Chapter 6

Conclusion

This thesis proposes a query refinement support method for a document retrieval system by mining ARs among keywords from a document database.

In very large databases, the task of refining an effective query is difficult in the sense that it requires users, without any knowledge about the collection of documents. The method of this thesis can support query refinement by displaying information about the collection of documents to users. Query refinement relates to data mining in databases because it is not only the process of retrieval but also the process of discovering knowledge that can be displayed to users as a guide for improving query.

The focus of ARs is based on finding relationships among data items. The method of ARs are similar with ones of related researches in information field, which are based on co-occurrence between keywords. The researches of information field use sym-

metric similarity to estimate relationship between keywords and choose the keywords with higher similarities as refinement candidates. This leads to bad effectiveness of screening and has the difficulties to differentiate between relevant and non-relevant documents. ARs use support and confidence estimate relationship between keywords and choose the keywords with higher support and confidence as refinement candidates. The support makes rules meet certain frequency and the confidence describes degree of relation. Unlike the similarity used in related researches, confidence of ARs is not symmetric. According to not symmetric confidence and a new threshold of maximum confidence used in this research instead of minimum confidence of usual ARs, refinement candidates have better effectiveness of screening.

One bottle neck of using ARs for query refinement is that too much ARs are generated, because user will have to take the responsibility of browsing and choosing from a large number of refinement candidates. To address this problem, this thesis introduced the concept of *stem rule*. By stem rule, the number of rule are reduced in generating and storing rules.

The technique of *stepwise refinement* by using stem rules effectively displays refinement candidates for user to browse easily them and reduces the time spent on displaying rules because non-stem rules need not be derived.

The method of stem rule reduces the number of ARs under the condition of certain minimum support. Minimum support is still a major factor that influences the number of ARs. Setting minimum support to a large value can reduce the number of ARs, but in the meantime causes the problem of coverage. The concepts of *minimum coverage* was introduced, in order to reduce the number of refinement candidates, instead of minimum support used in stem rule and usual ARs. The experiment shows that minimum coverage may reduce refinement candidates. That is, we greatly reduce the refinement candidates a user can choose to refine his/her query without lost any documents he/she wants to access.

A query refinement space based on coverage was defined in this thesis. All concept of query, query evaluation, and query refinement can be described within the query refinement space. An algorithm generating query refinement space was proposed. The calculation cost is dramatically reduced by comparing this algorithm with usual methods [Agra93] which generates not only a large number intermediate rules but also causes the problem of coverage.

The prototype system confirms the effectiveness of reducing the number of refinement candidates and improving result of retrieval by the stem rules and the minimum coverage. By the user interface of the prototype system, an user's query can be refined in an

interactive way.

The experiment shows that coverage always has higher recall and precision than other approaches because this system guarantees the coverage to be 100% and a user can interactively choose what he/she wants in order to refine his/her query.

As future work, the effectiveness of various coverage on retrieval results will be investigated. A query may have more than one of coverages which are sets of refinement candidates. According to the coverages, the results retrieved by user who refers to refinement candidates based on these coverages are different. The coverages having better retrieval result will be chosen as refinement candidates. The number of refinement candidates and the effectiveness of these refinement candidates for the precision will also be investigated. Reducing the number of refinement candidates makes user browse easily them and the time spent on displaying refinement candidates will be reduced. But this may effect the retrieval result because user only refers part of information he/her want to have. The efficiency of the algorithm generating query refinement space will be reformed in future work.