link request pages send requests to the Official Site.
(2) Check engine sends web robots to collect information about the link request pages.
(3) If the decision tree has been constructed, go to (6).
(4) The administrator evaluates some pages.
(5) Construct a decision by OC1.
(6) Use the decision tree to classify unvalued pages.
(7) Links are created by evaluation engine.

This Flow starts regularly from (2) even if there is no request detected.

It will work from (4) if the administrator gives new evaluations.

3 Conclusion

We proposed an automatic link creation system base on decision tree. We focused on the links page that is a part of nearly all of the official sites. Our system gathers information about the link request pages, evaluates them and creates links to them. We are preparing an evaluation experiment to inspect the effectiveness of our system. More evaluation features will be added at that time. Further discussion about our evaluation metrics need to be performed.

References

1. Perkowitz, M., Etizoni, O.: Toward Adaptive Web Sites: Concept and Case Study. Artificial Intelligence Journal, vol.118 (2000) 245-275
2. Courtenage, S.,Williams, S.: Automatic Hyperlink Creation Using P2P and Publish/Subscribe. Workshop on Peer-to-Peer and Agent Infrastructures for Knowledge Management (PAIKM) (2005)
3. Chakrabarti, S., Dom, B., Raghavan, P., Rajagopalan, S., Gibson, D, Kleinberg, J.: Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text. Proceedings of the Seventh International World Wide Web Conference (WWW7) (1998)
4. Page, L., Brin, S., Motwani, R., Winograd, T.: The Pagerank citation ranking: Bringing order to the Web. Technical report, Stanford University (1998)
5. Kleinberg, J.: Authoritative sources in a hyperlinked environment. In Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms (1998) 668—677
6. Phelan, D., Kushmerick, N.: A descendant-based link analysis algorithm for web search (2002)
7. Murthy, S., Kasif, S., Salzberg, S., Beigel, R.: Randomized induction of oblique decision trees. In Proceedings of the Eleventh National Conference on Artificial Intelligence (1993) 322-327