基于粗糙集的个性化 Web 搜索系统

时 雷[1] 席 磊[1] 段其国[2]

(河南农业大学信息与管理科学学院 郑州 450002)[1] (同济大学计算机科学与工程系 上海 201804)[2]

摘 要 本文提出了一种基于粗糙集理论的个性化 Web 搜索系统。用户偏好文件中对关键字进行分组以表示用户兴趣类别。利用粗糙集理论处理自然语言的内在含糊性,根据用户偏好文件对查询条件进行扩展。搜索组件使用扩展后的查询条件搜索相关信息。为了进一步排除不相关信息,排序组件计算查询条件和搜索结果之间的相似程度,根据计算值对搜索结果进行排序。与传统搜索引擎进行了比较,实验结果表明,该系统有效地提高了搜索结果的精度,满足了用户的个性化需求。

关键词 Web 检索,粗糙集,个性化

Personalized Web Search System Based on Rough Sets

SHI Lei[1] XI Lei[1] DUAN Qi-Guo[2]

(College of Information and Management Science, Henan Agricultural University, Zhengzhou 450002)[1]

(Department of Computer Science and Technology, Tongji University, Shanghai 201804)[2]

Abstract In this paper, a novel rough set based approach is proposed to create a personalized Web search system. Firstly, user profiles which consist of categories of user's interests by grouping related keywords are designed. Rough set theory is used to deal with inherent ambiguities of natural language and refine query according to user profiles. Then refined query is submitted to search component. To further filter out irrelevant documents for the user, retrieved results are re-ranked according to rough similarity measures between refined query and documents by ranking component. Experiments compared with traditional search engine are presented and experimental results indicate the precision of Web retrieval is greatly improved and system are suitable for individual usage.

Keywords Web retrieval, Rough sets, Personalization

1 引言

近年来随着网络和信息技术的发展,Web 上的信息量迅速增加。如何快速、准确地从浩瀚的信息资源中寻找到所需信息已经成为信息检索、Web 挖掘等领域的重要研究内容。Web 搜索引擎在一定程度上解决了这一问题,但传统的搜索引擎没有考虑用户的兴趣爱好。用户在使用搜索引擎时,常常是搜索到的网页成千上万,而真正需要的信息却又寥寥无几。在这种背景下,个性化的 Web 信息检索技术引起了越来越多的研究者的关注。个性化的 Web 检索有效地利用了用户的兴趣和偏好,可以帮助用户更加精确地检索信息。

随着个性化的 Web 检索研究的开展,研究成果和应用系统纷纷出现,如 Syskill&Webertt、WebWatcher 等[1~4]。这些系统大多采用了传统的信息检索模型,即布尔模型、向量模型和概率模型。然而,自然语言文档在本质上是基于上下文相关的,因此严格的二值逻辑无法处理自然语言中的细微差别。粗糙集理论是处理不精确、不确定和模糊信息的有力工具,它提供了一种近似地表示概念的方法。因为孤立地考虑文档中的词是无法对文档进行准确分类的,所以可以利用粗糙集对文档空间和查询项空间进行扩展,使其包括概念上相关的其它词。

本文提出了一种基于粗糙集的个性化 Web 检索系统。该系统利用粗糙集理论处理自然语言的内在含糊性,根据用户偏好对查询条件进行扩展。搜索组件使用扩展后的查询条件搜索相关信息。为了进一步排除不相关信息,排序组件计算查询条件和搜索结果之间的相似程度,根据计算值对搜索结果进行排序。

2 粗糙集

粗糙集理论是一种处理不完全或者不精确信息的工具[5,6],它已经在数据挖掘、机器学习和模式识别等领域中得到了广泛的应用。

时 雷 硕士,助教,研究方向:模式识别,数据挖掘;席 磊 硕士,讲师,研究方向:模式识别;段其国 博士生,研究方向:数据挖掘。

但是这几种修正方法都不一定是最科学的,如何根据具体的应用环境,让修正方式更加合理,是我们下一步的一个努力方向;进一步地,如何把修正矩阵从 VSM 推广到概率模型[5]和信度网[4]中,是我们需要进一步研究的重点。

参 考 文 献

1 Turney P D. Similarity of semantic relations. Computational Linguistics, 2006, 32 (3): 379~416

2 Zhu Y W, Hu Y M. Enhancing search performance on gnutella-like P2P systems. IEEE Transactions on Parallel and Distributed Systems, 2006, 17 (12): 1482~1495

3 Wang H M, Rajman M, Guo Y, et al. New PR-combining TFIDF with Pagerank. Artificial Neural Networks-icann, PT2 Lecture Notes in Comuter Science, 2006, 4132: 932~942

4 沈一栋,邢永康. 一种新的知识表达模型——信度网[J]. 计算机科学, 2000,9(27):40~43

5 邢永康,马少平. 信息检索的概率模型[J]. 计算机科学,2003,30(8):13~17

6 孙建军,成颖,等. 信息检索技术. 科学出版社, 204. 56~58

7 许忠锡. 查全率和查准率关系辨析. 上海高校图书情报研究, 2004,4:21~23

8 宋斌,方小璐. 基于网页特征的 TFIDF 改进算法. 微机算机应用,2002, 23(1):18~20

9 许建潮,胡明. 中文 Web 文本的特征获取与分类. 计算机工程, 2005,31(8):24~25,39

定义 1 在粗糙集理论中，一个信息表知识表达系统 S 可以定义为：

$$S=\langle U,A,V,f\rangle \tag{1}$$

其中 U 是对象的有限集合，也称为论域，$A=C\cup D$ 是属性集合，C 是条件属性集，D 是决策属性集，V 是属性值的集合，V_r 表示属性 $r\in A$ 的属性值范围，即属性 r 的值域，$f:U\times A\rightarrow V$ 是一个信息函数，它指定了 U 中每个对象 x 的属性值。

定义 2 对于每个属性子集 $R\subseteq A$，可以定义一个在 U 上的等价关系，也可称为不可分辨的二元关系，即：

$$IND(R)=\{(x,y)\in U^2 \mid \forall a\in Ra(x)=a(y)\} \tag{2}$$

如果 $(x,y)\in IND(R)$，则称 x 和 y 相对于属性集 R 是不可分辨的。$IND(R)$ 的所有等价关系簇记为 $U/IND(R)$ 或者简写为 U/R；$IND(R)$ 包含对象 x 的等价类，记为 $[x]_R$。对于任何概念 $X\subseteq U$ 和属性集合 $R\subseteq A$，X 可以由上近似集合和下近似集合来近似。

定义 3 X 的下近似就是那些对于知识 R 能完全确定地归入集合 X 的对象的集合，定义为：

$$\underline{R}X=\{x\in U\mid [x]_R\subseteq X\} \tag{3}$$

定义 4 X 的上近似是由那些对于知识 R 不能排除它们属于 X 的可能性的对象构成，定义为：

$$\overline{R}X=\{x\in U\mid [x]_R\cap X\neq\varnothing\} \tag{4}$$

X 上近似和下近似如图 1 所示。

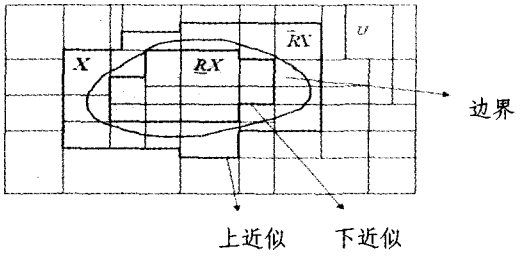

图 1 上近似和下近似示意图

在本文中，粗糙集理论用于查询条件的扩展。其中，论域 U 定义为词汇集，等价关系 R 定义为词汇在概念上的同义关系。

3 Web 检索系统的设计与实现

本文提出的个性化 Web 搜索系统的结构如图 2 所示。

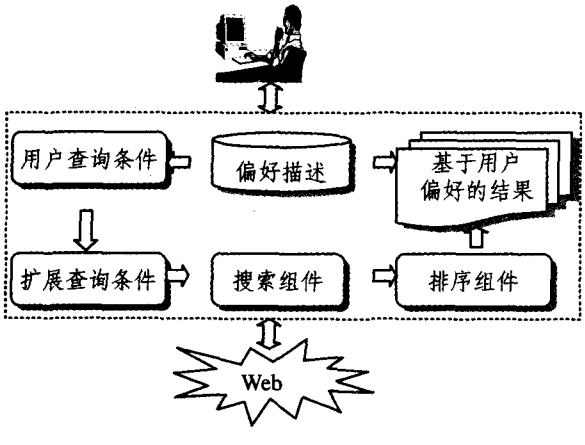

图 2 个性化 Web 搜索 Agent 系统的示意图

该系统主要包括了偏好描述文件、扩展查询条件、搜索组件和排序组件等。

(1)偏好描述文件：用于记录用户兴趣。偏好文件中对关键字进行分类，每一类关键字共同表示用户的一个兴趣。例如，关键字“粗糙集”，“分类规则”和“约简算法”等分为一类，表示用户的兴趣是针对于粗糙集的理论和算法。

(2)扩展查询条件：对于用户的初始查询条件，通过用户偏好文件中的关键词的同义关系，根据粗糙集理论创建一个初始查询条件的上近似。初始查询条件的上近似有效地扩展了用户的查询条件，能够更加准确地表达用户的检索意图。

(3)搜索组件：此组件调用搜索引擎，利用扩展查询条件搜索信息，获取相关的 URL 并下载提交给排序组件进行分析。

(4)排序组件：此组件根据搜索组件提供的网页 URL 对网页进行特征抽取，形成结构化的网页属性。通过计算网页与查询条件的粗糙相似度，把相似度过小的网页滤除，按照粗糙相似度对网页进行排序，最终将排序后的网页推荐给用户。排序组件的另外一个主要功能是接受用户反馈。用户在阅读系统推荐的网页后，对系统的检索结果进行评价。排序组件通过分析用户的反馈信息，编辑偏好描述文件，增加或者删除用于描述用户兴趣类别的关键字。

基于粗糙集的个性化 Web 检索系统中的关键算法主要包括查询条件的扩展算法、搜索算法和排序算法。

查询条件的扩展算法如 Alg. 1 所示。

Algorithm 1 查询条件扩展算法.

Input：$A=(U,R)$，查询条件 Q.
Output：扩展的查询条件 Q.
1. 计算 U 关于 R 的等价类：$\{E_1,E_2,\cdots,E_m\}$，$0<m<|U|$
2. While($i\leqslant m$)
　计算 Q 的上近似：
　$\overline{R}(Q)=U\{E_i:E_i\cap Q\neq\varnothing\}$，$i=i+1$
3. 返回 $\overline{R}(Q)$

搜索算法如 Alg. 2 所示。

Algorithm 2 搜索算法.

Input：候选文档集合 C_d，扩展的查询条件 $\overline{R}(Q)$，相似性阈值 S_T.
Output：检索结果集合 R_d.
1. 候选文档的总数目 $N=|C_d|$，$R_d=\varnothing$，$i=1$
2. While($i\leqslant N$)
　计算 $\overline{R}(Q)$ 与检索候选文档 $D_i\in C_d$ 之间的相似性 S：

$$S=\cos(D_i,\overline{R}(Q))=\frac{\sum_{j=1}^{n}d_{ik}\cdot d_{jk}}{\sqrt{\sum_{i=1}^{n}(d_{ik})^2\cdot(d_{jk})^2}}$$

3. 如果($S\geqslant S_T$)，将 D_i 加入到当前的检索结果集合 R_d 中，$R_d=R_d+D_i$
4. 返回 R_d

对于检索得到的检索结果，使用排序算法对检索结果进行排序，排序算法如 Alg. 3 所示。

Algorithm 3 排序算法.

Input：检索结果集合 R_d，查询条件 Q，等价关系 R.
Output：排序的检索结果集合 R_d'.
1. $N=|R_d|$ 表示检索结果集合中文档的数目；
2. 计算 Q 和文档 $D_i\in R_d(1\leqslant i\leqslant N)$ 之间的粗糙相似度：
$SIM=SIM(Q,D_i)=\overline{SIM}(Q,D_i)+\underline{SIM}(Q,D_i)$
　此处：

$$\overline{SIM}(Q,D_i)=\frac{|\overline{R}(Q)\cap\overline{R}(D_i)|}{|\overline{R}(Q)\cup\overline{R}(D_i)|},$$

$$\underline{SIM}(Q,D_i)=\frac{|\underline{R}(Q)\cap\underline{R}(D_i)|}{|\underline{R}(Q)\cup\underline{R}(D_i)|}.$$

3. 返回 R_d'

4 实验分析

本文设计并实现了一个基于粗糙集的个性化 Web 检索系统。系统的软件开发平台为 Window XP，开发工具是 Java。为了评价本文提出的个性化检索系统的有效性，对本文系统检索得到的结果与搜索引擎 Yahoo! 搜索得到的结果进行分析和比较。本文通过计算相关程度值来评价检索结果的有效性。相关程度值 *releance_deg* 定义为：

（下转第 249 页）

另外，该算法认为，如图 3 的情况是切矢方向改变最大的情况，即所有对原切矢方向的偏移不能超过 PI/4。因此引入一个中间矢量 V5。V5 的几何含义为 V3 向 V4 的方向旋转 PI/4。

具体的算法如下：

```
if(|v4<lengthThreshold)
        then Γ=v3
else
    if(angle(v3,v4)<angleThreshold)
        then v6=v3
        else
        if(angle(v3,v4)>π/4)
            then v6=between(v3,v5)
        else
            v6=between(v4,v5)
            Γ=between(v6,v3)
```

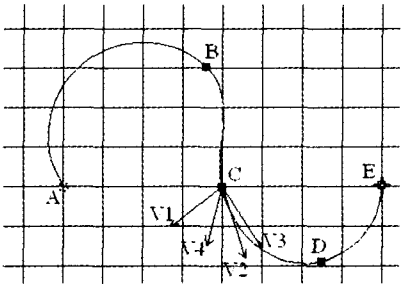

图 6　切矢

其中 Γ 为切矢，between(a,b)表示 a 和 b 中间的某值，通过函数因子算得。

在该算法中，为了简单起见，我们选用的函数因子均为一次函数，实验表明用一次函数构造的曲线能够保证变化的连续性，而且在视觉上有较好的变化均匀性。

4.2.3　控制点模长的确定算法

对于前面介绍的圆弧构造法，控制点相对于输入点的模长为 $L/(1+2*\cos(\theta/2))$，又因为需求一 b 中要求当三点成的圆弧为劣弧时，所形成的封闭曲线应为类椭圆形。我们对此提出的解决方案是以圆心角为 PI 分界，如果某一曲线片断的圆心角大于 PI，则该曲线片断的 θ 取值为 2 * PI-θ。

结束语　本文介绍了三次有理 Bezier 在刺绣工业打版 CAD 系统中的应用，并根据刺绣工业对曲线的特殊要求，给出了一种基于函数约束的 Bézier 造型方法。该方法已经应用到具体的系统中，应用效果良好。同时，本文提出的对有特殊造型需求曲线的构造方法对于其他类似的曲线应用具有很大的实用性。

参 考 文 献

1　王崇骏，于文涤，陈世福. 智能化刺绣 CAD 系统的绣法推理技术的研究. 计算机科学，2004，31 (8)
2　于文涤. 智能刺绣 CAD 系统中若干关键技术的研究：[南京大学硕士毕业论文]. 2005
3　李俊，张华，王崇骏，等. 智能化 CAD 系统中的工作流技术研究. 计算机科学，2005，32(3)：97～100
4　de Casteljau P. Outillages methods calcul：[Technical Report]. A. Citioen，Paris，1959
5　施法中. 计算机辅助几何设计于非均匀有理 B 样条. 北京：北京航空航天大学出版社，1994

(上接第 229 页)

$$releance\_deg=\sum_{i=1}^{n}\frac{(n-i+1)}{n}\times w_i \qquad (5)$$

其中，n 表示检索返回结果 URL 列表的前 n 个 Web 页面，i 表示检索返回结果中的第 i 个页面，w_i 表示第 i 个页面的权值，1 表示与检索主题相关，0 表示与检索主题不相关。因为考虑到大多数的用户通常只浏览全部检索结果的前 20 到 30 个页面，本文在实验中对检索结果的前 50 个页面进行分析和比较。本文系统的检索结果和传统搜索引擎 Yahoo! 的检索结果的相关程度的比较结果如表 1 所示。

表 1　查询结果的相关性程度值比较

查询项	前 50 个 URL 中文档的相关程度值	
	Yahoo	本文系统
Web 分类	9.12	23.46
粗糙集	12.30	21.26
Java	7.90	20.03
数据挖掘	13.53	26.69
农业专家系统	5.41	18.72

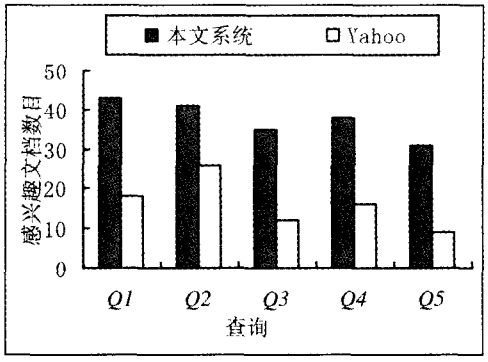

图 3　前 50 个 URL 中用户感兴趣的文档数目比较

本文系统的查询结果和传统的搜索引擎 Yahoo! 的查询结果中，用户感兴趣的文档数目的比较如图 3 所示。其中，查询项“Web 分类”、“粗糙集”、“Java”、“数据挖掘”和“农业专家系统”分别由 Q1、Q2、Q3、Q4 和 Q5 表示。

实验数据表明，对于本文系统检索结果的前 50 个 URL 而言，增加了用户感兴趣的文档数目，因此检索结果的相关性程度得到了显著的提高。

结论　本文中提出了一种基于粗糙集理论的个性化 Web 搜索系统。系统中采用了新的算法来提高检索过程的个性化，使得检索结果更贴近于用户的需求。随着 Web 的发展，个性化的 Web 检索系统一定会有广阔的应用前景。

参 考 文 献

1　Pazzani M，Muramatsu J，Billsus D. Syskill&Webert：Identifying Interesting Websites. In：Proc. Nat'l Conf. AI，AAAI，1996. 51～61
2　Joachims T，Freitag D，Mitchell T. WebWatcher：A Tour Guide for the World Wide Web. In：Proceeding of IJCA197，1997，08
3　Helmy T，Amamiya S，Amamiya M. Collaborative Kodama Agents with Automated Learning and Adapting for Personalized Web Searching. In：Proc. of the 13th Inter. Conference on Innovative Applications of AI (IAAI/IJCAI-2001)，2001. 65～72
4　Srinivasan P，Ruiz M E，Kraft D H，Chen J，Kundu S. Vocabulary Mining for Information Retrieval：Rough Sets and Fuzzy Sets. Information Processing and Management，2001，37：15～38
5　Pawlak Z. Rough sets：Theoretical aspects of reasoning about data. Kluwer Dordrecht，1991
6　王珏，苗夺谦. 关于 Rough Set 理论与应用的综述. 模式识别与人工智能，1996，9(4)：337～344