Abstract

In this thesis, a data mining approach for query refinement is proposed using Association Rules (ARs) among keywords being extracted from a document database. When a query is underspecified or contains ambiguous keywords, a set of association rules will be displayed to assist the user to choose additional keywords in order to refine his/her original query. To the best of our knowledge, no reported study has discussed on how to screen the number of documents being retrieved using ARs. The issues concerned in this thesis are as follows. First, an AR, $X \Rightarrow Y$, with high confidence will intend to show that the number of documents that contain both sets of keywords X and Y is large. Therefore, the effectiveness of using minimum support and minimum confidence to screen documents can be little. To address this issue, maximum confidence is used. Second, a large number of rules will be stored in a rule base, and will be displayed at run time in response to a user query. In order to reduce the number of rules, this thesis introduces two co-related concepts: "stem rule" and "coverage". The stem rules are the rules by which other rules can be derived. A set of keywords is said to be a coverage of a set of documents if these documents can be retrieved using the same set of keywords. A minimum coverage can reduce the number of keywords to cover a certain number of documents, and therefore can assist to reduce the number of rules to be managed. In order to demonstrate the applicability of the proposed method, a prototype system has been built, and a medium-sized document database is maintained. The effectiveness of using ARs to screen will be addressed in this thesis as well.