文章编号:1001 - 9081(2009)06 - 1575 - 03

# 基于关键词语义扩展的检索策略

杜金洋[1,3],易 禾[2],杨 春[1,3]
(1. 四川师范大学 计算机科学学院,成都 610101; 2. 四川师范大学 计算机信息网络中心,成都 610066;
3. 四川师范大学 可视化计算与虚拟现实四川省重点实验室,成都 610066)
(dujinyang@gmail.com; ycycyc@263.net)

**摘 要**:传统基于关键字匹配检索策略由于关键字的机械字符匹配和一词多义问题很容易造成漏检和错检。为此,从领域本体概念入手,结合关键字检索的特点,提出一种基于关键词语义扩展的检索策略。策略使用领域本体描述的语义结构扩展关键字匹配范围,避免完全机械的字符匹配造成的漏检,从而提高检索的查全率。在此基础上,利用领域本体中语义相关度过滤检索无关结果,以提高检索的查准率,并根据检索结果与本体语义相关度算法排序。

**关键词**:领域本体;扩展;查全率;查准率
**中图分类号**: TP391.3 **文献标志码**:A

# Retrieval strategy based on semantic expansion of keywords

DU Jin-yang[1,3], YI He[2], YANG Chun[1,3]
(1. *School of Computer Science, Sichuan Normal University, Chengdu Sichuan* 610101, *China*;
2. *Computer Network Information Center, Sichuan Normal University, Chengdu Sichuan* 610066, *China*;
3. *Sichuan Key Laboratory of Visualization Computing and Virtual Reality, Sichuan Normal University, Chengdu Sichuan* 610066, *China*)

**Abstract:** Traditional retrieval strategy based on keyword-matching is facing unavoidable bottleneck because of mechanical characters matching and polysemy problems. A retrieval strategy based on domain ontology and semantic expansion of keywords was proposed. On the one hand, the domain ontology concept structure was used to expand the retrieval area; on the other hand, a method was presented to filter the irrelevant results. In this way, both the recall ratio and the precision ratio could be improved simultaneously. After that, an algorithm of sorting the retrieval results by the similarity between ontology and result was given.

**Key words:** domain ontology; expansion; recall ratio; precision ratio

## 0 引言

目前最常用的文本检索主要是基于关键字匹配的方法。尽管这种检索简单且易学易用,但是由于关键字机械的字符匹配方式和一词多义,关键词语言难以反映词间相关关系等原因,检索噪声大,采用简单的关键词检索方法容易造成检索结果过多,查全率和查准率都无法满足用户的需求[1],往往需要用户经过多次过滤和重复检索才能得到需要的检索结果。

在信息检索领域,检索扩展早已经发展成为一个研究方向,其基本思想是利用关键词的相关词对检索进行修正,以找到相关文档,提高查全率。如文献[2]中利用概念树的层次形式对查询语义进行了扩展,检索范围扩大到父子概念,有效提高了检索的查全率,但文中使用的方法和实验都表现其对查准率的提高作用不大。针对这一问题,本文给出一种基于关键字语义的检索策略,目的是在检索扩展中提高查全率同时保证查准率。

## 1 检索质量标准分析

查全率和查准率是衡量检索质量的重要标准。

**定义 1** 查全率。

$$\text{查全率} = \frac{\text{命中相关记录数}}{\text{数据库中全部相关记录数}} \times 100\%$$

**定义 2** 查准率。

$$\text{查准率} = \frac{\text{命中相关记录数}}{\text{命中全部记录数}} \times 100\%$$

设 $R$ 为查全率,$P$ 为查准率,$a$ 为命中相关记录数,$N$ 为数据库中全部相关记录数,$x$ 为命中全部记录数。则:

$$R = a/N \tag{1}$$

$$P = a/x \tag{2}$$

关键字匹配将检索范围局限于一个较小的范围,将同义词匹配和相关词匹配排除在外。因此如果检索关键词的同义词和相关词必定会将检索范围相对扩大。相当于在式(1)中,$N$ 不变,如果 $a$ 增大,必然会使 $R$ 增大。然而事实上并非这么简单,因为单纯的检索关键词的同义词和相关词,增大的只是式(2)中的 $x$;

如果 $x$ 增大的同时,$a$ 同时增大会有以下情况发生:

1) $a$ 增大,$P$ 增大。

2) $a$ 增大,$P$ 不变。

3) $a$ 增大,$P$ 减小。

结果取决于 $a$ 与 $x$ 增加速度之比。由文献[3]中关于查全率和查准率顺变关系的研究结论可知,当由于检索策略的变化,使得检索到的相关记录的变化量与全部命中记录的变化量之比大于相关记录数与命中记录数之比时,查全率与查准率呈现顺变关系。

为了提高检索质量,必须从检索策略入手,且同时兼顾查全率和查准率的改变。即是说,一方面要扩大检索范围,避免漏检,从而提高查全率 $R$;另一方面又要保证检索精度,尽量

**收稿日期**:2008 - 12 - 22;**修回日期**:2009 - 03 - 01。 **基金项目**:国家 973 计划项目(2008CB317110)。

**作者简介**:杜金洋(1982 - ),男,四川广元人,硕士研究生,主要研究方向:网络数据库、知识工程; 易禾(1968 - ),男,四川成都人,实验师,主要研究方向:计算机网络; 杨春(1970 - ),男,四川绵阳人,教授,博士,主要研究方向:分布式计算、并行计算。

避免误检以提高查准率 $P$。

## 2 基于关键词语义扩展检索策略

本文提出的基于关键词语义扩展检索如图1所示。

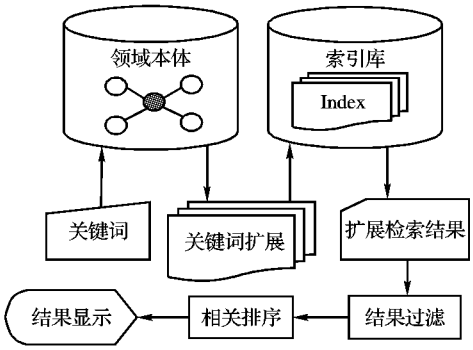


图1 基于关键词语义扩展的检索模型

基本思想：

1)用领域本体对检索关键词语义进行扩展，扩展匹配范围包括关键词及其相关词，以提高检索查全率。

2)对扩展概念检索得到的结果，利用过滤原则以及相关度排除无关数据，降低误检率，提高查准率。

3)计算检索结果与关键词语义相关度，对结果排序。

### 2.1 语义扩展策略

**定义3** 领域本体。领域本体是用于描述指定领域知识的一种专门本体，它给出概念间的相互关系、领域活动，以及该领域所具有的特性和规律的一种形式化描述。[4]

领域本体是由领域内对象、属性、关系和实例组成的。本体中要素间的关系可以用概念图 $O(C,R)$ 来进行表示，其中 $C$ 表示领域的概念集，它包括领域内的类，属性及实例；$R$ 表示领域内概念间关系的集合，并且 $R \subseteq C \times C \times W$，其中 $W$ 表示概念对象 $C$ 之间的相关度或语义距离。

**定义4** 概念语义相关度。领域本体中概念之间语义相关程度，对应于关系权值 $W$。

利用领域本体对关键词语义扩展，在领域本体 $O(C,R)$ 中 $R$ 包含了多种关系，如同义关系(same-as)，上下位关系(subclass-of)，整体与部分关系(part-of)，实例与概念关系(instance-of)，同类关系(same-kind-with)等。通过这些关系的扩展，每一个概念 $C_i$ 被扩展为以 $C_i$ 为中心的星形树，且 $C_i$ 与其邻接节点 $C_j$ 分别对应关系 $R_{i,j}$，即是说 $R_{i,j}$ 代表节点 $C_i$ 与 $C_j$ 之间的语义距离，对应于定义4中的概念语义相关度。

扩展检索思路：以关键词为中心，同时检索关键词邻接关系词，检索结果扩展为由关键词与其邻接关系词共同决定。由于不同邻接关系表示概念间的语义相关度，不同相关词对检索相关度的贡献也不一样。

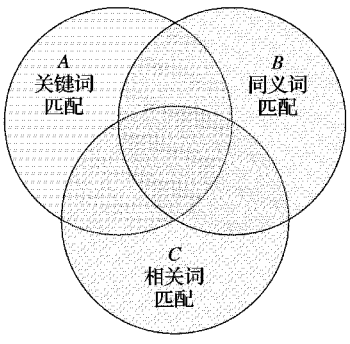


图2 关键词扩展匹配概念图

按照贡献值差异将检索对象分成三项，即关键词匹配，同义词匹配和相关词匹配。如图2所示，图中关键词匹配、同义词匹配、相关词匹配分别由 $A$、$B$、$C$ 三个圆圈表示，它们的公共部分表示满足多种匹配结果的交叠。

显然交叠部分与需要结果的匹配度更高，然而并非所有交叠部分都是我们需要的结果，例如一篇主题是关键词相关词概念的文章，文中与相关词概念的相关度会远大于关键词，因此需要对扩展检索的结果进一步过滤。

### 2.2 过滤策略

检索结果相关度与关键词匹配、同义词匹配、相关词匹配之间有如下关系：

关键词匹配 > 同义词匹配 > 相关词匹配

$w_{\text{keyword}} > w_{\text{synonyme}} > w_{\text{related}}$

从一般检索得出的经验如图3所示。

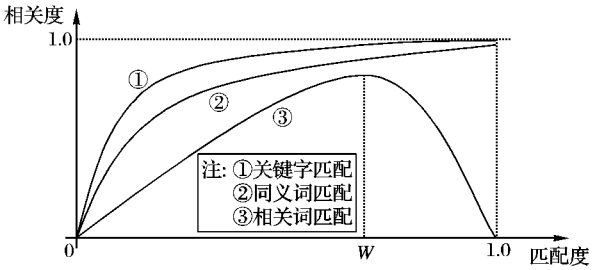


图3 匹配度与内容相关度变化图

经验1 与关键词匹配度越高，与需要结果相关度越高。

经验2 与关键词的同义词匹配度越高，与需要结果相关度越高。

经验3 在 $[0,W]$ 上，与关键词的相关词匹配度越高，与需要结果相关度越高；在 $[W,1]$ 上，与关键词的相关词匹配度越高，与需要结果相关度越低。其中 $W$ 即定义的语义关系权值。

因此检索的重点仍然是关键词匹配以及关键词的同义词的匹配，在此基础上，利用相关词的匹配度来衡量所检索关键词与领域的相关度，避免与相关词匹配相关度高于与关键字匹配相关度。以此作为结果过滤标准，以降低出现关键词而并非该领域相关内容被检索出的几率，即降低误检率。提高检索采样的精度，从而相对地提高查准率。

对检索结果进行过滤，规则如下：

1)只有发生关键词匹配或同义词匹配两事件之一，才可能与结果相关。

2)同时发生关键词匹配和同义词匹配事件，与需要结果必然相关。

3)同时满足关键词匹配与相关词匹配，并且关键词出现次数与任何相关词出现次数之比不小于对应相关词在本体中与关键词之间的相关度，认为与需要结果相关；否则，认为与结果无关。

4)同时满足同义词匹配与相关词匹配，并且同义词出现次数与任何相关词出现次数之比不小于对应相关词在本体中与关键词之间的相关度，认为与需要结果相关。否则，认为与结果无关。

5)只满足关键词匹配或同义词匹配之一，并且关键词或同义词出现次数少于某个常数值时，认为与需要结果不相关；否则认为与结果相关。

显然，经过检索结果过滤后，将排除以下的误检情况：

1)没有发生关键词匹配或同义词匹配事件，即图2中 $\overline{A \cup B}$ 所有结果。

2)发生关键词匹配或同义词匹配之一，未发生相关词匹配事件，并且关键词或同义词出现次数少于某个正常数。

3)发生关键词匹配或同义词匹配之一,且发生相关词匹配事件,且相关词匹配度远远大于关键词匹配度。

### 2.3　排序策略

对检索结果排序是检索至关重要的一步,把与结果相关度高的结果排在检索结果的前面,才能让用户更方便地找到。量化结果与需要的相关度就是排序的关键。

经过"关键词—领域本体—文档"的一系列相对映射后,结果与需要的相关度不仅由关键词词语与文档的相似度决定,还与关键词与领域本体中的概念语义相关度联系紧密。

把文档看成是由多个词项组成的向量,这些词项构成向量的维,将文档集看作是由文档组成的向量空间,可以表示为 $\boldsymbol{VSM} = \langle \boldsymbol{D}_1, \boldsymbol{D}_2, \cdots, \boldsymbol{D}_i, \cdots, \boldsymbol{D}_n \rangle$。在向量空间模型中,每个文档可以用一个向量表示 $\boldsymbol{D}_i = \langle t_{i1}, t_{i2}, \cdots, t_{ik}, \cdots, t_{in} \rangle$,其中 $1 \leqslant k \leqslant n$,且每个词 $t_{ik}$ 都对应一个权重 $w_{ik}$,表示该词与文档的相似程度(匹配度),所以文档又可以表示为 $\boldsymbol{D}_i = \langle (t_{i1}, w_{i1}), (t_{i2}, w_{i2}), \cdots, (t_{ik}, w_{ik}), \cdots, (t_{in}, w_{in}) \rangle$,其中 $1 \leqslant k \leqslant n$。

**定义 5**　词项匹配度。基于向量空间模型中词项在文档中出现的频度以及文档集情况,计算词项的特征权重,权重越大,则词项对文档贡献越大,词项匹配度(相似度)越大;反之,越小。

如定义 5 所述,匹配度的基础是词频,即词项在文档中出现的频度。词频计算公式:$f_{ik} = n_k / n_i$,其中 $n_k$ 是词 $t_k$ 在文档 $\boldsymbol{D}_i$ 中出现的次数,$n_i$ 是文档 $\boldsymbol{D}_i$ 中所有词向量总数。

根据著名的计算特征权重的方法——LTC[5-6]计算权重法,计算词项 $t_k$ 的特征权重:

$$w_{ik} = \frac{\log(f_{ik} + 1) \times \log(\frac{N}{N_k})}{\sqrt{\sum_{j=1}^{M} \left[ \log(f_{ij} + 1) \times \log(\frac{N}{N_j}) \right]^2}} \tag{3}$$

其中:$f_{ik}$ 是词 $t_k$ 在文档向量 $\boldsymbol{D}_i$ 中出现的词频,$N_k$ 是词 $t_k$ 在文档向量空间 $\boldsymbol{VSM}$ 中出现的总次数,$N$ 是文档向量空间 $\boldsymbol{VSM}$ 中文档的数量,$M$ 是文档向量空间 $\boldsymbol{VSM}$ 中所有有效词项(去除停用词后)的总数,$w_{ik} \in [0,1]$。

前面已经解释了领域本体中存在关系对象 $R \subseteq C \times C \times W$,$W$ 可以看作领域概念间的相关度。领域本体 $O_c$ 中对象间每个关系都对应一个权值 $w_{ck}$($w_{ck} \in [0,1]$),这个权值是本体构建过程中由领域专家赋予的,它表示各种关系对象间的相关度。借鉴文献[7]中给出的计算 query-document similarity 值所用公式,我们给出检索结果与关键词语义相关度的计算是由相关词的特征权重 $w_{ik}$ 与 语义关系对应权重 $w_{ck}$ 按下面公式求得:

$$\mathrm{sim}(D_i, O_c) = \sum_{T_k} (w_{ik} \times w_{ck}) \tag{4}$$

其中 $w_{ik}$ 是由式(3)中计算求得,$\boldsymbol{D}_i$ 表示文档向量,$\boldsymbol{T}_k$ 表示词 $t_k$ 及其所有相关词(包括同义词和相关词)向量集。

利用式(4)中 $\mathrm{sim}(D_i, O_c)$ 值的大小对检索结果进行排序。$\mathrm{sim}(D_i, O_c)$ 与词项匹配度和本体概念相关度有着密切的关系,式(4)利用了关键词及其相关词在文档集向量空间中的特征权重,同时配合领域本体中定义的语义距离,使计算结果从语义上更接近检索者需求。

## 3　实验数据分析

在 AMES 材料适应性评价系统中,材料体系、材料系列与材料牌号为树形结构,按其内容属性可分为多个层次。经过分析,材料牌号与材料具有同义关系,材料体系、材料系列与材料具有上下位相关关系,材料其他内容属性如性能、评价、浓度和基值等与材料是相关关系。设定概念语义权值 $W$ 如表 1 中所示。

**表 1　语义权值定义**

| 关系类型($R$) | 权值($W$) |
|---|---|
| 同义关系 | 0.950 |
| 整体与部分关系 | 0.500 |
| 上下位关系 | 0.500 |
| 同类关系 | 0.025 |
| 实例与概念关系 | 0.010 |
| 相关关系 | 0.010 |

在 AMES 材料适用性评价知识库文献中检索牌号为"XXXX"的结构钢,依次采用传统关键字检索方式和关键词语义扩展检索方式进行实验数据分析,结果如表 2。

**表 2　语义扩展前后检索查全率和查准率比较**

| 检索词 | 检索方式 | 查全率 | 查准率 |
|---|---|---|---|
| 结构钢 XXXX | 关键字检索 | 0.6014 | 0.4271 |
| | 语义扩展检索 | 0.8135 | 0.7475 |

实验中,关键字检索返回结果根据输入关键字不同,只返回文本中机械匹配检索关键词的检索结果,因此造成了漏检;而且只要是文本中包含了关键词的文本都被视为结果,而有的文档中只出现了关键词 1 ~ 2 次,很多情况并非文本主题,而是举例或文章过短而没有检索价值。这些问题在关键词语义扩展后能够被过滤策略排除。可见,检索策略改变后,检索结果比传统关键字匹配检索实现了更高的查全率和查准率,并且对检索结果的排序与检索目的更为接近。

在实验过程中,我们还发现检索策略改变后,检索结果的准确程度与检索文本长度及本体构建合理度关系较大。由于过滤策略与排序策略均建立在向量模型上,对文本过短的文本检索时效果不明显;本体构建的合理是关键词语义扩展的前提,构建合理有效的领域本体是检索有效的基础,特别是概念语义相关度的设置需要相关领域专家参与。

## 4　结语

本文在分析关键字检索的特点的基础上,结合领域本体概念提出对关键字语义扩展检索策略,利用本体结构的语义特点,扩展了关键字检索的范围,结合本体语义的过滤算法有效排除无关数据,并根据检索结果与关键词语义相关度进行排序,实验分析该策略有助于提高查全率和查准率。

**参考文献:**


[1] 张帆,朱红涛. 基于关键词的网络信息检索优化探索[J]. 情报科学,2005,23(6): 912 - 916.

[2] 张映海. 基于概念树扩展的中文文本检索研究[J]. 计算机工程与应用,2008,44(26):154 - 157.

[3] 马景娣. 查全率—查准率间存在顺变关系的数学证明[J]. 情报科学,2003,21(1):27 - 29.

[4] 陈刚,陆汝钤,金芝. 基于领域知识重用的虚拟领域本体构造[J]. 软件学报,2003,14(3):350 - 355.

[5] AAS J, EIKVIL L. Text Categorisation: A Survey, NR 941 [R]. Oslo: Norwegian Computing Center, 1999.

[6] BUCKLEY C, SALTON G, ALLAN J, *et al*. Automatic query expansion using SMART: TREC 3[EB/OL]. [2008 - 10 - 10]. http://trec.nist.gov/pubs/trec3/papers/cornell.new.ps.gz.

[7] SALTON G, BUCKLEY C. Term weighting approaches in automatic text retrieval [J]. Information Processing and Management, 1988, 24(5):513 - 523.