基于百科知识库的主题扩展研究[1]

闻彬[1,2] 何婷婷[1]

华中师范大学 计算机科学与技术系 湖北 武汉 430079[1]

咸宁学院 计算机学院 湖北 咸宁 437000[2]

E-mail:vanbrian@163.com

摘 要：由于目前的主题扩展方法不能从根本上改善信息检索的性能，提出一种基于百科知识库的主题扩展方法。百科知识库中对于主题进行了详细介绍，包含了大量的语义信息。本论文利用百科知识库扩展解决了传统主题扩展中没有融入语义信息的问题。实验证明，利用百科知识库得到的扩展词在检索上可以得到了很好的效果。

关键词：百科知识库；主题扩展；信息检索

Research on Topic Expansion Based on Encyclopedia Knowledge

Wen Bin[1,2] He Tingting[1]

Department of Computer Science, Huazhong Normal University, Wuhan 430079[1]

School of Computer, Xianning College, Xianning, 437000[2]

E-mail:vanbrian@163.com

Abstract: For the existed topic expansion can not effectively improve the performance of information retrieval, we proposed a new method based on encyclopedia knowledge. In encyclopedia knowledge, topics have been described in detail, included a large number of semantic information. In this paper, the problem of traditional topic expansion without semantic information had been solved by using encyclopedia knowledge. And the experimental results showed that the proposed approach is suitable for topic expansion.

Keywords: Encyclopedia Knowledge; Topic Expansion; Information Retrieval

1 引言

随着 web 的迅速发展和信息资源的高速膨胀，人们从互联网上可以获取越来越多的资料。同时，互联网上的巨大信息也给人们的生活产生了革命性的变化。但是，在利用传统的关键字匹配的搜索引擎查找信息过程中，往往出现关键字与文档信息表示不匹配的情况。因此需要引入主题扩展的概念以提高信息检索的效果。

主题扩展也即查询扩展，是解决词不匹配问题的有效技术手段，它以用户的初始查询为基础，通过研究获得策略加入一些相关的词，以提供更多有利于判断文档相关性的信息。目前的系统普遍存在检索精度和查全率不高的问题，因此在目前搜索腾飞的年代，主题扩展的研究显得尤为重要。例如我们常见的“计算机”、“电脑”和“PC”；“范跑跑”和“范美忠”都代表了同样的含义，但在检索的时候却无法同时被检索出来，这样就直接导致了重要信息的丢失。在不使用主题扩展的条件下，用户要精确找到所需要的信息就变得十分困难。

本文通过对关键词利用网络知识库进行分析，设计一种基于网络百科知识库的主题扩展算法。本文组织结构如下：第二节对查询扩展的相关研究现状进行简单的介绍；第三节对本文提出的基于百科知识库的方法进行详细论述；第四节为实验设计部分，介绍了实验的目的、数据以及结果；第五节对本文提出的方法进行了总结，并提出进一步的工作计划。

[1]国家自然科学基金重大研究计划 90920005；国家自然科学基金 60773167；国家十一五科技支撑计划课题 2006BAK11B03；973 国家重点基础研究发展计划 2007CB310804；教育部/国家外国专家局高等学校学科创新引智计划 B07042；湖北省自然科学基金计划项目资助 2009CDB145；武汉市晨光计划项目资助，项目编号：201050231067。

2 相关研究现状

传统的主题扩展技术分为基于全局分析和基于局部分析两类，下面分别对其进行简单介绍。

基于全局分析技术顾名思义就是利用整个文档集合的信息来进行扩展查询。现代全局分析的策略主要有两种：基于相似词典(Similarity thesaurus)[1]的主题扩展和基于统计词典(Statistical thesaurus)[2][3]的主题扩展。基于统计词典的方法首先将文档进行聚类，生成不同的簇，计算簇之间的相似度，然后选择将相似度较高的一对簇进行合并，循环合并，知道没有满足合并条件的为止，得到分层的簇，从分层的簇中按规则选取合适的词语。

基于局部分析的主题扩展利用初始检索结果中相关度较高的前若干篇文档来扩展词语，其中最常见的是伪相关反馈[4][5]扩展。伪相关反馈假设初始检索结果中排在前面的若干篇是相关文档，然后利用标准的相关反馈过程进行查询扩展。TREC 会议上证明了该方法是一种简单但十分有效的查询扩展技术[6]。

大量的研究表明，传统的主题扩展方法在技术上有了很大改进，但是不能够从根本上改善信息检索的性能。基于全局分析的主题扩展技术，由于其要考虑对象是整个文献集，因此系统复杂度比较高，效率比较低，不能很好的满足海量数据的检索；基于局部分析的主题扩展依赖于首次检索的结果，当首次检索结果与原查询文档相关度较低时，会对最后的检索结果产生强烈的干扰作用，严重降低了查准率。因此本文提出了基于百科知识库的主题扩展技术。

3 基于百科知识库的主题扩展研究

3.1 主题扩展检索流程图

本节设计系统验证算法有效性。该系统的处理由主题文档获取模块、预处理模块、TF-IDF 模块、文档建模模块和词语文档相关性模块，其系统结构图如图 1 所示。

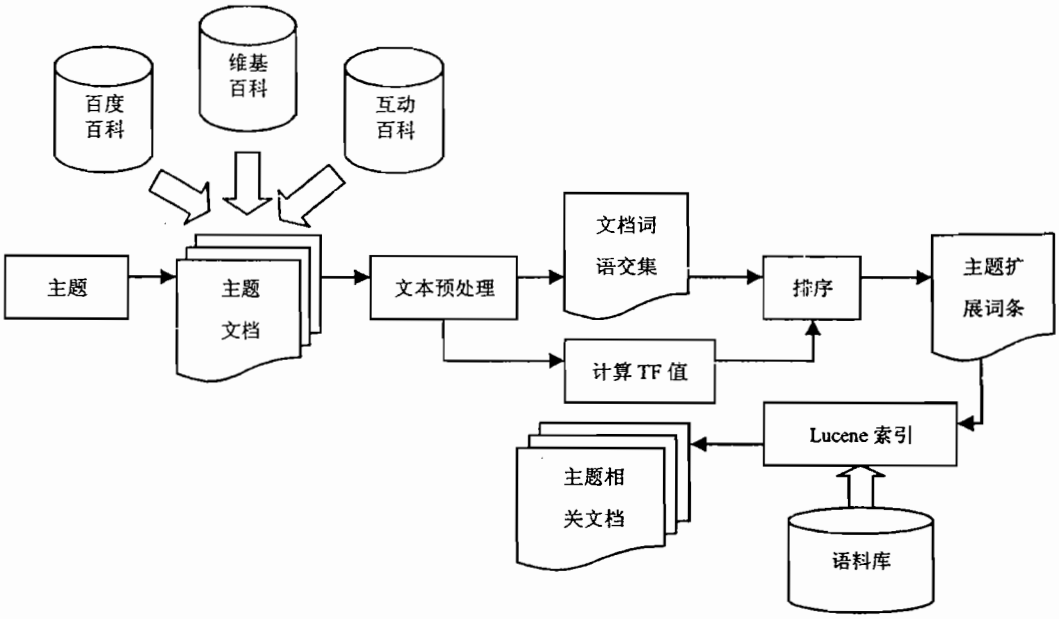

图 1 基于百科知识库的主题扩展流程图

3.2 网络资源简介

互联网上存在浩瀚的资源，那么以何种形式组织和获取资源成为了非常关键的问题。对于某个主题，例如“犀利哥”，如何找出与其相关的词条对原有查询进行扩展直接对得到的查询结果有很大的影响。鉴于近些年出现的百度百科、维基百科、互动百科等对于主题进行了详细的说明，对于主题的扩展有着非常大的帮助。例如：“犀利哥”事件亦席卷台湾、日本、新加坡。经证实，“犀利哥”是流浪在宁波市的一位有精神障碍的乞丐，本名程国荣，出生日期：1976年10月10日，户籍地：江西鄱阳县。其中许多的词语都对于主题给予了说明。

我们认为，对于某主题而言，与其相关的词条必然出现在知识库中，并且相关度越高的词条，在多个知识库中出现的概率也越高。本文使用来自于互联网的维基百科、百度百科和互动百科作为主题扩展的网络资源。

3.3 获取相关百科文档

百科知识库中对于社会现象、人物、事件等都条进行了详细的解释说明，几乎囊括了所有可解释性词语。知识库的使用对于信息开放的如今显得尤为重要，用户可以通过知识库获得准确的关于主题的各方面内容。针对社会现象和事件，可以获得该事件的解释、背景、状况等，例如“网络实名制”、“山西疫苗事件”、“创业板”等；针对人物方面，可以获得该人物的简介、走红的的起因、现况如何等。这些被已经被收录到知识库中的信息还持续被修改和完善。

鉴于百科知识库中知识的实时性，对于给定的主题，利用程序从不同的知识库中下载与主题对应的词条网页。

3.4 对文本进行预处理

在获取了相关百科文档之后，需要对相关的中文文档进行预处理，包括分词、词性标注、词法标注、删除停用词等。本节利用目前国内最好的分词软件 ICTCLAS[7]工具对文档进行相关处理。

3.5 获取相关扩展词

本文假设：在与主题对应的知识库文本中，如果某词语出现在越多的知识库文档里，说明该词语与主题之间有着极为重要的相关度；同时，如果该词语在文档中出现的次数越多，则说明该词语与文档之间有着更为重要的相关度。

$$P = A \cap B \cap C \tag{3.1}$$

其中：P 为词语集合；A、B、C 分别为从维基百科、互动百科和百度百科中得到的与主题对应的文档词语集合；那么集合 P 即为百科知识库中词语的交集。

从公式中我们可以看出集合 P 中的词语在三个文档中都有出现，但是考虑到交集 P 中的词语可能较多，需要对相关度进行排序以取其中排序靠前的部分词语。我们利用两个值来做排序工作：TF(TermFrequency)值和 TF-IDF(Term Frequency–Inverse Document Frequency)值。首先通过计算出的 TF 值对文本中的词语进行排序，再计算词语的 TF_IDF 值，结合 TF 值与 TF_IDF 值利用公式 3.2 选取前 8 个具有较高相关度的词语作为该主题的扩展词条。

$$Score(w_i) = TF(w_i) + TF\text{-}IDF(w_i) \tag{3.2}$$

3.6 文档与词语的相关度计算

索引文件[8]的建立对于提高文档的检索效果有很大的作用，本节使用以词为单位构建倒排索引文档表。

倒排索引(Inverted index)，也常被称为反向索引、反向档案，主要是用来存储在全文搜索下

某个单词在一个文档或者一组文档中的存储位置的映射。它是文档检索系统中最常用的数据结构。

利用倒排索引对文档建立索引之后，然后利用公式3.3计算词语文档的得分。

$$Score(w_i) = \sum_{t\ in\ q} tf(t\ in\ d) * idf(t) * boost(t.field\ in\ d) * lengthNorm(t.field\ in\ d) \quad (3.3)$$

其中：$Score(w_i)$ 为词语 w_i 与主题 $topic-w$ 相关的的文档得分；$boost(t.field\ in\ d)$ 表示对每个 $field$ 设置的一种激励因子。默认值为"1"；$lengthNorm(t.field\ in\ d)$ 为一个长度因子，由词条所在的 $field$ 的总长度决定。

3.7 相关文档得分

每个主题有若干个扩展词，每个扩展词 w_i 与文档可计算得分 $Score_{w_i}(file_k)$，利用公式3.4计算文档与主题的相关度得分。

$$Score(file_k) = \sum_{i=1}^{n} Score_{w_i}(file_k) \quad (3.4)$$

将得到的文档按 $Score(file_k)$ 非增序排列即为与主题相关的文档。

4 实验分析

本文从网络挑选较为热门的6个主题，利用3.5节算法对主题进行扩展，扩展结果见表1。

表1 扩展词实验结果

词语	扩展词
犀利哥	乞丐，救助，程国圣，老馋猫，程国荣，帅哥，混搭，恶搞
创业板	市场，上市，企业，证券，投资，发展，风险，融资
地王	土地，房价，市场，房地产，北京，住宅，地块，地价
淘宝	支付宝，网站，阿里巴巴，购物，商品，交易，卖家，买家
无厘头	文化，周星驰，香港，喜剧，夸张，风格，行为，表演
股指期货	股票，指数，交易，价格，市场，合约，风险，投资者

然后利用Google对此6个主题进行搜索，随机获取各个主题20篇文档，共120篇文档创建索引，具体的创建过程见3.6节。由于文档是利用搜索引擎获得的，因此都必然出现了主题关键字，例如在下载的"地王"的文档里，都出现了词语"地王"。考虑到必然会有相关性，所以实验中我们不考虑主题词与文档的相关性，而仅仅计算扩展词与所得文档的相关性。由于实验中词语进行检索后的数据结果庞大，本处仅列出"地王"扩展词的实验数据。如表2所示，对于扩展词与文档的相关性都进行了打分，得分越高，说明词语与主题相关性越高；反之则越低。

表2 "地王"扩展词检索结果

词语	相关文档	得分	词语	相关文档	得分	词语	相关文档	得分	词语	相关文档	得分
地价	diwang18.txt	0.39667222	地块	diwang17.txt	0.53211755	住宅	diwang19.txt	0.536516	房价	diwang15.txt	0.4714413
	diwang2.txt	0.29149336		diwang18.txt	0.4763216		diwnag20.txt	0.30105615		diwang18.txt	0.46588385
	diwang17.txt	0.29149336		diwang6.txt	0.45554543		diwang13.txt	0.28677988		diwang12.txt	0.29465082
	diwang16.txt	0.23800334		diwnag20.txt	0.3904675		diwang18.txt	0.25088012		diwang11.txt	0.27218676
	diwang11.txt	0.2243917		diwang7.txt	0.3615027		diwang8.txt	0.23175314		diwang1.txt	0.26897815
	diwang3.txt	0.19634274		diwang12.txt	0.30125225		diwang4.txt	0.20484276		diwang7.txt	0.2500195
	diwang9.txt	0.19634274		diwang8.txt	0.27828488		diwang7.txt	0.17381486		diwang16.txt	0.2500195
	diwang7.txt	0.16829377		diwang3.txt	0.24349926					diwang17.txt	0.2500195

				diwang2.txt	0.20871365					diwang4.txt	0.24058136
				diwang14.txt	0.20871365					diwang19.txt	0.23816341
				diwang11.txt	0.19677714					diwang5.txt	0.21518253
				diwang9.txt	0.17218					diwang13.txt	0.16840696
				diwang4.txt	0.122985706					diwnag20.txt	0.14434883
										diwang10.txt	0.12029068
词语	相关文档	得分	词语	相关文档	得分	词语	相关文档	得分	词语	相关文档	得分
房地产	diwang3.txt	0.43182823	市场	guzhiqihuo9	0.31826106	北京	diwang11.tx	0.47666937	土地	diwang11.txt	0.39016023
	diwang9.txt	0.43182823		chuangyeban1.txt	0.277183		diwnag20.txt	0.4378488		diwang3.txt	0.37397423
	diwang8.txt	0.3022168		chuangyeban20.txt	0.2525376		diwang6.txt	0.39014795		diwang8.txt	0.3489699
	diwang15.txt	0.3022168		chuangyeban18.txt	0.24004753		diwang8.txt	0.37684023		diwang15.txt	0.3022168
	diwang12.txt	0.28852737		chuangyeban14.txt	0.23721778		diwang4.txt	0.33308285		diwang2.txt	0.29262018
	diwang19.txt	0.2644397		guzhiqihuo18.txt	0.22083774		diwang1.txt	0.31599015		diwnag20.txt	0.29262018
	diwang14.txt	0.2266626		guzhiqihuo17.txt	0.22049774		diwang13.txt	0.29492414		diwang1.txt	0.24385014
	diwnag20.txt	0.2266626		chuangyeban3.txt	0.19599798		diwang7.txt	0.2189244		diwang4.txt	0.2181062
	diwang1.txt	0.2181062		guzhiqihuo10.txt	0.19368751		diwang9.txt	0.20854285		diwang9.txt	0.21591412
	diwang16.txt	0.18506923		diwang19.txt	0.19174086		taobao10.txt	0.20640391		diwang18.txt	0.1888855
	diwang5.txt	0.17448495		guzhiqihuo15.txt	0.18003565		taobao9.txt	0.182437		diwang7.txt	0.18506923
	diwang11.txt	0.17448495		guzhiqihuo16.txt	0.17323938		wulitou8.txt	0.17875102		diwang17.txt	0.18506923
	diwang4.txt	0.15422437		chuangyeban19.txt	0.17149824		diwang2.txt	0.17875102		diwang13.txt	0.15267433
	diwang18.txt	0.15422437		guzhiqihuo12.txt	0.16973923		chuangyeban2.txt	0.16852808		diwang19.txt	0.15267433
	diwang2.txt	0.13086371		taobao20.txt	0.1500297		diwang10.txt	0.14895917		diwang14.txt	0.13086371
	diwang7.txt	0.13086371		chuangyeban17.txt	0.1500297		chuangyeban16.txt	0.12639606		guzhiqihuo2.txt	0.13086371
	diwang17.txt	0.13086371		taobao15.txt	0.14852183		diwang14.txt	0.12639606		diwang5.txt	0.12337949
	chuangyeban5.txt	0.1090531		chuangyeban12.txt	0.14852183		guzhiqihuo2.txt	0.12639606		diwang12.txt	0.1090531
	chuangyeban6.txt	0.1090531		chuangyeban16.txt	0.1469985		diwang5.txt	0.11916734		guzhiqihuo1.txt	0.1090531
				diwnag20.txt	0.1469985		taobao20.txt	0.10533005			
				guzhiqihuo2.txt	0.1469985		guzhiqihuo1.txt	0.10533005			
				guzhiqihuo4.txt	0.1469985						
				guzhiqihuo11.txt	0.1469985						

实验中，我们对下载的不同类别的文档进行了不同的命名。6 个主题的 120 篇文档分别命名为 $\{diwang_i; taobao_i; chuangyeban_i; quzhiqihuo_i; wulitou_i; xiligei\}i = 1, .., 20$ 。从以上实验结果可以看出，大部分有得分的文档都属于人工下载的“地王”的类别，但是根据扩展词的不同依然会有一些其他类别的文档，将表 2 中的文档按照 3.7 节给的算法对文档进行综合评分，表 3 列出综合评分后所得到的排名前 20 篇文档。

表 3 词语总得分

主题	排名	相关文档	得分	主题	排名	相关文档	得分
地王	1	diwang20.txt	1.940003	地王	11	diwang15.txt	1.214466
	2	diwang18.txt	1.932868		12	diwang2.txt	1.102442
	3	diwang8.txt	1.538065		13	diwang1.txt	1.046952
	4	diwang7.txt	1.488488		14	diwang13.txt	0.902785
	5	diwang17.txt	1.389563		15	diwang5.txt	0.880806
	6	diwang19.txt	1.383534		16	diwang6.txt	0.845639
	7	diwang11.txt	1.258001		17	diwang14.txt	0.692636
	8	diwang3.txt	1.245644		18	diwang16.txt	0.673092
	9	diwang4.txt	1.243823		19	diwang11.txt	0.476669
	10	diwang9.txt	1.224808		20	guzhiqihuo9.txt	0.318261

从表 3 可以看出，在获得的综合排名前 20 篇文档中，其中 19 篇文档为人工挑选的分类正确的文档。而且这 19 篇相关文档排在了前 19 位，由于最后一篇文档中出现了较多的本主题相关词语，因此产生了较大的干扰性。

5 小结

本论文在传统的主题扩展的基础上，提出了一种基于百科知识库的主题扩展方法，利用知识库中的词条，对主题进行扩展。通过实验分析，该方法在获得相关文档方面有较高的准确性。

然而，如果给出的主题在知识库中没有对应的词条，则本方法无法完成主题扩展，进而检索的相关文档结果只能完全依赖于主题本身。另外，本文中作者选取的主题有较强的不相关性，因此不排除当主题之间本身有较强相关性时结果可能也会有较大偏差。需要在进一步的工作中研究。

参考文献

[1] Qiu Y. and Frei H. P. Concept based query expansion. In: proceedings of the 16th Annual International ACM Conference on Research and Development in Information Retrieval(SIGIR'93), 1993, 160-169.

[2] van Rijsbergen, C. J. Information Retrieval (2nd ed.). Butterworths, London, UK, 1979.

[3] Sparck Jones K. Automatic Keyword Classification for Information Retrieval. Butterworths, London, 1971.

[4] Buckley C. & Salton G. (1995). Automatic query expansion using SMART: TREC-3. In D. Harman (Ed.). Overview of the Third Text Retrieval Conference (TREC-3) (pp. 69-80). Gaithersburg, MD: NIST Special Publication 500-225.

[5] Xu, J. & Croft, B. (1996). Query expansion using local and global document analysis. In H.-P. Frei, & P. Schauble, Proceedings of the 19th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland (pp.4-11). New York, NY: ACM Press 18-22 August.

[6] Voorhees, E. and Harman, D. Overview of the Sixth Text Retrieval Conference. In: proceedings of the 6th Text Retrieval Conference(TREC-6), 1998.

[7] ICTCLAS. ICTCLAS's Home Page.http://www.ictclas.org/.

[8] A. Smeaton, and R. Wikinson. Spanish and Chinese document retrieval in TREC-5. Maryland, 1996.