第35卷 第10期 计 算 机 工 程 2009年5月
Vol.35 No.10 Computer Engineering May 2009
·软件技术与数据库· 文章编号：1000—3428(2009)10—0078—02 文献标识码：A 中图分类号：TP391

基于查询行为和关联规则的相关反馈查询扩展

黄名选[1]，张师超[2]，严小卫[2]
(1. 广西教育学院数学与计算机科学系，南宁 530023；2. 广西师范大学计算机科学与信息工程学院，桂林 541004)

摘 要：针对现有查询扩展缺陷，提出基于用户查询行为和词间完全加权关联规则挖掘的相关反馈查询扩展算法。在不改变用户查询信息习惯的前提下，无须用户参与，根据用户查询行为判断初检文档的相关性，提取相关的初检文档，挖掘与原查询相关的关联规则，构造规则库，从中提取与原查询相关的扩展词，实现查询扩展。实验结果表明，该算法能提高信息检索性能，具有很好的应用前景。

关键词：查询扩展；关联规则；相关反馈；信息检索

Query Expansion of Relevance Feedback Based on Users' Query Behaviors and Association Rules

HUANG Ming-xuan[1], ZHANG Shi-chao[2], YAN Xiao-wei[2]
(1. Department of Math and Computer Science, Guangxi College of Education, Nanning 530023;
2. College of Computer Science & Information Technology, Guangxi Normal University, Guilin 541004)

【Abstract】 Aiming at the limitations of existing query expansion, this paper proposes a novel query expansion algorithm of relevance feedback based on users' query behaviors, as well as the technique of item-all-weighted association rule mining in retrieved relevance documents. According to the duration of user's clicking and browsing, or the existence of some querying behaviors such as downloading, this algorithm is able to determine whether a document is related to users' query intentions and interests, automatically extract those item-all-weighted association rules related to original query from retrieved relevance documents to construct an association rules-based database, and collect terms related original query as expansion terms from the database. Experimental results show the retrieval performance of the algorithm is improved remarkably.

【Key words】 query expansion; association rules; relevance feedback; information retrieval

查询扩展是改善和提高信息检索性能的关键技术之一，它指利用计算机语言学、信息学等多种技术，把与原查询相关的语词添加到原查询中，得到新查询后再次检索文档，以改善信息检索系统的查全率和查准率，解决信息检索领域的词不匹配问题。传统的查询扩展技术主要有全局分析[1]、局部分析[2]以及基于用户查询日志[3]和基于关联规则挖掘的查询扩展[4]。在分析现有查询扩展的不足，并对用户查询信息的行为进行研究和探讨的基础上，本文提出基于用户查询行为和词间完全加权关联规则挖掘的相关反馈查询扩展算法。

1 面向查询扩展的完全加权词间关联挖掘算法

1.1 完全加权词间关联规则挖掘过程

完全加权词间关联规则挖掘[5]针对基于向量空间模型的文本数据库，在进行词间关联规则挖掘时充分考虑了各个特征词项在不同文档记录中有着不同权重。面向查询扩展的词间完全加权关联规则挖掘的基本思想是只挖掘含有查询词项的完全加权关联规则，其挖掘过程如图1所示。

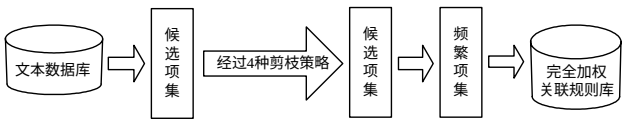

图1 面向查询扩展的完全加权关联规则挖掘过程

1.2 完全加权词间关联规则挖掘算法描述

算法 AWARMFQE

输入 文本数据库 *D*，最小完全加权支持度(*minawsupport*)和置信度阈值(*minawconf*)

输出 完全加权强关联规则

Begin

扫描数据库 D，找出可能的最大项目集的项目个数，事务记录总数和项目总数。

累加各个1-项集的支持数及权值，求其最大权值以及包含1_项集的2_权值阈值，产生 C_1(候选1_项集)，C_1^-(是不可能成为频繁2_项集的候选1_项集)，L_1(频繁1_项集)。

for (k=2; ; k++) {

由 C_{k-1} 连接生成 C_k ； // C_{k-1} 为候选 k-1_项集，C_k 为候选//k_//项集

if(k==2) 进行第1次剪枝：将不含原查询项的候选2_项集剪掉，只保留含有原查询项的候选2_项集；

if (C_{k-1}^- 和 C_k 都不空) 进行第2次剪枝：对于 C_k 的各个(k-1)_子项集，只要存在其某个子项集的权值之和值小于其 k_权值阈值[5]，根据文献[5]中的定理2，就可以删除该候选项集；

// C_{k-1}^- 为不可能成为频繁 k_项集的候选 k-1_项集。

for 每一个事务记录

累加候选项集 C_k 在数据库 D 中出现频度；

if (C_k 的出现频度为0) 进行第3次剪枝：剪掉频度为0的完全加权候选项集(C_k)；

基金项目：国家自然科学基金资助项目(60496327, 60463003)；广西教育厅科研基金资助项目(200808MS192)

作者简介：黄名选(1966－)，男，副教授、硕士，主研方向：数据挖掘，查询扩展；张师超、严小卫，教授、博士

收稿日期：2009-02-06 **E-mail：**mingxh05@163.com

for 每一个事务记录

遍历文本数据库 D，统计 C_k 中所有候选项集的权值之和和包含 C_k 的(k+1)-权值阈值；

进行第 4 次剪枝：据文献[5]中的定理 1，剪掉其权值之和值低于其相应的 k-权值阈值[5]的完全加权候选项集；

生成频繁项目集，并入库；

输出频繁项集；

if(C_k 为空或者 k 大于最大项目集的项目个数)退出循环；

}

输出完全加权强关联规则；

End

2 基于查询行为和词间加权关联挖掘的查询扩展

2.1 算法基本思想

本文提出的查询扩展算法的基本思想是：在不改变用户查询习惯的情况下，通过传统的向量空间模型检索算法(即 tf-idf 算法)对文档集进行初次检索，检索系统将初检文档返回给用户，然后根据用户点击浏览初检文档时间长短或文档是否存在被用户下载操作等查询行为来判断初检文档与原查询的相关性，采用 AWARMFQE 算法对相关的初检文档进行挖掘，提取含有原查询词的关联规则，构造规则库，从中提取与原查询相关的扩展词，实现查询扩展。

2.2 相关文档的确定及扩展词权重的计算

通常用户在查询信息时，如果某篇文档是用户感兴趣的，或是用户需要的，或用户觉得和原查询相关的，则其在这篇文档上停留的时间会长些，甚至有下载操作。否则，用户不会仔细浏览，也不会下载，停留的时间也很短。因此，可以将用户的点击浏览时间(*browse_time*)和下载(*download*)操作作为判断文档相关性的 2 个参数。相关文档的具体确定方法是：规定一个浏览时间阈值，如果某篇文档被用户点击浏览的时间超过了这个阈值，或存在下载操作，则认为该篇文档是用户感兴趣的，是与原查询相关的，应该提取出来作为关联规则挖掘的相关文档数据集。

在查询扩展中，原查询项永远是最重要的，最能反映用户查询意图，应该具有最高的权重。因此，本文规定原查询的各个查询项权重为 2，而扩展词的权重 W_{expt} 计算公式为

W_{expt}=(与扩展词相关的前件或后件的查询项个数/原查询中所有查询项总个数)×规则置信度

2.3 查询扩展算法描述

算法 基于查询行为和词间加权关联挖掘的查询扩展算法

输入 原查询 *Q*, *minawsupport*, *minawconf*, *browseTIME*(点击浏览时间阈值)，*download*(是否存在下载操作的参数)，*related_number*(通过捕捉查询行为获得的相关文档数参数)

输出 与原查询 Q 相关的扩展词集合和新查询

Begin

(1)搜索引擎对用户查询 Q 初检，并返回初检结果。

(2)捕捉用户对初检文档的处理行为。跟踪用户查询行为，若某文档被浏览的时间不低于 browseTIME，或存在被用户下载(download==ture)操作，则提取该文档，组成相关初检文档集，同时 related_number 记数，当 related_number 达到某个给定的常数时才往下进行。

(3)用 AWARMFQE 算法对相关文档集进行完全加权关联规则挖掘，提取与原查询项相关的强关联规则，构建规则库。

(4)从规则库中提取扩展词，计算其权重并排序，最后存入扩展词库。

(5)从扩展词库中提取其权值不低于 W_Threshold(扩展词权重阈值)的前列扩展词，或者提取前列 m 个扩展词，和原查询一起构成新查询，实现查询扩展。

(6)输出查询扩展后的新查询。

End

3 实验设计及结果分析

3.1 实验测试的文档集、查询集

由于搜索引擎的研究范围很广，因此本实验只是一个模拟实验，是在传统的基于向量空间模型的检索系统中完成的。实验测试文档集是从网上下载的 720 篇计算机方面的论文，同时，设计了 10 个实际的查询(Q_1, Q_2,…, Q_{10})作为查询集供实验用。在原始测试文档集中通过人工检索比较，获得这 10 个查询的相关文档篇数。对原始测试文档集和查询集进行语词切分、去掉停用词、抽取特征词、计算其权值等常规文档预处理。

3.2 实验结果分析

将本文算法与基于完全加权关联规则挖掘的局部反馈查询扩展算法(All-weighted-ARMing-based QE)、基于 Apriori 算法[6]的局部反馈查询扩展算法(Apriori-based QE)、基于局部上下文分析的查询扩展[2]算法(LCA-based QE)和没有查询扩展的向量空间模型检索算法(tf-idf)进行检索性能比较。实验参数设定如下：扩展词数量统一取 30，并进行权重规范化处理；*related_number* 取 15，*browse_time* 取 30 s；在挖掘时，最小完全加权支持度和置信度阈值都设为 0.01。5 种算法的实验结果如图 2 所示。

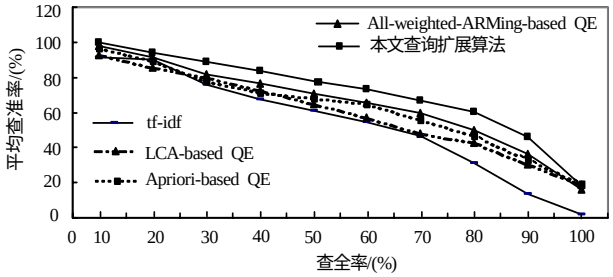

图 2 相关反馈查询扩展的查全率和查准率

对照组在相同查全率水平下，本文算法与 tf-idf 相比平均查准率提高 32.93%，与 All-weighted-ARMing- based QE 相比提高 9.69%，与 Apriori-based QE 相比提高 13.90%，与 LCA-based QE 相比提高 19.67%。

实验结果表明，本文扩展算法的检索性能最优，它克服了局部反馈和用户相关反馈的查询扩展的不足，在不改变用户查询信息习惯的情况下，通过捕捉用户查询行为获得的初检文档都是与原查询相关的或是用户感兴趣的，通过挖掘得到的扩展词噪音较少，即正相关的扩展词较多，而负相关或假相关的扩展词较少。

另外，在挖掘词间关联规则时，充分考虑了完全加权的项权值，使关联规则和从中获得的扩展词比较合理。而 All-weighted-ARMing-based QE 和 Apriori-based QE 的局部文档集中难以避免存在一定量的与原查询不相关或相关性很低的文档，其得到的扩展词噪音就多，使查询扩展效果变差。Apriori-based QE 采用的挖掘算法是 Apriori 算法，和 LCA-based QE 一样，只考虑从初检文档中选出与原查询词共现的扩展词，并没有考虑扩展词在不同事务文档中具有不同权重，因此，其检索性能不如本文算法。

(下转第 82 页)