DOI:10.11721/cqnuj20130422

信息检索技术中基于语义的扩展查询研究*

李 兴 春
(重庆文理学院 教学部,重庆 永川 402160)

摘要:用户查询与文档之间语义匹配但词法不匹配现象是影响信息检索效果的重要原因之一。鉴于语义检索受限于本体自身的质量,为了降低其对检索效果的影响,通过分析目前语义查询扩展的研究现状,在已有概念相似度计算算法研究基础上进行改进和融合,提出了一种基于本体的信息检索查询扩展方法,并主要对基于本体技术的概念相似度计算算法进行修正,得到了组合向量空间模型 $QCR(Q,C_i)=\sum_{k=1,\cdots,K} w_k * Sim\_Rel(q_K,C_i)$,作为引入查询扩展后的查询结果相关度评价方法。这种方法中,通过建立本体模型并计算本体中概念间的语义相似度来确定扩展查询词,它可以根据用户输入的名称,检索出相关文档并由用户自由设置相似度阈值,并将普通主题检索与语义检索合并,在本体乏力时返回普通检索结果,这在一定程度上弥补了垂直检索系统发展的不足。

关键词:信息检索;语义相似度;扩展查询
中图分类号:TP391.1　　**文献标志码**:A　　**文章编号**:1672-6693(2013)04-0113-04

目前搜索引擎的主要工作方式是:搜索服务提供公司抓取互联网上的网页、分析网页并针对其中内容建立索引,用户提交一定长度的检索请求,由搜索引擎服务器通过基于关键字匹配技术检索出相关的内容,并通过一定的排序算法呈现在用户面前。在这种基于关键词的检索系统中,只有当用户的查询词出现在文档中,这个文档才有可能被检索到。由于自然语言的复杂性,常常存在下面两种情况:一个概念可以有很多种不同的表达方式,相同概念在不同场景下往往有不同的含义,即自然语言中的同义词和多义词;两个概念存在语义上或逻辑上的关联,但仅仅基于关键字的查询系统无法找出这些隐含的关系。因此,关键词查询系统中查全率往往不尽人意,经常会出现与用户查询词在语义上匹配的信息无法被检索出来的情况,这时用户就不得不变换查询词来找到所需要的信息。查询扩展(Query expansion)正是解决该问题的有效方法之一。所谓查询扩展,简而言之,即是将与用户初始查询词相关的概念术语也一起并作查询概念词以形成最终详细的查询信息关键词集。基于本体的检索查询扩展主要是检索词的概念语义扩展,主要包括同义扩展、语义蕴涵、外延扩展及语义相关联想等一系列推理方式[1-3]。

通过查询扩展得到的概念关键词集合不仅提高了查全率,也一定程度上细化了用户搜索需求,从而提升了检索操作的用户体验。但由于扩展得到的关键词集合本身没有按照相关程度进行排序,因此不能完整真实地反映领域知识中的关联特点。这样需要利用语义相似度算法来进行计算和排序。但最终的排序,还必须结合检索页面与检索词的相似度来进行综合排序[4]。

按照来源的不同,语义查询扩展的方法主要分为两类[5]:一类是基于语义关系/语义结构的方法,它借助于已有的词典、本体,这样会对于检索词有所要求;另一类是基于大规模语料库的方法,这对语料库的要求较高,并且需要先找出所有可能的查询词,才能求出其相关扩展概念集和其"查询词语-概念"相关度大小,并存库,否则每次查询过程都重新计算会消耗不少时间和内存,导致检索效率低下。

目前,基于大规模通用本体 HowNet(即知网,中文的词汇语义网络)的查询扩展已取得了不少成绩。HowNet 是描述概念与概念间关系以及概念的属性与属性间关系的知识系统。现在最显著的就是基于知网研发的概念相似度计算软件和概念相关场计算软件。和知网不同的是,WordNet 在一开始概念定义的时候就采用了网状结构,对每个概念的定义中都同时标注了他的上下位关系词、同义词、反义词等,而 HowNet 对概念的定义则是完全孤立的,单纯的从概念的应用方法角度进行定义[6-7]。

知网的查询扩展借助于其基本单位——义原。由于一个整体的各个不同部分在整体中的作用是不同的,只有在整体中起相同作用的部分互相比较才有效,通过对这两个整体(词语)的各部分(义原)之间建立一一对应关

* **收稿日期**:2013-04-06　**修回日期**:2013-05-18　**网络出版时间**:2013-07-20 19:23
资助项目:重庆市自然科学基金(No. CSTC2011JJA40011);重庆市教委科技项目(No. KJ111216)
作者简介:李兴春,男,助理实验师,硕士,研究方向为计算机应用技术等,E-mail:w214513@163.com
网络出版地址:http://www.cnki.net/kcms/detail/50.1165.N.20130720.1923.201304.113_019.html

系,可以得出如下该两词语之间的相似度计算公式:

$$Sim(C_1,C_2)=\sum_{i=1}^{n}w_iSim(P_i),\ P_i=\begin{cases}<A_{1i},A_{2i}>,1\leqslant i\leqslant m\\<\varphi,A_{2i}>,m<i\leqslant n\end{cases}$$

$$w_i=\frac{d_i}{\sum_{i=1}^{m}d_i},(1\leqslant i\leqslant n,d_i=\min(\text{depth } A_{1i},\text{depth } A_{2i}))$$

在式中,C 表示概念词,$w_i(1\leqslant i\leqslant n)$表示是可调节的参数,即义原对的相似度对于总体相似度所起到的作用权重,且有:$w_1+w_2+w_3+\cdots+w_n=1$,$w_1\geqslant w_2\geqslant w_3\geqslant\cdots\geqslant w_n$;$d_i$ 是该两义原在义原层次体系中的最小深度,是一个正整数;A_{1i}表示第一个词语的第 i 个义原,两词的义原数分别为 m 和 n(假设 $m\leqslant n$),各义原对的相似度在知网中均已知,且任何义原与空值的相似度被定义为一个较小的常数。

基于知网的查询扩展具有一定的通用性,且由于其设有各种处理接口,具有良好的扩展性和二次开发性能[8]。但由于其并非免费开放接口,且义原覆盖所有领域,故专业性不强。本文提出 QuExT 模型,根据用户输入的名称,检索出相关文档并由用户自由设置相似度阈值。由于本文查询扩展的最终目的是尽可能找出语义相关的文档,且目标领域是计算机技术,故采用通过计算该领域本体的概念相似度来进行语义扩展。

1 基于本体的相似度计算研究现状

查询扩展的主要目的即是找出与指定概念相似度大于一定阈值的概念词集,并作为原概念的扩展概念予以传递给下一步处理程序。针对概念相似度计算,目前已有不少的相关研究[7-9]。主要有以下两种:基于概念之间的距离和基于概念所含的信息容量。概念距离的计算可分为两种:基于统计的方法和基于本体的方法。相对来说,基于统计的方法准确度相对要高些,但其需求的语料库较大,否则不符合统计原理,无说服力;另一方面,其对语料库质量的依赖较大,需和测试文档保持一致,否则容易造成错误判断;再者,前期处理阶段中基于全文(或文章片段)的分词和统计效率较低,而语料库需要定期更新,这将浪费较大的资源。而基于本体的相似度计算则更为方便,只需对本体进行定期维护,代价相对要低很多。目前,基于本体的相似度计算算法研究主要集中针对概念的上下文位关系,而忽略了其他有效关系。

2 基于本体技术的概念相似度计算算法修正

综合现有相似相关度计算的研究,这里对相似度的计算进行适当修正,最终的概念相似度计算算法如下:

鉴于两概念之间的关联度不仅与其概念之间定义的内在关系(包括公有属性等)有关,还与其在本体树中分布距离有关,也即由结构内和结构外两因素共同主导,前者简称为概念间相关度,后者叫做概念间的初始相似度。其计算算法可描述为:(1)输入:领域本体,两概念(已定义);(2)参数设置:相似_相关度阈值;(3)输出:两概念的最终相似_相关度。

2.1 初始相似度计算

假设:1) 若两概念的距离越大,则其之间的相似度反而越小;2) 当两组概念距离相同时,深度越大的组相似度越大;3) 两概念的深度差越大,相似度越小; 4) 若两概念间共同拥有的父节点个数越多,则其相似度愈越大。据此,可得出初始相似度的计算公式为:

$$sim(A,B)=\frac{\text{Min}(Depth(A),Depth(B))+1}{Distance(A,B)+\text{Min}(Depth(A),Depth(B))+1}*\frac{1}{|Depth(A)-Depth(B)|+1}*\frac{nodeList(A)\cap nodeList(B)}{nodeList(A)\cup nodeList(B)}$$

2.2 相关度计算

根据实际情况笔者做出以下假设:

1) 当两概念等价时,其相关度最大,设为 1,其他情况按照右边公式规定进行赋值;

$$Type(A,B)=\begin{cases}1,A,B\text{ 等价}\\0.9,A,B\text{ 继承}\\0.7,\text{其他关系}\\0.3,\text{无关联}\end{cases}$$

2) 两概念的属性重合度越大,即共同拥有的属性越多,相关度也越大。最后的相关度计算公式如下:

$$Rel(A,B)=Type(A,B)*\frac{Number(Attribute(A)\cap Attribute(B)}{Number(Attribute(A)\cup Attribute(B)}$$

概念间总关联度为两者的乘积:$Sim\_Rel(A,B)=Sim(A,B)*Rel(A,B)$

2.3 查询-概念相关度计算

由于查询并不只是单个词,更多的时候是句子或者词组,而且在查询语句中各自的权重不一样,因此,不能

统一地对所有词进行查询扩展再累加组合，这可能导致主题漂移现象，因此采用计算总关联度的办法来选取最终的扩展概念。

$$QCR(Q,C_i)=\sum_{k=1,\cdots,K} w_k * Sim\_Rel(q_K,C_i)$$

该公式计算查询向量与概念的关联度，其中 w_k 是查询词 q_K 所占的权重，可以采用多种方法进行确认，比较经典的有 Rochio 以及 Ide 方法等，这里只是简单地根据分词后得到的词性来进行权重标注。根据其大小，可以选择与查询最相关的 N 个概念，或者通过阈值来筛选。

2.4 计算总关联程度

计算最后查询词与文档的总关联程度，按其大小进行排列并返回用户查询相关文档。采用 Lucene 中的 TF-DIF 公式计算最后概念与文档的匹配程度，最后累计所有查询词的得分，即最终的查询-文档相似度，按大小进行排序，并返回相关检索文档。另外，进行全文匹配时，如果概念词在文档的标题、段首句、段尾句中出现，鉴于其特殊的主题代表性能，笔者可人为提升其权重值，从而提高最终的匹配程度。

主要实验伪码如下：

```
// concepts for similarity computing
Concept S1,S2;
//define a class
public class ontleaf {
String getUri();
String getLine();
// initialize the depth of the concept
int depth=-1;
…
}
int getDistance(ontleaf s,ontleaf t) {
int i = ontleaf (sharedParentNode). getDepth;
return (deepleaf. depth-i)+(shallowleaf. depth-i);
}
float getSim(ontleaf s,ontleaf t){
return (float)((shallowleaf. depth+1) * intersection) *
1/ ((distance+shallowleaf. depth+1) * (deepleaf. depth-shallowleaf. depth+1) * union);
}
float Rel(ontleaf s,ontleaf t)
{
…// ontology parser
float type= relationshipOfConcept(S1, S2);
Size=S. listDeclaredProperties();
attribute = (float)intersection/(double)(SizeOf S1+ sizeOf S2 - intersection);
return (type * attribute);
}
Return simi_rel (S1, S2) = Math. log(similarity (S1, S2) * relation (S1, S2));
```

3 查询扩展的实现

考虑到概念间相似度的计算直接影响扩展效果，而其受限于所依赖的本体，为了降低其影响力，笔者对扩展过程做适当调整，扩展算法流程伪码部分如下：

```
// suppose that most of instances have more than one label
If (query term is a class)
{
extendQuery = isClass(queryString);
queryString. append ( highSim _ Rel ( queryString). getClassName);
}
elseIf (query term is individual)
{
extendQuery = isInstance(queryString);
queryString. append ( lableNamesOf ( queryString) +ClassNameOf(queryString));
}
elseIf (query term is a label name)
{
extendQuery = isLableName(queryString);
queryString. append ( InstanceNameOf ( queryString) + otherLabelName);
}
```

由于最终笔者只需选取关联度为特定阈值内的概念，即对关联度的大小比较，为了避免人为确定系数的麻烦和误差，以上计算都略去各影响因素的权重因子，改用乘积替代，意在挑选相似度和相关度都相对较大的概念。为了放大相似度值之间的差异，笔者对原值进行取对数。最终查询扩展的筛选阈值可综合各区域值的比例以及召回率和召回精度来进行考虑。

当输入概念“office”时，根据上述思路和算法，系统能很好地返回该概念的相关概念词。这里，需要重点提出的是，如若没有事先界定为计算机领域，单纯“office”一词，很大可能返回的是工作室或者事务所之类的信息。

4 结语

本体技术运用于检索系统有利于挖掘网页文本之间的关联,在确保查询精度的条件下尽可能多地返回用户所需信息,而查询扩展成为其中的关键。为了更好地实现查询扩展,本文在已有概念相似度计算算法研究基础上进行改进和融合。鉴于语义检索受限于本体自身的质量,为了降低其对检索效果的影响,本文在查询扩展接口上做了改进,以达到将普通主题检索与语义检索的合并,在本体乏力时返回普通检索结果,一定程度上弥补了垂直检索系统发展的不足。本文最后成功实现了该扩展算法。随着本体技术的不断发展,其在搜索引擎领域的应用将愈加完善。

参考文献:

[1] Budanitsky A, Hirst G. Evaluating wordnetbased measures of lexical semantic relatedness[J]. Computational Linguistics, 2006, 32(1): 13-47.

[2] 蔡平,王志强,傅向华.基于语义的跨媒体信息检索技术研究[J].微电子学与计算机,2010,27(3):102-105.
Cai P, Wang Z Q, Fu X H. Research on the semantics-based cross-media information retrieval[J]. Microelectronics & Computer, 2010, 27(3): 102-105.

[3] 王刚.一种基于眼动轨迹的语义提取方法研究[J].重庆师范大学学报:自然科学版,2013,30(1):73-76.
Wang G. Study on method of semantic extraction based on eye tracking[J]. Journal of Chongqing Normal University: Natural Science, 2013, 30(1): 73-76.

[4] 赖文炜.一种改进的语义相似度计算模型[J].江西教育学院学报:综合,2012,33(6):53-56.
Lai W W. An improved semantic similarity calculation model[J]. Journal of Jiangxi Institute of Education: Comprehensive, 2012, 33(6): 53-56.

[5] 蒋溢,丁优,熊安萍,等.一种基于知网的词汇语义相似度改进计算方法[J].重庆邮电大学学报:自然科学版,2009,21(4):533-537.
Jiang Y, Ding Y, Xiong A P, et al. An improved computation method of word's semantic similarity based on HowNet[J]. Journal of Chongqing University of Posts and Telecommunications: Natural Science Edition, 2009, 21(4): 533-537.

[6] Wong A K Y, Ray P, Waran N P. Ontology mapping for the interoperability problem in network management[J]. IEEE Journal on Selected Areas in Communication, 2005, 23(10): 2058-2068.

[7] 罗俊丽,李慧娜,路凯. 基于词义消歧的语义查询扩展研究[J]. 微电子学与计算机,2012,29(1):71-75.
Luo J L, Li H N, Lu K. Semantic query expansion research based on word sence disambiguation[J]. Microelectronics & Computer, 2012, 29(1): 71-75.

[8] 何薇,徐伟华.信息检索的粗糙集方法[J].重庆理工大学学报:自然科学版,2010,24(9):84-88.
He W, Xu W H. Rough set approach to information retrieval[J]. Journal of Chongqing University of Technology: Natural Science, 2010, 24(9): 84-88.

[9] 朱鲲鹏,魏芳.基于用户日志挖掘的查询扩展方法[J]. 计算机应用与软件,2012,29(6):113-115.
Zhu K P, Wei F. A new query expansion method based on user logs mining[J]. Computer Applications and Software, 2012, 29(6): 113-115.

Research on the Semantic Query Expansion of Information Retrieval Technology

LI Xing-chun

(Teaching Department, Chongqing University of Arts and Sciences, Yongchuan Chongqing 402160, China)

Abstract: One of the most important reason to affect the information retrieval result is the phenomenon of semantic match of the user query and the file while syntactic mismatch. Because the semantic retrieval is limited by its quality, to decrease its influence on the retrieval result, through the analysis of the present situation of the semantic query expansion study, based on the improvement and integration of the algorithm study of the similarity of present concept, a new semantic query expansion of information retrieval technology was put forward, in which the integration was revised, and the mode $QCR(Q, C_i) = \sum_{k=1, \cdots, K} w_k * Sim\_Rel(q_K, C_i)$ was obtained and used as the evaluation method. In this method, by building the model and calculating the similarity of the concept similarity to determine the query word, it can search the related files and set up the threshold value by the user based on the query word, and the common subject retrieval and semantic query are combined, in which common retrieval results were returned when the noumenon couldn't give the results, which completed the development of vertical search system to some degree.

Key words: information retrieval; semantic similarity; query expansion

(责任编辑 游中胜)