Refining the Results of Automatic e-Textbook Construction by Clustering

Jing Chen\textsuperscript{1}, Qing Li\textsuperscript{1}, and Ling Feng\textsuperscript{2}

\textsuperscript{1} Department of Computer Engineering and Information Technology, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong
\{jerryjin, itqli\}@cityu.edu.hk
\textsuperscript{2} Department of Computer Science, University of Twente, PO Box 217, 7500 Enschede, The Netherlands
ling@cs.utwente.nl

Abstract. The abundance of knowledge-rich information on the World Wide Web makes compiling an online e-textbook both possible and necessary. The authors of [7] proposed an approach to automatically generate an e-textbook by mining the ranking lists of the search engine. However, the performance of the approach was degraded by Web pages that were relevant but not actually discussing the desired concept. In this paper, we extend the work in [7] by applying a clustering approach before the mining process. The clustering approach serves as a post-processing stage to the original results retrieved by the search engine, and aims to reach an optimum state in which all Web pages assigned to a concept are discussing that exact concept.

1 Introduction

The World Wide Web has evolved into one of the largest information repositories. It now becomes feasible for a learner to access both professional and amateurish information about any interested subject. Professional information often includes compiled online dictionaries and glossaries, course syllabus provided by teachers, tutorials of scientific software, overview of research areas by faculties from research institutes, etc. Discussion boards sometimes offer intuitive description of the interested subjects, beneficial for students or beginning learners. All these resources greatly enrich and supplement the existing printed learning material. The abundance of knowledge-rich information makes compiling an online e-textbook both possible and necessary.

The most common way of learning through the Web is by resorting to a search engine to find relevant information. However, search engines are designed to meet the most general requirements for a regular user of the Web information. Use Google ([1]) as an example. The relevance of a Web page is determined by a mixture of the popularity of the page and textual match between the query and the document ([2]). Despite its worldwide success, the combined ranking strategy still has to face several problems, such as ambiguous terms and spamming. In the case of learning, it becomes even harder for the search engine to satisfy the need of finding instructional
information, since the ranking strategy cannot take into account the needs of a particular user group, such as the learners.

1.1 Background of Research

Recently, many approaches have been proposed to improve the appearance of Web search engine results. A popular solution is clustering, providing users with a more structured means to browse through the search engine results. Clustering mainly aims at solving the ambiguous search terms problem. When the search engine is not able to determine what the user’s true intention is, it returns all Web pages that seem relevant to the query. The retrieved results could cover widely different topics. For example, a query ‘kingdom’ actually referring to biological categories could result in thousands of pages related to the United Kingdom. Clustering these results by whole pages or their snippets is the most commonly used approach to address this problem ([3][4][5]). However, the structure of the hierarchy presented is usually determined on-the-fly. Cluster names and their organized structure are selected according to the content of the retrieved Web pages and the distribution of different topics within the results. The challenge here is how to select meaningful names and organize them into a sensible hierarchy. Vivisimo [6] is an existing real-life demonstration of this attempt.

The clustering approach works well to meet the needs of a regular user. But when the application is narrowed down to an educational learning assistant, it is possible to provide the learners with more ‘suitable’ Web pages that satisfy their needs in the pursuit of knowledge. Users seeking for educational resources prefer Web pages with a higher quality of content. Such Web pages often satisfy the criteria of being self-contained, descriptive and authoritative [7]. Limited work has been done to distinguish higher quality data from the Web. An important one is [8], where the authors attempt to mine concept definitions on the Web. They rely on an interactive way for the user to choose a topic and the system automatically discovers related salient concepts and descriptive Web pages, which they call informative pages. They not only proposed a practical system that successfully identified informative pages, but also more importantly pointed out a novel task of compiling a book on the Web.

In [7], the authors proposed an approach to automatically construct an e-textbook on the Web. They extend Liu et al.’s work by adding a concept hierarchy that outlines the user-specified topic. In the concept hierarchy, also called a concept tree, each node corresponds to a concept and the ancestor relationship of nodes represents the containing relation of the concepts. The use of the concept tree is essential and benefits the learning experience to a great extent. The concept tree is used to gather Web pages that are more likely to be of learning importance. It also readily serves as a table-of-content for the final e-textbook. It is easier for the users to understand compared with the cluster hierarchy generated on-the-fly, thus saves time for browsing. The approach is described concisely in the following:

1. Dataset collection: The concepts in the concept tree are used to generate a query phrase for each node. The query terms indicate the relationship of concepts. Web pages that cover more concepts in the query are more likely to be ranked high in the list.