Resumo

A teoria dos conjuntos difusos e a teoria dos conjuntos aproximativos não são tópicos novos na área da recuperação de informação. No entanto, a combinação das duas teorias não tem sido objecto de um estudo aprofundado na área. A motivação desta pesquisa foi a exploração desta combinação num sistema experimental de recuperação de informação no qual foram aplicadas, em diferentes fases do processo, as teorias dos conjuntos difusos e dos conjuntos aproximativos. Uma estratégia de indexação difusa foi usada para a descrição de documentos, baseada num tesauro criado a partir da colecção de documentos com recurso a uma medida estatística. Os documentos recuperados foram agrupados em clusters usando uma abordagem no domínio da teoria dos conjuntos aproximativos. Os algoritmos foram testados com recurso a colecções de teste. Defende-se nesta dissertação que as teorias citadas são aplicáveis à resolução dos problemas de recuperação em vários dos seus estádios. Apresentam-se os fundamentos das duas teorias em estudo, uma análise dos modelos de recuperação actuais, a arquitectura de um sistema de recuperação proposto, os detalhes da sua implementação, os resultados obtidos e as conclusões para as diferentes abordagens.

Os resultados obtidos não foram inteiramente satisfatórios, tendo sido efectuados desenvolvimentos sobre o protótipo inicial que conduziram a uma melhoria na precisão obtida. No entanto, esta continua a ser muito baixa comparativamente a outros métodos. As conclusões apontam para a necessidade de serem efectuados testes suplementares com recurso a outras colecções de teste que apresentem menor uniformidade temática e a outras medidas que produzem uma melhor definição da semântica dos termos. Foi estabelecido que as teorias dos conjuntos difusos e aproximativos são aplicáveis na recuperação de informação, devendo um sistema de recuperação fundamentado nestas teorias apresentar uma maior eficácia quando usado em colecções em que documentos de dimensão adequada sejam descritos por termos suficientemente heterógeneos.

Abstract

Rough sets and fuzzy sets are not a new topic in Information Retrieval (IR) and have been applied in different ways in the field, in order to understand their appropriateness to IR models. However, the combinations of these two theories in this domain have not been extensively studied. The motivation of this research is to explore one such combination in an experimental IR system. An information retrieval system has been designed and implemented applying the theories of fuzzy sets and rough sets in different parts of the retrieval process. A fuzzy indexing strategy has been used to describe documents, based on a fuzzy thesaurus generated from the document collection using a statistical measure; retrieved documents were clustered using a rough set approach. The algorithms were tested using standard information retrieval test collections. This thesis presents an outline of both theories,

the current logical models proposed in the IR field, the architecture of the system, implementation details, test results, and conclusions for the different approaches.

The conclusions point out that further testing should be made using different measures for the fuzzy indexing scheme and different testing collections, in order to induce a more adequate term semantic, since the results obtained were not completely satisfactory, due to a low precision induced by a relatively high number of retrieved documents. Improvements have been made into the original prototype that induced a higher precision. Despite the effort, the precision obtained is still very low compared to other methods, namely the space vector model implemented on the SMART system. However, a better effectiveness can be establish using the combinations of both theories if using proper documents description by means of usage of more heterogeneous terms.