ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.19, No.8, August 2008, pp.2043−2053
DOI: 10.3724/SP.J.1001.2008.02043

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

语义查询扩展中词语-概念相关度的计算*

田 萱[1,2,3], 杜小勇[1,2+], 李海华[1,2]

[1](教育部数据工程与知识工程重点实验室,北京 100872)

[2](中国人民大学 信息学院,北京 100872)

[3](北京林业大学 信息学院,北京 100083)

Computing Term-Concept Association in Semantic-Based Query Expansion

TIAN Xuan[1,2,3], DU Xiao-Yong[1,2+], LI Hai-Hua[1,2]

[1](Key Laboratory of Data Engineer and Knowledge Engineer for the Ministry of Education, Renmin University of China, Beijing 100872, China)

[2](School of Information, Renmin University of China, Beijing 100872, China)

[3](School of Information Science & Technology, Beijing Forestry University, Beijing 100083, China)

+ Corresponding author: E-mail: duyong@ruc.edu.cn

Tian X, Du XY, Li HH. Computing term-concept association in semantic-based query expansion. *Journal of Software*, 2008,19(8):2043−2053. http://www.jos.org.cn/1000-9825/19/2043.htm

Abstract: In semantic-based query expansion, computing term-concept association is a key step in finding associated concepts to describe the needed query. A method called K2CM (keyword to concept method) is proposed to compute the term-concept association. In K2CM, the attaching relationship among term, document and concept together with term-concept co-occurrence relationship are introduced to compute term-concept association. The attaching relationship derives from the fact that a term is attached to some concepts in annotated corpus, where a term is in some documents and the documents are labeled with some concepts. For term-concept co-occurrence relationship, it is enhanced by the text distance and the distribution feature of term-concept pair in corpus. Experimental results of semantic-based search on three different corpuses show that compared with classical methods, semantic-based query expansion on the basis of K2CM can improve search effectiveness.

Key words: semantic-based query expansion; concept; ontology; term-concept association

摘 要: 在基于语义的查询扩展中,为了找到描述查询需求语义的相关概念,词语-概念相关度的计算是语义查询扩展中的关键一步.针对词语-概念相关度的计算,提出一种 K2CM(keyword to concept method)方法.K2CM 方法从词语-文档-概念所属程度和词语-概念共现程度两个方面来计算词语-概念相关度.词语-文档-概念所属程度来源于标注的文档集中词语对概念的所属关系,即词语出现在若干文档中而文档被标注了若干概念.词语-概念共现程度是在词语概念对的共现性基础上增加了词语概念对的文本距离和文档分布特征的考虑.3 种不同类型数据集上的语义检索实验结果表明,与传统方法相比,基于 K2CM 的语义查询扩展可以提高查询效果.

关键词: 语义查询扩展;概念;本体;词语-概念相关度

* Supported by the National Natural Science Foundation of China under Grant Nos.60496325, 60573092 (国家自然科学基金)
Received 2007-02-14; Accepted 2007-08-24

中图法分类号: TP311 文献标识码: A

在信息检索领域,查询扩展(query expansion,简称QE)早在20世纪60年代以前就有人提出[1],是公认的能够有效提高查全率的技术之一.其基本思想是利用与查询关键词相关的词语对查询进行修正,以找到更多相关文档,提高查全率.然而,基于关键词的传统查询扩展方式常常会带来许多语义理解错误,文献[2]中称其为词语问题(vocabulary problems),如,同义词问题(synonyms)、歧义问题(polysemy)、异体问题(lemmas)、准同义问题(quasi-synonyms)等,在提高查全率的同时难以保证查准率.

产生词语问题的根本原因在于,人们在现实生活中描述同样的对象或事件的用词存在着多样性,例如,单车和脚踏车都是对自行车这一概念的称谓.为解决这个问题,人们提出了基于概念的语义查询扩展(semantic-based QE),用概念来描述查询主旨,找到与查询语义相关的概念对查询进行扩展[3,4],因为概念是专门用来描述现实世界对象的,概念、词语和现实世界对象三者具有如图1所示的对应关系.基于概念,可以消除现实世界中人们对同一真实对象的不同表达方式之间的差异.

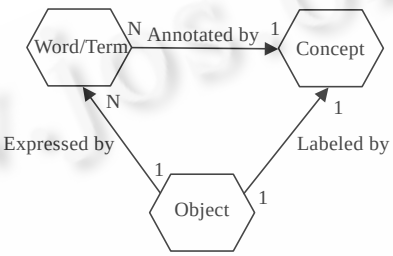

Fig.1 Relationships of words/terms, concepts and objects

图1 词语、概念和对象三者之间的关系

在基于概念的语义查询扩展中,查询通常采用自然语言来描述,因此,如何找到语义相关的概念集描述查询主旨成为一个主要问题.词语-概念相关度词典(association thesaurus)[5]是一个以词语(word/term)和概念(concept)语义相关程度为元素的矩阵,是找到描述查询主旨相关概念的基础.其中,词语-概念语义相关程度的确立是一个中心环节,这就是语义查询扩展中词语-概念相关度的计算问题.针对这个问题,本文提出一种K2CM(keyword- to-concept method)方法,从两个角度描述词语-概念相关度:一方面,针对当前基于本体对文档进行标注的资源组织形式,考虑词语通过文档和概念构成的所属关系,利用词语-文档-概念所属程度描述词语-概念相关度;另一方面,在利用传统词语-概念共现程度描述词语-概念相关度时,考虑词语概念对(term-concept pair)的文本距离和分布特征的影响,以找到与查询语义主旨匹配的概念.

本文第1节介绍相关研究工作.第2节介绍K2CM方法.第3节将K2CM与其他词语-概念相关度计算方法进行对比分析.第4节描述基于K2CM的语义查询扩展在概念检索中的具体实现.第5节给出实验结果和分析.第6节对本研究工作进行总结.

1 相关研究工作

1.1 语义查询扩展

按照来源的不同,语义查询扩展的方法主要分为两类,一类是基于语义关系/语义结构的方法,另一类是基于大规模语料库的方法.

基于语义关系/语义结构的方法常常依据已有的词典/本体,如WordNet(http://wordnet.princeton.edu/),HowNet(http://www.keenage.com/),以及领域词典/本体,如医学领域的MeSH(medical subject headings, http://www.nlm.nih.gov/mesh/)、计算机科学领域的ACMCCS(ACM computer classification system, http://www.acm.org/class/)等.这类方法忽略了用户使用查询关键词存在多样性,查询关键词可能在词典/本体之

外这一事实,前提条件假设查询关键词都来源于词典/本体,即都属于受控词(controlled words)范畴.

基于语料库的方法不存在上面的问题,因为用户查询关键词和语料库词源都来源于现实生活,因而可以认为来源一致.语料库方法的基本思想源于语料库中共现性大的词语往往相关性也很大[6].共现性分析往往基于对文档整体或对文档片断(snippet)的分析,主要分为 3 种:局部分析(local analysis)、局部上下文分析(local context analysis,简称LCA)和全局分析(global document analysis)[4,7].局部分析是从检索结果集的top-k文档中找到出现频率最大的词语(the most frequent term)作为扩展词语;局部上下文分析是从检索结果集的top-k文档中找到与查询词语共现度最大的top-n个词语作为扩展词语;全局分析是从检索文档集中找到与查询共现程度大的词语作为扩展词语.

局部分析和局部上下文分析同属于伪反馈(pseudo feedback)的情况,即假设检索结果集的top-k文档与查询相关.然而,若这个假设条件并不满足,则查询扩展的结果往往有很大偏差,会带来“查询漂移”(query drift)问题,即查询扩展后的主旨偏离了原本的查询意图[8].

全局分析方法计算量较大,但可以在预处理阶段完成,因此并不影响检索效率.本文提出的 K2CM 方法即是一种全局分析方法.

另外,用户反馈是一种公认的效果较好的查询扩展来源[9],如点击浏览过的文档、保存打印的网页、查询日志[10]等,尤其是用户对检索结果相关程度的判断.

1.2 词语-概念相关度的计算

为了找到相关概念,局部上下文分析和全局分析都需要计算词语-概念相关度[5,7].有的研究则是利用相似性来代替相关性[11,12].必须指明的是,查询扩展的目的是尽可能同时提高查全率和查准率,利用与查询语义相关度大的概念是主要途径,而相似度大的概念往往相关度也很大,因此也有助于提高查全率和查准率.计算相似性的方法有很多,如余弦相似度、Dice相似度等,这些方法的前提假设是词语之间是完全独立的.

相关度计算的主要途径是利用文档集中词语间共现性的统计数据.这种方法来源于这样一种直觉,即在语料库中经常共同出现的词语往往相关度很大.分析共现性时,可以采用词语粒度、短语粒度[7]、概念粒度[4,13]等.在语义查询扩展中,概念粒度是最为常见的方式.概念可以来源于文档中的词语聚类[3,4,14]或是本体上已有的概念[15].

另外,信息熵[16]、句法上下文[17,18]等也是相关度计算的依据.然而,这些方法大都是利用文档或文档片断中包含的内容信息,忽略了从文档外部观察文档、概念、词语之间的关系.例如,经过语义标注后的文档(如语义Web)成为概念的实例,词语通过文档和概念构成一种所属关系,这种所属关系从另一个角度说明了词语与概念的相关程度.

2 K2CM 方法

随着语义Web和本体技术的发展,大家普遍认为按照本体标注和组织资源可以方便计算机之间基于语义的交换和处理[19].人们根据本体为越来越多的文档资源添加语义信息,对文档内容中的概念进行标注,以及把文档标注到 1 个或多个概念类别下作为实例是其中最为常见的操作[12,20].针对经过标注的这类文档,K2CM方法在计算词语-概念相关度时从两个角度考虑,一方面基于词语-文档-概念的所属关系;一方面基于有效窗口的局部共现性,用两者的相互作用来衡量词语-概念相关度.

为便于下面进行说明,先进行以下假设表示:设文档集为D,其中的文档数目为M;用来标注该文档集的本体概念集合为C,其中有N个概念.$d_j(j=1,\ldots,M)$表示文档集D中第j个文档,$c_i(i=1,\ldots,N)$表示概念集合中第i个概念.$Q=(q_1,\ldots,q_K)$表示给定的查询,$q_k(k=1,\ldots,K)$表示查询中的一个关键词.

2.1 基于词语-文档-概念的所属关系的考虑

经过本体标注后,文档被标注到 1 个或多个概念类别下,成为本体概念下的实例.这时,文档到概念存在所属关系,同时文档中的词语到概念也存在所属关系,这种所属关系蕴含着词语-概念的相关关系,如图 2 所示.

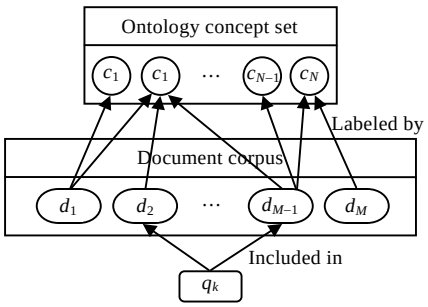

Fig.2 Attaching relationships of words/terms, documents and concepts

图 2 词语、文档、概念之间的所属关系

在词语-文档-概念的所属关系中,一个词语可能被包含在多个文档中,而每个文档又属于 1 个或多个概念类别,通过统计包含词语的文档所属的概念类别,可以统计出这个词语对不同概念类别的所属程度,这种所属程度从另外一个角度说明了词语-概念间的相关关系.

下面我们给出 3 个直觉假设,来说明一个词语对一个概念的所属关系.

假设 1:一个词语通过文档映射到的概念个数越多,它对单个概念的所属程度越低.

假设 2:一个词语在一个概念下属文档中的词频越高,它对这个概念的所属程度越高.

假设 3:一个词语在一个概念下属的越多文档中存在,它对这个概念的所属程度越高.

假设 1 是从词语在概念空间的分布情况来分析.根据直觉,一个词语与越多的概念关联,它对概念的区分性就越不明显,它与概念的关联程度也就越低.

假设 2 是从词语在一个概念下的文档空间中出现的频率来分析.这里,我们选择基于词语的统计粒度,而不是基于文档的统计粒度,即把词语在一个概念下的文档空间的词频作为统计量,而不是把一个概念下的文档空间出现该词语的文档数目作为统计量.这样考虑的原因在于,如果只考虑文档数目,粒度太粗,词语对概念的所属程度区分性不强;而如果按照词语在该概念下的文档空间的词频统计,粒度细,区分性强,则可以更准确地刻画这个词对概念的所属程度.例如,如图 3 所示,两个词语k_1和k_2,它们属于概念c_1的文档数目相等(都是 2 个),如果按文档粒度统计,它们对这个概念的所属程度相同;但如果按照词频统计,k_1的词频比k_2大,则k_1对c_1的所属程度比k_2对c_1的所属程度要大,这样,不同词语对相同概念的所属程度更具区分性.

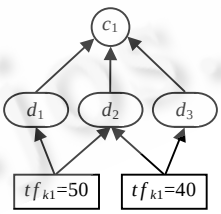

Fig.3 An example of two terms in the same amount documents of a concept

图 3 一个概念下属于相同数目文档的两个词语举例

假设 3 从词语在一个概念下的文档空间中的分布情况来分析.直觉上,词语在概念下属的越多文档中存在,说明它在这个概念中分布得越均匀,也就是说,它与概念的所属关系被越多的文档承认,因而它对这个概念的所属程度也就越高.

基于以上 3 个假设,我们给出词语对概念所属程度权值(attaching weight)的计算方法.

设$d_m \propto c_i$表示文档d_m是概念c_i的实例;D_i表示概念c_i下的文档空间,即$D_i=\{d_m | d_m \in D \wedge d_m \propto c_i\}$;$count(q_k;d_m)$表示词语$q_k$在文档$d_m$中出现的次数;$len(d_m)$表示文档$d_m$的长度;$tf_{k,i}=\sum_{d_m \in D_i} \frac{count(q_k;d_m)}{len(d_m)}$表示词语$q_k$通过文档

映射到概念c_i的词频统计量;n_k表示q_k根据文档-概念关系映射到概念上的概念数目;l_i表示c_i概念下文档的数目,即c_i概念下的文档空间大小;$l_{k,i}$表示词语q_k出现在c_i概念下文档空间中的文档数目.根据上面 3 个假设,我们给出公式(1)来计算词语q_k对概念c_i所属程度的权值$aw_{k,i}$.

$$aw_{k,i}=\log\left(\frac{N}{n_k}+1.0\right)\cdot tf_{k,i}\cdot\log\left(\frac{l_{k,i}}{l_i}+1.0\right) \tag{1}$$

2.2 基于有效窗口内共现性的考虑

有许多基于共现性发现词语-概念相关度的研究,如在每篇文档中的共现性,段落中的共现性以及句子中窗口中的共现性等[21].这里,我们把整个文档集看作一篇大的文档,在尽量短而又有意义的窗口(文献[22]称之为有效窗口)内扫描词语-概念在这篇文档中的共现性.有效窗口的确定可以采用文献[22]的建议,即以词语为核心,汉语语料库中位置为[–8,9]和英语语料库中位置为[–16,13]的窗口.

用$cw_{k,i}$表示词语q_k和概念c_i的共现程度权值(co-occurrence weight).根据直觉,给出以下 3 个假设来计算$cw_{k,i}$.

假设 4:一个词语-概念对在整个文档集中出现的次数越多,它们的共现程度越强.

假设 5:一个词语-概念对在文档中出现的物理位置越近,它们的共现程度越强.

假设 6:一个词语-概念对分布于越多的不同文档中,它们的共现程度越强.

根据这 3 个假设,我们给出公式(2).

$$cw_{k,i}=\frac{tpf_{k,i}\cdot\log\left(\frac{m_{k,i}}{M}+1.0\right)}{\log(avgdist_{k,i}+1.0)} \tag{2}$$

其中,$tpf_{k,i}=\dfrac{count(q_k,c_i;W)}{\mathrm{Max}_{j=1...N}(count(q_k,c_j;W))}$,表示词语-概念对$(q_k,c_i)$在文档集中出现的频率(term-concept pair frequency),$count(q_k, c_j; W)$表示词语-概念对(q_k,c_i)在文档集中按W大小的窗口扫描出现的次数;$m_{k,i}$表示词语-概念对(q_k,c_i)在文档集中出现的文档数目;$avgdist_{k,i}$表示词语-概念对在W大小窗口中位置距离的平均值.

然而,文档是由词语构成的,并不是概念.如果单纯统计概念c_i自身的词语,会产生许多误差.例如,“他昨天在中关村买了台PC机”这句话中实际也包含了词语“中关村”和概念“计算机”的共现关系(假设某本体中PC机不是概念).因此,在统计词语-概念共现度时不能只包含概念c_i自身的词语,要尽可能包含概念c_i的同义词、入口词(entry item)等.据此,对公式(2)进行如下改进:

假设Γ_i为概念c_i的同义词集,$t_{ij}\in\Gamma_i,j=1,...,|\Gamma_i|$,其中$t_{ij}$是概念$c_i$的同义词或入口词(entry item).词语$q_k$和概念$c_i$的共现程度权值$cw_{k,i}$可按如下公式进行计算:

$$cw_{k,i}=\frac{1.0}{|\Gamma_i|}\sum_{j=1,...,|\Gamma_i|}\frac{tpf_{k,ij}\cdot\log\left(\frac{m_{k,ij}}{M}+1.0\right)}{\log(avgdist_{k,ij}+1.0)} \tag{3}$$

其中,$tpf_{k,ij}=\dfrac{count(q_k,t_{ij};W)}{\mathrm{Max}_{\substack{h=1,...,N,\\ p=1,...,|\Gamma_h|}}(count(q_k,t_{hp};W))}$,表示$(q_k,t_{ij})$在文档集中出现的频率(term-concept pair frequency),$count(q_k,t_{ij};W)$表示(q_k,t_{ij})在文档集中按W大小的窗口扫描出现的次数;$m_{k,ij}$表示(q_k,t_{ij})在文档集中出现的文档数目;$avgdist_{k,ij}$表示词语-概念对(q_k,t_{ij})在W大小窗口中位置距离的平均值.

2.3 词语-概念相关度的计算

对于一个词语q_k和一个概念c_i的相关度,$aw_{k,i}$从q_k经过文档映射到c_i的所属关系中分析,$cw_{k,i}$从两者在有效窗口的共现性角度分析.下面,我们把这两个因素综合起来评价词语-概念相关度.

这里我们忽略共现性与所属性对词语-概念相关度的贡献性差别,即认为两者对词语-概念相关度具有相同的影响力.同时,考虑到共现性和所属性都是对相关性的一种描述,两种因素相互作用的结果更能有效说明词语

与概念相关程度的强弱.所以,我们采用两种因素直接相乘的方式来定义词语-概念相关度(term-concept association)$tca_{k,i}$,如公式(4)所示.

$$tca_{k,i} = cw_{k,i} \cdot aw_{k,i} \tag{4}$$

值得强调的是,计算词语-概念相关度的目的是找到那些与查询关键词语义相关的概念,从而用来描述整个查询的语义主旨.因此,采用两个因素相乘的方法可以突出和查询关键词语义相关度大的那些概念,即关键词和它们的共现度与所属度都相对较大的概念.

3 K2CM 和其他方法的对比

表 1 对 K2CM 和其他几种具有代表性的词语-概念相关度计算方法进行了对比分析.与其他方法相比,K2CM 最显著的特点是,一方面根据被标注文档中存在的词语-文档-概念所属关系,一方面根据有效窗口的全局共现性,把两者结合起来计算词语-概念相关度.

Table 1 Comparisons of K2CM with other methods of computing term-concept association

表 1 K2CM 与其他词语-概念相关度计算方法的对比

Title of referring paper	Expansion source	Granularity of analysis	Expansion method
Concept based query expansion[3]	Whole document sets	Co-Occurrence of documents "within" terms	Constructing term-term similarity thesaurus
Improving the effectiveness of information retrieval with local context analysis[7,23]	Top-ranked documents	Local co-occurrence of terms and clusters (concepts)	Constructing term-cluster co-occurrence thesaurus
Query reformulation using automatically generated query concepts from a document space[4]	Whole document sets, retrieved documents	Term clusters in sentences	Constructing orthogonal similarity of term-cluster by cosine method
This paper	Whole annotated document sets, ontology	Global co-occurrence of term-concept pair, attaching relationship of term-document-concept	Constructing association thesaurus of term-concept

4 查询-文档相关度的计算

语义查询扩展的目的是找到与查询整体语义主旨相关的概念进行扩展[3].K2CM方法计算的是一个词语和本体中一个概念的语义相关度,目的是构建出词语-概念相关度词典(association thesaurus)[5],应用于语义查询扩展中.这里,我们还必须要考虑到如何根据词语-概念相关度词典计算查询整体和概念的语义相关程度(即查询-概念相关度).因为如果只单纯把一个词语的相关概念引入查询扩展,而忽略查询中其他词语的影响,反而会使查询需求描述变得模糊,发生查询漂移之类的问题.例如,在英语中,"program"与"computer"紧密相关,但对"TV program"这一有关电视节目查询,就不适合把"computer"加入到查询扩展中.

检索过程是文档与查询一一匹配的过程.概念检索中,文档按照概念索引,同时查询用概念来描述.因此,如何基于概念计算查询与文档相关的程度(即查询-文档相关度)是检索过程中的主要环节.需要指出的是,传统向量空间模型中的余弦相似度方法并不适合基于概念的查询-文档相关度计算,因为向量空间模型的前提假设在于词语之间是正交无关的,而概念的最大特点是概念之间具有语义相关性.

下面介绍语义查询扩展在概念检索中的实现过程,主要包括查询-概念相关度(query-concept relevance)的计算和查询-文档相关度(query-document relevance)的计算.

4.1 查询-概念相关度的计算

根据K2CM建立的词语-概念相关度词典,对查询$Q=(q_1,\ldots,q_K)$,可以得到相应的一个查询关键词-概念相关度矩阵(keyword-concept correlation matrix,简称KC-CM),如公式(5)所示.其中,一个行向量(keyword-concept correlation vector,简称KC-CV)描述了一个查询关键词和本体中所有概念的相关度,$tca_{i,j}$描述了关键词q_i与概念c_j的相关度.

$$KC\text{-}CM=\begin{pmatrix} tca_{1,1} & & tca_{1,N} \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ tca_{k,1} & & tca_{k,N} \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ tca_{K,1} & & tca_{K,N} \end{pmatrix} \tag{5}$$

对查询-概念相关度(query-concept relevance)$qcr(Q,c_i)$的计算,可以按照线性组合把查询看成一个整体,表示成各个词语的概念向量的线性组合[3,7],如公式(6)所示.

$$qcr(Q,c_i)=\sum_{k=1..K} w_k tca_{k,i} \tag{6}$$

其中,w_k表示关键词q_k在查询Q中的权重.设置w_k的方法有许多,如经典的Rochio方法、Ide方法等[1].本文实验中采用了Rochio方法.

根据查询与N个概念的查询-概念相关度$qcr(Q,c_i)$,可选择出top-n个概念作为对查询Q的概念描述.对n的确定、按阈值选择或指定个数是研究中常采用的方法[4,24],本文实验中采用指定个数的方法.

4.2 查询-文档相关度的计算

由于本体中概念之间具有语义相关性,所以,基于概念计算查询-文档相关度时必须考虑到概念两两之间的相关性.假设已知概念c_i和c_j之间相关度(concept-concept association)$cca(c_i,c_j)$,查询Q由N维概念向量$(c_1,c_2,\ldots,c_N)$描述,文档d_m由N维概念向量$(c'_1,\ldots,c'_2,\ldots,c'_N)$描述.我们按照公式(7)计算查询-文档相关度(query-document relevance)$qdr(Q,d_m)$.

$$qdr(Q,d_m)=\sum_{i=1,\ldots,N}\sum_{j=1,\ldots,N}\gamma_i\eta_j cca(c_i,c'_j) \tag{7}$$

其中,γ_i表示概念c_i在查询Q中的权重,本文实验中令$\gamma_I=qcr(Q,c_i)$.η_j表示概念c'_j在文档d_m中的权重,可以按照传统的求词语在文档中权重的方法(如TF×IDF,Okapi BM250方法)来获得[1].但计算η_j的前提是文档已经按照本体概念建立概念索引,或文档内容已经按照本体概念被标注过.本文实验中按照本体中概念词自身、同义词和入口词来标注文档内容,建立概念索引,η_j的设置采用了经典的TF×IDF方法.

这种基于概念相关度计算查询-文档相关度方法的关键前提是需要事先知道本体上任意两个概念c_i和c_j之间的相关度$cca(c_i,c_j)$.关于这个问题有许多公开方法和相关应用,如文献[15,25–27],鉴于这不是本文研究的重点,这里不再进行介绍,本文实验中采用了文献[27]的方法.

5 实验评价

5.1 实验设置

为了便于考察不同本体规模对 K2CM 的影响,实验中采用了 3 个不同规模的本体,一个是通用本体 WordNet,另外两个是领域本体——经济学领域本体 EO(economic ontology)和计算机科学领域本体 ACMCCS98(ACM Computer Classification 98).WordNet 是为大家所广泛采纳的通用本体,WordNet 的读取采用了 SourceForge 开放源码社区提供的 JWNL 接口(http://sourceforge.net/projects/jwordnet);EO 是 NSFC 资助项目“通用网上知识编辑器及示范主题语义网研究”的一部分成果,基本涵盖了经济学领域的重要概念和关系;ACMCCS98 (http://www.acm.org/class/1998/)是 ACM 推荐的计算机科学领域的分类体系(是经历了 ACMCCS64,ACMCCS91 后的最新版本,ACM 2006 年推荐版本),也包括了基本的同义和关联关系,因此,这里把它作为计算机科学的领域本体.3 个本体的统计信息详见表 2.

Table 2 Statistic data of the three ontologies

表 2 3 个本体的统计数据

Ontology Name	# of Concepts
WordNet	145 104
EO	9 470
ACMCCS98	1 433

对应于 3 种不同的本体,相应采用的检索文档集/语料库有 3 个,一个是美国国家标准技术局(National Institute of Standards and Technology,简称 NIST)于 2004 年公开发布的 TREC2001 Filtering Track 中使用的 Reuters 数据集(http://www.jmlr.org/papers/volume5/lewis04a/lyrl2004_rcv1v2_README.htm),另一个是来源于人民大学数字图书馆中部分被标注的有关经济学的文档资源(只有论文标题),还有一个是经过 ACM 数字图书馆元数据(http://portal.acm.org/portal.cfm)标注后的 DBLP 数据库(http://dblp.uni-trier.de/xml/)的资源(只有论文摘要).3 个文档集的统计信息详见表 3.

Table 3 Statistic data of the three corpuses

表 3 3 个文档集的统计数据

Corpus name	# of documents	Average # of words per Doc	Average # of annotated concepts per Doc	Language
Reuters RCV1-v2/LYRL2004	804 414	45.9	19.4	English
DLPers	785 426	8.2	5.3	Chinese
DBLP	19 229	169.7	80.5	English

我们在 3 个数据集上分别构建了 10 个查询进行检索,人工对返回的前 20 个结果文档的相关性进行评价,评价原则和结果的评测指标参见第 5.3 节.

5.2 实验对比方法

鉴于K2CM方法适用于基于概念的语义检索,因此,我们选择未经查询扩展的概念检索作为基线对比方法(baseline).另外,我们选择LCA作为另一种对比方法.LCA是一种经典的语义查询扩展方法,由Xu和Croft提出来[7],它最主要的贡献在于通过计算初始检索结果集的top-k文档中词语和查询中词语的共现度,以找到查询相关词语作为扩展词语[7].两个词语共现度的计算如公式(8)所示.

$$co\_degree(c,w)=\log(co(c,w)+1.0)\times idf(c)/\log(n) \tag{8}$$

其中, $co(c,w)=\sum_{d\in S} tf(c,d)tf(w,d)$, $idf(n)=\min(1.0,\log(N/N_c)/5.0)$ ·$tf(c,d)$表示词语c在文档d中出现的频率,S是初始检索结果集中的top-k文档集,N_c是文档集中含有词语c的文档数目.

在查询扩展过程中,加入的词语/概念个数、查询扩展后的结构以及查询扩展词语/概念的权值分配都是需要研究的重要问题.这里,对查询扩展加入的词语个数采用常见的个数限定法,即限定查询扩展中加入的词语个数.20~30 是许多研究中给出的个数限定范围,因此,实验中统一把查询扩展的词语/概念最大个数限定为 30.

5.3 实验评测标准和实验结果分析

主要从查询准确率方面进行评价,分别采用*Precision@n*和*AP@k*来衡量.*Precision@n*是前n个结果文档中查询准确率,常用来衡量大多数用户关注的前n个结果文档的准确率[1].由于实验中采用多个本体进行测试,标准TREC数据集难以支撑,所以,我们自己选择了本体对应的文档集.根据文档规模,要从人力上判断文档集中所有文档与查询的相关性是难以完成的.因此借鉴标准指标MAP(mean average precision)定义了*AP@k*,目的是描述前k个结果文档中相关文档的准确率平均值,以衡量前k个结果文档中相关文档的排序情况.这样,*AP@k*和*Precision@n*在一起能够更全面地对top-k检索结果进行评价,而这也符合大多数检索用户的习惯,因为大多数用户在检索过程中主要关注top-k检索结果.

Precision@n 的计算方式是:*Precision@n*=# of relevant docs in top-n retrieved/n,其中 n 表示前 n 个结果文档. *AP* 的计算方式是: $AP@k=\frac{1}{r}\sum_{rank_j\le k}\frac{j}{rank_j}$,其中,$r$ 表示前 k 个结果文档中相关文档的个数,j 表示前 k 个结果文档

中第j个相关文档;$rank_j$表示第j个相关文档在结果文档中的排序.考虑到用户往往只关注前 20 个检索结果,这里我们取 n=k=20.

在判断查询结果相关性时,为了避免人工判断的倾向性,我们将每种方法得到的前 20 个查询结果混合在一起,经过去重后,再人工判断每个结果是否相关,最后统计出每种方法下的相关查询结果的数量.图 4 中给出了不同方法在不同数据集上的检索效果,表 4 中列出了 3 种方法在 3 个文档集上的检索结果的具体指标值.

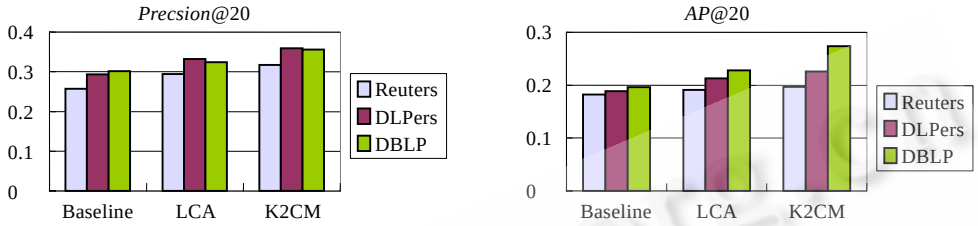

Fig.4 Comparisons on different methods

图 4 不同方法的比较

Table 4 Search result evaluation on the three corpuses

表 4 3 个文档集上的检索结果评价

Metric	Document set	Baseline	LCA	K2CM
Precision@20	Reuters	0.257	0.294(↑14.40%)	0.317(↑23.35%)
	DLPers	0.293	0.332(↑13.31%)	0.359(↑22.53%)
	DBLP	0.301	0.324(↑7.64%)	0.356(↑18.27%)
AP@20	Reuters	0.182	0.191(↑4.95%)	0.197(↑8.24%)
	DLPers	0.189	0.213(↑12.70%)	0.226(↑19.58%)
	DBLP	0.196	0.228(↑16.33%)	0.274(↑39.80%)

我们先从总体上分析本体规模和构建粒度对语义查询扩展的影响.从图 4 中的基本数据可以看出,虽然Reuters数据集中文档数量最多,采用的也是含有最多概念的WordNet本体,但其整体的检索效果比其他两种采用领域本体的数据集都差.因为文档数量对检索效果几乎没有影响[28],因此可知通用本体对查询扩展的作用不如领域本体.这个结论与其他一些研究的结论一致[29].另外,通过表 4,可以观察到DBLP上*Precision@n*的提高水平(7.64%,18.27%)比Reuters和DLPers上相应要低,这说明领域本体ACMCCS在提高检索准确率方面不如其他本体.分析其原因,由于ACMCCS本质上是个计算机科学领域的分类体系,粒度比较粗,因此查询扩展过程中概念扩展的粒度也比较粗,导致检索效果的提高有限.

我们再来分析 K2CM 方法在检索效果中表现的特点.首先,可以看出,K2CM 在两个指标上提高的程度都比 LCA 要大,这一点原因比较明显,与单纯基于共现关系的 LCA 相比,K2CM 通过语料库分析找到的相关概念更准确,因而对查询扩展更有效.其次,可以看出,K2CM 在 DBLP 数据集的 *AP*@20 指标上提高得尤其明显(↑39.80%),究其原因,DBLP 数据集的平均文档长度最长,说明 K2CM 在表达长文档的语义主旨方面更有效,因而特别有助于提升相关文档的排序效果.从这一点也可以推测出,越长的查询,K2CM 的检索效果应该越好.有关这方面的研究和实验将在我们下一步的工作中进行.

6 总 结

现实生活中存在的查询表达多样性常常给基于关键词的传统查询扩展带来许多语义理解错误——词语问题,基于概念的语义检索已经是一种公认的解决方法.在基于概念的语义查询扩展中,词语-概念语义相关度的确立是个中心环节.随着语义 Web 和本体技术的发展,越来越多的文档按照本体来标注和组织.针对这类文档,本文提出一种 K2CM 方法,从两个角度描述词语-概念相关度,一方面基于词语-文档-概念所属程度,一方面基于词语-概念共现程度,以找到与查询语义主旨匹配的概念,进而提高查询效果.实验结果表明,与未经查询扩展的概念检索方法和基于 LCA 的概念检索方法相比,本文提出的 K2CM 方法更有助于提高查询效果.

References:

[1] Baeza-Yates R, Ribeiro-Neto B. Modern Information Retrieval. New York: Addison-Wesley-Longman, 1999.

[2] Furnas GW, Landauer TK, Gomez LM, Dumais ST. The vocabulary problem in Human-System communication. Communications of the ACM, 1987,30(11):964−971.

[3] Qiu YG, Frei HP. Concept based query expansion. In: Korfhage R, Rasmussen E, Willett P, eds. Proc. of the 16th annual Int'l ACM SIGIR Conf. on research and development in information retrieval. Pittsburgh: ACM Press, 1993. 160−169.

[4] Chang Y, Ounis I, Kim M. Query reformulation using automatically generated query concepts from a document space. Information Processing and Management, 2006,42:453−468.

[5] Jing YF, Croft WB. An association thesaurus for information retrieval. Technical Report, UM-CS-1994-017, Amherst: University of Massachusetts, 1994.

[6] van Rijsbergen CJ. Information retrieval. Department of Computing Science, University of Glasgow, 1979. http://www.dcs.gla.ac.uk/Keith/Preface.html.

[7] Xu JX, Croft WB. Improving the effectiveness of information retrieval with local context analysis. ACM Trans. on Information Systems, 2000,18(1):79−112.

[8] Mitra M, Singhal A, Buckley C. Improving automatic query expansion. In: Croft W B, Moffat A, Wilkinson R, Zobel J, eds. Proc. of the 21st Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Melbourne: ACM Press, 1998. 206−214.

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

[10] Cui H, Wen JR, Li MQ. A statistical query expansion model based on query logs. Journal of Software, 2003,14(9):1593−1599 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/19/9.htm

[11] Ido D, Lillian L, Fernando CNP. Similarity-Based models of word cooccurrence probabilities. Machine Learning, 1999,34(1-3): 43−69.

[12] Zazo ÁF, Figuerola CG, Berrocal JLA, Rodriguez E. Reformulation of queries using similarity thesauri. Information Processing and Management, 2005,41(5):1163−1173.

[13] Zhang M, Song RH, Ma SP. Document Refinternet based on semantic query expansion. Chinese Journal of Computers, 2004, 27(10):1395−1401 (in Chinese with English abstract).

[14] Lin DK. Automatic retrieval and clustering of similar words. In: Boitet C, Whitelock P, eds. Proc. of the 17th Int'l Conf. on Computational Linguistics. Montreal: Association for Computational Linguistics, 1998. 79−112.

[15] Kim JW, Candan KS. CP/CV: Concept similarity mining without frequency information from domain describing taxonomies. In: Yu PS, Tsotras VJ, Fox EA, Liu B, eds. Proc. of the 15th ACM Int'l Conf. on Information And Knowledge Management. Arlington: ACM Press, 2006. 483−492.

[16] Jang MG, Myaeng SH, Park SY. Using mutual information to resolve query translation ambiguities and query term weighting. In: Dale R, Church K, eds. Proc. of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics. College Park: Association for Computational Linguistics, 1999. 223−229.

[17] Gao JF, Zhou M, Nie JY, He HZ, Chen WJ. Resolving query translation ambiguity using a decaying Co-Occurrence model and syntactic dependence relations. In: Järvelin K, Chairs P, Baeza-Yates R, Myaeng SH, eds. Proc. of the 25th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Tampere: ACM Press, 2002. 183−190.

[18] Gregory G. Use of syntactic context to produce term association lists for text retrieval. In: Belkin N, Ingwersen P, Pejtersen AM, eds. Proc. of the 15th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Copenhagen: ACM Press, 1992. 89−97.

[19] Loh S, Wives LK, de Oliveira JPM. Concept-Based knowledge discovery in texts extracted from the Web. ACM SIGKDD Explorations Newsletter, 2000,2(1):29−39.

[20] Fraenkel AS, Klein ST. Information retrieval from annotated texts. Journal of the American Society for Information Science, 1999, 50(10):845−854.

[21] Sun RX, Ong CH, Chua TS. Mining dependency relations for query expansion in passage retrieval. In: Efthimiadis EN, Dumais ST,

Hawking D, Järvelin K, eds. Proc. of the 29th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Seattle: ACM Press, 2006. 382−389.

[22] Lu S, Bai S. Quantitative analysis of context field in natural language Processing. Chinese Journal of Computers, 2001,24(7): 742−747 (in Chinese with English abstract).

[23] Xu JX, Croft WB. Query expansion using local and global document analysis. In: Frei HP, Harman D, Schäble P, Wilkinson R, eds. Proc. of the 19th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Zürich: ACM Press, 1996. 4−11.

[24] Martin T, Ralf S, Gerhard W. Efficient and self-tuning incremental query expansion for Top-*K* query Processing. In: Baeza-Yates R, Ziviani N, eds. Proc. of the 28th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Salvador: ACM Press, 2005. 242−249.

[25] Green SJ. Building hypertext links by computing semantic similarity. IEEE Trans. on Knowledge and Data Engineering, 1999,11(5): 713−730.

[26] Kandola JS, Shawe-Taylor J, Cristianini N. Learning semantic similarity. In: Becker S, Thrun S, Obermayer K, eds. Advances in Neural Information Processing Systems 15 (Neural Information Processing Systems, NIPS 2002). Vancouver: MIT Press, 2002. 657−664.

[27] Varelas G, Voutsakis E, Raftopoulou P, Petrakis EGM, Milios EE. Semantic similarity methods in WordNet and Their Application to Information Retrieval on the Web. In: Bonifati A, Lee D, eds. Proc. of the 7th Annual ACM Int'l Workshop on Web Information and Data Management. Bremen: ACM Press, 2005. 10−16.

[28] Fang H, Tao T, Zhai CX. A formal study of information retrieval heuristics. In: Sanderson M, Järvelin K, Allan J, Bruza P, eds. Proc. of the 27th annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Sheffield: ACM Press, 2004. 49−56.

[29] Lin J, Demner-Fushman D. The role of knowledge in conceptual retrieval: A study in the domain of clinical medicine. In: Efthimiadis EN, Dumais ST, Hawking D, Järvelin K, eds. Proc. of the 29th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Seattle: ACM Press, 2006. 99−106.

附中文参考文献:

[10] 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):1593−1599. http://www.jos.org.cn/1000-9825/14/1593.htm

[13] 张敏,宋睿华,马少平.基于语义关系查询扩展的文档重构.计算机学报,2004,27(10):1395−1401.

[22] 鲁松,白硕.自然语言处理中词语上下文有效范围的定量描述.计算机学报,2001,24(7):742−747.

田萱(1976－),女,山东济宁人,博士生,CCF 学生会员,主要研究领域为智能信息检索,知识工程,高性能数据库.

李海华(1971－),女,博士生,CCF 学生会员,主要研究领域为语义 Web,知识工程,智能信息检索.

杜小勇(1963－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为高性能数据库,智能信息检索,知识工程.