Review of “Data Mining for Hypertext: A Tutorial Survey”

Xiaoguang Qi

Starting from the models used in web mining, the author reviewed the research that applying techniques from data mining to mining the data on the web.

In general, this is a well-written paper. It covers most outstanding contributions that have been made in web mining. In addition, the material is well-organized.

In my opinion, however, the paper could be better if following improvements are made.

As an important semi-supervised learning method, co-training may be mentioned in s5.

When talking about PageRank, although it is minor issue, it might be better if the paper also mentions about how to deal with the situation that a page has no outgoing links.

When talking about Google search engine in s6.1.1, it is possible that readers get the idea that Google ranks pages solely based on PageRank. It is known that Google’s ranking scheme is based on many factors, with PagrRank being one of them.

Given that the survey was written in 2000, it could not cover recent research in this area. Otherwise, recent progress in this area such as “Topic-Sensitive PageRank” could have been included.

A typo in the last line of the second paragraph of s3, where “maintining” should be “maintaining”.

In conclusion, this paper is an informative introduction in web mining as well as a good “hub” which points to many authoritative papers.