Abstract

The World Wide Web (WWW) provides vast resource for information of almost all types. Users commonly use search engine or follow related links to retrieve the information. However, searching information using search engine is not effective because it will provide tedious data and so many related links which are wasting time to read it one by one, even sometimes the result is not related at all to what have user entered. After experiment, it's discovered that web pages that have same information will have same structure too, moreover due to loose standard of web page publishing, different authors can use different wordings (labels) which describe the same information.

This thesis builds a system that can classify web pages by class using label discovery algorithm (LDA). First, LDA will find labels or words that represent class of web pages, so that would be obtained the structure of class, finally the structure will be used for classifying web pages.

Testing results show that LDA can be used for finding labels or words that represent class of web pages and the structure obtained by this method can classify web pages accurately.

Keywords: *label discovery, structure, classification, class, web pages, similarity function*