

基于关联规则挖掘的查询扩展检索性能研究

黄名选¹, 陈燕红², 张师超³

(1. 广西教育学院 数学与计算机系, 广西南宁 530023;

2. 广西大学 物理学院, 广西南宁 530004;

3. 广西师范大学 计算机学院, 广西桂林 541004)

摘要:介绍了基于关联规则的局部反馈查询扩展基本思想, 重点研究关联规则支持度、置信度和扩展词数量对查询扩展检索性能的影响。实验结果表明, 这种查询扩展的检索性能对其支持度、置信度以及扩展词数量比较敏感; 从关联规则获得的扩展词可以分为两类, 即与原查询正相关的扩展词和与原查询负相关或者假相关的扩展词(即噪音), 前者可以提高和改善查询扩展的检索性能, 而后者只能降低其检索性能。

关键词:查询扩展; 支持度; 置信度; 关联规则

中图分类号: TP391.3

文献标识码: A

文章编号: 1673-629X(2008)10-0103-03

Studies on Retrieval Performance of Query Expansion Based on Association Rules Mining

HUANG Ming-xuan¹, CHEN Yan-hong², ZHANG Shi-chao³

(1. Department of Math and Computer Science, Guangxi College of Education, Nanning 530023, China;

2. College of Physical Science, Guangxi University, Nanning 530004, China;

3. College of Computer Science, Guangxi Normal University, Guilin 541004, China)

Abstract: The basic idea of the query expansion of local feedback based on association rules is given. And then the impact of support and confidence of association rules, and the number of expansion terms on its retrieval performance are researched principally. Experimental results indicate that the its retrieval performance is sensitive to support and confidence of association rules, and the number of expansion terms, and expansion terms from association rules can be classified into two categories: expansion terms that positive correlation of original query and expansion terms that fake or negative correlation related to original query (namely noises), the former can improve retrieval performance of query expansion, while the later makes its retrieval performance to fall.

Key words: query expansion; support; confidence; association rules

0 引言

Web信息的急剧膨胀导致人们查询信息时出现信息过载和词不匹配等难以克服的问题。查询扩展(Query Expansion)是解决信息过载和词不匹配问题的关键技术之一, 它指的是利用计算机多种技术, 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, 以便更完整地描述原查询所隐含的语义或者主题, 改善和提高信息检索系统的查全率和查

准率。传统的查询扩展方法^[1]主要有全局分析的、局部分析的以及基于用户查询日志的和基于关联规则的查询扩展。

基于关联规则的查询扩展是从数据挖掘的角度对查询扩展进行研究, 是一种崭新的查询扩展方法, 近年来得到较多学者的关注和研究, 文献[2~4]从不同的角度和策略研究了基于关联规则挖掘的查询扩展, 取得了令人鼓舞的研究成果。现有的研究中, 很少重视研究关联规则的挖掘技术及其规则的质量对查询扩展性能的影响, 更没有研究关联规则的支持度和置信度对查询扩展检索性能的影响。针对这个问题, 文中从实验的角度研究了关联规则支持度和置信度以及扩展词数量对查询扩展检索性能的影响。实验结果表明, 支持度、置信度和扩展词数量对查询扩展检索性能有

收稿日期: 2008-01-21

基金项目: 国家自然科学基金资助项目(60496327, 60463003)

作者简介: 黄名选(1966-), 男, 广西乐业人, 工程师, 硕士, 研究方向为数据挖掘和查询扩展; 陈燕红, 高级实验师, 研究方向为文本挖掘、信息检索和网络安全; 张师超, 博士, 教授, 主要从事人工智能、数据挖掘等研究。

较大的影响。通过关联规则挖掘获得的扩展词存在一定的噪音,这些噪音使查询扩展检索性能降低,也就是说,在查询扩展时会同时存在两类扩展词,即与原查询正相关的扩展词和与原查询负相关或者假相关的扩展词,即噪音。前者可以大大地提高查询扩展的检索性能,而后者只能降低其检索性能,甚至低于没进行查询扩展时的情况。

1 基于关联规则的局部反馈查询扩展简介

基于局部反馈的查询扩展是目前应用最广泛的查询扩展方法之一,它通过两次检索的方法解决查询扩展问题,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。基于关联规则的局部反馈查询扩展^[5]是一种有效的查询扩展方法,它的基本思想是:首先对用户查询采用传统的向量空间模型检索算法(即 $tf-idf$ 算法)对文档集初检,对初检相似度值排序,然后,对前列 n 篇初检文档进行词间关联规则挖掘,提取含有原查询项的关联规则构建规则库,从库中提取扩展词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二次检索。文献的实验结果表明,此方法有效,能改善和提高信息检索的性能。

2 支持度和置信度以及扩展词数量对查询扩展检索性能影响

2.1 实验设计及结果

为了研究支持度、置信度及扩展词数量对查询扩展检索性能的影响,以基于完全加权关联规则挖掘的局部反馈查询扩展算法作为实验算法,编写了源程序进行实验。从网上下载了 720 篇计算机方面的论文作为实验用的原始测试文档集。首先经过 $tf-idf$ 检索算法对原查询进行初检,提取前列初检文档,运用完全加权词间关联规则挖掘算法^[5](即 AWARM 算法)挖掘出与原查询项相关的关联规则入库,构建规则库(取 $sup=0.01$, $conf=0.01$),然后进行下面 3 个实验,实验中采用的查询扩展模型是: $Q_i \rightarrow T_j(sup, conf)$, 其中, sup 是规则支持度, $conf$ 是规则置信度,规则前件 Q_i 是查询项集合,后件 T_j 是扩展项集合,若同一个扩展词重复出现在多个关联规则中,则在其支持度不低于所给的支持度阈值时取置信度值最大的关联规则,将其置信度作为该扩展词的置信度。实验中的基准是指没有进行查询扩展的向量空间模型检索算法的检索结果值。

实验 1 置信度不变,改变支持度的值进行查询扩展,考察其查询扩展检索性能的变化。置信度 $conf$

$=0.01$, 而支持度 sup 取值分别为 0.01, 0.05, 0.07, 0.1, 0.13, 0.15, 0.20, 0.22, 其实验结果见图 1。

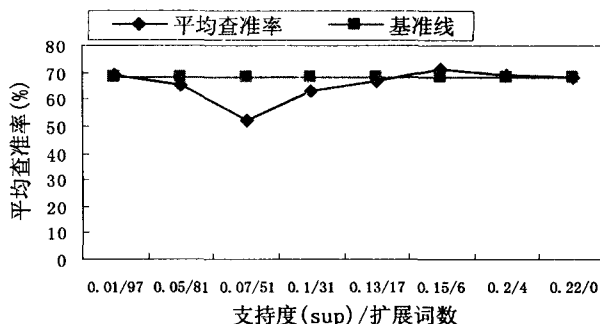


图 1 不同的支持度对应的平均查准率曲线图
($conf=0.01$)

图 1 表明,查询扩展的检索性能对支持度(sup)是很敏感的。当 sup 为 0.01、0.2 和 0.22 时,其平均查准率和基线相当,它们的扩展词数差别很大,分别是 97、4 和 0,说明扩展词并没有明显改善检索性能;而 sup 为 0.05, 0.07, 0.1, 0.13 时,其平均查准率低于基线,说明其扩展词不仅没有使检索性能提高,反而使其下降了,当 sup 为 0.07 时,平均查准率降到最低点;只有当 sup 为 0.15 时,平均查准率才明显地高于基线,说明此支持度下的扩展词能使检索性能提高。

实验 2 支持度不变,改变完全加权置信度的值进行查询扩展,考察其查询扩展检索性能的变化。支持度取 $sup=0.01$, 置信度 $conf$ 取值分别为 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.45, 0.5, 其实验结果见图 2。

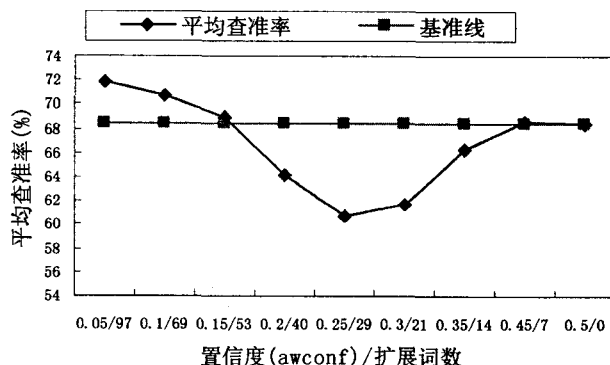


图 2 不同置信度对应的平均查准率曲线图
($sup=0.01$)

图 2 表明了查询扩展的检索性能对权重置信度($conf$)也是很敏感的。当 $conf$ 为 0.05, 0.1 时,其平均查准率均高于基线值,扩展词对检索性能具有明显的改善和提高; $conf$ 在 0.15, 0.45 时,其平均查准率和基线差不多,扩展词并没有改善和提高其检索性能;而在 0.2, 0.25, 0.3, 0.35 时,其检索性能下降了,扩展词对检索性能并没有改善作用。

实验 3 支持度和置信度都不变,改变扩展词的数量进行查询扩展,考察其查询扩展检索性能的变化。实验时完全加权支持度 $\text{sup} = 0.01$, 置信度 $\text{conf} = 0.05$, 扩展词数量在 7, 17, 27, 37, 47, 57, 67, 77, 87, 97, 实验结果见图 3。

图 3 表明,扩展词数量的改变对查询扩展检索性能有很大的影响。当扩展词数量在 7 和 57 时,平均查准率和基线几乎相同,检索性能没有得到改善;而扩展词数量为 17, 27, 37, 47 时,平均查准率低于基线,检索性能下降了,在 27 时,平均查准率下到最低点,当超过 57 后,检索性能有所改善和提高,到 97 时,平均查准率达到最高点。

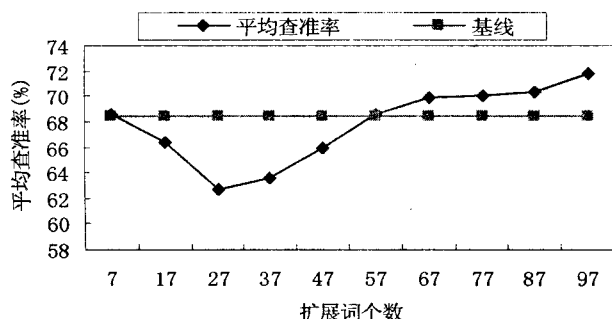


图 3 不同的扩展词的数量对应的平均查准率曲线图

2.2 实验结果分析

实验结果表明了在基于完全加权关联规则挖掘的局部反馈查询扩展中,完全加权支持度、置信度和扩展词数量对查询扩展检索性能都有较大的影响,其检索性能是受多方面因素综合影响的,并且对支持度、置信度比较敏感。

对实验结果的分析 and 研究发现,通过关联规则挖掘所获得的扩展词会存在与原查询“假相关”或弱相关甚至是负相关的扩展词,即存在一定的噪音,使查询扩展检索性能降低。因此,从关联规则获得的扩展词可以分为两类:与原查询正相关的扩展词和与原查询负相关或者假相关的扩展词,即噪音,前者可以提高和改善查询扩展的检索性能,而后者只能降低其检索性能,甚至低于没进行查询扩展时的情况。也就是说,在基于关联规则挖掘的局部反馈查询扩展中同时存在这两类扩展词,当正、负相关的扩展词数量相当时,它们对查询扩展检索性能的影响互相抵消,此时并没有使检索性能得到改善,例如,图 1 的 sup 为 0.01 和 0.2 时,图 2 的 conf 在 0.15, 0.45 时的情况。如果与原查询负相关或者假相关的扩展词数量多于正相关的扩展词,则其检索性能不但没有改善和提高,反而下降,例如,图 1 的 sup 为 0.05, 0.07, 0.1, 0.13 时,图 2 的 conf 在

0.2, 0.25, 0.3, 0.35 时的情况。只有当正相关的扩展词数量占绝大多数时,查询扩展的检索性能才能得到充分改善和提高。另外,扩展词数量的变化对查询扩展检索性能有很大的影响,事实证明只有与原查询词正相关的扩展词数量越多,检索性能才得到充分的改善和提高。

3 结束语

首先简单介绍基于关联规则挖掘的局部反馈查询扩展基本思想,然后重点研究关联规则支持度、置信度和扩展词数量对查询扩展检索性能的影响。实验结果表明,基于关联规则挖掘的局部反馈查询扩展检索性能对其支持度、置信度以及扩展词数量比较敏感;在基于关联规则挖掘的局部反馈查询扩展中同时存在两类扩展词:与原查询正相关的扩展词和与原查询负相关或者假相关的扩展词,即噪音,前者可以提高和改善查询扩展的检索性能,而后者只能降低其检索性能,甚至低于没有查询扩展的情况。对于查询扩展而言,正相关的扩展词数量越多越好,越能促进查询扩展检索性能的提高,而负相关的扩展词数量越少越好。因此,如何区分正、负相关的扩展词以及在查询扩展中引入负关联规则挖掘是一个很有意义和富有挑战的下一步研究课题。

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[J]. 计算机应用与软件, 2007, 24(11): 1-4.
- [2] Zhang Chengqi, Qin Zhenxing, Yan Xiaowei. Association-Based Segmentation for Chinese-Crossed Query Expansion[J]. IEEE Intelligent Informatics Bulletin, 2005, 5(1): 18-25.
- [3] Qin Zhenxing, Liu Li, Zhang Shichao. Mining Term Association Rules for Heuristic Query Construction[C]//Proceedings of 8th Pacific-Asia Conference, PAKDD 2004. Sydney, Australia; [s. n.], 2004: 145-154.
- [4] Wei Jie, Qin Zhenxing, Bressan S, et al. Mining Term Association Rules for Automatic Global Query Expansion: A Case Study with Topic 202 from TREC4[C]//Anne H H. Web Information Systems Engineering, Springer. Proceedings of the First International Conference. [s. l.]: Springer, 2000: 366-373.
- [5] 黄名选,严小卫,张师超. 基于文本数据库的完全加权词间关联规则挖掘算法[J]. 广西师范大学学报:自然科学版, 2007, 25(4): 24-27.