基于多知识库和局部反馈的查询扩展研究*

付剑锋[1,2] 刘宗田[2] 刘念祖[1]

(1. 上海立信会计学院数学与信息学院 上海 201620;
2. 上海大学计算机工程与科学学院 上海 200027)

摘 要 查询扩展是优化信息查询的一种重要手段。提出了一种基于多知识库和局部反馈的查询扩展方法,该方法首先融合了领域本体与同义词词林两种不同类型的知识库对查询关键词进行扩展,然后再用局部反馈方法对扩展结果进行筛选。实验表明,该方法可以有效提高查询性能。

关键词 查询扩展 多知识库 领域本体 同义词词林 局部反馈

中图分类号 TP391 **文献标识码** A **文章编号** 1002-1965(2013)02-0103-04

Research on Query Expansion Based on Multi-knowledge Base and Local Feedback

Fu Jianfeng[1,2] Liu Zongtian[2] Liu Nianzu[1]

(1. School of Mathematics and Information, Shanghai Lixin University of Commerce, Shanghai 201620
2. School of Computer Engineering & Science, Shanghai University, Shanghai 200072)

Abstract Query expansion is an important means to optimize information inquiries. This paper presents a method of query expansion based on multi-Knowledge Base and local feedback. Two different types of Knowledge Base, domain ontology and Tongyici Cilin, are combined to expand keyword query in the method used, and then the local feedback algorithm is used to filter the expansion words. Experimental results show that the query performance of the method is improved effectively compared to other methods.

Key words query expansion multi-Knowledge Base domain ontology Tongyici Cilin local feedback

0 引 言

搜索引擎是互联网用户查找和获取信息的重要工具,用户通过查询返回了大量的网页,但是其中往往包含大量的与搜索主题无关或者相关度很低的信息。为了提升用户的检索体验,众多搜索引擎提供了网页内容片段,让用户不必点击链接打开网页就可以快速浏览检索网页中的部分内容,但是这种内容片段只是关键词的上下文,并不能准确地反映用户的真实需求,也无法从本质上优化查询结果。查询扩展[1](Query Expansion,QE)是优化信息查询的一种重要手段,可以解决搜索请求中表达差异的问题。通过在原查询词的基础上加入与原查询相关的词或者词组,组成新的、更准确的查询词序列,以便更完整、更准确地描述原查询所隐含的查询语义或主题,尽可能以较小的遗漏检索出候选文档,提高信息检索系统的查全率和查准率。

1 相关研究

查询扩展中的关键技术之一就在于扩展词表的构造。目前扩展词表的构造通常有两种方式:一种是基于统计的方法[2-6],主要研究利用大规模通用语料库的统计信息(例如同现概率、互信息等)构造扩展词表;另一种是基于语义的查询扩展词表构造方法[7-11],主要研究从已有的语义知识词典、领域本体或语义概念网络中选取扩展词。其中,常用的语义知识词典包括 WordNet、HowNet、同义词林等。

收稿日期:2012-09-27 修回日期:2012-12-07

基金项目:国家自然科学基金"事件本体模型与应用技术"(编号:60975033);上海高校青年教师培养资助计划"面向突发事件的语义知识获取"(编号:SLX11010);上海市信息委项目"中共上海市委统战部统战信息处理系统"(编号:XZ20088002)的阶段性研究成果。

作者简介:付剑锋(1978-),男,博士,讲师,研究方向:语义 Web 的研究;刘宗田(1946-),男,博士生导师,教授,研究方向:人工智能、软件工程的研究;刘念祖(1955-),男,教授,研究方向:智能信息系统的研究。

基于统计的查询扩展的基本思想:对文档集中的语词进行相关分性析(如语词共现分析),得到每对语词的关联程度(如共现率),构造候选扩展词表,再从候选扩展词表中选取与原查询关联程度较高的词作为扩展词进行查询扩展。基于统计的查询扩展又可以分为局部分析和全局分析。局部分析方法利用初始检索结果前面相关度最高的若干篇文档从中提取扩展词;全局分析法则认为语料库中相关性高的词汇倾向于在该语料库的文档中共同出现,可以利用整个文档集合的信息来扩展查询,采用的技术包括全局聚类、相似性叙词表、潜在语义索引等。

基于语义的查询扩展的基本思想是把用户提交的原始查询当作一个概念而不是字符串,从已构建的概念语义空间中扩展查询概念的语义。概念语义空间可以是语义知识词典、领域本体或语义概念网络。查询概念的语义扩展主要包括同义扩展、细化/子概念扩展、泛化/父概念扩展、实例化扩展、抽象化扩展等五种。基于语义的查询扩展可以有效克服传统查询扩展技术中产生的查询偏移、计算量过大等种种不足,并且能有效提高信息检索性能,近年来已经成为一个新的研究热[1]。

本文将基于语义和基于统计的两种查询扩展方法相结合,提出一种基于多知识库和局部反馈的查询扩展方法。该方法融合了领域本体与同义词词林两种不同类型的知识库,领域本体可以有效扩展专业范围内的语义相关词汇,同义词林则用于扩充领域本体无法覆盖到的一般词汇。对从两种知识库中扩展出来的候选词,再用局部反馈法对其进行筛选,去除其中的噪音词汇,避免出现查询主题漂移问题。

2 基于多知识库和局部反馈的查询扩展

2.1 领域本体和同义词林 本体(Ontology)最初是一个哲学上的概念,后来被引入计算机领域中,目前普遍接受的对于本体的定义是 Gruber[12] 给出的"本体是概念模型的明确的规范说明"。它将特定领域有关的对象、概念以及概念之间的关系以形式化的说明来严格规定,明确的描述了概念的含义以及概念之间的语义关系。领域本体描述的是某一专业领域之内的特定知识,在专业领域内具有针对性强、概念描述准确等特点。本文采用 protégé 作为本体构建工具,以统战部工作所涵盖的领域构建了统战部领域本体。该本体以中央统战部网站的栏目分类作为分类体系,同时参考部分内部工作文档进行适当的补充。统战部领域本体主要分为六大类,分别包括:多党合作、少数民族、中国的宗教、非公有制经济人士、党外知识分子、港澳台等,每一个分类下面又包含各自的子类,共包含了 3125 个概念,如图 1 所示。

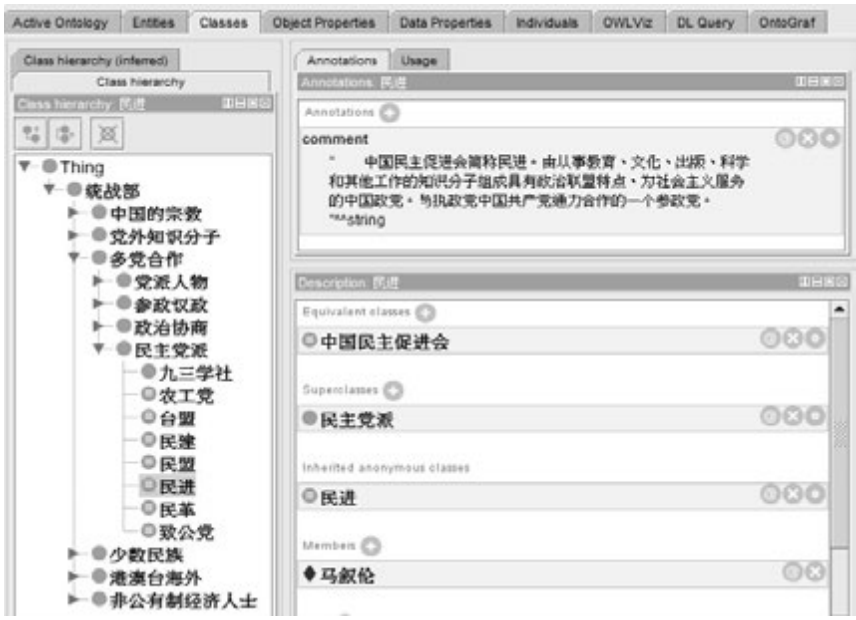

图 1 统战部领域本体

《同义词词林》由梅家驹等人于 1983 年编纂而成,初衷是希望提供较多的同义词语,对创作和翻译工作有所帮助。哈工大《同义词词林(扩展版)》是在梅家驹等人的工作基础之上进一步收集、整理和扩展而成的一部关于同义词的电子词典。该词典不仅包括了一个词语的同义词,同时也包含了一定数量的同类词,即广义的相关词。扩展后的《同义词词林》共收录了 77 343条词语,含有比较丰富的语义信息。《同义词词林(扩展版)》采用了 5 级编码格式将词义按照由上到下、由宽泛到具体词义的语义分类体系,将所收录的词语组织在树状的层级结构中,如图 2 所示。

2.2 多知识库的查询扩展 研究表明[11,13],在专业范围内领域本体通常能够取得比较好的查询扩展效果。但是领域本体也存在规模较少,扩展能力不足的问题。如果采用多知识库的扩展技术,将领域本体与通用词典结合在一起进行关键词扩展,既能弥补领域本体概念数量较少、覆盖面有限的不足,又能突出专业领域内的扩展能力。

本体和同义词词林都是以树状形式进行组织,任意两个概念/词语之间均可达。概念/词语之间语义关系的紧密程度通过语义相似度来度量,在树中两个概念/词语的距离越近则相似度越高,反之则越小。设 $c1$ 和 $c2$ 是树中的任意两个概念/词语,它们之间的语

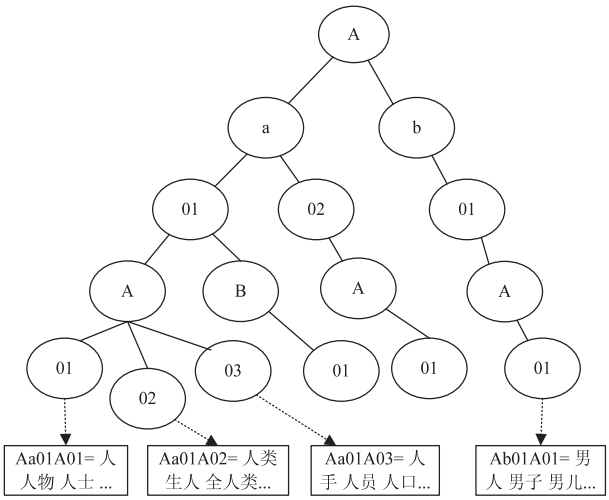

图 2　同义词词林的树状组织结构(片段)

义相似度计算公式如下:

$$\mathrm{sim}(c1,c2)=\frac{1}{(\mathrm{Dist}(c_1,c_2))^2+\alpha} \tag{1}$$

其中, $\mathrm{Dist}(c_1,c_2)$ 是概念/词语 $c1$ 和 $c2$ 之间的最短距离, α 为调节参数,本文设置 $\alpha=1$。

借助领域本体和同义词词林两种知识库,把用户的原始查询映射成知识库中的概念,利用知识库中的语义信息,可实现用户原始查询的语义层次扩展。针对多知识库的特点,本文在传统本体查询扩展算法[11]的基础上进行了改进,具体步骤如下:

假设用户输入查询请求 S,对其进行预处理(分词、去停用词等),得到查询向量 $Q=\{q_1, q_2,\cdots, q_k\}$,其中 $q_i(i=1,2,\ldots,k)$ 表示向量中的第 i 个查询关键词。

设 $C=\{c_1, c_2,\cdots, c_m\}$ 表示本体中概念的集合, $c_j(j=1,2,\cdots,m)$ 表示概念集合中的第 j 个概念,将查询向量中的词 q_i 映射为本体中的概念 $q_i\rightarrow c_{q_i}$,并根据公式(1)扩展出与 c_{q_i} 相似度大于阈值 λ_1 的概念子集 $C_{q_i}=\{c'_1, c'_2,\cdots, c'_g\}$,并将所有概念子集合并得到查询向量 Q 在本体中的扩展集 $C_Q=\bigcup_{i=1}^{k} C_{q_i}$ 。

设 $W=\{w_1, w_2,\ldots, w_n\}$ 表示同义词词林中词汇的集合, $w_j(j=1,2,\cdots,n)$ 表示词汇集合中的第 j 个词汇,将查询向量中的词 q_i 映射到同义词词林中 $q_i\rightarrow w_{q_i}$,并根据公式(1)扩展出与 w_{q_i} 相似度大于阈值 λ_2 的词汇子集 $W_{q_i}=\{w'_1, w'_2,\cdots, w'_h\}$,并将所有词汇子集合并得到查询向量 Q 在同义词词典的扩展集 $W_Q=\bigcup_{i=1}^{k} W_{q_i}$

将 C_Q 与 W_Q 合并得到查询向量 Q 的扩展集合 E_Q 。

2.3　用局部反馈法筛选扩展结果　在查询扩展方法中,过多的扩展词将会影响查询的准确率,同时也会带来“主题漂移”问题[1],使扩展后的查询主题偏离原查询意图,因此必须从扩展集合中去除与查询主题相关度较低的词汇。本文采用局部反馈法[14]对扩展结果进行筛选,具体步骤如下:

将用户的查询向量 $Q=\{q_1, q_2,\cdots, q_k\}$ 提交给信息检索系统进行查询,从查询结果中取前 m 篇文档作为文档集 $D=\{d_1, d_2,\cdots, d_n\}$;

在文档集中计算原查询 $q_i(i=1,2,\cdots,k)$ 与扩展词 $w(w\in E_Q)$ 的共现频数,计算公式为:

$$co(q_i,w)=\sum_{d\in D} tf(q_i,d)\,tf(w,d) \tag{2}$$

其中, $tf(q_i,d)$ 为查询词 q_i 在文档 d 中出现的频数, $tf(w,d)$ 为查询扩展词 w 在文档 d 中出现的频数。

最后,将 $co(q_i,w)$, $(i=1,2,\cdots,k,w\in E_Q)$ 从高到低进行排序,取得分排在前 n 位的查询扩展词 w 。

3　实验结果及分析

用 JAVA 语言编码实现了本文算法,其中全文检索引擎采用了开源工具包 Lucene,分词工具采用了中科院的 ICTCLAS(JAVA 开源版),实验的具体流程如图 3 所示。

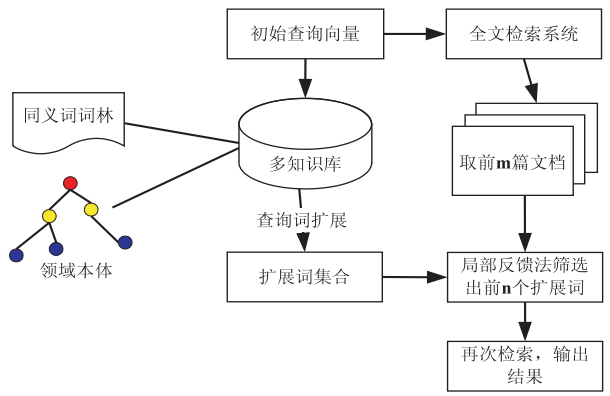

图 3　基于多知识库和局部反馈的查询扩展实验流程图

传统信息检索的评测,一般是在面向通用领域的标准数据集(如:TREC、NTCIR、CLEF 以及 CWIRF 等)上测试,这些数据集很少涉及到统战领域。本文主要针对统战部工作所涉及的领域设计了一个领域本体,实验采用上海大学语义智能实验室的网页抓取工具从互联网上抓取了统战领域新闻网页,经过正文抽取、排重等处理后,取 800 篇作为测试集。实验设计了 10 个查询主题,在测试文档集中通过人工检索比较,获得这 6 个查询的相关文档篇数,见表 1。并将检索结果分别与基准查询(即 *tf – idf* 算法)、基于领域本体(Domain Ontology Based,DOB)的查询扩展、基于多知识库(Muilt– KnowledgeBase Based,MKBB)的查询扩展以及本文方法(Multi – KnowledgeBase and Local Feedback Based,MKBLFB)进行比较,实验结果采用国

际通用的信息检索评价方式查准率 P、查全率 R 和 F 度量作为评价标准,其中 $F=2PR/(P+R)$,实验结果如表 2 所示。

表 1 实验查询主题及其相关文档数

编号	查询主题	相关文档数
Q1	农工党	138
Q2	参政议政	159
Q3	中国道教协会	36
Q4	民族政策	186
Q5	光彩事业	78
Q6	民革中央主席	32

表 2 不同查询算法实验结果比较

查询算法	P	R	F
Baseline($tf-idf$)	0.761	0.413	0.535
DOB($\lambda_1=0.75$)	0.785	0.638	0.704
MKBB($\lambda_1=0.65,\lambda_2=0.75$)	0.798	0.659	0.722
MKBLFB($\lambda_1=0.65,\lambda_2=0.75,m=10,n=8$)	0.847	0.753	0.797

从表 2 可以看出,与基准查询相比较,三种查询扩展方法在查全率和查准率两方面均有提高,这是因为采用查询扩展之后,可以更加完整和准确的描述原查询的语义或主题,比如:对于 Q1 农工党,可以扩展出“中国农工民主党”,而采用机械式关键词匹配的 $tf-idf$ 算法,则无法检索到仅包含“中国农工民主党”的相关文章。在三种查询扩展方法中,MKBLFB 方法可以有效地利用前 10 篇文档进行扩展词的筛选,消除了查询漂移的影响,因此在查准率和查全率两方面比 DOB 方法和 MKBB 方法有了进一步的提升。总体而言,本文提出的 MKBLFB 查询扩展方法可以取得更好的检索性能。

4 总 结

本文提出了一种基于多知识库和局部反馈的查询扩展方法,先将领域本体与同义词词林两种不同类型的知识库对查询词进行扩展,为了防止扩展后带来的查询主题漂移问题,再用局部反馈方法对扩展结果进行筛选。实验结果验证了本文方法的有效性。在将来的研究工作中,我们将继续扩充和完善统战领域本体中的概念,研究更加有效的查询扩展词筛选算法,进一步提高检索性能。

参 考 文 献

[1] C Carpineto, G Romano. A Survey of Automatic Query Expansion in Information Retrieval[J]. ACM Computing Surveys (CSUR), 2012 (44): 1-56

[2] J Xu, W B Croft. Query Expansion Using Local and Global Document Analysis[A]. in Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1996: 4-11

[3] 丁国栋,白 硕,王 斌. 一种基于局部共现的查询扩展方法[J]. 中文信息学报,2006(3): 84-91

[4] 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(20): 1854-1865

[5] 李卫疆,赵铁军,王宪刚. 基于上下文的查询扩展[J]. 计算机研究与发展,2010(2):300-304

[6] 田 萱,杜小勇,李海华. 语义查询扩展中词语-概念相关度的计算[J]. 软件学报,2008,19: 2043-2053

[7] R Bodner, F Song. Knowledge-based Approaches to Query Expansion in Information Retrieval[A]. Lecture Notes in Computer Science, 1996:146-158

[8] 黄名选,严小卫,张师超等. 关联语义的概念查询扩展模型[J]. 情报杂志,2007,26: 92-95

[9] 桑艳艳,刘培刚,李 勇. 基于语义计算的查询扩展优化研究[J]. 情报学报,2007,26:704-710

[10] 王瑞琴,孔繁胜. 基于无导词义消歧的语义查询扩展[J]. 情报学报,2011,30:131-137

[11] 杨清琳,李陶深,农 健. 基于领域本体知识库的语义查询扩展[J]. 计算机工程与设计,2011,31:3853-3856

[12] T R Gruber. A Translation Approach to Portable Ontology Specifications[J]. Knowledge Acquisition, 1993(5):199-220

[13] J Tuominen, T Kauppinen, K Viljanen et al. Ontology-based query Expansion Widget for Information Retrieval[A]. in 6th European Semantic Web Conference (ESWC 2009), Heraklion, Greece, 2009

[14] R Attar, A Fraenkel. Local Feedback in Full-text Retrieval Systems[J]. Journal of the ACM (JACM), 1977(24):397-417

(责编:贺小利)

(上接第 146 页)

馆工作与研究,2011(6):58-60,87

[14] 澳大利亚政府信息公开的有关情况[EB/OL]. http://www.mofcom.gov.cn/aarticle/co/cp/200711/ 200711 05220889.html

[15] 张正禄. 英国政府信息增值利用机制探讨[J]. 图书馆学研究,2010(1):86-88

[16] 陈兰杰. 政府信息商业性再利用的概念、方式与流程研究[J]. 图书馆学研究,2011(12):78-82

[17] 陈传夫,冉从敬. 欧美政府信息增值开发制度及其对我国的启示[J]. 情报资料工作,2008(4):39-43

[18] 傅秀兰. 公共图书馆在政府信息增值服务中的社会效用——以知识管理为视角[J]. 图书馆论坛,2011(2):111-113

[19] 张高蓉. 浅谈图书馆在政府信息公开制度中的作用[J]. 四川图书馆学报,2004 (2):79- 81

[20] 陈 仪. 澳大利亚信息公开的例外事项及对我国的立法启示—兼评我国《政府信息公开条例》的相关规定[J]. 淮海工学院学报:社会科学版,2009(12):40-43

[21] 商淑玲,陈兰杰. 政府信息商业性再利用管理模式初探[J]. 情报资料工作,2011(6):55-59

(责编:贺小利)