University of Macau

Abstract

WEB IMAGE CLUSTERING AND RETRIEVAL

by U Leong Hou

Thesis Supervisor: Associate Professor, Gong Zhiguo
E-Commerce Master

This thesis presents the techniques for web image information retrieval with Web mining technologies. The whole system includes several modules, such as an Image Search Engine, a clustering models (clustering by text) and a link analysis model (based on the web page segmentation). Image Search Engine includes several components such as a crawler, a preprocessor, a semantic extractor, an indexer, a knowledge learner and a query engine. The semantic extractor is fundamental and it extracts the text representations for Web images which can be used for Web image clustering. Semantic extractor extracts the distance and the importance for each semantic block to the corresponding image, and each block can represent the relationship between the text and the image. Based on this knowledge, it helps much in the later steps.

We know that automatic categorization of all objects on the web is a goal of information retrieval system. Many researches work on this purpose. Such as document classification techniques and document clustering techniques. Our clustering models provide a new method of a successful web image clustering. In our first model, we use the HTML content (text) and structure to classify web image into several clusters. The clustering algorithm is based on a wide famous algorithm – *CHAMELEON*. Our model includes the following parts: data representation model, clustering model and analysis interface. Data representation model can convert the web documents to a set of useful information, help the later process. Clustering model use *CHAMELEON* that split the large set of data into several sub-clusters, and assign

the web images into corresponding clusters based on the similarity measurement. The combination of these parts creates an underlying model for robust and accurate similarity calculation that leads much improved results in web image clustering over traditional methods.

In the last part of my research, we use the web link structure to increase my system performance. There are various link-based ranking strategies developed in the recent years for improving Web-search query results. And HITS and PageRank are two typical algorithms for link structure analysis. Actually, the successful HITS and PageRank are based on two assumptions: (1) two pages are similar if there is a link between them, and (2) two pages are similar if they have common co-cited pages. However, it is a common situation to find a Web page with multiple topics in the explosive amount of the Web. And the information consumer's needs be well covered by only the part (other than the whole page) of the pages with the target topic. After survey on some existing researches, we found that block based link analysis is better than traditional page based link analysis. In this research, we provided a new web page segmentation method based on the text semantic cohesion. This new method shows a high performance both on accurate and processing time. After finished the block based link analysis, I apply the block based link analysis to the system to increase the quality of web-search query results.