基于概念图的用户兴趣查询扩展模型的研究

牟力科,张 蕾,张晓孪
MOU Li-ke,ZHANG Lei,ZHANG Xiao-luan

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

MOU Li-ke,ZHANG Lei,ZHANG Xiao-luan.Query expansion model of users' profile based on conceptual graphs. Computer Engineering and Applications,2008,44(6):184-186.

Abstract: Query expansion is an important part of information retrieval systems.Neglection of user' profile influent the precision and recall of information retrieval in unity model,the reason resulting in the problem is analyzed and the idea of a query expansion model of user' profile based on conceptual graphs.Experiments show that the new idea can improve the precision and recall of information retrieval effectively compared with convention ones.

Key words: query expansion;conceptual graphs;profile

摘 要: 查询扩展是信息检索技术研究的一个重要组成部分。目前的查询扩展是基于统一的用户模型,没有考虑到用户的个人兴趣,这对查询扩展的精确度造成了一定的影响。分析了产生这种问题的原因,提出了基于概念图的用户兴趣扩展模型,通过该模型来有效提高查询扩展的精确度。实验显示,该方法能有效提高查询的查全率和查准率。

关键词: 查询扩展;概念图;用户兴趣
文章编号: 1002-8331(2008)06-0184-03 **文献标识码:** A **中图分类号:** TP391

1 引言

信息已成为人们生活中的重要资源,以搜索引擎为代表的信息检索系统是人们从网上获取信息的主要工具。随着网上信息量的急剧增加,人们对信息的获取方式、获取技术提出了更高的要求。个性化的查询扩展是实现这种要求的重要组成部分。

以往的查询扩展方法主要包括全局分析方法和局部分析方法[1]。全局分析方法的基本思想是对全部文档中的词或词组进行相关性分析,计算每对词或词组间的关联程度。当一个新的查询到来时,则根据预先计算好的词间相关关系,将与查询用词关联程度最高的词及词组加入原查询以生成新的查询。这种方法可以最大限度地探求词间关系,并在词间关系词典建立之后以较高的效率进行查询扩展。但是,当文档集合非常大时,建立全局的词间关系词典在时间和空间上往往是不可行的,并且在文档集合改变后的更新代价巨大。局部分析的思想是将初次查询的前 N 篇文章认为是相关文章, 并以此为依据对查询扩展进行扩展。这种方法在目前的应用最为广泛,并在一些实际的信息检索系统中得以使用。但是,当初次查询后排在前面的文档与原查询相关度不大时,局部分析会把大量无关的词加入查询,从而严重降低查询精度,甚至低于不做扩展优化的情形。

针对上述查询扩展方法存在的不足,本文提出了基于概念图[3]的用户兴趣查询扩展方法,该方法以概念图为基础,以用户需求为中心,做基于概念图的个人兴趣查询扩展研究。并结合用户的主观调节,充分提高了查全率以及查准率。

2 基于概念图的用户兴趣查询扩展框架

2.1 基本思想

(1)用户向信息检索系统提交代表其信息需求的查询式,系统对需求查询式进行分析,并进行概念提取,形成预扩展的概念。

(2)对预扩展的概念在知网[4]中进行相近词,同义词的扩充。同时结合概念图库形成基于概念图的个人兴趣查询扩展。

(3)过程(1)、(2)可以循环进行,允许用户手动扩展以获得精确的查询,并得到较为满意的检索结果。

(4)在用户的访问信息和查询信息中收集个人兴趣信息。在个人兴趣信息中抽取相关概念,从而形成概念库,在此基础上结合知网建立概念图库。

2.2 模型

模型的核心在于对用户个性化兴趣信息的获取和基于概念图的个性化的查询式的扩展上,见图 1。模型分为以下 4 模块:

(1)兴趣收集模块:对用户的查询历史进行整理并且对用户的浏览页面计算页面兴趣度,按照页面兴趣度对浏览过的页面排队。在此基础上对用户感兴趣的概念提取出来,建立概念库。

(2)建立概念图库:用上面建立的概念库和知网建立概念图库。

(3)个性化扩展模块:用户将查询式提交检索模块之前,首先对查询式进行概念提取,形成原始概念。其次应用概念图库进行个人兴趣度的概念图语义相似度的计算,找到相似度近的

作者简介: 牟力科,男,硕士研究生,研究方向:人工智能及自然语言理解;张蕾,女,硕士生导师,博士,教授,研究方向:人工智能及自然语言理解;张晓孪,女,硕士研究生,研究方向:人工智能及自然语言理解。

收稿日期: 2007-06-13 **修回日期:** 2007-08-29

概念进行扩展,形成预扩展概念。最后在知网中对预扩展概念查询同义词和近义词进行扩展,形成新的查询式。

(4)检索模块:采用某一检索模型或搜索引擎方式进行信息检索。

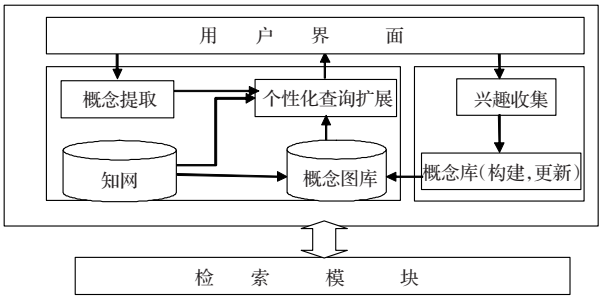

图 1　基于概念图的用户兴趣查询扩展模型

3　查询扩展方法

以基于概念图的用户兴趣查询扩展为基础的查询扩展方法,将概念库中的概念作为扩展词,并且充分考虑用户的个性化兴趣[6]以提高扩展的精确度,具体从以下两个方面对查询扩展方法进行介绍。

3.1　用户兴趣的获取

3.1.1　用户访问信息的收集

在基于概念图的用户兴趣查询扩展的体系结构中,用户查询历史是用户兴趣挖掘的重要依据。用户在信息检索的过程中会产生下面 4 种信息:

(1)用户输入信息检索的关键词。

(2)用户的浏览行为,包括用户在每个页面上的停留时间,对每个页面进行操作(如选择喜欢的页面进行的保存、打印等操作)等。

(3)服务器 Web 日志,用户对系统的访问行为会被服务器记录下来,形成服务器 Web 日志,包括访问时间、访问的页面、页面的大小等信息。

(4)用户下载、保存的页面和其它信息等。

3.1.2　用户页面兴趣度

用户页面兴趣度是用户建模和个性化扩展中的一个重要信息来源。页面兴趣度[1]是指一个用户对某个页面感兴趣的程度,可以用公式(1)计算后获得。

$$I_p=F(t)*G(h) \tag{1}$$

其中 $F(t)$是指用户在页面 p 上的停留时间 t 的函数,$G(t)$是用户对页面喜好程度的函数。

$$F(t)=\eta\times\ln\frac{z(t)}{T_1} \tag{2}$$

$$z(t)=\begin{cases}1 & t<T_1\\ t & T_1\leqslant t\leqslant T_2\\ 300 & t>T_2\end{cases} \tag{3}$$

$Z(t)$是一个去噪函数,把过大和过小的 t 去掉,使参数居于 1~300 之间。一般来说,对于只浏览标题的网页,$t<T_1$ 秒,这说明用户对此页面不感兴趣。浏览标题、关键词和摘要的,t 一般在 T_2 秒以内。浏览全文的网页,$t>T_2$ 秒。

$$G(t)=\begin{cases}t^2 & t>T_1\\ t & T_1>t>T_2\\ \dfrac{\mu}{t} & t<T_2\end{cases} \tag{4}$$

其中 $G(t)$依据用户在该页面的停留时间求得。

3.1.3　概念的获取

对收集的兴趣度大于阈值的页面,进行概念的抽取。因为这部分信息表明了用户的兴趣,是基于概念图的个性化[7]兴趣查询扩展的关键。

在概念提取中用到了 $TF\cdot IDF$。TF 指概念的频度(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。

$$w_{ik}=\frac{TF_{ik}\cdot\log(\dfrac{N}{n_k})}{\sqrt{\sum_j^t(TF_{ij})^2\cdot(\log(\dfrac{N}{n_k}))^2}} \tag{5}$$

其中,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 $TF\cdot IDF$ 权值,TF_{ik} 为 T_k 相在 D_i 中出现的频度,N 是所有的文档数目,t 为所有概念的总数目,n_k 为给特定用户所有浏览网页中包含有概念 T 的文档数。对网页中的概念提取出来形成概念图。

3.2　概念图的个性化兴趣查询扩展

概念图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,该理论建立在谓词逻辑上,能完全与自然语言相互翻译,表示出自然语言的语义。概念图由概念和关系组成,常用的表示方式有图形方式和线性方式。为了能够准确表达用户感兴趣的概念,在概念图的基础上考虑到个人兴趣,提出并建立了概念图的个性化兴趣查询扩展。相关定义如下。

定义 1　如果参数 δ 和 γ 表示概念图中的特化(specialization)和泛化(generalization)[4],那么对具有上下位关系的具体概念 x,y 来说,当 $I(t,c)>0$ 时,参数 $\delta_{t\to c}$ 和 $\gamma_{t\to c}$ 的确定如下:

$$\delta_{x\to y}=\mu\log\frac{p(t/c)}{p(t)} \tag{6}$$

$$\gamma_{x\to y}=\beta\log\frac{p(c/t)}{p(c)} \tag{7}$$

其中

$$I(t,c)\approx\log\frac{p(t,c)}{p(t)p(c)} \tag{8}$$

式中 $p(t,c)$为概念 t 和概念 c 同时出现在紧密相连的一组(在这里用 5 个紧密相连的概念为一组)中时(或查询词组)的概率,$p(t)$为概念 t 出现而概念 c 不出现在同一组的概率,$p(c)$为概念 c 出现而概念 t 不出现在同一组的概率,$p(t/c)$和 $p(c/t)$为条件概率 μ 和 β 为可调整参数。

定义 2　设有概念 T,则它的扩展为:

$$T+=1/T+sim(T,T_1)/T_1+\cdots+sim(T,T_n)/T_n \tag{9}$$

式(9)的+不是表示相加与求和,$sim(T,T_n)/T_n$ 也不是分数,它们只是表示概念 T 扩展而引用的符号。公式中 $sim(T,T_n)$表示概念 T 和 T_n 概念的相似度,其中的相关定义与计算参考文献[5]。在扩展时找出与 T 的相似度大的概念进行扩展。

例 1　如图 2 所示概念图,概念分别为中文信息处理、文本分类和贝叶斯方法,试表示概念中文信息处理和文本分类的扩展。

则依据式(9)扩展为如下的表示:

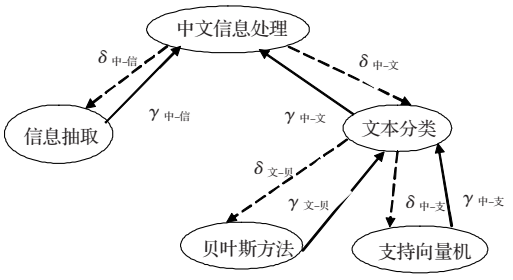

图 2 基于概念图的用户兴趣查询扩展

$$文本分类+=\frac{1}{文本分类}+\frac{\gamma_{中-文}}{中文信息处理}+\frac{\delta_{文-贝}}{贝叶斯方法}+\frac{\delta_{中-支}}{支持向量机}$$

$$中文信息处理+=\frac{1}{中文信息处理}+\frac{\delta_{中-信}}{信息抽取}+\frac{\delta_{中-文}}{文本分类}+\frac{\delta_{中-文}\times\delta_{中-贝}}{贝叶斯方法}+\frac{\delta_{中-文}\times\delta_{中-S}}{Svm}$$

3.3 算法描述

```
CGraphs Rqmodel(CGraphs&CGq,CGraphs CGr[])
{
    //有向图 CGq 表示从查询式中获得的信息
    //有向图 CGr 表示的是概念图数组中所有的弧的方向都向上的概念图
        Rvalue=Find(CGraphs&CGq,CGraphsCGr[]);
        If(Rvalue 是有效值)
          If(Similarity(CGq,Rvalue)>K)  revise Query;
        //K 为常数,表示相似度的阈值
        //函数 Similarity(CGraphs&CGq,CGraphs&CGr)判断两个概念之间的相似度
        Return CGq;}
CGraphs Find(CGraphs&CGa,CGraphs CGb[])
{
    //该模块查找图库中与查询信息匹配的概念图
    For(i=0;i≤n;i++)
    If (CGb[i]的值和 CGb[i]的值相等) return CGb[i];
}
```

4 实验与结果分析

实验数据(见表 1)取自己的访问 log 文件,总的数据量是两个月的访问记录,由于受到数据量的限制,设定的参数值均不高。经过词性标注后去掉一些特殊词性的词和低频词,抽取概念建立了概念库。在此基础上在扩展系统上进行实验,以“文本分类”、“信息抽取”和“切分词”三个关键词为例进行扩展。在扩展中取扩展后的前三个扩展概念作为最终扩展并最终形成查询信息进行检索。在 Google 和百度两个搜索引擎上用准确率和召回率对基于概念图的用户兴趣扩展进行检验。与查询相关的一组文档记为{Relevant},被系统检索出来的文档记为{Retreved},即相关有被检索出的一组文档记为 {Relevant}∩{Retreved}。

$$precision=\frac{|\{relevant\cap retrieved\}|}{|relevant|} \tag{10}$$

$$Recall=\frac{|relevant\cap retrieved|}{|retrieved|} \tag{11}$$

表 1 基于几种不同的搜索引擎技术的实验结果

搜索技术	准确率/%	召回率/%
基于统计的搜索	73.59	74.25
基于个人兴趣概念图的搜索	75.38	76.58

实验显示基于概念图的用户兴趣查询扩展当用户查询相对集中时取得的查询结果比局部扩展有较好的准确度。这主要是因为本文的方法在消除查询词的歧义性方面取得了一定的效果。在实验中选取具有明显的歧义词,查询结果表明采用了基于概念图的用户兴趣查询扩展后的精确度比局部扩展的精确度明显要高。但当浏览信息涉及领域广泛且兴趣差别不大时,本系统的扩展查询精确度结果回到局部扩展查询的精确度甚至在某些情况下低于局部扩展查询,经分析认为造成这种情况的主要原因是当浏览信息不集中时对扩展结果有造成影响。

5 结束语

文章提出了基于概念图的用户兴趣扩展模型的研究,是在语义层次上考虑到用户个人兴趣扩展的尝试,同时用概念图的方法描述了用户感兴趣的概念之间的语义信息,在查询扩展方面取得了一定得效果。然而当对词典中从未出现的数据进行搜索时,由于不存在与查询词相关的概念图,因此扩展结果的受到了影响,这是今后继续努力的方向。

参考文献:

[1] 张选平,蒋宇.一种基于概念的信息检索查询扩展[J].微电子学与计算机,2006,4(23):110-114.

[2] Vries P H de.Representation of scientific texts in knowledge graphs[D].Rijksuniversiteit Groningen,Groningen,The Netherlands,1989.

[3] Sowa J F.Conceptual structures:information processing in mind and machine[M].UK:Addison-Welsley,1984.

[4] 董振东,董强.知网—知网简介[EB/OL].http://www.keenage.com.

[5] 殷亚玲,张蕾,李海军.基于概念图的相关反馈技术研究[J].计算机工程与应用,2006,42(3):164-167.

[6] 于戈.面向智能信息检索的 Web 挖掘关键技术研究[D].沈阳:东北大学,2006.

[7] Zeng C,Xing C X,Zhou L Z.A survey of personalization[J].Journal of Software,2002,13(10):1952-1961.