第 34 卷 第 20 期 计 算 机 工 程 2008 年 10 月
Vol.34 No.20 Computer Engineering October 2008
·开发研究与设计技术· 文章编号：1000—3428(2008)20—0277—03 文献标识码：A 中图分类号：N945

基于本体的证据分析工具研究与实现

曹茂诚[1,2]，王英龙[2]，王金栋[2]，张淑慧[2]
(1. 山东轻工业学院信息科学与技术学院，济南 250353；2. 山东省计算中心，济南 250014)

摘 要：在研究传统证据分析工具不足的基础上，结合模糊本体概念，提出一种基于模糊本体的证据分析方法。该方法利用模糊本体中的隶属度概念，对查询语句和文档关键词向量空间模型进行模糊本体概念映射及概念相似度计算。通过数值实验对该方法进行性能分析。实验证明了该方法的有效性。

关键词：信息检索；证据分析；本体；模糊本体

Research and Implementation of Evidence Analysis Tool Based on Ontology

CAO Mao-cheng[1,2], WANG Ying-long[2], WANG Jin-dong[2], ZHANG SHU-hui[2]
(1. College of Information Science and Technology, Shandong Institute of Light Industry, Jinan 250353;
2. Shandong Computer Science Centre, Jinan 250014)

【Abstract】Based on the deficiency of traditional evidence analysis tools and the concept of fuzzy ontology, this paper presents a novel evidence analysis method. In this method, concept mapping and similarity calculating are made based on the membership degree concept of fuzzy ontology, which makes it possible to analyze documents on concept level. The experiments are carried out to test its performance, and the result shows its efficiency.

【Key words】information retrieval; evidence analysis; ontology; fuzzy ontology

1 概述

目前常用的证据分析工具大多是基于人工分类目录或关键词匹配。前者对海量信息资源的揭示效率不高、深度有限；后者在信息的语义揭示上有局限性，缺乏知识处理和理解能力。如何构造一个能更准、更快、更全地查找到所需信息的证据分析工具成为计算机取证领域急需解决的问题。本体的提出使文本分析技术从基于关键词层面提升到语义概念层面。国外学者在本体领域进行了各种深入探索，并取得一些成果。文献[1]在基于应用本体的基础上，提出一种从非结构化文档中提取信息的方法；文献[2]提出一种本体自动学习的方法；文献[3]构造一个名为 S-CREAM 的系统框架，该框架能在特定领域内完成本体的自动学习。国内也有学者在研究如何将本体应用于信息检索领域。我国学者徐振宁等人把本体作为信息检索系统的核心，通过构造形式化的领域本体，提出一种将知识表示和知识处理引入互联网信息处理的方法，为互联网上半结构化数据和关系数据库提供统一的语义模型；重庆大学的张海英等人在基于关键词和概念分析的基础上，提出一种基于语义概念模型检索的向量空间模型；文献[4]对基于模糊背景生成模糊本体进行研究，并提出一种基于模糊聚类技术来生成模糊本体的方法。

本文基于模糊逻辑，提出一种基于模糊本体的证据分析工具方法，利用领域本体的语义信息对磁盘文本和用户检索条件进行分析推理，得出较为准确的结果。

2 证据分析工具的基本框架

本文通过模糊本体，构建一个证据分析系统，系统框架如图 1 所示。

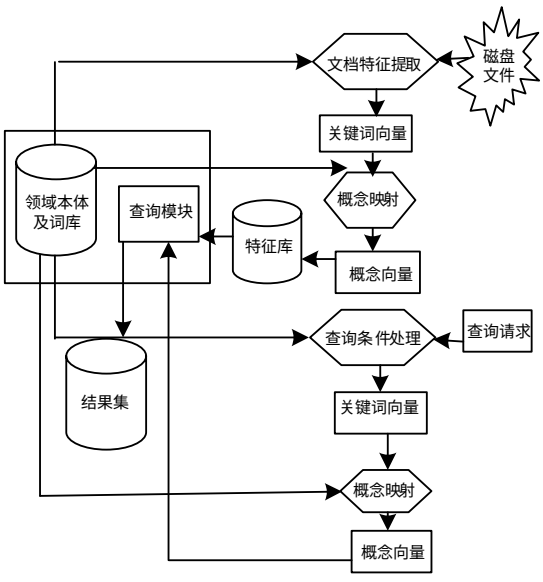

图 1 证据分析工具的体系结构

基金项目：山东省自然科学基金资助项目(Q2005G0)
作者简介：曹茂诚(1981－)，男，硕士研究生，主研方向：计算机网络，计算机取证；王英龙，研究员、博士生导师；王金栋，副研究员；张淑慧，助理研究员
收稿日期：2007-11-25 **E-mail：**caomch@keylab.net

知识库中的分词词库采用已有词典，计算机取证领域模糊本体，由熟悉该领域的专家构建。在此基础上，本系统的具体工作流程如下：

(1)通过与知识库中词库的交互，对磁盘文件进行文档特征提取、分词、频率统计，最终形成关键词向量；

(2)与领域模糊本体交互，实现由关键词向量向概念向量的概念映射转换，并提交给特征库，最后提交给查询模块；

(3)用户或应用软件由查询用户接口输入查询条件，查询条件处理模块通过与领域模糊本体的映射，将各种查询条件转换为结构化查询条件，并通过模糊本体实现关键词向量的概念映射，并提交给查询模块；

(4)查询模块完成概念向量的比对工作，并返回一系列满足条件查询的实例，这些实例经过结果处理模块处理后，存放到结果集，并最终返回给用户。

3 技术实现

3.1 知识库的构建

知识库是关于实体形式化知识的实现，它在整个系统中占有核心地位[5]。

在知识库中，本体的构建是关键。模糊本体和一般本体最大的区别在于，模糊本体具有隶属度，在处理不确切的信息上，更具弹性和优势。

定义 1 模糊背景是一个 3 元组 $K=(O,D,I)$，其中，O 是对象集；D 是属性集；I 是域 $G\times M$。每对关系$(o,d)\in I$ 有隶属度值 $I(o,d)$，值落在[0,1]中。

定义 2 设(O_1,D_1), (O_2,D_2)是模糊形式背景(O,D,I)的 2 个模糊概念，当且仅当 $O_1 \subseteq O_2$，则有 $D_2 \subseteq D_1$，则(O_1,D_1)是(O_2,D_2)的子概念，(O_2,D_2)是(O_1,D_1)的超概念。

定义 3 模糊形式概念(O_1,D_1)和其父、子概念(O_2,D_2)的相似度计算为

$$E_{\text{概念相似度}}(c_1,c_2)=\frac{\delta(o_2)}{\delta(o_1)}$$

$$\delta(O_i)=\sum_{o\in O_i}\mu(o,D_i),\mu(o,D_i)=\min_{d\in D}I(o,d)$$

在文献[2]的基础上，对模糊本体产生构架进行改进，具体流程如下：

Step1 在资料库中获得信息，进行模糊形式概念分析，构建模糊概念。

Step2 进行概念相似度计算以及模糊概念聚类，得到模糊概念层次。

Step3 由模糊概念层次映射产生模糊本体。

Step4 将模糊本体转换为 OWL 语言，使本体可被其他应用程序使用，进而构建语义网构架。

步骤 1 中的模糊形式概念分析，包含从不确定数据的数据库构造模糊形式背景、从模糊形式背景构造模糊概念；步骤 2 中的模糊概念聚类，采用一定的聚类技术、依据节点间的相似度，对模糊概念上的概念进行聚类，生成模糊概念聚类集合，进而生成模糊概念层次；步骤 3 中的映射又分为：

(1)给模糊概念层次中的概念节点一个标识，每个标识名对应模糊本体中的一个类名，模糊概念之间的层次关系对应本体中相应类之间的分类关系；

(2)在本体中，每个类对应的属性由模糊概念层次中相应模糊概念内涵对应的模糊语言变量值来表示，属性的值对应形式背景中模糊隶属度值；

(3)在本体中，类的实例即为模糊形式背景的对象。

3.2 概念映射

磁盘文件进行文档特征提取，形成传统的关键词向量空间模型。它将文档中的词条看作孤立的个体单元，不考虑词条间的语义层次关系和同义、歧义现象，孤立的词条单元通过该模型计算后，会与原文档有较大的语意差距。本文在文献[6]的基础上，结合模糊逻辑的相关知识，提出一种基于模糊概念层次的向量空间模型。该向量空间模型简述如下：

关键词概念集序列$(Q_1,Q_2,\cdots,Q_m)$对模糊概念层次遍历后可展开为$((q^1,\ q_1{}^F,\ q_1{}^1,\ q_1{}^2,\ \cdots,\ q_1{}^{Si}),\ (q^2,\ q_2{}^F,\ q_2{}^1,\ q_2{}^2,\cdots,\ q_2{}^{Si}),\cdots,\ (q^m,\ q_m{}^F,\ q_m{}^1,\ q_m{}^2,\cdots,\ q_m{}^{Si})$，其中，$q_i{}^F$为$q^i$的父概念，设$q^i$有$S_i$个子概念，$q_i{}^j$为$q_i$的第$j$个子概念。

设q^i与文档d中的第k个词条t_k对应，则$Q_i(q^1,\ q_i{}^F,\ q_i{}^1,\ q_i{}^2,\cdots,\ q_i{}^n)$在文档中映射为$(t_k,\ t_k{}^F,\ t_k{}^1,\ t_k{}^2,\cdots,\ t_k{}^S)$，合并为$C_k$，权重记为$W$。

本文结合模糊本体中隶属度的概念，模糊概念权重计算公式为

$$W=t^f\times idf\times I=t^{f_k}\times(\text{lb}(N/n_k)+1)\times I$$

其中，N代表文档集合中的文档数量；n_k代表在文档集合中出现特征项t_k的文档数目；I为概念间隶属度。从上式可知，I越大，W值也越大。

设文档d经概念词条合并调整后，有$u(u\le m)$个合并模糊概念集，v个非合并概念词条。定义C_k的合并权重 $W_{c_k}(d)$ 为

$$W_{c_k}(d)=W_{t_k}(d)+E_{\text{概念相似度}}(t_k,t_k^F)W_{t_k}^F(d)+\sum_{i=1}^{s}E_{\text{概念相似度}}(t_k,t_k^j)W_{t_k}^j(d)$$

同理，对查询语句分析后，可定义检索关键词 q_i 映射为模糊概念的权重 $W_{q_i}(q)$为

$$W_{q_i}(q)=1+E_{\text{概念相似度}}(q_i,q_i^F)+\sum_{i=1}^{s_i}E_{\text{概念相似度}}(q_i,q_i^j)$$

其中，查询语句没有特别标识说明，一般将各关键词概念词条及父子概念词条视为具有相同的权重 1。

3.3 查询比对

用户输入查询语句q后，得到关键词序列$(q_1,q_2,\cdots,q_m)$，关键词向量空间模型与领域模糊本体进行模糊概念映射，并转换为模糊概念向量空间模型，提交给特征库进行概念相似度比较，查询语句q与文档d的相似度 $s_{cs}(d,q)$ 定义为

$$S_{cs}=(d,q)=\frac{\sum_{i=1}^{m}(W_{c_i}(d)W_{q_i}(q))}{\sqrt{(\sum_{j=1}^{v}W_{t_j}^2)(d)+\sum_{k=1}^{u}W_{c_k}^2(d))\sum_{i=1}^{m}W_{q_i}^2(q)}}$$

4 数值实验

该工具的领域模糊本体是根据相关部门对某个人崇拜组织资料检索的需求而构建的。

因此，实验选取 300 篇与某个人崇拜组织相关的文档组成集合，对模糊本体模型和关键词向量空间模型进行比对实验，其中包含某个人崇拜事件相关文档 100 篇，某个人崇拜组织相关文档 100 篇，某个人崇拜成员相关文档 100 篇。

在实验中，取查准率、查全率作为性能测试标准，并采用文献[5]中的公式。

输入查询语句为“个人崇拜组织”，实验结果如图 2、图 3 所示。输入查询语句为“个人崇拜成员”，实验结果如图 4、图 5 所示。

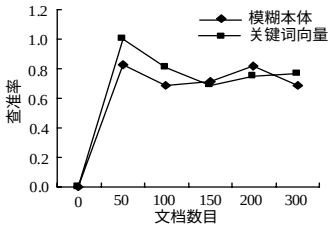

图 2　输入“个人崇拜组织”时的查准率

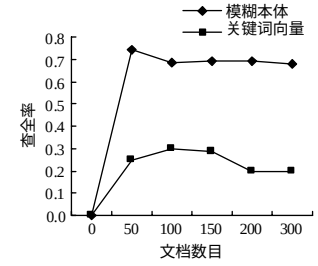

图 3　输入“个人崇拜组织”时的查全率

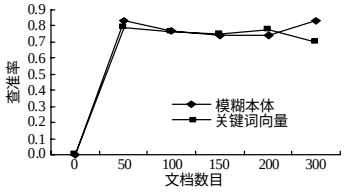

图 4　输入“个人崇拜成员”时的查准率

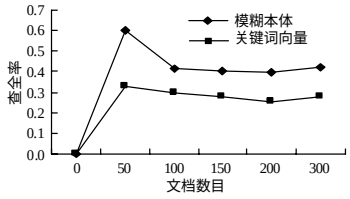

图 5　输入“个人崇拜成员”时的查全率

查看对比图可以看到，模糊本体模型在数据查全率上的性能比关键词向量空间模型高，并且随文档数目的增加，模糊本体模型在数据查全率上的性能趋于稳定；在查准率上的性能比关键词向量空间模型稳定。模糊本体模型充分利用隶属度概念，具有更大的弹性和优势。

5　结束语

计算机取证学的研究始于 20 世纪 80 年代，并随计算机和网络技术的普及而变得越来越重要。本文对传统证据分析技术进行研究，提出一种基于模糊本体的证据分析系统方法，并通过具体的数值实验对其进行性能分析。实验结果表明，该系统在数据查准率和查全率上具有良好性能。

参考文献

[1] Navigli R, Velardi P. Ontology Learning and Its Application to Automated Terminology Translation[J]. IEEE Intell. Syst., 2003, 18(1): 22-31.

[2] Embley D, Campbell R. Ontology-based Extraction and Structuring of Information from Data-rich Unstructured Documents[C]//Proc. of ACM Conf. on Inf. Knowledge Manage. [S. l.]: ACM Press, 1998.

[3] Handschuh S, Staab S. S-cream-semi-automatic Creation of Metadata[C]//Proc. of the 13th Int. Conf. on Knowledge Eng. Manage. [S. l.]: IEEE Press, 2002.

[4] 强　宇, 刘宗田, 李　旭. 一种基于模糊聚类的模糊本体生成方法[J]. 计算机科学, 2006, 33(4): 148-150.

[5] Jian Zhiwei, Huang Linkai. A Fuzzy Ontology and Its Application to News Summarization[J]. IEEE Transctions on Systems, 2005, 35(5): 528-535.

[6] 张映海. 基于关键词与语义概念结合的信息检索研究[J]. 计算机应用, 2006, 26(12): 2964-2966.

(上接第 276 页)

(2)对于给定的 n, p，当 b/w 由 0.2 增加到 0.4 时，$H(m)$ 的减少不超过 2%，所以，不妨通过付给被挤掉的消费者较多的赔偿金来提高声誉度。

(3)综合考虑经济效益和声誉度，可给定 $P_5(m)<0.3$, $P_{10}(m)<0.1$，由图 1、图 2 可以看出，对于 $n=400$，若估计 $p=0.05$，则 $m=424$。

6　结束语

本文研究了如何保证在网格上的远程教育服务质量，资源预留是一个可行的方法，但在网格预留问题上，已有的预留机制有其局限性。提出一种基于计算经济的资源预留机制，它能综合考虑教育服务提供方的经济利益和声誉度，确定预订服务的教育消费者数量的最佳限额。在以后的研究中将进一步考虑消费者的诚信，根据不同客源制定不同的资源预留策略。

参考文献

[1] Foster I, Kesselman C, Lee C, et al. A Distributed Resource Management Architecture That Supports Advance Reservations and Co-Allocation[C]//Proc. of the International Workshop on Quality of Service. [S. l.]: IEEE Computer Society, 1999.

[2] Buyya R, Abramson D, Venugopal S. The Grid Economy[Z]. 2005.

[3] Buyya R, Abramson D, Giddy J. An Economy Driven Resource Management Architecture for Global Computational Power Grids[C]//Proc. of the 2000 International Conference on Parallel and Distributed Processing Techniques and Applications. Las Vegas, USA: CSREA Press, 2000.

[4] Burchard L, Heiss H, Rose C. Performance Issues of Bandwidth Reservations for Grid Computing[C]//Proc. of the 15th Symposium on Computer Architecture and High Performance Computing. [S. l.]: IEEE Computer Society, 2003.

[5] Medvinsky G, Neuman C. NetCash: A Design for Practical Electronic Currency on the Internet[C]//Proc. of the 1st ACM Conference on Computer and Communication Security. Berlin, Germany: ACM Press, 1993.