分类号_____ 学 号 M070962

U D C_____ 密 级_____

学 位 论 文

基于语义的协作信息检索研究

盛艳

指导教师姓名：李云教授，扬州大学，江苏扬州，225009

申请学位级别：硕 士 学科专业名称：计算机应用技术

论文提交日期：2010.4 论文答辩日期：2010.5

学位授予单位：扬 州 大 学 学位授予日期：_____

答辩委员会主席：_____

论 文 评 阅 人：_____

_____

2010年 4月

Research on collaborative information retrieval based on semantic

Thesis submitted to Yangzhou University

In partial fullfilment of the requirements for the
Master Degree of Computer Science

By
Yan Sheng

Department of Computer Science, School of Information Engineering,
Yangzhou University

Supervisor: Professor Yun Li
April 2010

摘　要

形式概念分析自 1982 年由德国的 Wille 教授提出以后，近年来被广泛用于软件工程、知识发现、信息检索等领域。形式概念分析中的核心数据结构概念格通过 Hasse 图来表现出概念之间的层次关系。概念作为人的思想和知识的基本单元，一直以来，深受哲学界和科学界的重视，很自然地，也就成为了人工智能学科的重要研究对象，这主要体现在知识表示和机器学习等领域。将概念格理论和其他理论，如粗糙集理论、语义本体理论和文本特征提取方法结合起来，并将它们应用到基于语义的协作信息检索中，能更好的提取出概念之间的本质联系。在分布式环境下，使用 FCA 理论和技术进行协作信息检索，这可大大提高检索时间和效率。本文主要是基于概念格和相关理论知识来解决语义信息检索问题。主要研究工作包括：

1) 结合人类语言中普遍存在的 Zipf 定律，对大量文本集进行局部和全局的特征提取。根据文本各个特征自身和类别信息的统计特性，提取更少的特征来尽可能的表达出文本蕴含的信息，得到一种有效的文本特征提取方法，其中也对文本特征进行约简从而得到少而精的特征信息。
2) 基于概念格的数学理论基础，结合粗糙集理论和语义本体理论，提出了一个形式概念相似度计算方法，改进了基本 Tversky 相似度模型。从形式概念的结构相似和语义相似两个层次上衡量概念间的相似度，并将两者结合起来构成最终的形式概念相似度衡量方法以尽可能地体现形式概念的真实相似度。
3) 针对分布式环境，提出并构建一个协作信息检索框架，实现从结构和语义两个层次上的概念匹配，获得更符合用户需求的检索结果。
4) 在对检索结果排序过程中，利用语义本体 WordNet 对用户历史检索记录进行分析，从而构建用户兴趣模型，以便更精确的表达出用户兴趣所在，并按照此模型对初始检索结果重新排序。

关键字：形式概念分析，粗糙集，语义本体，特征提取，协作信息检索

Abstract

The formal concept analysis has been greatly applied to many fields, such as software engineering, scientific discovery and information retrieval in recent years since it was presented by professor Wille. Concept lattice, which can uncover the relationship between concepts through Hasse Diagram, is the core of formal concept analysis. As the basic unit of human's thinking and knowledge, the concept has been deeply focused by philosophers, scientists and become an important subject of artificial intelligence. Combining FCA with Rough set, Semantic Ontology and other theory， we can obtain the nature relationships of concepts better in semantic information retrieval. The paper is focused on the semantic information retrieval on the basis of FCA and relative theory. The main work includes:

1) In order to extract fewer features to express the information in the text as much as possible, the paper analysis the various features' statistical properties and to extract the global features according to Zipf's law; and then, based on the statistical analysis of the features' classified information, the efficient feature is extracted by computing the contribute of a feature; After that, the traditional TF-IDF formula is improved using category frequencies named by TF-IDF-CF for calculating the feature weight; Finally the text classification method is proposed.
2) Based on the basic concept lattice mathematical theory, combining the rough set theory and semantic ontology, this paper advanced a formal concept similarity measurement by improving the original Tversky similarity model. The method used the hierarchical information of concept lattice in the lattice structure and its own semantic meaning to measure the similarity of from the structure information of the

concepts and semantic information respectively, which reflect the true similarity of formal concept to some extent.

3) In Collaborative Information Retrieval, we construct formal concept lattices for a number of sub-text databases by the progressive lattice-built algorithm. Then, respectively, we obtained the new formal concepts by merging the temporary formal concepts which are from the different lattices by measuring similarity between formal concepts and next, matching the new concepts to return the results which are to meet the user needs ultimately. The whole framework is easy to deploy in a distributed environment and match the query words in an imprecise way. It reflects the needs of humanity better.

4) Personalized search is more suitable to return search results for user based on different user's search history and improve the search efficiency and user's retrieval experience. The paper put forward a search results ranking method that uses User Interest Model based on a semantic ontology WordNet, so that it returns the search results in accordance with user interestingness. Thereinto, different types of words are given by different score used different scoring mechanism, in order to obtain user interest model more accurately then provide a basis for measurement for ranking the original search results. The method advanced in this paper is proved to be practical and feasible. It improves the search effect and user's search experience to some extent.

Keywords: Formal Concept Analysis, Rough set, Semantic Ontology, Feature Selection, Collaborative Information Retrieval

目 录