个性化搜索中的用户兴趣模型研究

宋 毅[1],徐志明[2]
(1.哈尔滨工业大学 华德应用技术学院 计算机应用技术系,黑龙江 哈尔滨 150025;
2.哈尔滨工业大学 计算机科学与技术学院,黑龙江 哈尔滨 150001)

摘 要:研究目的是挖掘搜索引擎中用户兴趣偏好,实现个性化搜索引擎技术。研究方法采用识别用户输入查询串,通过查询进行挖掘用户兴趣类别,但有时用户输入查询串短,或者出现查询词歧义等。由于查询会返回一系列文档,将相关文档分类处理,能够更清晰识别用户兴趣偏好。结果显示通过文档关系矩阵,将用户查询映射到对应类别,发现用户兴趣爱好。对于兼类查询等问题可以通过扩展查询解决。结论是该模型通过查询串和相关文档之间关系,进而实现用户偏好的辨别。该技术为搜索引擎信息推荐等技术打下良好基础。

关键词:搜索引擎;矩阵;类别;挖掘
中图分类号:TP39 **文献标识码**:A **文章编号**:1673-629X(2011)11-0153-03

Research of User Profile Model in Personalized Search

SONG Yi[1],XU Zhi-ming[2]
(1. Dept. of Computer Application and Technology, Huade School, Harbin University of Technology, Harbin 150025, China;
2. Computer Application and Technology Institute, Harbin University of Technology, Harbin 150001, China)

Abstract: The goal of the researches is digging user interest and realizing personalized search. The method of research is finding user query. Digging user class by query, but sometime the query is short or query is ambiguity. Query will return some documents and then class the document finding user interest. The result shows query is mapped to the class by document matrix. Query expansion effectively solves the question of query short and ambiguity class. The result is that the input of the model is user query and the document, the output is user interest that provides the foundation for sorting technology.

Key words: search engine; matrix; class; dig

0 引 言

个性化服务技术就是针对这个问题而提出的,它为不同用户提供不同的服务,以满足不同的需求。传统搜索引擎未能满足个性化搜索,个性化搜索引擎以方便用户需求为前提,分析用户上网行为特点,进而发现用户的兴趣爱好,个性化服务通过收集和分析用户信息来学习用户的兴趣偏好,从而实现主动推荐的目的。根据用户搜索历史中构建的用户模型,发现用户兴趣偏好,即用户兴趣挖掘,也就成为了个性化服务核心和关键技术[1~3],启发我们将用户查询和相关文档相结合,采用查询扩展及其文本分类等相关技术进行挖掘用户兴趣类别偏好,能够很好识别交叉类别兴趣,具有良好的可扩展性。

收稿日期:2011-04-25;修回日期:2011-07-27
基金项目:国家自然科学基金项目(60736044,60773070)
作者简介:宋 毅(1981-),女,吉林德惠人,硕士研究生,主要研究领域为自然语言处理、搜索引擎技术;徐志明,教授,博士,主要研究领域为自然语言处理、搜索引擎技术。

1 用户兴趣模型

1.1 用户兴趣模型表示

基于web内容的挖掘和基于web日志的挖掘[4~8],促使通过构建用户兴趣模型来表示用户搜索历史和用户兴趣的关系。Fan研究了查询分类问题[9],提出了通过映射查询到类别进行个性化搜索,但兼类查询的兴趣识别不够明显[10]。通过加入查询扩展改进了用户兴趣模型,不仅可以识别用户兴趣类别偏好,而且有效识别了兼类的查询。用户兴趣模型是描述用户兴趣偏好的特征空间。用户模型表示方法采用五元组结构,$M=[Q,C,R(q_i,c_j),T,D]$,具体实现在下面会详细说明。

1.2 系统结构

通过对挖掘相关文档和用户所关注的内容,形成查询和类别特征矩阵M_{QC},相关文档通过分类技术形

成文档类别特征矩阵 M_{DC}。通过 Rocchio 最终形成挖掘模型。挖掘系统具体内容如图 1 所示。

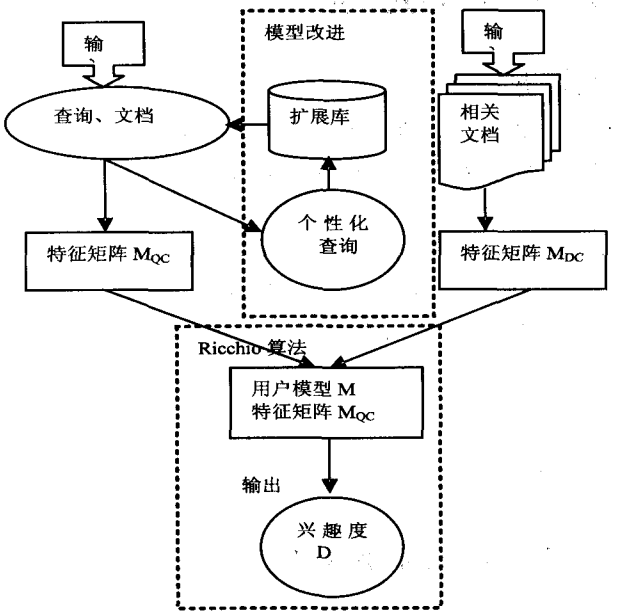

图 1 挖掘系统

1.2.1 文档和查询特征矩阵 M_{DQ}

文档和查询特征矩阵 M_{DQ}，行意味着文档，列意味着查询。例如查询“笔记本”在该文档中权重为 0.31。

1.2.2 文档和类别特征矩阵 M_{DC}

文本分类有很多分类技术，例如：支持向量机等。SVM 方法的优点是使用很少的训练集，计算量小、分类准确度高[11]，所以采用 SVM 分类器进行文本分类，例如体育分类结果为 0.92，教育分类结果为 0.16。分类语料中训练语料 31000 篇，14000 篇测试。

1.2.3 查询和类别特征矩阵 M_{QC}

表示用户搜索历史查询和类别特征矩阵 M_{QC}。

(1)特征矩阵降维。

向量空间(VSM)很庞大了，计算速度很慢，用奇异值分解(SVD)法，通过 SVD 方法将 M_{DQ} 分解成 $U * R * V^T$，计算特征矩阵：

$$M'_{DQ} = U * R^+ * V^T \tag{1}$$

式中 R^+ 为 R 的倒置；

(2)查询和类别特征矩阵 M_{QC}。

自学习机制的核心思想是通过反馈更新和改进当前用户模型的特征空间。目前较好的策略包括 LR 和 Rocchio 等。

$$D^t = M_{QC} = \frac{Ni^{t-1}}{Ni}\sum_{k=1}^{m} D^{t-1} + \frac{1}{Ni}\sum_{k=1}^{m} M^T_{DC(k,i)} M'_{QD(k,j)} \tag{2}$$

式中 D^t—在 t 时刻用户兴趣度；

N_i^{t-1}—从 0 时刻累加到 $t-1$ 时刻和第 i 类相关文档的数量；

D^{t-1}—在 $t-1$ 时刻的兴趣度；

Ni—第 i 类相关文档数量。查询“火箭”在体育类权重为 0.78，兴趣度在体育类为 0.86。说明查询“火箭”是查询体育的火箭队。而不是军事中的火箭信息。

(3)用户兴趣模型。

查询和类别特征矩阵 M_{QC} 中一行代表一类，用户在体育、军事、教育、汽车、旅游、IT 六类下的兴趣，每一列为用户输入 query，每行求出平均值，可以表示出用户对该类的兴趣度 d，矩阵映射成向量空间表示形式，$M = [Q, C, R(q_i, c_j), T, D]$ 其中，Q：用户需求的 query 的表示；C：表示类别；R：用户所关注问题；T：表示访问周期；D：兴趣爱好程度。将类别按兴趣 d 的权值大小排序，可以挖掘。通过 $R(q_i, c_j)$ 得出对应类别为体育的兴趣度值 0.97，表示用户对体育兴趣度相对较高。

2 实 验

用户搜索历史表示为查询和浏览相关文档，实验得出结果见图 2，挖掘发现变化情况如下：

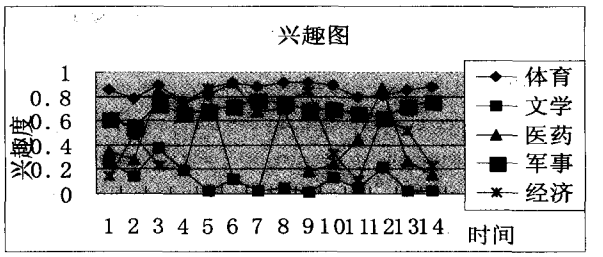

图 2 兴趣趋势折线图

将类别按兴趣度排序，权值越高的表明用户偏爱程度大，将兴趣度大的类别推荐给用户。

3 模型改进

通过上面兴趣挖掘中，发现存在用户查询是兼类，查询扩展目的是解决查询属于兼类的用户兴趣类别偏好。查询扩展方法采用基于词典和 Rocchio 相关反馈相结合的方法进行扩展查询。

(1)基于搜狗词典进行查询扩展。

搜狗拼音输入法可以覆盖几乎所有的中文词汇，所以词典采用搜狗细胞词库。

(2)基于 Rocchio 算法[12]识别用户兴趣算法。

扩展方法是从与该查询类别相关的文档中特征词进行提取，相似度高的加入查询扩展词库，抽取相关类文档中的特征词，进行查询扩展，阈值高的放入扩展词库，阈值低的进行过滤处理[9,13,14]。

$$q_m = \alpha q_0 + \beta \frac{1}{Ni}\sum_{d_j \in Ni} d_j \tag{3}$$

$$W' = W_{qd} * \sum_{q' \neq q, q' \in DS} \text{cooc}(q, q') \quad (4)$$

式中 w_{qd} ——查询在文档中的权重;

cooc (q,q') —— 查询 q 扩展 q' 的可信度。

经过查询扩展技术,将扩展库中的扩展词放入用户模型中重新计算用户兴趣类别偏好,得到实验结果见图 3。

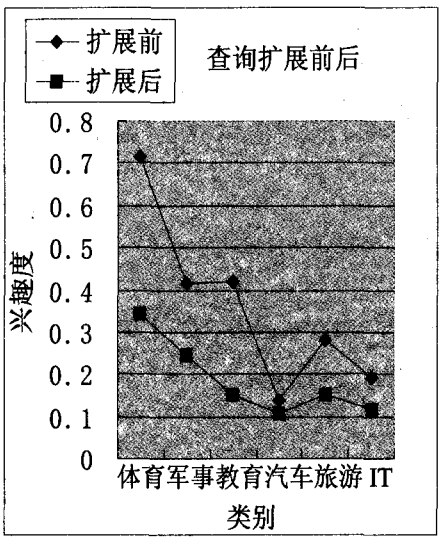

图 3　扩展后兴趣类别偏好

经过查询扩展后有效识别了用户兴趣类别偏好。

4　结束语

研究并且挖掘偏好建模, $M = [Q, C, R(q_i, c_j), T, D]$,实现了挖掘技术,在实现用户兴趣模型中,用户兴趣模型特点是:通过挖掘用户相关信息,每天用户行为内容发现用户的特点。对于兼类查询问题处理效果良好,具有很好的扩展性。当然,该方法有不足之处,随着时间变化,用户兴趣也会变化。这就要求用户兴趣模型需要学习和更新[15,16]。

参考文献:

[1] Fragoudis D. User Modeling in Information Discovery: An Overview[C]//Proceedings of Advanced Course on Artificial Intelligence. [s. l.]:[s. n.],1999:17-43.

[2] 张　炜. 个性化推荐系统中基于本体的用户兴趣挖掘研究[D]. 南京:南京理工大学,2007.

[3] Claypool M, Le P, Waseda M, et al. Implicit interest indicators [C]//In: Proceedings of Intelligent User Interfaces. [s. l.]: [s. n.],2001:30-40.

[4] Qiu Feng, Cho J. Automatic Identification of User Interest For Personalized Search [C]//International World Wide Web Conference Committee. Edinbufh, Scotland: ACM, 2006: 23-26.

[5] Chen L, Sycara K. Webmate: a personal agent for browsing and searching[C]//Proc. 2nd Int. Conf. on Autonomous Agents and Multiagent Systems. [s. l.]:[s. n.],1998:132-139.

[6] Pretschner A, Gauch S. Ontology based personalized search [C]//ICTAI. [s. l.]:[s. n.],1999:391-398.

[7] Sugiyama K, Hatano K, Yoshikawa M. Adaptive web search based on user profile constructed without any effort from users [C]//Proceedings of the Thirteenth Int World Wide Web Conf. [s. l.]:[s. n.],2004:5-8.

[8] 郭　岩,白　硕. 网络规模日志分析和用户兴趣挖掘[J]. 计算机学报,2005,28(9):1-3.

[9] 彭彬彬,金翔宇,徐晓刚,等. 基于 svm 增量学习的用户适应性研究[J]. 计算机科学,2003(30):75-76.

[10] 李村合,杨献峰,张培颖. 基于 web 挖掘与相关反馈的多层次用户兴趣挖掘算法[J]. 微计算机应用,2007,28(9):1-2. [11] Salton G, Buckley C. Improving retrieval performance by retrieval feedback[J]。Journal of the American Society for Information Science,1990(4):288-297.

[12] Liu Fan, Yu C, Meng Weiyi. Personalized Web Search by Mapping User Queries to Categories[C]//Conference on Information and Knowledge Management. Mclean, Virginia, USA: ACM,2002:4-9.

[13] 曾　春,邢春晓,周立柱. 个性化服务技术综述[J]. 软件学报,2002(13):1953-1955.

[14] 陈　媛,苟光磊. 个性化服务用户模型研究[J]. 计算机工程设计, 2008,29(9):2413-2416.

[15] 岳　文,陈治平,林亚平. 基于查询扩展和分类的信息检索算法[J]. 系统仿真,2006,18(9):1926-1929.

[16] 费洪晓,穆王君,刘　正. 基于文本聚类和权重调整的用户兴趣建模算法[J]. 计算机技术与发展,2007,17(2):128-129.

(上接第 152 页)

大学出版社, 2005:360-364.

[9] Bhatt R B, Gopal M. On fuzzy-rough sets approach to feature selection[J]. Pattern Recognition Letters,2005,26:965-975.

[10] Hu Q H, Pedrycz W, Yu D R, et al. Selecting Discrete and Continuous Features Based on Neighborhood Decision Error Minimization[J]. IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics,2010,40(1): 137-150.

[11] Hu Q H, Yu D R, Liu J F, et al. Neighborhood rough set based heterogeneous feature subset selection [J]. Information Sciences,2008,178:3577-3594.

[12] Hu Q H, Yu D R, Xie Z X. Neighborhood classifiers[J]. Expert Systems with Applications,2008,34:866-876.

[13] 刘宗田. 属性最小约简的增量式算法[J]. 电子学报,1999,27(11):96-98.

[14] 林俊伟,叶东毅. 基于邻域辨识矩阵的属性约简增量式算法[J]. 计算机应用,2009,29(6): 119-121.

[15] 夏富春,苗夺谦,李道国. 信息系统属性增量约简算法的设计与实现[J]. 计算机工程与应用,2006,21:149-152.