基于语词抽取与负关联规则挖掘的信息检索

黄名选[1],冯 平[2],谢统义[1]
(1.广西教育学院 科研处,广西 南宁 530023;
2.广西工学院 电控系,广西 柳州 545006)

摘 要:将语词抽取、负关联规则挖掘和查询扩展技术应用于信息检索,提出一种基于语词抽取与负关联规则挖掘融合的信息检索系统模型及其算法。详细论述模型的设计思想、各模块的功能,以及模型的理论分析和检索算法。该模型能够将语词抽取、负关联规则挖掘和查询扩展三种技术融合,对初检文档集进行有效地处理,得到高质量的与原查询词相关的扩展词,和原查询组合成新查询,再进行二次检索,有效地解决了词不匹配的问题。实验结果表明,该模型有效,能改善和提高信息检索性能。

关键词:查询扩展;信息检索;模型;负关联规则
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2012)05-0157-04

Information Retrieval Based on Terms Extraction and Negative Association Rules Mining

HUANG Ming-xuan[1], FENG Ping[2], XIE Tong-yi[1]
(1. Scientific Research Office, Guangxi College of Education, Nanning 530023, China;
2. Electronic Information and Control Engineering Department, Guangxi University of Tech., Liuzhou 545006, China)

Abstract: In order to apply terms extraction and negative association rules mining, as well as query expansion technique to information retrieval, a novel information retrieval system model and algorithm is introduced based on terms extraction and negative association rules mining. Its design conception and the function of each module is expounded. Its theoretical analysis for implementation and searching algorithm is also expatiated. The model can integrate terms extraction and negative association rules mining as well as query expansion three kinds of technologies and deal with effectively the top-ranked retrieved local documents to obtain high-quality expansion terms related to the original query terms. And then, the expansion terms are combined with original query to carry on the second retrieval again and the term mismatch issue of existing information retrieval is solved availably. The results of the experiment show that the proposed model can effectively improve and enhance the information retrieval performance.

Key words: query expansion; information retrieval; model; negative association rule

0 引 言

在信息社会的今天,随着计算机技术的广泛应用和网络通信技术的迅速普及,人们对信息需求以及信息的查询结果有更高的期望,信息检索研究因此受到众多专家学者的关注和重视,取得了丰硕的研究成果。这些成果除了传统的经典信息检索模型(布尔模型、向量空间模型和概率模型)以外,大致归结为基于自然语言处理的信息检索[1~3]、基于粗糙集和模糊集的信息检索[4,5]、基于本体的信息检索模型[6~8]和基于数据挖掘技术的信息检索[9,10]等等。然而,迄今为止,现有的信息检索模型还没有从根本上完全解决查全率和查准率问题,其信息查询结果更没有完全让用户满意,其主要问题是当前的信息检索系统都是基于关键词查询机制,难以避免词不匹配、信息过载等问题,例如,对于某个查询,web信息检索系统(如搜索引擎)会产生数十万以上的搜索结果,而在这些返回的结果中,真正用户关注和需要的信息并不多。

针对上述问题,文中在文献[11]研究基础上,提出一种基于语词抽取与负规则挖掘融合的信息检索系统模型及其算法。该模型将语词抽取、负关联规则挖掘和查询扩展三种技术融合,对初检文档集进行有效地处理,得到质量较高的与原查询词相关的扩展词,和原查询组合成新查询,再进行二次检索,有效地解决了

收稿日期:2011-10-10;**修回日期**:2012-01-12
基金项目:广西教育科研项目(200808MS191,201106LX388);广西高校优秀人才资助计划项目(桂教人[2011]40号);广西教育学院2010年度院级重点课题(桂教院科研[2010]7号)
作者简介:黄名选(1966-),男,硕士,副教授,CCF会员,研究方向为信息检索、查询扩展和文本挖掘。

词不匹配的问题。实验结果表明,该模型有效,能改善和提高信息检索性能。

1 基于语词抽取与词间负关联规则挖掘的信息检索

1.1 检索系统模型的基本思想

基于语词抽取与词间负关联规则挖掘融合的信息检索模型基本思想是:首先用向量空间模型检索算法对每一个用户查询在文档集中进行初次检索,将初检前列 m 篇文档提取,作为局部相关文档集。然后,将语词抽取、负规则挖掘和查询扩展三种技术融合,对初检文档集进行有效地处理,得到与原查询词相关的高质量扩展词,和原查询组合成新查询,再进行二次检索,最后,将二次检索结果返回给用户。具体处理过程是:首先对局部相关文档集进行语词抽取,构建语词数据库,然后以语词数据库中的前 k 个语词作为项集的集合,对局部相关文档集挖掘频繁项集和非频繁项集,建立频繁项集库和非频繁项集库,从频繁项集库和非频繁项集库挖掘负关联规则,建立负关联词库,最后从语词库中删除和负关联词库相同的负关联语词,将语词库中余下的语词作为扩展词实现查询扩展。

1.2 检索系统模型结构图及其模块和数据库组成

根据基于语词抽取与词间负关联规则挖掘的信息检索模型基本思想,设计出其检索模型结构图,如图 1 所示。该模型由文档预处理、传统检索算法、局部相关文档提取、语词抽取、项集挖掘、词间负关联规则挖掘、负关联词提取、正扩展词提取、查询扩展和最终检索信息返回等功能模块组成;同时设计了原始文档库、初检文档库、语词库、频繁项集库、非频繁项集库、负关联规则库、负关联词库和正扩展词库等数据库。

1.3 检索系统模型中各模块功能描述

①文档预处理模块:该模块主要是预处理用户查询和各类文档集,包括切分语词和停用词排除,同时提取特征词等工作,建立基于向量空间模型的原始文档库和初检文档库。

②传统检索算法模块:使用基于向量空间模型的检索算法对全部文档集进行检索,即计算查询与文档集中每一篇文档的相似余弦值,并按降序排列。

③局部相关文档提取模块:该模块主要功能是前列 m 篇初检文档提取,作为初检局部文档集,生成局部相关文档库。

④语词抽取模块:对初检文档数据库进行语词抽取,统计每一个语词在局部相关文档集中的出现频度并排降序,建立语词库。

⑤项集挖掘模块:以语词数据库中的前 k 个语词作为项集的集合,对局部相关文档集挖掘频繁项集和非频繁项集,建立频繁项集库和非频繁项集库。

⑥词间负关联规则挖掘模块:从频繁项集库和非频繁项集库挖掘词间负关联规则,并且提取负规则前件是原查询词项和后件是非原查询词项的负关联规则建立负关联规则库。

⑦负关联词提取模块:从负关联规则库中提取规则的后件,作为负关联词,建立负关联词库。计算每一个负关联词与整个查询词项的相关性,相关性的计算方法详见文献[12]。

⑧正扩展词提取模块:删除语词库中其相关性不大于 1 的负关联词,将语词数据库中余下的词作为最终扩展词,将其频度作为扩展词权值,并作规范化处理,建立正扩展词库。

⑨查询扩展模块:该模块的功能是将来自正扩展词库中的正扩展词和原查询重新组合,生成新查询,实现查询扩展,并对新查询进行二次检索。

⑩最终检索信息返回模块:对第二次检索信息结果进行相关的处理,根据用户的输出格式要求推送到用户界面。

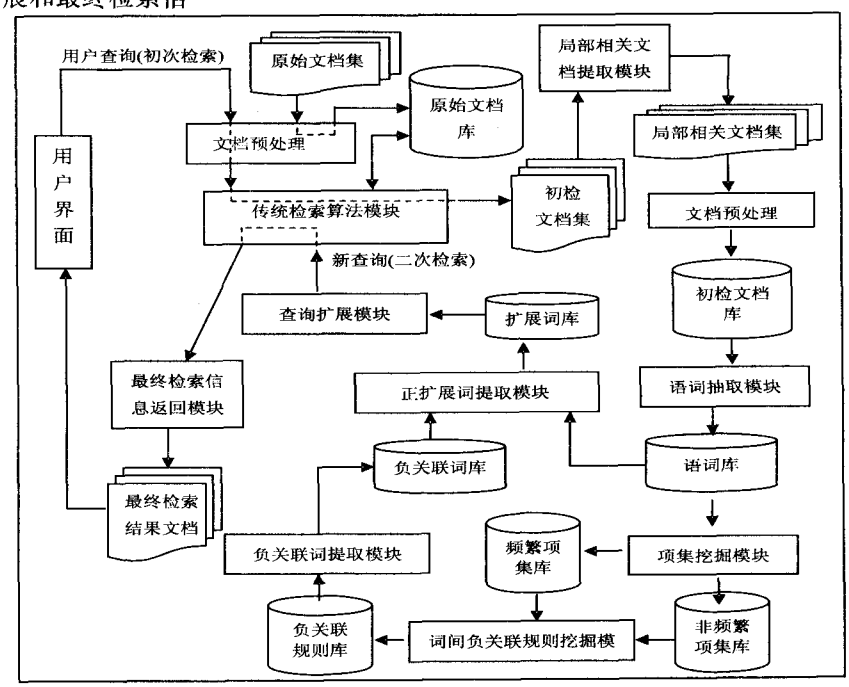

图 1 基于语词抽取与词间负关联规则挖掘的信息检索模型结构图

1.4 信息检索系统模型的理论分析

文中信息检索系统模型的理论分析如下:如何提高和改善信息检索查询性能是信息检索的核心问题,而查询扩展是提高和改善信息检索查询性能的关键技术,扩展词的来源及其权值设置是查询扩展的核心问题。

文中信息检索模型采用的是局部反馈的查询扩展技术。局部反馈技术假设排在最前列的 m 篇初检文档被认为都与原查询相关。根据该假设,文中将查询与初检文档的余弦值规范化后不小于0.2[13]的所有初检前列局部文档(此时的初检文档篇数就是 m 值)作为扩展词的来源。普遍认为,一个特征语词在一篇文档中出现的次数越多,它与这篇文档对应的主题越相关。

由此可见,利用语词抽取技术从初检局部文档集中抽取的特征语词按其出现的频次排降序,排在前面的语词就应该与原查询更相关,故文中选取包含原查询词的前列 k 个特征词作为候选关联词。然而,这种相关并非都是正相关,有可能是负相关或假相关。对于查询扩展来说,最关心的是与原查询词项成正相关的扩展词,而那些有可能是负相关或假相关的语词必须排除。以候选关联语词为挖掘的项集,利用负关联规则挖掘技术对初检文档集挖掘前件是原查询项而后件是非查询项的负关联规则,可以发现与原查询成负相关或者假相关的负关联语词。为了保证这些负关联语词是真正的负相关词,还必须计算其与全部查询词项的相关性,只有其相关性不大于1的负关联语词才是真正与原查询负相关或假相关。

1.5 信息检索系统的关键技术分析

查询扩展和负关联规则挖掘技术是文中信息检索系统的关键技术。

通过语词抽取技术从初检文档集中抽取出的语词,有可能存在与原查询成负相关或假相关,而负关联规则挖掘技术是发现这些负相关或者假相关语词的关键技术,因此,负关联规则挖掘技术对文中检索模型是至关重要的,是获得高质量扩展词的保证。文中只考虑 $q \Rightarrow \neg\ t$ 形式负关联规则,其中 q 代表原查询项集及其原查询项子集,t 代表非查询项语词项集。$q \Rightarrow \neg\ t$ 形式负关联规则的挖掘过程详见文献[14]。

查询扩展是指将计算机技术与语言学、信息学结合,补充与原查询相关的词或者词组到原查询中,建立新查询,然后再次检索,以提高信息检索系统查询性能。在文中的检索模型中,将来自于初检文档的特征语词作为候选扩展词,同时通过负关联规则挖掘技术可以发现候选扩展词中负相关的词,从而提高了扩展词的质量。扩展词权值设置详见文献[11]。

2 基于语词抽取与负关联规则挖掘融合的信息检索算法

算法:基于语词抽取与负规则挖掘的信息检索算法(Information Retrieval Based on Terms Extraction and Negative Rules Mining, 简称 TE&NR_IR)

输入:用户查询,前列特征语词数量(k),频繁项集支持度阈值(FIsup),负关联规则支持度阈值(Nsup),负关联规则置信度阈值(Nconf)。

输出:最终检索信息结果。

Begin

1)对用户查询利用基于向量空间模型检索算法对原始文档集进行初检,并提取前列排序文档组成局部前列文档集,建立初检文档库。

2)提取初检文档库中的特征词,统计其在局部前列文档集中的频度,按其频度排降序,抽取前列 k 个特征词(包含所有原查询词),建立语词库。

3)以2)步获得的 k 个特征语词为挖掘的项集,按照给定的频繁项集支持度阈值(FIsup)对初检局部文档集挖掘同时含有查询项和非查询项的频繁项集和非频繁项集。

4)按照给定的负关联规则的支持度和置信度阈值(Nsup,Nconf),从频繁项集和非频繁项集挖掘前件是查询项的负关联规则而后件是非查询项的负关联规则,建立负关联规则库。

5)从负关联规则库中提取规则的后件,作为负关联词,计算负特征词与整个查询词项的相关度,建立负关联词库。

6)删除特征语词数据库中相关性不大于1的负关联词和原查询语词,将余下的特征语词作为最终扩展词,将其频度作规范化处理,以此作为扩展词的权值。

7)将原查询词和6)步获得的扩展词组合成新查询,实现查询扩展,同时将原查询语词的权值设置为2。

8)将新查询对原始文档集进行第二次检索,返回其结果给用户。

End

3 模型的原型系统设计及其实验结果分析

3.1 模型的原型系统设计

根据上述所给的信息检索模型算法,设计了模型的原型系统。同时,构建了一个小型的原始文档测试集,以便测试文中提出的信息检索模型(即 TE&NARRetrieval 算法)的检索性能。该测试集由720篇有关计算机领域方面的论文组成。对测试集进行预处理,建立原始文档库。设计6个原查询($Q_1, Q_2, \cdots, Q_6$),经

人工检索获得原始查询在文档测试集中的相关文档篇数。预处理每一个查询的初检局部文档集，建立初检文档库。对初检文档库进行语词抽取，统计每一个语词在局部相关文档集中的出现频度并排降序，构建语词库，从语词中提取前列 $k=50$ 个语词（包含全部原查询词）作为挖掘用的项集。使用 MAP（Mean of Average Precision）作为文中检索系统模型的检索性能评测指标。

3.2 实验结果分析

以传统向量空间模型算法（即 tf-idf 算法）实验结果为基线，在相同的测试文档集中，将文中检索系统的 TE&NR_IR 算法和 PRFQEofFTE&C[11] 算法分别对所设计的用户原查询进行检索，统计其 MAP 值，实验结果如表 1 所示（PRFQEofFTE&C 算法的实验结果详见文献[11]）。

表 1 查询性能比较（Nsup=0.1，Nconf =0.1）

原查询	MAP 值			TE&NARRetrieval 试验参数	
	tf-idf	PRFQEof FTE&C 算法[11]	TE&NR_IR 算法	Sup	扩展词数
Q_1	0.754	0.846 (+12.26%)	0.817(+8.36%)	0.4	28
Q_2	0.652	0.809 (+24.10%)	0.814(+24.85%)	0.4	32
Q_3	0.641	0.654 (+2.12%)	0.625(-2.50%)	0.4	34
Q_4	0.521	0.668 (+28.17 %)	0.665(+27.64 %)	0.4	40
Q_5	0.495	0.462 (-6.67 %)	0.461(-6.87 %)	0.4	33
Q_6	0.333	0.401 (+20.58%)	0.474(+42.34%)	0.1	34
平均值	0.566	0.640(+13. 07%)	0.643(+13. 60%)		

从表 1 可以看出，与传统向量空间模型算法（tf-idf 算法）相比，TE&NR_IR 算法和 PRFQEofFTE&C 算法的 MAP 值分别平均提高 13.60% 和 13.07%，而文中 TE&NR_IR 算法的单个查询提高的幅度最高可达 42.34%（即 Q_6），文献[11]的单个查询提高的幅度最高只是 24.1%。实验表明，文中 TE&NR_IR 算法和文献[11]的 PRFQEofFTE&C 算法的检索性能都取得了较好的效果，其 MAP 获得了明显的提高，而 TE&NR_IR 算法比 PRFQEofFTE&C 算法的检索效果略好些，主要原因是文中信息检索系统模型将语词抽取和负关联规则挖掘运用于查询扩展，使得检索系统获得与原查询相关的高质量扩展词，改善和提高了信息查询性能。

从表 1 可以发现，文中 TE&NR_IR 算法的 MAP 值对于各个用户查询提高的幅度是不相同的，有些并没有得到改善和提高，反而下降了，例如 Q_3 和 Q_5 的 MAP 分别下降 2.50% 和 6.87%。表明 TE&NR_IR 算法还存在一些局限性，主要原因是文中检索模型采用的是无加权负关联规则挖掘技术，所有负关联词并没有完全被挖掘出来，事实上，文档库中每一个特征词在不同的文档中都有着不同的重要性，应该有各自的权值。只有考虑了特征词权值才能比较准确、全面地挖掘出词间关联性，这些问题有待于进一步研究和探讨。

4 结束语

文中尝试在信息检索系统中首先将语词抽取和负规则挖掘运用于查询扩展，然后再将查询扩展技术应用于信息检索系统，提出基于语词抽取与负规则挖掘的信息检索系统模型及其算法，取得了良好的效果。该模型由 9 个功能模块和 8 个数据库组成。模型对用户查询的初检文档采用了语词抽取和负关联规则挖掘技术，获取与原查询相关的扩充词，使原查询得到进一步扩展和优化，提高了信息检索性能。实验表明，该检索系统模型有效，能提高和改善信息检索系统的检索性能。下一步重点研究如何将该成果运用于实际的信息检索系统，以解决现有检索系统存在的问题。

参考文献：

[1] 李卫君，赵铁军，臧文茂. 基于文摘的信息检索模型[J]. 软件学报，2008，19(9)：2329-2338.

[2] 刘 宁，柴雅凌. 自然语言在智能信息检索中的应用[J]. 图书与情报，2006(1)：91-95.

[3] 李晓光，王大玲，于 戈. 基于统计语言模型的信息检索[J]. 计算机科学，2005，32(8)：124-127.

[4] 黄治国，朱承学，薛 凡，等. 基于概率粗糙集模型的信息检索[J]. 计算机工程，2008，34(23)：193-195.

[5] Nowacka K，Zadrożny S，Kacprzyk J. A new fuzzy logic based information retrieval model[C]//Proceedings of IPMU'08. [s.l.]：[s.n.]，2008：1749-1756.

[6] Jimeno-Yepes A，Berlanga-Llavori R. Ontology refinement for improved information retrieval[J]. Information Processing & Management，2009(7)：45-50.

[7] 高 琳，夏清国，王黎明. 基于本体的智能信息检索系统的构建方法[J]. 计算机工程与设计，2008，29(24)：6309-6311.

[8] 陈 锐，张 蕾，胡艳华. 基于语义的信息检索模型[J]. 计算机工程与应用，2009，45(26)：141-143.

[9] 黄名选，严小卫，张师超. 基于完全加权关联规则挖掘和查询扩展的信息检索[J]. 计算机应用与软件，2009(8)：26-28.

[10] 陈小华，赵捧未. 基于关联规则的个性化信息检索系统研究[J]. 情报科学，2006，24(6)：915-918.

[11] 冯 平，黄名选. 特征词抽取和相关性融合的伪相关反馈查询扩展[J]. 现代图书情报技术，2011(1)：52-56.

[12] Han Jiawei，Kamber M. Data Mining：Concepts and Techniques[M]. [s.l.]：Morgan Kaufmann Publishers，Inc.，2001.

[13] 黄名选，严小卫，张师超. 基于矩阵加权关联规则挖掘的伪相关反馈查询扩展[J]. 软件学报，2009(7)：1854-1865.

[14] 黄名选，余 如. 基于负关联规则与频繁项集挖掘的信息检索系统[J]. 现代图书情报技术，2011(7)：91-95.