基于语义的信息检索模型

陈 锐,张 蕾,胡艳华
CHEN Rui,ZHANG Lei,HU Yan-hua

西北大学 信息科学与技术学院,西安 710127
School of Information Science & Technology,Northwest University,Xi'an 710127,China
E-mail:vb_study@126.com

CHEN Rui,ZHANG Lei,HU Yan-hua.Model based on semantic information retrieval.Computer Engineering and Applications,2009,45(26):141-143.

Abstract: Words mismatch between queries words and documents lead to a number of related doucuments can not be successfully retrieved in information retrieval,which affects the effectiveness of retrieval results.This paper proposes and realizes a model based on the conceptual graphs and HowNet,and proposes a kind of relevance feedback algorithm,which achieves the similarity of new documents and query words with re-weighting word items by vector space model,and proposes semantic retrieval model.Experiments show that the new idea proves this method is effective.

Key words: information retrieval;similarity;vector space model;HowNet;relevance feedback

摘 要: 由于查询与文档中词语的不匹配现象导致一些相关的文档不能被成功地检索出来,在信息检索的研究与实现中,这是影响检索效果的一个很关键的问题。把概念图和知网结合起来,提出对应的相关反馈算法,重新计算词项权重,利用向量空间模型和语义相似度进行语义检索,并给出了语义检索模型。实验结果显示该方法取得了良好的效果。

关键词: 信息检索;相似度;向量空间模型;知网;相关反馈
DOI: 10.3778/j.issn.1002-8331.2009.26.041 **文章编号:** 1002-8331(2009)26-0141-03 **文献标识码:** A **中图分类号:** TP18

1 引言

目前的信息检索系统普遍存在检索精度和召回率不高的问题。信息检索系统在应用中,用户的真实信息需求到用户提交的查询请求之间,查询请求到系统对查询请求的理解之间存在一定的偏差。例如,在信息检索系统中,通常用户输入的查询中只是包含了几个关键词语,因此可能出现查询词语与文档集合中的词语不匹配,从而导致相关文档不能被检索出来[1]。

查询扩展[2]是提高信息检索性能的有效技术手段之一,能极大地改善系统性能,减少这些偏差对系统造成的影响。目前的查询扩展方法分为三类:基于语义知识词典的方法、全局分析方法和局部分析方法。全局分析方法对整个文档集合进行分析,计算查询词与文档中词语的相似度进行扩展;局部分析方法通过相关反馈之后,然后进行扩展;而基于词典的方法是语义层面的方法,主要利用语义资源直接扩展[2]。采用类似于相关反馈技术,利用第一次查找出来的前 50 篇文档提取相关的特征词语,重新计算权重,利用基于向量空间的模型,得到查询与文档相似度,从而达到提高召回率的目的,然后进行语义检索,提高检索的准确率。

针对目前信息检索技术的不足,提出基于语义的信息检索模型查询的方法。该方法以向量空间模型为基础,通过"相关反馈"技术,扩充相关概念;以概念图和知网为基础,通过知识表示把自然语言表示成计算机能识别的计算模型,以概念图的形式实现语义检索,充分提高查询过程中的召回率和准确率。

2 背景

2.1 概念图

概念图[3]是一种描述复杂对象结构的知识表示工具,其思想来源于 C.S.Pierce 的存在图和菲尔墨的语义网络,其理论建立在谓词逻辑上,能完全与自然语言相互翻译,表达出自然语言的语义。概念图一般由概念和关系组成,有线性表示法和图形表示法两种[3]。概念图中的概念结点和关系结点用弧连接,而概念和概念之间,关系和关系之间没有弧相连接。

2.2 知网

知网[4-5]是董振东教授 1998 年发布的一个知识资源,是一个以汉语和英语的词语所代表的概念为描述对象,以揭示概念与概念之间,以及概念所具有的属性之间的关系为基本内容,并能由计算机处理的网状的知识系统。知网认为,概念或词语是汉语最基本的语义和语法的单位,义原是最基本的、不易于再分割的意义的最小单位。知网使用了一种知识词典的描述语言(KDML)对所有概念进行了定义,从而使概念定义形式化,有效地保证了概念的语义复杂度和一致性高度统一。知网中的

作者简介: 陈锐(1979-),男,硕士研究生,研究方向:人工智能及自然语言理解;张蕾(1964-),女,博士,教授,硕士生导师,研究方向:人工智能及自然语言理解;胡艳华(1980-),女,硕士研究生,主要研究方向:无线网络、网格计算、人工智能。

收稿日期: 2008-05-29 **修回日期:** 2008-08-13

义原分类树把各个义原及它们之间的联系以树的形式组织在一起,这是进行语义相似度计算的基础[4-5]。

3 基于语义的信息检索模型框架

基于语义的信息检索模型如图1所示。

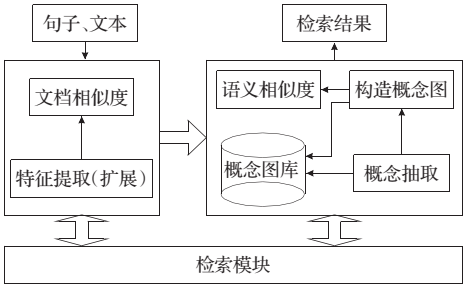

图1 基于语义的检索模型框架

(1)特征提取模块:利用概念特征提取对文档中的特征词进行提取,根据知网重新计算词项权重;

(2)文档相似度模块:根据特征提取之后得到的权重,利用向量空间模型计算文档相似度,得到第二次检索结果;

(3)构造概念图模块:根据第二次检索的结果,抽取概念,生成概念图,计算概念图相似度,得到最终查询结果。

4 语义相似度计算

我国的李彬等提出了基于语义依存的汉语句子相似度计算,但是这种方法只是考虑到了概念关联,没有考虑到概念间关系的关联。通常情况下,理解自然语言往往把字组成词,然后把词连接成句子,只有这样才能理解整个句子的含义,词语才能体现出其实际意义。因此,进行自然语言理解不能孤立地分析词语,而应该把它们联系起来考虑。采用概念图和知网相结合的方法进行语义相似度的计算。利用文献[6]先计算概念图中词语的相似度,然后根据公式(1)计算这个概念图的相似度。

4.1 概念图的相似度

概念图相似度[6-7]运算是一个递归的定义,其运算如公式(1)所示[6-7]。

$$SoG(c_Q,c_R)=w(c_Q,c)*sim_c(c_Q,c_R)+\max_{foreverycombination}\left\{\sum_i w(c_Q,i)*sim_r(r_Q^i,r_R^i)*\left[SoG(c_Q^{r_Q^i},c_R^{r_R^i})\right]\right\}$$

$$w(c_Q,c)+\sum_j w(c_Q,i)=1 \tag{1}$$

这里,sim_c是概念相似度,sim_r是关系相似度。$c_Q^{r_Q^i}$和$c_R^{r_R^i}$分别是对应子图的入口点,r_Q^i和r_R^i分别是查询图和资源图中的第i条关系。$w(c_Q,c)$和$w(c_Q,i)$分别是对应入口的权重和与对应入口相关的第i条关系的权重,c_Q和c_R分别表示查询图和资源图的入口点。

4.2 词语相似度

在《知网》中,并不是将每一个概念(词语)对应于一个树状概念层次体系中的一个结点,而是通过用一系列的义原来描述一个概念,概念的相似度计算公式(2)所示[8]。

$$Sim(C_1,C_2)=\max_{i=1..n,j=1..m} Sim(S_{1i},S_{2j}) \tag{2}$$

其中,$S_{11},S_{12},\cdots,S_{1n}$是$C_1$的$n$个义项(概念),$S_{21},S_{22},\cdots,S_{2m}$是$C_2$的$m$个义项(概念)。这样,就把两个概念之间的相似度问题归结到了义原相似度计算。

5 该文算法

5.1 文档的概念提取

在概念提取中,通过计算TF*IDF。TF是概念的频度(term frequency),即概念在一篇文章中出现的次数,IDF为文档频度的倒数(inverse document frequency),概念的文档频度是指所有文章中包含该概念的文章数目,如公式(3)所示[9-10]。

$$w_{ik}=\frac{TF_{ik}\cdot\log(\frac{N}{n_k})}{\sqrt{\sum_j^i (TF_{ij})^2}\cdot(\log(\frac{N}{n_k}))^2} \tag{3}$$

其中,n_k为给定用户所有浏览网页中包含概念T的文档数,t为所有概念的总数目,N是所有的文档数目,w_{ik}是当第k个概念T_k相对于特定用户在第i个文章D_i的TF*IDF权值,TF_{ik}为T_k在D_i中出现的频度。

5.2 扩展算法

通过利用Giannis对查询关键词的项频度和对文档项频度重新计算,然后利用向量空间模型计算文档和向量的相似度。其词语的项频度计算如公式(4)所示[11-12]。

$$tf_{ij}=tf_{ij}+\sum_{sim(i,q)\geq\delta}^{i\neq q} ksim(i,q) \tag{4}$$

其中,tf_{ij}是词项i在文档j中的频度,k是特征词的个数,δ是用户定义的阈值(δ=0.85)。通过计算特征词的相似度,把相似度大于阈值的词语进行扩展,然后归一化处理。

然后根据文献[10]的改进权重计算公式计算词项的权重,如公式(5)所示[10]。

$$W_{ij}=(0.5+0.5*\frac{tf_{ij}}{\max tf_i})*\log(\frac{n_i}{len_j}) \tag{5}$$

其中,tf_{ij}是在文档d_j中查询词t_i的出现频率,$\max tf_i$为查询词的最高出现频率,n_i是查询词t_i出现的总频度,len_j是文档中出现的所有查询词数[10]。

算法描述:

```
void QueryExpansion()
{ j=1;
 While(j≤N)
 {
  //w_{i,q}为第i个查询关键词词条对应的原始权重
  计算文档j中所有特征词的权重w_{i,q};
  w=w_{i,q};
  j=j+1;
 }
 for(i=1;i≤N;i++)
 {
  根据公式(4)提取特征词c_i;
  for(j=1;j≤N;j++)
  { if(sim(c_i,q≥δ)&& c_i∈c_Q)   //c_i是旧的特征词且大于阈值
     w=w+ksim(c_i,q);   //计算文档j的第i个特征词c_i的权重
   else if(sim(c_i,q≥δ)&& c_i∉c_Q)   //c_i是新的特征词且大于阈值
     w=ksim(c_i,q);
  }
```

}

归一化处理并计算查询与文档的相似度；

}

其中：用户查询 q，文档 d_j，文档总数 N，β_1，β_2 根据具体情况进行调节。通过利用知网进行相似度计算，得到与特征词意义相近的词语，如果相似度大于阈值 δ，则大于进行扩展。

5.3　句法结构依赖强度

针对句法分析结果进行检验并处理，主要依据来源于文献[14]和李海军提出的汉语短语结构消歧方法，这里认为这种汉语短语结构的消歧处理主要理论依据是汉语本身的句法结构之间的相互依赖强度，句子中语法成分不同其依赖强度是不一样的。虽然以上短语消歧方法具有一定的效果，但是这里却认为，语法成分之间的内在关系的依赖强度不能仅仅依据相似度衡量，而最贴切的是依赖于相关度来衡量，相似度是衡量词语之间的相近程度，而相关度是衡量词语之间的关联程度、粘合程度。例如，“漂亮的中学语文老师”，通过知网计算“中学”与“语文”、“语文”与“老师”、“中学”与“老师”的相关程度是不一样的，显然这三者都是相关的，但是要判断“中学”是修饰“语文”还是修饰“老师”，需要通过它们之间的依赖强度判断，看它们之间的粘合程度哪个比较大，以此来进行句法结构的有效性检验。采用相关度计算公式来衡量句法结构依赖强度。w_1 和 w_2 为句法结构成分，其相关度计算如公式(6)所示[13-14]。

$$Rel(w_1,w_2)=P_1*R_1+P_2*(R_2+R_3)+P_3*R_4 \quad (6)$$

其中，R_1，R_2，R_3，R_4 表示义原的相同、关联、同位、包含关系。

6　实验与结果分析

采用准确率和召回率在 webinfo 实验系统对基于语义的信息检索模型进行有效性分析，与查询相关的一组文档记为{*Relevant*}，被系统检索出的一组文档记为{*Retrieved*}，既相关又被检索出的一组文档记为{*Relevant*}∩{*Retrieved*}[9]。实验将二者结合起来作为实验结果的评估方式，实验数据见表 1。

表 1　基于几种不同的检索模型的实验结果

搜索技术	准确率/(%)	召回率/(%)
基于向量空间的检索	73.59	74.25
基于语义的检索模型	75.72	75.53

它们的形式定义[9]如下：

$$precision=\frac{|relevant\cap retrieved|}{|\{retrieved\}|} \quad (7)$$

$$recall=\frac{|relevant\cap retrieved|}{|\{relevant\}|} \quad (8)$$

召回率和准确率的关系基本是成反比关系，其效果通过图 2 反映出来。

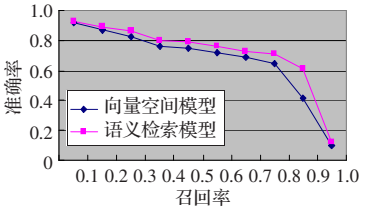

图 2　基于语义的信息检索模型和空间向量模型的比较

实验显示基于语义的信息检索查询比基于向量空间的检索有较好的准确率，并模拟了召回率和准确率的关系(图 2)，这主要是该文的方法在基于语义的相似度计算方法取得了一定的效果。在实验中选取了第一次检索出的前 50 个文档，从中抽取了相近词语，并进行重新计算权重，得到文档相似度，重新查询，查询结果表明采用“相关反馈”的查询扩展的召回率比基于向量空间的方法的召回率明显要高。但是系统在查询时，由于计算量比较大，造成检索速度较慢。

7　结束语

提出了基于语义的信息检索模型的研究，这是在语义层次上的检索的尝试，同时利用“相关反馈”技术提取概念，进行扩展，提高了召回率，并用概念图和知网进行结合的方法，计算语义相似度，进行语义匹配，实验表明该方法取得了一定的效果。

参考文献：

[1] 高珊，何婷婷.信息检索中的查询扩展相关技术研究[D].武汉：华中师范大学，2008.

[2] 王树梅，吴慧中.信息检索相关技术研究[D].南京：南京理工大学，2007.

[3] 张蕾，李学良.概念结构及其应用[D].西安：西北工业大学，2001-05.

[4] 周舫，郑逢斌.汉语句子相似度计算方法及其应用的研究[D].郑州：河南大学，2005-05.

[5] 董震东，董强.知网的理论发现[J].中文信息学报，2007(7).

[6] Zhong Ji-wei，Zhu Hai-ping.Conceptual graph matching for semantic search supported by IBM China research laboratory[C]//Stuckenschmidt H.18th International Joint Conference on Artificial Intelligence，Ontologies and Distributed Systems，IJACI 2006，Acapulco，Mexico，2006：53-59.

[7] 朱海平，俞勇.基于概念图匹配的语义搜索[D].上海：上海交通大学，2006-10.

[8] 刘群，李素建.基于《知网》的词汇语义相似度的计算[C]//第三届汉语词汇语义学研讨会，2002.

[9] 牟力科，张蕾，张晓孪.基于概念图的用户兴趣查询扩展模型的研究[J].计算机工程与应用，2008，44(6)：184-186.

[10] 王秀娟，郭军.文本检索中若干问题研究[D].北京：北京邮电大学，2005-05.

[11] Varelas G，Voutsakis E，Raftopoulou P.Semantic similarity methods in wordnet and their application to information retrieval on the Web[C]//Proceedings of the 7th Annual ACM International Workshop on Web Information and Data Management，Bremen，Germany，2005：10-16.

[12] Natsev A P，Haubold A.Semantic concept-based query expansion and re-ranking for multimedia retrieval[C]//Proceedings of the 15th International Conference on Multimedia，Augsburg，Germany，2007：991-1000.

[13] 郑旭玲，李堂秋，杨晓峰，等.基于语义规则的汉语短语结构分析排歧初探[C]//全国第六届计算语言学联合学术会议论文集，2001：219-226.

[14] 闫蓉，张蕾.基于语义的汉语词义消歧方法研究[D].西安：西北大学，2006-05.

[15] 胡珍新，丁恽，王明文.面向用户的查询扩展研究与实现[D].南昌：江西师范大学，2004-07.