

基于概率粗糙集模型的信息检索

黄治国¹, 朱承学², 薛凡¹, 王加阳³

(1. 黄淮学院国际学院, 驻马店 463000; 2. 湖南第一师范学院信息技术系, 长沙 410002; 3. 中南大学信息科学与工程学院, 长沙 410083)

摘要: 针对经典粗糙集模型难以分类标引空间以及体现类间关联的缺陷, 将条件概率关系结合粗糙集理论引入信息检索, 提出一种基于概率粗糙集的信息检索模型。定义标引词空间的条件概率关系, 自动挖掘概念相似类形成概念空间。定义文档与查询、文档与文档间语义贴近度的计算方法。根据贴近度实现检索匹配结果的排序输出。仿真实例表明了该方法的可行性和有效性。

关键词: 粗糙集; 信息检索; 条件概率关系; 语义贴近度

Information Retrieval Based on Probability Rough Set Model

HUANG Zhi-guo¹, ZHU Cheng-xue², XUE Fan¹, WANG Jia-yang³

(1. International College, Huanghuai University, Zhumadian 463000; 2. Department of Information and Technology, Hunan First Normal College, Changsha 410002; 3. School of Information Science and Engineering, Central South University, Changsha 410083)

【Abstract】 Aiming at the disadvantage of classical rough set theory on identifying the conceptually similar terms and the relationships between classes, this paper proposes a novel information retrieval model based on conditional probability relation and rough set. Conception space is formed by defining conditional probability relation in index words space to mine conception similar class automatically. A method is designed to calculate semantic distance between a document and a query, as well as documents. And the ordered outputs of retrieval result are acquired. The simulation instance shows that this algorithm is feasible and effective in practice.

【Key words】 rough set; information retrieval; conditional probability relation; semantic distance

1 概述

信息检索是对信息集合与用户查询式之间的相关性进行匹配与选择的过程。迄今为止, 信息检索理论研究者已提出了多种信息检索数学模型, 其中被广泛认可的有布尔模型、向量空间模型、概率模型等。布尔模型的突出优点在于运算简单易行, 且其结构化提问方式符合用户思维习惯; 但布尔检索本身所赖以建立的二值逻辑和集合理论缺乏必要的灵活性, 不能给出部分满足用户要求的结果。向量模型允许对标引项进行加权, 易于实现检索结果的排序; 但存在提问式缺乏结构性、存储和计算量大, 难以反映标引项间关系等一些缺陷。概率模型考虑到文档与查询式间的相关性, 体现了更为先进的检索思想, 从客观上使检索更趋合理; 然而, 一篇文档是否相关的可能性大小是一个随机事件, 此事件的随机性难以确定。因此, 在可靠的参数估计方法问题得到解决之前, 概率检索模型较难应用于实际。

自粗糙集理论^[1]诞生以来, 已有较多学者将其与信息检索结合进行了一定的研究。文献[2]首次提出信息检索的粗糙集模型, 认为粗糙集检索模型以标引词空间的不可分辨关系为基础形成概念类, 从而考虑了标引词间的语义关系, 实际上是隐含并扩展了布尔模型; 允许对标引项进行加权, 并且为检索设置不同的匹配等级, 同等级内又定义了检索结果与查询间的贴近度, 因而可方便有效地实现检索结果与查询式间相似度排序。同时也指出其有待进一步研究的难题, 即如何设计分类方法, 以及如何体现类间关联。文献[3]建造一种 Rough 模糊集, 其检索算法复杂性不随文献资料数量的增加而增加, 但该算法人为规定标引词关于文档的隶属度, 选取 C_0^3 的标引词组合构造 Rough 模糊集, 并没有从本质上考虑各

标引词间的相关性。文献[4]引入模糊相似关系提出广义模糊粗糙集模型, 此模型极大程度地反映了对象间的相关性, 但并没就信息检索方法作进一步研究。因此, 为粗糙集检索模型设计一种自动的、合理体现类间关联的分类方法就成为必要和可能。本文将条件概率关系与粗糙集理论结合应用于信息检索, 并给出概率粗糙集信息检索模型的定义。其策略是引入条件概率关系分类数据库中的标引词, 即构造标引词空间的相似关系, 以自动聚合标引词相似概念类, 所生成相似概念空间就使得文档与文档、文档与查询间的比较成为可能。

2 概率粗糙集模型

Rough Set 理论是由波兰科学家 Z.Pawlak 在 1982 年提出的一种处理含糊和不精确问题的新型数学工具。这一理论引入代数学中的等价关系讨论知识, 把知识看作是论域的划分。虽然粗糙集理论易于分析数据, 但是不一定能反映实际应用中元素间关系的现实视图。与基于划分的标准粗糙集理论相比, 基于论域覆盖的模型更具现实意义, 因为实际应用中数据对象间关系不一定严格满足对称性与传递性。本节将条件概率关系与粗糙集理论相结合, 以表示对象间关联, 并给出概率粗糙集模型描述。

定义 1 U 为一非空有限论域, 一个条件概率相似关系是

基金项目: 河南省教育科学“十一五”规划课题基金资助项目(2008-JKGGHAGH-413); 湖南省自然科学基金资助项目(06JJ20075); 湖南省教育厅科研基金资助项目(08C015)

作者简介: 黄治国(1978-), 男, 硕士, 主研方向: 数据挖掘, 决策支持; 朱承学, 副教授、硕士; 薛凡, 硕士; 王加阳, 教授、博士

收稿日期: 2008-04-06 **E-mail:** huangzhiguo2001@tom.com

一个映射 $R:U \times U \rightarrow [0,1]$, R 对 $\forall x,y \in U$ 满足

$$R(x,y) = P(x|y) = P(y \rightarrow x) = \frac{|x \cap y|}{|y|}$$

定义 2 对于 $x,y \in U$, μ_x, μ_y 为 x,y 关于属性集 A 的模糊集, 一个模糊条件概率关系是一个映射 $R:U \times U \rightarrow [0,1]$, R 对 $\forall x,y \in U$ 满足

$$R(x,y) = \frac{\sum_{a \in A} \min\{\mu_x(a), \mu_y(a)\}}{\sum_{a \in A} \mu_y(a)}$$

其中, $\mu_x(a)$ 为 x 关于 a 的隶属函数。

条件概率关系与模糊条件概率关系均用来表示对象间相似关系。且模糊条件概率关系代表了更一般化情形。

定义 3 U 为一非空有限论域, R 为 U 上一条件概率关系。对 $\forall x \in U$, 其 a -被支持集与 a -支持集分别定义为

$$R_S^\alpha(x) = \{y | y \in U \wedge R(x,y) \geq \alpha\}$$

$$R_P^\alpha(x) = \{y | y \in U \wedge R(y,x) \geq \alpha\}$$

其中, $\alpha \in [0,1]$; $R_S^\alpha(x)$ 为支持 x 的对象集; $R_P^\alpha(x)$ 为被 x 支持的对象集。条件概率关系满足自反性, 因此 $\{R_S^\alpha(x) | x \in U\}$ 与 $\{R_P^\alpha(x) | x \in U\}$ 均构成论域 U 上的一个覆盖。以下仅讨论 $R_S^\alpha(x)$, 关于 $R_P^\alpha(x)$ 通过类似方法推导可得。

定义 4 U 为一非空有限论域, R 为 U 上一条件概率关系。对于论域 U 的任意子集 $X \subseteq U$, 其下近似集与上近似集分别定义为

$$\underline{R}_S^\alpha(X) = \{x | R_S^\alpha(x) \subseteq X\}$$

$$\overline{R}_S^\alpha(X) = \{x | R_S^\alpha(x) \cap X \neq \emptyset\}$$

下近似集 $\underline{R}_S^\alpha(X)$ 由所有为 X 子集的对象 x 构成, 上近似集 $\overline{R}_S^\alpha(X)$ 由所有与 X 相交不为空的对象 x 构成。

3 基于概率粗糙集的信息检索模型

文献[2]指出: 如何设计分类方法是将粗糙集模型应用于信息检索的关键难题之一, 具体表现在如何确定相似关系以及如何控制由此关系导致的粒度大小。此后较多学者对此进行了一定的研究^[3-6], 但均没有很好地解决这些问题。本节讨论在将文档自动标引后, 如何应用概率粗糙集模型进行信息检索。

在传统的计算机信息检索中, 对每篇文档抽取若干标引词, 用这些词条的集合来代表原文, 近似表示原文的语义, 从而实现按原文语义进行检索。假设 m 个文档构成文档集 $D = \{d_1, d_2, \dots, d_m\}$, 其标引词空间 $T = \{t_1, t_2, \dots, t_n\}$, 文档 $d_j (1 \leq j \leq m)$ 形式化表示为 $d_j = \{t_{1j}, t_{2j}, \dots, t_{nj}\}$ 。可定义 $t_{ij} (1 \leq i \leq m, 1 \leq j \leq n)$ 为布尔取值, 此时 d_j 为文档的精确标引词空间表示; 其取值定义为区间 $[0,1]$ 更符合当前信息检索的一般方法时, d_j 为文档的模糊标引词空间表示。针对文档的精确表示和模糊表示进行信息检索, 为自动挖掘相似概念类, 须分别构造标引词空间的条件概率关系和模糊条件概率关系。

定义 5 标引词空间 $T = \{t_1, t_2, \dots, t_n\}$ 上的条件概率关系是一个映射 $R: T \times T \rightarrow [0,1]$, 使得对

$$\forall t_i, t_j \in T, R(t_i, t_j) = \frac{|S(t_i) \cap S(t_j)|}{|S(t_j)|}$$

其中, $S(t_i)$ 为含有标引词 t_i 的文档集; $S(t_i \wedge t_j)$ 为同时含有标引词 t_i 与 t_j 的文档集。

定义 6 标引词空间 $T = \{t_1, t_2, \dots, t_n\}$ 上的模糊条件概率关系

是一个映射 $R: T \times T \rightarrow [0,1]$, 使得对

$$\forall t_i, t_j \in T, R(t_i, t_j) = \frac{\sum_{d \in D} \min\{\mu_d(t_i), \mu_d(t_j)\}}{\sum_{d \in D} \mu_d(t_j)}$$

其中, $\mu_d(t_i)$ 为标引词 t_i 关于 d 的隶属度。

条件概率关系实质上是模糊条件概率关系的特例, 对应于标引词隶属度为逻辑取值情形。以下仅讨论一般化情形——模糊条件概率关系即可。条件概率关系与模糊条件概率关系体现了这样一个事实: 若 2 标引词趋向同时出现在文档对象中, 则认为此 2 个标引词相互依赖, 属于同一相似概念。既然模糊概率关系对应在区间 $[0,1]$ 内取值, 那么当然就可以在此关系基础上, 通过设置一阈值 α 以自动挖掘各标引词的相似概念类。

定义 7 设 R 是标引词空间 $T = \{t_1, t_2, \dots, t_n\}$ 上的模糊条件概率关系, 对 $\forall t_i \in T$, 分别定义其 a -被支持集和 a -支持集如下:

$$R_S^\alpha(t_i) = \{t_j | t_j \in T \wedge R(t_i, t_j) \geq \alpha\}$$

$$R_P^\alpha(t_i) = \{t_j | t_j \in T \wedge R(t_j, t_i) \geq \alpha\}$$

定义 8 假设有 m 个文档构成文档集 $D = \{d_1, d_2, \dots, d_m\}$, R 是标引词空间 $T = \{t_1, t_2, \dots, t_n\}$ 上的模糊条件概率关系, 对 $\forall d \in D$, 分别定义其关于 R 的 a -下近似集与 a -上近似集如下:

$$\underline{R}_S^\alpha(d) = \{t_i | R_S^\alpha(t_i) \subseteq d\}$$

$$\overline{R}_S^\alpha(d) = \{t_i | R_S^\alpha(t_i) \cap d \neq \emptyset\}$$

$R_S^\alpha(t_i)$ 与 $R_P^\alpha(t_i)$ 用于在标引词空间挖掘概念类形成类空间, α 越大则分类所导致粒度越小, 相反 α 越小则分类所导致粒度越大, 因此需根据分类结果选择一合适的 α 值。然后在此基础上根据定义 9 可求取文档集中任意对象的 a -下近似集与上近似集, 以便于下一步的贴近度计算。同时, 针对文档的模糊表示, 为进一步得到其下、上近似模糊集, 还须定义标引词关于文档 a -下、上近似集隶属度计算方法。

定义 9 设 R 是标引词空间 $T = \{t_1, t_2, \dots, t_n\}$ 上的模糊条件概率关系, 文档 d 的模糊表示对应文档论域上的一个模糊集, 标引词 $t_i (t_i \in T)$ 关于模糊集 d 的下、上近似隶属度分别定义为

$$\mu_{\underline{R}_S^\alpha(d)}(t_i) = \text{Inf}\{\mu_d(t_j) | t_j \in T \wedge t_j \in R_S^\alpha(t_i)\}$$

$$\mu_{\overline{R}_S^\alpha(d)}(t_i) = \text{Sup}\{\mu_d(t_j) | t_j \in T \wedge t_j \in R_S^\alpha(t_i)\}$$

其中, Inf 表示取下确界; Sup 表示取上确界。这样就可首先以文档论域为基础根据 $R_S^\alpha(t_i)$ 形成标引词概念空间, 然后计算文档对象的近似集, 再依据近似集隶属度定义计算每一标引词关于文档对象近似集的隶属度, 从而得到文档近似集的模糊表示。查询式下、上近似集及隶属度计算方法同文档对象计算方法。

定义 10 文档与查询间语义贴近度定义如下:

$$\text{SIM}(Q_i, d_j) = \text{SIM}(Q_i, d_j) + \overline{\text{SIM}}(Q_i, d_j)$$

$$\text{SIM}(Q_i, d_j) = \frac{|R_S^\alpha(Q_i) \wedge R_S^\alpha(d_j)|}{|R_S^\alpha(Q_i) \vee R_S^\alpha(d_j)|}$$

$$\overline{\text{SIM}}(Q_i, d_j) = \frac{|R_S^\alpha(Q_i) \wedge \overline{R}_S^\alpha(d_j)|}{|R_S^\alpha(Q_i) \vee \overline{R}_S^\alpha(d_j)|}$$

得到查询或文档下、上近似集模糊表示后, 即可应用贴近度公式计算文档与查询间以及文档与文档间语义贴近度, 最终根据贴近度值实现检索匹配结果的排序输出。

4 仿真实例

本文以 NPL 语料库 (包含 11 429 篇文档, 7 491 个标引词) 为实验数据。设 T 为文档集 D 中所有标引词构成的标引词空

间。文档 $d_j(1 \leq j \leq m)$ 形式化表示为 $d_j=(t_{1j}, t_{2j}, \dots, t_{nj})$; $t_{ij} \in T$ 为标引词 t_i 关于文档 d_j 的权重。 D 中各文档对象标引词权重为 1, 文档集表示见表 1。

表 1 文档集表示

文档对象	文档对象的标引词表示
d_1	$t_7, t_{25}, t_{80}, t_{224}, t_{346}, t_{578}, t_{647}, t_{921}, t_{1181}, t_{1592}, t_{1711}, t_{1934}, t_{2058}$
d_2	$t_6, t_{17}, t_{25}, t_{36}, t_{38}, t_{49}, t_{71}, t_{72}, t_{92}, t_{117}, t_{192}, t_{268}, t_{280}, t_{369}, t_{427}$
d_3	$t_3, t_6, t_{10}, t_{25}, t_{34}, t_{36}, t_{83}, t_{145}, t_{162}, t_{179}, t_{237}, t_{541}, t_{542}, t_{551}, t_{714}, t_{987}, t_{1094}, t_{1504}, t_{1849}, t_{2395}$
d_4	$t_{71}, t_{103}, t_{577}, t_{693}, t_{1022}, t_{1255}, t_{1501}, t_{1552}$
d_5	$t_1, t_3, t_7, t_{25}, t_{51}, t_{61}, t_{71}, t_{85}, t_{135}, t_{145}, t_{206}, t_{210}, t_{224}, t_{256}, t_{335}, t_{495}, t_{518}, t_{591}, t_{626}, t_{702}, t_{963}, t_{1227}, t_{1681}, t_{1855}, t_{3699}, t_{6577}$
d_6	$t_1, t_3, t_7, t_{16}, t_{51}, t_{176}, t_{402}, t_{431}, t_{546}, t_{572}, t_{626}, t_{742}, t_{897}, t_{1021}, t_{1761}$
d_7	$t_1, t_3, t_6, t_{20}, t_{47}, t_{64}, t_{86}, t_{114}, t_{116}, t_{119}, t_{162}, t_{176}, t_{199}, t_{236}, t_{304}, t_{344}, t_{401}, t_{474}, t_{556}, t_{558}, t_{915}, t_{1068}, t_{2152}, t_{3898}, t_{6883}$
d_8	$t_3, t_{14}, t_{42}, t_{67}, t_{85}, t_{119}, t_{139}, t_{209}, t_{213}, t_{241}, t_{245}, t_{277}, t_{302}, t_{337}, t_{365}, t_{379}, t_{415}, t_{424}, t_{456}, t_{808}, t_{940}, t_{1117}$
d_9	$t_1, t_3, t_9, t_{23}, t_{28}, t_{48}, t_{51}, t_{81}, t_{119}, t_{124}, t_{129}, t_{159}, t_{210}, t_{211}, t_{213}, t_{252}, t_{483}, t_{546}, t_{626}, t_{1627}, t_{1645}, t_{1855}, t_{1915}, t_{2004}, t_{6228}, t_{6893}$
d_{10}	$t_{88}, t_{127}, t_{213}, t_{412}, t_{3787}$
...	...

根据定义 8 对 $\forall t_i \in T$ 可求得其 a -被支持集 $R_S^a(t_i)$ 。令 $a=0.8$, 得

$$R_S^a(t_{103}) = \{t_{103}, t_{577}, t_{693}, t_{1022}, t_{1255}, t_{1501}, t_{1552}\}, R_S^a(t_{1021}) = \{t_{16}, t_{402}, t_{431}, t_{572}, t_{742}, t_{897}, t_{1021}, t_{1761}\}, R_S^a(t_{1761}) = \{t_{103}, t_{577}, t_{693}, t_{1022}, t_{1255}, t_{1501}, t_{1552}\}, \dots$$

由定义 9 求取各文档的下上近似集见表 2。

表 2 文档下、上近似集表示

文档对象	文档下近似	文档上近似
d_1	$t_{80}, t_{346}, t_{578}, t_{647}, t_{921}, t_{1181}, t_{1592}, t_{1711}, t_{1934}, t_{2058}$	$t_{80}, t_{346}, t_{578}, t_{647}, t_{921}, t_{1181}, t_{1592}, t_{1711}, t_{1934}, t_{2058}, t_{224}, t_{25}, t_7$
d_2	$t_{17}, t_{38}, t_{72}, t_{92}, t_{117}, t_{192}, t_{268}, t_{280}, t_{369}, t_{427}$	$t_{17}, t_{38}, t_{72}, t_{92}, t_{117}, t_{192}, t_{268}, t_{280}, t_{369}, t_{427}, t_{25}, t_{36}, t_{49}, t_6, t_{71}$
d_3	$t_{34}, t_{83}, t_{179}, t_{237}, t_{541}, t_{542}, t_{551}, t_{714}, t_{987}, t_{1094}, t_{1504}, t_{1849}, t_{2395}$	$t_{10}, t_{34}, t_{83}, t_{179}, t_{237}, t_{541}, t_{542}, t_{551}, t_{714}, t_{987}, t_{1094}, t_{1504}, t_{1849}, t_{2395}, t_{145}, t_{162}, t_{25}, t_{36}, t_3, t_6$
d_4	$t_{103}, t_{577}, t_{693}, t_{1022}, t_{1255}, t_{1501}, t_{1552}$	$t_{103}, t_{577}, t_{693}, t_{1022}, t_{1255}, t_{1501}, t_{1552}, t_{71}$
d_5	$t_{61}, t_{135}, t_{206}, t_{256}, t_{335}, t_{495}, t_{518}, t_{591}, t_{702}, t_{963}, t_{1227}, t_{1681}, t_{1855}, t_{3699}, t_{6577}$	$t_1, t_{51}, t_{61}, t_{135}, t_{206}, t_{210}, t_{256}, t_{335}, t_{495}, t_{518}, t_{591}, t_{626}, t_{702}, t_{963}, t_{1227}, t_{1681}, t_{1855}, t_{3699}, t_{6577}, t_{145}, t_{224}, t_{25}, t_3, t_{85}, t_7, t_{71}$
d_6	$t_{16}, t_{402}, t_{431}, t_{572}, t_{742}, t_{897}, t_{1021}, t_{1761}$	$t_1, t_{51}, t_{626}, t_{16}, t_{176}, t_{402}, t_{431}, t_{546}, t_{572}, t_{742}, t_{897}, t_{1021}, t_{1761}, t_3, t_7$
d_7	$t_{47}, t_{64}, t_{86}, t_{114}, t_{116}, t_{119}, t_{162}, t_{176}, t_{199}, t_{236}, t_{304}, t_{344}, t_{401}, t_{474}, t_{556}, t_{558}, t_{915}, t_{1068}, t_{2152}, t_{3898}, t_{6883}$	$t_1, t_{51}, t_{176}, t_{20}, t_{47}, t_{64}, t_{86}, t_{114}, t_{116}, t_{119}, t_{162}, t_{176}, t_{199}, t_{236}, t_{304}, t_{344}, t_{401}, t_{474}, t_{556}, t_{558}, t_{915}, t_{1068}, t_{2152}, t_{3898}, t_{6883}, t_{119}, t_{162}, t_{199}, t_3, t_6$
d_8	$t_{67}, t_{139}, t_{209}, t_{241}, t_{245}, t_{277}, t_{302}, t_{337}, t_{365}, t_{379}, t_{415}, t_{424}, t_{456}, t_{808}, t_{940}, t_{1117}$	$t_{67}, t_{139}, t_{209}, t_{241}, t_{245}, t_{277}, t_{302}, t_{337}, t_{365}, t_{379}, t_{415}, t_{424}, t_{456}, t_{808}, t_{940}, t_{1117}, t_{119}, t_{14}, t_{213}, t_3, t_{85}, t_{42}$
d_9	$t_9, t_{28}, t_{48}, t_{81}, t_{124}, t_{129}, t_{159}, t_{211}, t_{252}, t_{483}, t_{1627}, t_{1645}, t_{1915}, t_{2004}, t_{6228}, t_{6893}$	$t_1, t_{51}, t_{210}, t_{626}, t_{1855}, t_{546}, t_{923}, t_{28}, t_{48}, t_{81}, t_{124}, t_{129}, t_{159}, t_{211}, t_{252}, t_{483}, t_{1627}, t_{1645}, t_{1915}, t_{2004}, t_{6228}, t_{6893}, t_{119}, t_{213}, t_3$
d_{10}	$t_{88}, t_{127}, t_{412}, t_{3787}$	$t_{88}, t_{127}, t_{412}, t_{3787}, t_{213}$
...

求取查询和文档下、上近似集表示后, 即可进行语义贴程度计算。若此时有一查询 $Q_i = \{t_{78}, t_{1902}, t_{191}, t_{591}, t_{1021}\}$, 各标引词指示与对应标引词分别为: device/t78, transmission/t1902,

(上接第 189 页)

参考文献

[1] 张伟, 石纯一. Agent 的组织承诺和小组承诺[J]. 软件学报, 2003, 14(3): 473-478.
 [2] Ferber J, Gutknecht O. A Meta-model for the Analysis and Design of Organization in Multi-agent Systems[C]//Proceedings of the 3rd International Conference on MAS. Paris, France: [s. n.], 1998.

ray/t191, signal/t591, data/t1021。与 Q_i 匹配的结果文档集按语义贴程度排序结果见表 3。

表 3 文档与查询间语义贴程度

近似集	d_{11298}	d_{3496}	d_{4370}	d_{4765}	d_{47570}	d_{2312}	d_{2705}	d_{9322}	d_{1705}	d_{6677}	...
$SIM(Q_i, d_j)$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	...
$\overline{SIM}(Q_i, d_j)$	0.750	0.744	0.727	0.727	0.706	0.686	0.653	0.649	0.636	0.632	...
$SIM(Q_i, d_j)$	0.750	0.744	0.727	0.727	0.706	0.686	0.653	0.649	0.636	0.632	...

由表 3 的计算结果可知, 当查询为 $Q_i = \{t_{78}, t_{1902}, t_{191}, t_{591}, t_{1021}\}$, $a=0.8$ 时, 与查询 Q_i 最接近的前 3 个文档为 $d_{11298}, d_{3496}, d_{4370}$ 。选择不同的支持度 a 会导致不同粒度的概念相似类, 支持度大则相似类粒度小, 反之则粒度大。文档对象与查询的下上近似集对应不同的相似类也会产生变化。所以, 当支持度 a 不同时, 文档与查询间相似度也会不同。因此, 可依据相关反馈调整 a 值, 以取得理想的检索效果。

5 结束语

本文基于模糊相似关系构造条件概率关系, 并将其结合粗糙集理论引入信息检索, 构造基于概率粗糙集的信息检索模型。首先定义标引词空间的条件概率关系, 以自动挖掘概念类形成概念空间; 然后定义文档与查询、文档与文档间的语义贴程度计算方法; 最终根据贴程度值实现检索匹配结果的排序输出。在标引词空间定义条件概率关系, 不仅能充分挖掘和利用标引词间相似关系, 而且可调节 a 取值, 以选择合适的相似概念类粒度大小。实例分析表明了该方法的可行性和有效性。

参考文献

[1] Pawlak Z. Rough Sets[J]. International Journal of Computer and Information Sciences, 1982, 11(5): 341-356.
 [2] Das-Gupta P. Rough Sets and Information Retrieval[C]//Proc. of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Grenoble, France: [s. n.], 1988: 567-581.
 [3] 李龙澍, 张霞. 基于 Rough 集的情报检索研究[J]. 情报学报, 2002, 21(1): 7-11.
 [4] Intan R, Mukaidono M. Generalized Fuzzy Rough Sets by Conditional Probability Relations[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2002, 16(7): 865-881.
 [5] Miyamoto S. Application of Rough Sets to Information Retrieval[J]. Journal of the American Society for Information Science, 1998, 49(3): 195-205.
 [6] Srinivasan P, Ruiz M, Kraft D H, et al. Vocabulary Mining for Information Retrieval: Rough Sets and Fuzzy Sets[J]. Information Processing and Management, 2001, 37(1): 15-38.

[3] 张伟, 石纯一. Agent 组织的一种递归模型[J]. 软件学报, 2002, 13(11): 2149-2154.
 [4] 陈圣磊, 吴慧中, 韩祥兰, 等. 基于信念型承诺的 Agent 协作机制研究与应用[J]. 计算机工程, 2006, 32(11): 37-39.
 [5] 兰少华, 吴慧中, 顾一禾. 基于 BDI 的 Agent 合同网实现[J]. 小型微型计算机系统, 2001, 22(12): 1471-1474.