ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.20, No.7, July 2009, pp.1854−1865
doi: 10.3724/SP.J.1001.2009.03368

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于矩阵加权关联规则挖掘的伪相关反馈查询扩展*

黄名选[1+], 严小卫[3], 张师超[2,3]

[1](广西教育学院 数学与计算机科学系,广西 南宁 530023)

[2](中山大学 逻辑与认知研究所,广东 广州 510275)

[3](广西师范大学 计算机科学与信息工程学院,广西 桂林 541004)

Query Expansion of Pseudo Relevance Feedback Based on Matrix-Weighted Association Rules Mining

HUANG Ming-Xuan[1+], YAN Xiao-Wei[3], ZHANG Shi-Chao[2,3]

[1](Department of Math and Computer Science, Guangxi College of Education, Nanning 530023, China)

[2](Institute of Logic and Cognition, SUN YAT-SEN University, Guangzhou 510275, China)

[3](College of Computer Science and Information Engineering, Guangxi Normal University, Guilin 541004, China)

+ Corresponding author: E-mail: huangmx@mailbox.gxnu.edu.cn

Huang MX, Yan XW, Zhang SC. Query expansion of pseudo relevance feedback based on matrix-weighted association rules mining. *Journal of Software*, 2009,20(7):1854−1865. http://www.jos.org.cn/1000-9825/3368.htm

Abstract: An algorithm of matrix-weighted association rule mining for query expansion is presented based on the quadruple pruning, and a related theorem and its proof are given. This method can tremendously enhance the mining efficiency. Experimental results demonstrate that its mining time is averagely reduced by 87.84%, compared to that of the original one. And a query expansion algorithm of pseudo relevance feedback is proposed based on matrix-weighted association rule mining, which combines the association rules mining technique with the query expansion. The algorithm can automatically mine those matrix-weighted association rules related to the original query in the top-ranked retrieved documents to construct an association rules-based database, and extract expansion terms related to the original query from the database for query expansion. At the same time, a new computing method for weights of expansion terms is given. It makes the weighted value of an expansion term more reasonable. Experimental results show that this method is better than traditional ones in average precision.

Key words: information retrieval; pseudo relevance feedback; query expansion; association rule; matrix-weighted

摘 要: 提出一种面向查询扩展的矩阵加权关联规则挖掘算法,给出与其相关的定理及其证明过程.该算法采用 4 种剪枝策略,挖掘效率得到极大提高.实验结果表明,其挖掘时间比原来的平均时间减少 87.84%.针对现有查询扩展

* Supported by the National Natural Science Foundation of China under Grant No.90718020 (国家自然科学基金); the National Basic Research Program of China under Grant No.2008CB317108 (国家重点基础研究发展计划(973)); the Australian Research Council Discovery under Grant No.DP0667060 (澳大利亚 ARC 项目); the MOE Project of Key Research Institute of Humanities and Social Sciences at Universities of China under Grant No.07JJD720044 (教育部人文重点研究基地重大项目)

Received 2007-10-10; Revised 2008-01-10; Accepted 2008-04-15

的缺陷,将矩阵加权关联规则挖掘技术应用于查询扩展,提出新的查询扩展模型和更合理的扩展词权重计算方法.在此基础上提出一种伪相关反馈查询扩展算法——基于矩阵加权关联规则挖掘的伪相关反馈查询扩展算法,该算法能够自动地从前列 n 篇初检文档中挖掘与原查询相关的矩阵加权关联规则,构建规则库,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,该算法的检索性能确实得到了很好的改善.与现有查询扩展算法相比,在相同的查全率水平级下,其平均查准率有了明显的提高.

关键词: 信息检索;伪相关反馈;查询扩展;关联规则;矩阵加权

中图法分类号: TP311 **文献标识码**: A

查询扩展(query expansion)是提高信息检索系统性能的关键技术.它利用计算机语言学、信息学等多种技术,以用户原查询为基础,把与原查询相关的词或者词组添加到原查询,得到比原查询更长的新查询,以便更完整、更准确地描述原查询所隐含的语义或主题,帮助信息检索系统提供更多有利于判断文档相关性的信息,达到改善和提高信息检索系统的查全率和查准率.其核心问题是如何设计和利用扩展词的来源.

查询扩展主要分为全局分析的、局部分析的、基于用户查询日志的和基于关联规则挖掘的查询扩展 4 类方法[1].基于全局分析的查询扩展是对全部文献集合中的词或词组进行相关分析,计算每对词或词组间的关联程度,将与用户查询关联程度较高的词或者词组加入原查询,其主要技术有聚类算法[2]、潜在语义索引(latent semantic indexing,简称 LSI)[3]和相似性词典[4]等等;基于局部分析的查询扩展是利用初检出的与原查询最相关的 n 篇文档作为扩展词的来源,其主要技术有伪相关反馈(pseudo relevance feedback)(也称局部反馈)[5]、用户相关反馈[6]和局部上下文分析[7]等等.而基于伪相关反馈的查询扩展是通过假定初检前列文档与原查询相关来模拟用户相关反馈的方法;基于用户查询日志的查询扩展[8,9]其基本思想是,充分利用用户的查询日志分析词间各种关联,自动选择与原查询高度相关的词或词组实现查询扩展;基于关联规则挖掘的查询扩展是近年来兴起的新研究方向,已经得到了广泛的关注.它的核心思想是,通过数据挖掘技术挖掘词间关联规则,将关联规则的后件或者前件作为扩展词的来源.基于关联规则挖掘的查询扩展主要有以下 3 种方法:

第 1 种方法[8−10]是从搜索引擎查询日志中提取与原查询相关的词间关联规则作为扩展词的来源,实现查询扩展.文献[8,9]的实验表明,这种扩展方法对于查询短小、文档集内容比较分散的情形尤其适用,可以极大地提高查询精度和查全率.不足之处是首先须有大量的用户查询日志存在,需要有一个积累的过程,而且基本上要求大量用户有共同的兴趣,还需要在服务器端实现.

第 2 种方法是[11−16]从全局分析的角度对整个文档集进行词间关联规则挖掘,构造规则库,从规则库提取扩展词实现查询扩展.然而,基于全局分析的方法要推广到实际应用层次,在目前的技术环境下,难度很大,因为全局分析下的文档集必然很大,由于频繁项集的数量是随着数据库中数据项数目的增加呈指数增长,在全局分析下的文本数据库中数据项一般都有数千,甚至数万,因此,即使采取各种剪枝策略,要处理的候选项集和频繁项集的数量还是非常多,致使挖掘文档词间关联规则的效率和时间无法使用户接受,而用户查询信息时都追求速度快、信息全而准.

第 3 种方法是[17]从局部分析的角度对初检局部文档集进行词间关联规则挖掘,从中发现与原查询相关的扩展词实现查询扩展.文献[17]中的实验表明,这种方法只针对局部相关文档进行词间挖掘,极大地提高了查询扩展效率,是一种应用前景较好的查询扩展方法.

但是,这些研究通常不重视关联规则的挖掘技术及其质量对查询扩展检索性能的影响,也不考虑在挖掘词间关联规则时其特征词在不同的事务文档记录中有不同的重要性.针对这些问题,本文设计了一种面向查询扩展的矩阵加权关联规则挖掘算法,有效地将矩阵加权关联规则挖掘技术运用于伪相关反馈查询扩展中,它具有以下优点:

(1) 与传统的向量空间模型检索算法(tf-idf 算法)的实验结果比较表明,在相同的查全率水平级下,其平均查准率提高了 21.19%;与基于局部上下文分析的查询扩展[7]算法相比,其平均查准率提高了 9.10%;与基于无加权的 Apriori 算法[18]的局部反馈查询扩展相比,其平均查准率也有明显的提高.

(2) 采用本文提出的面向查询扩展的矩阵加权关联规则挖掘算法不仅能够获得更加实际、合理的相关扩展词,而且提高了查询扩展效率.

本文第 1 节阐述面向查询扩展的矩阵加权关联规则挖掘策略及其算法.第 2 节详细论述基于矩阵加权关联规则挖掘的伪相关反馈查询扩展算法.第 3 节是实验及其结果分析.最后给出全文的总结,并指出进一步研究的方向.

1 面向查询扩展的矩阵加权关联规则挖掘算法

1.1 相关概念和定理

定义 1. 矩阵加权关联规则模型:设 $T=\{t_1,t_2,\ldots,t_n\}$是一个文本数据库,t_j 表示 T 中的第 j 个记录(文档),$I=\{i_1,i_2,\ldots,i_m\}$表示文本数据库的特征词项集,$W=[W[t_j][i_p]]_{j\times p}$ 表示特征词项 i_p 在文档 t_j 中的矩阵权值($1\le j\le n,1\le p\le m$).其中,如果 i_p 不在 t_j 中,则 $W[t_j][i_p]=0$.

令 $X\subset I,Y\subset I$ 且 $X\cap Y=\varnothing$,则有如下定义:

定义 2. 特征词项集(X,Y)的矩阵加权关联规则可以表示为 $X\to Y$.

定义 3. 矩阵加权关联规则支持度:

$$mwsupport(X,Y)=\frac{1}{n\times k}\left(\sum_{t_j\in T}\sum_{i_p\in(X\cup Y)}W[t_j][i_p]\right) \tag{1}$$

其中,k 为项集$\{X\cup Y\}$的项目数,n 为数据库的记录数.

定义 4. 矩阵加权关联规则置信度:

$$mwconf(X,Y)=\frac{mwsupport(X\cup Y)}{mwsupport(X)} \tag{2}$$

定义 5. 矩阵加权频繁项集(X,Y).即满足 $mwsupport(X,Y)\ge\min mwsupport$ 的项集.其中,$\min mwsupport$ 为最小矩阵加权支持度阈值.

定义 6. 矩阵加权强关联规则$(X\to Y)$:即满足,$mwsupport(X,Y)\ge\min mwsupport$ 和 $mwconf(X,Y)\ge\min mwconf$ 的关联规则.其中,$\min mwconf$ 为最小矩阵加权置信度阈值.

定义 7. k-权值阈值(k-item weighted threshold,简称 KIWT).

令 $I_1\subset I$ 是 q-项集,且 $q\le k(k\le m)$.在$(I-I_1)$项集中,记前$(k-q)$个权值最大的项为 $t_{r_1},t_{r_2},\ldots,t_{r_{k-q}}$,其相应的权值为 $w_{r_1},w_{r_2},\ldots,w_{r_{k-q}}$,项集 I_1 在文本数据库中的出现次数为 $SC(I_1)$,假设包含 q-项集 I_1 的 k-项集是频繁的,则很明显,包含项集 I_1 的 k-项集最大可能权值之和应为 $\mathrm{Max}w(I_1,k)$[19],即

$$\mathrm{Max}w(I_1,k)=\sum_{t_j\in T}\sum_{i_p\in(I_1)}w[t_j][i_p]+SC(I_1)\sum_{l=1}^{k-q}w_{r_l} \tag{3}$$

$$\frac{\mathrm{Max}w(I_1,k)}{n\times k}\ge\min mwsupport \tag{4}$$

$$\Rightarrow\sum_{t_j\in T}\sum_{i_p\in(I_1)}w[t_j][i_p]\ge n\times k\times\min mwsupport-SC(I_1)\sum_{l=1}^{k-q}W_{r_l} \tag{5}$$

称公式(5)右边部分为包含 q-项集 I_1 的 k_项集权值阈值($KIWT(I_1,k)$),简称 k-权值阈值,即

$$KIWT(I_1,k)=n\times k\times\min mwsupport-SC(I_1)\sum_{l=1}^{k-q}W_{r_l} \tag{6}$$

则

$$\sum_{t_j\in T}\sum_{i_p\in(I_1)}w[t_j][i_p]\ge KIWT(I_1,k) \tag{7}$$

公式(7)表明,如果 q-项集 I_1 的权值之和不低于 k-权值阈值,那么包含 I_1 的 k-项集很有可能是频繁项集.

定理 1. 如果矩阵加权 q-项集 I_1 的权值之和小于 k-权值阈值,那么包含 I_1 的矩阵加权 k-项集一定是非频繁

项集.

证明:由题设可知,$\sum_{t_j\in T}\sum_{i_p\in(I_1)} w[t_j][i_p] < KIWT(I_1,k)$,则

$$\sum_{t_j\in T}\sum_{i_p\in(I_1)} w[t_j][i_p] < n\times k\times \min mwsupport - SC(I_1)\sum_{l=1}^{k-q} W_{\eta_l} \tag{8}$$

结合公式(3),公式(8)可以转换为

$$\frac{\mathrm{Max}w(I_1,k)}{n\times k} < \min mwsupport \tag{9}$$

公式(9)左边部分正好是包含完全加权 q-项集 I_1 的 k-项集最大可能支持度.由此可知,包含 I_1 的完全加权 k-项集一定是非频繁项集.证毕. □

定理 2. 对于矩阵加权 k-项集的任何子集,只要至少存在一个子集的权值之和小于其 k-权值阈值,则该 k-项集一定是非频繁项集.

证明:根据题设,对于矩阵加权 k-项集的任何子集,至少存在一个子集的权值之和小于其 k-权值阈值,不妨令该子集为 T_{sub},则由定理 1 可知,包含 T_{sub} 的 k-项集一定是非频繁项集.证毕. □

1.2 剪枝策略

1.2.1 4 种剪枝策略

面向查询扩展的矩阵加权关联规则挖掘算法采用了 4 种剪枝策略,极大地提高了挖掘效率,具体是:

第 1 种剪枝策略:当挖掘到矩阵加权候选 2_项集时,只保留含有原查询项的矩阵加权候选 2_项集,而将不含有原查询项的矩阵加权候选 2_项集剪掉.

第 2 种剪枝策略:对于各个矩阵加权候选项集 C_k,检查其各个(k–1)_子集,只要存在它的某个子集出现在 $\overline{C}_{k-1}$ ($\overline{C}_{k-1}$ 是指不可能成为频繁 k-项集的(k–1)-项集集合)中,则根据定理 2,可以从矩阵加权候选项集集合中删除该候选项集 C_k.

第 3 种剪枝策略:将频度为 0 的候选项集剪掉,因为频度为 0 的候选项集不可能成为矩阵加权频繁项集.

第 4 种剪枝策略:根据定理 1,对每个矩阵加权候选项集 C_k 进行其权值之和与其相应的 k-权值阈值的比较:如果其权值之和不低于其相应的 k-权值阈值,则保留;否则,将其保存到 $\overline{C}_k$ 中($\overline{C}_k$ 是指不可能成为频繁(k+1)-项集的 k-项集集合),并从矩阵加权候选项集集合中删除该候选项集 C_k.

1.2.2 实验及其结果分析

上述剪枝策略不仅不影响查询扩展的效果,反而大大提高了挖掘速度,是一种很好的策略.其中,第 1 种剪枝策略对扩展词的挖掘效率影响最大.下面首先从理论上对第 1 种剪枝策略进行分析,然后通过实验加以说明.

(1) 理论分析

由于查询扩展的核心问题是如何设计和利用与原查询项相关的扩展词的来源,因此,从当矩阵加权关联规则中提取扩展词时,那些含有原查询项的矩阵加权候选项集、频繁项集和关联规则才对查询扩展有实际意义.相对于整个数据库的特征项,原查询项的特征项是极少数.因此,只挖掘含有原查询项的矩阵加权关联规则能够节省很多挖掘时间.选择在矩阵加权候选 2_项集进行剪枝,可以保证在后续挖掘中产生的矩阵加权候选项集都是含有原查询项的候选项集.虽然这种剪枝策略也会剪掉一部分频繁项集和关联规则,但剪掉的频繁项集和规则绝大部分都不含有原查询项,即使存在少量的含有查询项的关联规则,也不会影响查询扩展的效果.因为查询项和扩展项往往都重复出现在各个频繁项集和规则中.但是,按照所给定的查询扩展模型,所需的扩展词都能从现有的关联规则中提取,实现查询扩展.

(2) 实验分析

为了验证第 1 种剪枝策略既能极大地提高挖掘效率又不影响其查询扩展效果,我们编写了源程序,以进行对比实验.具体的实验方法叙述如下:

- 实验材料:从网上下载 52 篇有关数据挖掘的论文作为实验文档集,对该文档集进行文档预处理.实验中,

假设用户查询为“文本挖掘”,即原查询 Q={文本,挖掘},特征词数量分别取 5,10,20,30,40,50;

- 实验方法:采用 MWARM(matrix-weighted association rules mining)算法(详见第 1.4 节),在给定的查询扩展模型下(详见第 2.4 节)分两种情况进行实验,比较其挖掘时间、频繁项集、关联规则和所获得的扩展词个数及其权值:一种是挖掘时采用第 1 种剪枝策略;另一种是不采用第 1 种剪枝策略.

实验结果:如图 1 和表 1 所示的实验结果表明,挖掘时采用第 1 种剪枝策略,其挖掘时间平均减少 87.84%,最大时可以减少 99.13%,候选项集比原来的平均值减少了 96.20%,频繁项集数量比原来的平均值减少了 94.29%,关联规则数量也比原来的平均值减少 93.04%,而两种挖掘方式中所获得的扩展词及其权重完全相同.因此,采用第 1 种剪枝策略不但可以极大地提高挖掘效率,而且不会影响其查询扩展效果(扩展词及其权重不变),是一种很好的策略.

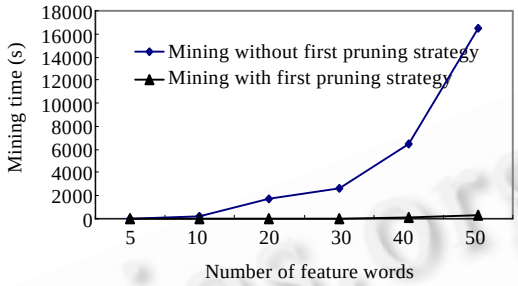

Fig.1 Comparison of mining time (s) of two mining techniques (*mwsupport*=0.1, *mwconf*=0.01)

图 1 两种情况的挖掘时间比较(*mwsupport*=0.1,*mwconf*=0.01)

Table 1 Comparison of experimental results of two mining techniques (*mwsupport*=0.11,*mwconf*=0.03)

表 1 两种挖掘方式的实验结果比较表(*mwsupport*=0.11,*mwconf*=0.03)

Number of feature words	Number of frequent itemsets		Number of association rules		Number of expansion terms		Comparison of expansion terms and its weights obtained
	Without first pruning	With first pruning	Without first pruning	With first pruning	without first pruning	With first pruning	
5	23	15	52	32	3	3	The same
10	197	29	514	82	8	8	The same
20	1 090	59	2 606	174	18	18	The same
30	1 671	93	4 058	286	27	27	The same
40	3 320	111	7 784	342	34	34	The same
50	6 597	430	14 954	1 170	42	42	The same

Without first pruning—mining without first pruning strategy). With first pruning—mining with first pruning strategy

1.3 算法基本思想

面向查询扩展的矩阵加权关联规则挖掘算法的基本思想是,首先根据包含(k–1)-项集的 k-权值阈值找出在后续遍历中有可能生成频繁 k-项集的(k–1)-项集,组成候选(k–1)-项集 C_{k-1},经过 4 次剪枝后,矩阵加权候选项集最大限度地减少,只产生含原查询项的矩阵加权候选项集 C_{k-1},再由 C_{k-1} 产生矩阵加权频繁(k–1)-项集 L_{k-1}.同时,通过连接 C_{k-1} 生成矩阵加权候选 k-项集 C_k,重复运用 k-权值阈值逐层迭代和 4 种剪枝策略产生矩阵加权频繁项集,直到矩阵加权候选项集集合为空才结束挖掘.最后,根据矩阵加权置信度由矩阵加权频繁项集生成矩阵加权强关联规则.

1.4 算法描述

算法. Matrix-Weighted association rules mining(简称 MWARM 算法).

输入:文本数据库 D,min*mwsupport* 和 min*mwconf*;

输出:矩阵加权强关联规则.

Begin

1) $L=\varnothing$;
2) ($TransactionCount$,$Itemsets\_Maxsize$,$ItemCount$)=$Search(D)$;
3) for (i=1;i<=$Itemsets\_Maxsize$;i++)
4) {$C_i=\varnothing$;$L_i=\varnothing$;} //清空 C_1,C_2,…和 L_1,L_2,…等
5) for (j=1;j<$ItemCount$;j++)
6) ($SC(C_1)$,Max$Weight(C_1)$,$SumWeight(C_1)$,$KIWT(C_1,2)$,C_1, $\overline{C}_1$,L_1)=$Counting$ (D);
7) for (k=2;k++) {
8) $L=L\cup L_{k-1}$;
9) $C_k=Join(C_{k-1})$;
10) if (k==2) $C_2=FirstPrune(k)$;
11) if (($\overline{C}_{k-1}$!=$\varnothing$) && (C_k!=$\varnothing$)) $C_k=SecondPrune(C_k, \overline{C}_{k-1})$;
12) for each Transaction Record
13) ($SC(C_k)$)=$SumCount(D)$;
14) if ($SC(C_k)$==0) $C_k=ThirdPrune(C_k,k)$;
15) for each Transaction Record
16) ($SumWeight(C_k)$,$KIWT(C_k,k+1)$)=$check(D)$;
17) $C_k=FourthPrune(C_k, \overline{C}_k, KIWT(C_k,k+1))$;
18) $L_k=Gen\_LargeSets(C_k,minawsupport)$;
19) $Out\_LargeItemSets(L_k)$;
20) if (C_k==$\varnothing$) $break$;
21) if (k>$Itemsets\_Maxsize$) $break$;
22) }
23) $R=Rules\_Gen(L,minawconf)$;

End

主要的子程序说明如下:

(1) $search(D)$:扫描数据库 D,找出可能的最大项目集的项目个数($Itemsets\_Maxsize$),事务记录总数($TransactionCount$)和项目总数($ItemCount$).

(2) $Counting(D)$:扫描数据库 D,累加各个 1-项集的支持数($SC(C_1)$)和权值($SumWeight(C_1)$)、找出各 1_项集的最大权值(Max$Weight[C_1]$)以及计算包含 1-项集的 2-权值阈值($KIWT(C_1,2)$),最后产生 C_1, $\overline{C}_1$,L_1.

(3) $Join(C_{k-1})$:由 C_{k-1} 连接生成 C_k,连接方法与 Apriori 算法类似.

(4) $FirstPrune(k)$:进行第 1 种剪枝;

(5) $SecondPrune(C_k, \overline{C}_{k-1})$:进行第 2 种剪枝;

(6) $SumCount(D)$:累加候选项集 C_k 在数据库 D 中出现的频度;

(7) $ThirdPrune(C_k,k)$:进行第 3 种剪枝;

(8) $check(D)$:遍历事务数据库 D,统计 C_k 中所有候选项集的权值之和($SumWeight(C_k)$)和包含 C_k 的(k+1)-权值阈值($KIWT(C_k,k+1)$);

(9) $FourthPrune(C_k, \overline{C}_k, KIWT(C_k,k+1))$:进行第 4 种剪枝;

(10) $Gen\_LargeSets(C_k,minawsupport)$:生成频繁项目集,并入库;

(11) $Out\_LargeItemSets(L_k)$:输出频繁项集;

(12) $Rules\_Gen(L,minawconf)$:产生强关联规则.

2 基于矩阵加权关联规则挖掘的伪相关反馈查询扩展算法

2.1 基本思想

基于矩阵加权关联规则挖掘的伪相关反馈查询扩展算法的基本思想是,首先对用户查询采用 tf-idf 算法对文档集初检,然后运用 MWARM 算法对前列 n 篇初检文档进行矩阵加权关联规则挖掘,提取含有原查询项的矩阵加权关联规则,构建规则库,根据所给的查询扩展模型和扩展词权重计算方法从规则库中提取与原查询相关的扩展词,计算其权重,并添加到原查询中构建新查询,实现查询扩展,最后对新查询进行第 2 次检索.

2.2 初检文档的预处理及其数量的确定

初检文档的预处理指的是对前 n 篇初检文档进行结构化处理,抽取代表其特征的特征词.即,对前列初检文档进行语词切分,去掉停用词,抽取特征词,计算其权值,然后求出特征词权值之和,按其权值总和值降序排列,建立特征词库和以每篇文档为记录的向量空间模型文本数据库.

基于伪相关反馈的查询扩展的前提条件是假设初检前列 n 篇文档都与原查询相关.实际上,前列文档中会存在一些与原查询不相关的文档.因此,初检前列 n 篇文档的选择是值得重视的,其 n 的选择至关重要:若 n 过大,即初检文档过多,则不相关文档也多,由于噪音多而与原查询的相关性就会下降;若 n 过小,初检文档太少,就会遗漏一些相关的反馈信息.为了探讨合适的 n 值,本文进行了如下的实验:

- 实验材料:从网上下载 500 篇有关计算机方面的论文作为实验文档集,对文档集进行预处理;
- 实验方法:设计 10 个查询(详见第 3.1 节),利用 tf-idf 算法对上述构建的测试文档集进行检索,统计 10 个查询的调和平均值(harmonic mean)[20].
- 实验结果:如图 2 所示的实验结果表明,当查询与文档的相似度为 0.2 时,其调和平均值 F 达到最大,表明这一点是查全率和查准率之间的最大可能折衷,检索性能较好.因此,在本文的后续实验中,初检文档与查询的相似度值取 0.2,此时获得的初检文档数量就是 n 值.然而,上述的实验研究仅仅是一种探讨,其结果并非是最佳值.至于最合适的 n 值,还有待于进一步研究.

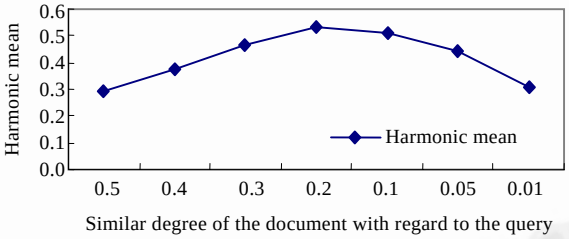

Fig.2 Harmonic mean with various similar degree

图 2 不同的相似度对应的调和平均值

2.3 前列特征词数量的确定

初检文档预处理后,得到按其权值总和值排序的特征词库,这些特征词是供后面挖掘词间关联规则所用的特征项.一般来说,特征词库中的特征项数量是较多的,太大的特征项数量会使挖掘词间关联规则效率变得很低,挖掘时间长,这不符合信息检索的要求,因为用户查询信息时都要求查询快、准、全.事实上,权值总和值较大的特征词最能反映整个前列 n 篇初检文档的语义内容,而权值总和值较小的特征词对整个前列 n 篇初检文档的语义内容贡献较少.因此,为了提高挖掘效率,减少挖掘时的特征词个数和噪声,提高关联规则的质量,需要根据特征词权值之和过滤特征词,即去掉权值总和值较小的特征词.为此,本文规定一个特征词权值总和阈值 w_m,取前列特征词权值总和不低于 w_m 的特征词进行词间关联规则挖掘.另外,为保证都能挖掘出与每一个查询项相关的扩展词,必须使所取的前列特征词中都包含全部查询项.因此,w_m 的值应等于特征词库里所有查询项(词)权值总和值中的最小者.

综上所述,过多的特征词不仅使挖掘效率低下,而且还会增加扩展词的噪声,查询扩展性能变差;而特征词数量太少,也会漏掉有趣的扩展词.显然,前列特征词的数量对词间关联规则的挖掘效率和查询扩展性能都有较大的影响.然而,前列特征词的数量取多少才是最佳,目前还没有定论,这也是下一步值得研究的课题.本文后面的实验中,前列特征词个数最少取 50 是属于实验上的需要,并非是最佳值.

2.4 查询扩展模型

基于矩阵加权关联规则挖掘的伪相关反馈查询扩展模型为 $Q_i \rightarrow T_j(mwsupport,mwconf)$.

该模型是一种矩阵加权关联规则,规则前件 Q_i 代表第 i 个 1 个以上的多查询项组成的集合,后件 T_j 代表第 j 个 1 个以上的多扩展项组成的集合.如果规定了 *mwsupport* 和 *mwconf* 的具体值,则可以得到与原查询项关联的扩展项.将之添加到原查询中,实现查询扩展.

2.5 扩展词权重的计算方法

1. 扩展词权重应该体现以下原则:

① 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;而扩展词指的是与原查询相关的语词,是原查询的语义补充和完善,其重要性不会高于原查询语词;

② 在基于矩阵加权关联规则挖掘的伪相关反馈查询扩展中,由于矩阵加权关联规则的置信度表明了扩展词与查询词的关联程度,所以扩展词的权重可以用其所在的矩阵加权关联规则置信度(*mwconf*)来充当;

③ 扩展词与整个查询的相关程度是不同的,存在全相关和部分相关的关联.若扩展词只与部分查询项相关,则称为部分相关;若与原查询中所有查询项都相关,则称为全相关.在进行查询扩展时,扩展词应体现与整个查询的相关程度,与整个查询相关程度越高,其权重应该越大.

2. 查询扩展时原查询和扩展词权重计算方法:根据上述原则,本文给出进行查询扩展时原查询和扩展词权重计算方法,即:

① 原查询的各个查询项权重设为 2;

② 当扩展词重复出现在不同的矩阵加权关联规则中,并且存在不同的矩阵加权置信度(*mwconf*)时,在其支持度不低于所给的最小支持度阈值的情况下,选择其置信度值最高的关联规则,并将其置信度作为该扩展词的置信度;

③ 扩展词权重 W_{expt_i} 计算公式如下:

$$W_{expt_i}=\frac{n_{relation}}{n_{total}}\times mwconf \tag{10}$$

其中,$n_{relation}$ 代表与扩展词相关的前件或后件的查询项个数,n_{total} 代表原查询中所有查询项的总个数.

2.6 查询扩展算法描述

算法. Query expansion of pseudo relevance feedback based on matrix-weighted association rules mining(简称 MWARMining-Based QE 算法).

输入:原查询 Q,min*mwsupport*,min*mwconf*,m(前列扩展词个数)和 *sim*(文档与查询的相似度);

输出:与原查询 Q 相关的扩展词集合及其对应的权重,查询扩展后的检索结果.

算法描述:

Begin

1) q=*Pretreat*(Q);
2) *First_Retried_Document*=*VSMRetrieval*(q);
3) *Top_Ranked_Document*=*GetTopRankedDocument*(q,*sim*);
4) *MWA_rules*=*MWARM_Mining*(*Top_Ranked_Document*,min*mwsupport*,min*mwconf*);
5) *ExpansionTerms*=*ExtractExpansionTerms*(*MWA_rules*,m);

6) q_{exp} =*Assemble*(*q*,*ExpansionTerms*);

End

算法中子程序说明如下:

(1) *Pretrea* (*Q*):将原查询 *Q* 进行预处理,生成查询向量 *q*.

(2) *VSMRetrieval*(*q*):利用向量空间模型(vector space model,简称 VSM)检索算法和原始查询向量 *q* 对文档集初检.

(3) *GetTopRankedDocument*(*q*,*sim*):从初检文档中提取其相似度值不低于 *sim* 的前列排序文档组成局部前列文档集,本文 *sim* 取 0.2.

(4) *MWARM_Mining*(*Top_Ranked_Document*,min*mwsupport*,min*mwconf*):用 MWARM 算法对前列文档集进行矩阵加权关联规则挖掘,挖掘到(查询项数+1)_频繁项集,提取含有原查询项的频繁项集生成矩阵加权关联规则库.

(5) *ExtractExpansionTerms*(*MWA_rules*,*m*):根据所给的查询扩展模型,从矩阵加权关联规则库中提取与原查询相关的扩展词,计算其权重,并排序,存入扩展词库中,本文 *m* 取 30.

(6) *Assemble*(*q*,*ExpansionTerms*):将原查询词与扩展词组合构建新查询 q_{exp} ,实现查询扩展.

3 实验设计及其结果分析

3.1 实验测试文档集、查询集及其语料预处理

为了测试 MWARMining-Based QE 算法的检索性能,从网上下载 720 篇有关计算机方面的文档作为实验用的原始测试文档集.设计了 10 个实际的查询(Q_1,Q_2,…,Q_{10})作为查询集,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,见表 2.对原始测试文档集进行预处理,构建基于向量空间模型的文本数据库和整个文档集的总特征词库.其中,平均每篇文档的特征词数为 77 个,共获得 3 531 个特征词.对查询集中的 10 个查询也作类似的预处理,得到查询向量形式.为了评估算法的检索性能,将查全率(recall)和查准率(precision)作为评估标准,采用 *t*-检验作为对比实验结果的显著性验证.*t*-检验用于信息检索性能分析在文献[21]中已有详细的描述.

Table 2 Queries and its relevance document number in test corpus

表 2 查询集及其在原始测试文档集中的相关文档数

Query No.	Queries	Relevance number in test corpus
Q_1	Data mining technology	154
Q_2	Computer network technology	211
Q_3	Knowledge discovery	160
Q_4	Information security and encryption	63
Q_5	Relational database design and application	166
Q_6	Graphics and image processing	68
Q_7	Information retrieval	59
Q_8	Text mining	40
Q_9	Internet technology	172
Q_{10}	Network security	52

3.2 对比算法

编写了实验源程序,使用标准的向量空间模型(VSM)作为检索算法,将本文提出的 MWARMining-Based QE 算法与基于 Apriori 算法的局部反馈查询扩展(即 Apriori-Based QE)以及基于局部上下文分析的查询扩展[7](即 LCA-Based QE)进行检索性能比较.对比实验的基线由不加任何查询扩展的基于 VSM 的检索系统(即 tf-idf 算法)得到.实验中的参数设定如下:(1) 扩展词数量:按照扩展词权重降序排列,取前列 30 个扩展词加入到原查询中,即扩展词个数统一为 30;(2) 扩展词权重规范化:先将前列 30 个扩展词中的最高权重赋给变量 maxw_{exp},规范化权重值等于将其原权重值除以 maxw_{exp};(3) 挖掘时,最低支持度阈值设为 0.01,最低置信度阈值设为 0.01;

(4) 挖掘时,特征词选择及其数量为:至少取前列 50 个特征词,其中应包含全部的查询项特征词.对于本文算法,按照特征词权重的总和值降序排列;而对 Apriori-Based QE 算法,则按特征词频度降序排列.LCA-Based QE 算法的实验采用与文献[7]中的实验相同的参数.

3.3 实验结果及其分析

将本文的 MWARMining-Based QE 算法与 Apriori-Based QE,LCA-Based QE 以及 tf-idf 算法分别就所设计的 10 个查询在相同的测试文档集中进行检索,统计这 10 个查询在相同的查全率水平级下其平均查准率,进行综合比较,实验结果见表 3 并如图 3 所示.本文算法与其他算法的 *t*-检验值见表 4.

从表 3 和图 3 可以看出,在测试文档集上,本文算法确实获得了相当好的结果.在相同查全率水平级下,本文算法的平均查准率比 LCA-Based QE 算法和 Apriori-Based QE 算法分别提高了 9.10%和 3.84%;相对于传统的 tf-idf 算法,MWARMining-Based QE 算法、Apriori-Based QE 算法和 LCA-Based QE 算法在相同查全率水平级下其平均查准率都有明显的提高,分别提高了 21.19%,16.71%和 11.08%,其中本文算法提高的幅度最大,而 LCA-Based QE 算法提高的幅度最小.表 4 表明,本文算法与对照算法的实验结果之间在统计上存在显著性差异,在自由度为 9 的情况下,其显著性水平明显高于临界值α为 0.01 时的 *t* 检验值,表明本文算法检索性能的提高在统计上是有一定意义的.

Table 3 Comparison of retrieval performance

表 3 查询性能比较

Recall (%)	Average precision (%)			
	tf-idf (baseline)	MWARMining-Based QE (*awsup*=0.05, *awconf*=0.03)	Apriori-Based QE (*sup*=0.11, *conf*=0.03)	LCA-Based QE
10	91.42	98.08	96.85	92.85
20	90.14	91.15	89.42	85.75
30	75.93	81.99	77.60	79.59
40	67.43	76.70	71.17	72.37
50	60.85	70.91	67.89	64.72
60	54.44	65.48	64.98	56.85
70	46.86	59.37	55.51	47.99
80	31.01	49.92	46.46	42.60
90	13.55	36.18	32.98	30.66
100	1.63	16.42	19.43	18.93
Total averaged	53.32	64.62 (+21.19%)	62.23 (+16.71%)	59.23 (+11.08%)

Table 4 Comparison of *t*-test values of MWARMining-based QE and the existing algorithms

表 4 MWARMining-Based QE 算法与其他算法之间的 *t*-检验值比较

	MWARMining-Based QE vs. tf-idf	MWARMining-Based QE vs. Apriori-Based QE	MWARMining-Based QE vs. LCA-Based QE
t-test values	5.64	6.29	6.86
Level of significance	Most remarkable	Most remarkable	Most remarkable

$(t_{0.05})_9$=2.262, $(t_{0.01})_9$=3.25, $(t_{0.001})_9$=4.781

综上所述,本文提出的 MWARMining-Based QE 算法是有效的,能够改善和提高信息检索性能.与现有算法比较,检索性能获得了明显的提高.主要原因分析如下:查询扩展机制使得具有明显歧义性的短查询词通过扩展词可以达到消歧作用,同时还能检索到原始短查询中所不能检索到的文档,所以本文算法比没有进行查询扩展的 tf-idf 算法在相同查全率水平级下其平均查准率有了明显的提高.另外,本文算法采用矩阵加权词间关联规则挖掘算法,在前列 *n* 篇初检文档中挖掘与原查询相关的扩展词,充分考虑了各个特征项及其项集在不同的事务文档记录中具有不同的重要性.引入矩阵加权项权值,因而挖掘出的词间关联规则比 Apriori 算法挖掘出的更加合理,获得的扩展词更能准确地体现与原查询的语义关系,查询扩展的检索效果变得更好.而 Apriori 算法只考虑特征词项的出现频度,基于局部上下文分析的查询扩展算法只考虑从初检文档中选出与原查询词共现的扩展词,它们都没有考虑特征词项在不同事务文档记录中具有不同的权重问题.这些不足使其查询扩展性能比本文算法要差,实验结果表明了上述结论.

4 结论和下一步研究方向

本文首先提出了一种面向查询扩展的矩阵加权关联规则挖掘算法,采用 4 种剪枝策略,挖掘效率得到极大提高.然后,针对现有查询扩展存在的缺陷,将矩阵加权关联规则挖掘技术和查询扩展相结合,提出新的查询扩展模型和更合理的扩展词权重计算方法.最后,提出一种新的伪相关反馈查询扩展算法——基于矩阵加权关联规则挖掘的伪相关反馈查询扩展算法,重点描述其查询扩展的思想及算法.实验结果表明,本文提出的查询扩展算法确实是有效的.与现有算法相比,在相同的查全率水平级下,其平均查准率有了较大的提高.

本文重点研究了基于矩阵加权关联规则挖掘的伪相关反馈查询扩展算法,而相关的其他问题还处于探讨之中.因此,在本文工作的基础上,还有以下一些问题需要进一步地分析、研究:

(1) 初检前列文档和前列特征词数量的确定问题;

(2) 扩展词数量,min*mwsupport* 和 min*mwconf* 等参数设置对查询扩展性能的影响问题;

(3) 本文的研究还处于基础性研究阶段,下一步的研究重点应该是如何把查询扩展技术应用到实际的信息检索系统(如现有的 Web 搜索引擎)中,开发出具有实际应用价值的信息检索系统.因此,其研究内容和需要考虑的因素还会很多,涉及的内容也很多,例如检索算法的时间、空间性能问题等等.

References:

[1] Huang MX, Yan XW, Zhang SC. Review and perspective of query expansion techniques. Computer Applications and Software, 2007,24(11):1−4 (in Chinese with English abstract).

[2] Voorhees EM. The effectiveness and efficiency of agglomerative hierarchic clustering in document retrieval [Ph.D. Thesis]. Cornell University, 1986.

[3] Furnas GW, Deerwester S, Dumais ST, Landauer TK, Harshman RA. Information retrieval using a singular value decomposition model of latent semantic structure. In: Chiaramella Y, ed. Proc. of the 11th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM Press, 1988. 465−480.

[4] Jing Y, Croft W. An association thesaurus for Information retrieval. Technical Report, UM-CS-1994-017, Amherst: University of Massachusetts, 1994..

[5] Attar R, Fraenkel AS. Local feedback in full-text retrieval systems. Journal of the ACM, 1977,24(3):397−417.

[6] Salton G, Buckley C. Improving retrieval performance by relevance feedback. Journal of the American Society for Information Science, 1990,41(4):288−297.

[7] Xu J, Croft WB. Query expansion using local and global document analysis. In: Frei HP, Harman D, Schaübie P, Wilkinson R, eds. Proc. of the 19th Annual Int'l ACM SIGIR Conf. on Research and Development in Information. New York: ACM Press, 1996. 4−11. http://eprints.kfupm.edu.sa/60386/1/60386.pdf

[8] Cui H, Wen JR, Nie JY, Ma WY. Query expansion by mining user logs. IEEE Trans. on Knowledge and Data Engineering (TKDE), 2003,15(4):829−839.

[9] Cui H, Wen JR, Li MQ. A statistical query expansion model based on query logs. Journal of Software, 2003,14(9):1594−1599 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/14/1593.htm

[10] Fonseca BM, Golgher PB, Moura ES, de Possas B, Ziviani N. Discovering search engine related query using association rules. Journal of Web Engineering, 2003,2(4):215−227.

[11] Latiri CC, Yahin SB, Chevallet JP, Jaouaa A. Query expansion using fuzzy association rules between terms. In: Proc. of the Conf. on Journées Informatique Messine (JIM 2003). Metz, 2003. 3−6. http://www-mrim.imag.fr/publications/2003/Jim2003.pdf.

[12] Martin-Bantista MJ, Sanchez D, Chamorro-Martinez J, Serrano JM, Vila MA. Mining Web documents to find additional query terms using fuzzy association rules. Fuzzy Sets and Systems, 2004,148(1):85−104.

[13] Zhang CQ, Qin ZX, Yan XW. Association-Based segmentation for Chinese-crossed query expansion. IEEE Intelligent Informatics Bulletin, 2005,5(1):18−25.

[14] Qin ZX, Liu L, Zhang SC. Mining term association rules for heuristic query construction. In: Dai HH, Srikant R, Zhang CQ, eds. Proc. of the 8th Pacific-Asia Conf. (PAKDD 2004). London: Springer-Verlag, 2004. 145−154.

[15] Géry M, Haddad MH. Knowledge discovery for automatic query expansion on the world-wide Web. In: Peter P, David W, Jacques K, eds. Proc. of the Workshops on Evolution and Change in Data Management, Reverse Engineering in Information Systems, and the World Wide Web and Conceptual Modeling. London: Springer-Verlag, 1999. 334−347.

[16] Wei J, Bressan S, Ooi BC. Mining term association rules for automatic global query expansion: Methodology and preliminary results. In: Zhou X, Fong J, Jia X, Kambayashi Y, eds. Proc. of the 1st Int'l Conf. on Web Information Systems Engineering. Los Alamitos: IEEE Computer Society Press, 2000. 366−373.

[17] Song M, Song IY, Hu XH, Allen RB. Integration of association rules and ontology for semantic-based query expansion. In: Tjoa AM, Trujillo J, eds. Proc. of the 7th Int'l Conf. on Data Warehousing and Knowledge Discovery. New York: Springer-Verlag, 2005. 326−335. http://www.ischool.drexel.edu/dmbio/publication/song_dek_2006.pdf

[18] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large database. In: Buneman P, Jajodia S, eds. Proc. of the '93 ACM SIGMOD Int'l Conf. on Management of Data. New York: ACM Press, 1993. 207−216.

[19] Tan YH, Lin YP, Li MQ. Mining all-weighted association rules from vector space model. Computer Engineering and Applications, 2003,39(13):208−211 (in Chinese with English abstract).

[20] Jr Shaw WM, Burgin R, Howell P. Performance standards and evaluations in IR test collections: Cluster-Based retrieval models. Information Processing and Management, 1997,33(1):1−14.

[21] Hull D. Using statistical testing in the evaluation of retrieval experiments. In: Korfhage R, Rasmussen EM, Willett P, eds. Proc. of the 16th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM Press, 1993. 329−338.

附中文参考文献:

[1] 黄名选,严小卫,张师超.查询扩展技术进展与展望.计算机应用与软件,2007,24(11):1−4.

[9] 崔航,文继荣,李敏强.基于用户日志的查询扩展统计模型.软件学报,2003,14(9):1594−1599. http://www.jos.org.cn/1000-9825/14/1594.htm

[19] 谭义红,林亚平.向量空间模型中完全加权关联规则的挖掘.计算机工程与应用,2003,39(13):208−211.

黄名选(1966－),男,广西南宁人,副教授,主要研究领域为数据挖掘,信息检索,查询扩展.

张师超(1962－),男,博士,教授,博士生导师,主要研究领域为数据挖掘,机器学习.

严小卫(1961－),男,博士,教授,博士生导师,主要研究领域为人工智能,数据挖掘.