基于概念树扩展的中文文本检索研究

张映海
ZHANG Ying-hai

武警广州指挥学院 计算机教研室,广州 510440
Computer Office,Guangzhou Commanding Institute of the Armed Police of China,Guangzhou 510440,China
E-mail:yhzhang76@163.com

ZHANG Ying-hai.Research on Chinese text retrieval based on expansion of concept tree.Computer Engineering and Applications,2008,44(26):154-157.

Abstract: The expansion of semantic levels of concept is analyzed,the relationship between father and son concept in concept tree is translated by the similarity of words.Two methods,which are based on computing retrieval concept weight,and one method,which is based on computing text concept weight based on retrieval concept,are proposed.These methods are used in Chinese text retrieval.So,two text retrieval models are constructed based on concept tree expansion.Experiments show that the precision of the two retrieval models can be in line with keyword retrieval model,but the recall rate is improved greatly.
Key words: Chinese text retrieval;concept tree;concept weight;retrieval model

摘 要: 分析了概念在语义层次上的扩展,将概念树中的父子概念关系用词语的相似度进行量化。提出了检索概念权重计算的两种方法和一种基于检索概念的文本概念权重计算方法,并将这些方法用于中文文本检索,因此,构建了基于概念树扩展的两个文本检索模型。实验显示,这两个检索模型的精确率与关键词检索模型保持基本一致,召回率却得到较大提高。
关键词: 中文文本检索;概念树;概念权重;检索模型
DOI: 10.3778/j.issn.1002-8331.2008.26.047 **文章编号:** 1002-8331(2008)26-0154-04 **文献标识码:** A **中图分类号:** TP391

1 引言

基于关键词的文本检索遇到的一个重要问题就是词汇不匹配,即一个概念包含了树状网络语义,可能会产生由于用户的背景不同,所输入的词汇查找不到相应文本的情况。依据查询词在其概念树中的位置,对查询词进行上下的语义层次扩展后,再进行检索,一定程度上解决了关键词检索中,检索文本只匹配查询词而导致查询召回率低的状况,提高了检索系统的整体性能。

2 概念树

2.1 概念树的结构

概念树是整个语义网络的核心,定义了特定领域中的概念及概念之间关系。其语义网络为分层次的树状结构,根节点为该特定领域概念集合的总概念, 叶子节点为最小或较小子概念。上层概念是对其所有子概念属性的概括,子概念是从不同的角度对父概念的细化,同一个父概念的所有子概念之间形成平等关系,将它们称为兄弟概念。

概念树的构造将遵循以下规则[1]:

(1)概念树的结构是层次型的,它的框架将由手动给出,它的完善和优化将半自动完成;

(2)父概念比子概念更一般,即父结点概念包含子结点概念;

(3)子结点的概念必须是覆盖父结点概念的某一领域;

(4)由这种方式构成的层次结构,并不严格要求子女仅有一个父母,因此它可能会是一个有向无环图。

2.2 概念树的构成及扩展

概念语义层次扩展包括概念语义蕴含扩展和概念语义外延扩展,概述如下:

(1)构建概念树

构建概念树的词条一般是名词或名词实体,一个领域根据需要可以有一棵或多棵概念树。领域专家将文本集合中的文本浓缩成能较好体现该领域文本内容的概念词条集(ct_1,ct_2,…,ct_n),用分类树的方法建立起概念词条之间的上下层关系,给出一个虚概念作为这个分类层次体系树(概念树)的根节点(如“计算机领域”),记为第0层。代表具体概念意义的子树构建如下[2]:第1层是最具概括性的一些概念,每一概念节点都是该领域内相对独立的主题领域,如计算机、计算机软件、计算机硬件、计算机网络等;第2层,是对第1层概念的进一步划分,如“计算机软件”又可以分为“操作系统”、“程序设计语言”等;在第2层下面可以再划分第3层,如“操作系统”又可划分为“DOS”、“Windows”等,分层的数量根据需要确定。

基金项目: 国家自然科学基金(the National Natural Science Foundation of China under Grant No.60173060)。
作者简介: 张映海(1976-),男,讲师,研究方向:自然语言处理。
收稿日期: 2007-11-06 **修回日期:** 2008-02-18

(2)标识概念

在概念树中,给每一个概念节点赋子一个标识代码,利用代码可以反映出概念在分类树中的位置,也可以体现出概念之间的层次关系,所有代码组成了概念树的完整描述。

(3)扩展匹配

语句 S 经分词去除停用词后,得到词条序列($st_1,st_2,\cdots,st_m$)并输入知识网络,知识系统遍历各自的概念树获取并分析词条序列的分类代码,进一步遍历它们的父概念和子概念(遍历父概念的扩展称为概念语义外延扩展,遍历子概念的扩展称为概念语义蕴含扩展),得到关键词概念集序列($Q_1,Q_2,\cdots,Q_m$),Q_i 中包括 q_i、q_i 的父概念和子概念。若 q_i 在概念树中没有父概念,则只遍历自身及子概念;若没有子概念,则只遍历自身及父概念;孤立节点概念只遍历自身,根节点虚概念不参与遍历扩展。

3 检索概念权重计算的两种方法

现存检索模型的检索词权重基本都采用布尔模型[3],各检索词条的权重要么是1,要么是0,没有区分彼此间的轻重关系,也没有考虑各检索概念的父子概念。为此,利用概念间的语义相似度,提出检索概念权重的两种计算方法,分别结合不同的文本向量。

由于概念树中存在概念节点之间间隔多层的现象,概念节点层级间隔越远,相关性越小,检索概念在扩展时只考虑直接父子概念,为了描述方便,设所有检索概念都只有一个直接父概念(下同)。

3.1 合并增大法

基本思想:检索概念的父子概念越多,且与本身越相似,检索概念权重就越大。

设检索语句 q 包括 m 个检索概念 $q_1,q_2,\cdots,q_m$,q_i(i=1,2,$\cdots,m$)在概念树遍历到父概念 q_i^F、子概念 $q_i^1,q_i^2,\cdots,q_i^{S_i}$,各概念及其父子概念的单独权重仍采用布尔框架,在文本中出现为1,不出现为0。将 q_i 的父子概念权重经语义相似度计算映射到 q_i 并与 q_i 的权重合并后删除父子概念,于是得到 q_i 合并后的权重 $w_{c_i}(q)$:

$$w_{c_i}(q)=w_{q_i}(q)+D(q_i,q_i^F)w_{q_i^F}(q)+\sum_{l=1}^{S_i}D(q_i,q_i^l)w_{q_i^l}(q)$$

$w_A(q)$为 A 的布尔权重,$D(X,Y)$为 X、Y 之间的语义距离。

因该方法将检索概念的父子概念与检索概念本身合并,父子概念的权重通过语义相似映射累加到检索概念,作者将其命名为合并增大法(Merger Increase:MI)。

MI 的具体算法如下:

算法:MI(检索概念权重计算方法)//解决检索概念的权重计算//

输入:检索(查询)语句 q;

输出:各检索概念的权重。

过程:

(1)将 q 分词(去除停用词后)为 m 个检索概念 $q_1,q_2,\cdots,q_m$;

(2)在概念树中遍历 q_i 的父概念 q_i^F、子概念 $q_i^1,q_i^2,\cdots,q_i^{S_i}$,($i$=1,2,$\cdots,m$);

(3)输出 q_i 的权重 $w_{c_i}(q)$:

$$w_{c_i}(q)=w_{q_i}(q)+D(q_i,q_i^F)w_{q_i^F}(q)+\sum_{l=1}^{S_i}D(q_i,q_i^l)w_{q_i^l}(q)$$

$w_A(q)$为 A 的布尔权重,$D(X,Y)$为 X、Y 之间的语义距离。

(4)重复(2)、(3),直到 m 个检索概念的权重全部输出。

3.2 展开降低法

基本思想:检索概念的父子概念也直接参与检索,但权重小于检索概念。

设检索语句 q 包括 m 个检索概念 $q_1,q_2,\cdots,q_m$,q_i(i=1,2,$\cdots,m$)在概念树遍历到父概念 q_i^F、子概念 $q_i^1,q_i^2,\cdots,q_i^{S_i}$,各概念及其父子概念的单独权重仍采用布尔框架,$q_i$ 及其父子概念均直接参与检索。由于用户主要是检索 q_i,所以其父子概念的权重应小于 q_i,这里仍采用语义相似度进行映射,得到 q_i 及其扩展检索概念的权重为:

$$w_{q_i}(q),D(q_i,q_i^F)w_{q_i^F}(q),D(q_i,q_i^1)w_{q_i^1}(q),\cdots,D(q_i,q_i^{S_i})w_{q_i^{S_i}}(q)$$

$w_A(q)$为 A 的布尔权重,$D(X,Y)$为 X、Y 之间的语义距离。

该方法将检索概念的父子概念也展开为检索项,因父子概念是联想出来作为辅助检索的,其权重应小于原检索概念的权重,采取乘语义相似度因子降低权重,也就是与原概念越相似的父子概念,越值得作为检索概念,赋予的权重越大。因各父子概念的权重(假设在文本中出现,与布尔权重比较)都有所降低,故作者将该方法称为展开降低法(Expansion Decrease,ED)。

ED 的具体算法如下:

算法:ED(检索概念权重计算方法)//解决检索概念的权重计算//

输入:检索(查询)语句 q;

输出:各检索概念及其父子概念的权重。

过程:

(1)将 q 分词(去除停用词后)为 m 个检索概念 $q_1,q_2,\cdots,q_m$。

(2)在概念树中遍历 q_i 的父概念 q_i^F、子概念 $q_i^1,q_i^2,\cdots,q_i^{S_i}$,($i$=1,2,$\cdots,m$)。

(3)输出 q_i 及其父子概念的权重:

$$w_{q_i}(q),D(q_i,q_i^F)w_{q_i^F}(q),D(q_i,q_i^1)w_{q_i^1}(q),\cdots,D(q_i,q_i^{S_i})w_{q_i^{S_i}}(q)$$

$w_A(q)$为 A 的布尔权重,$D(X,Y)$为 X、Y 之间的语义距离。

(4)重复(2)、(3),直到 m 个检索概念及其父子概念的权重全部输出。

4 基于检索概念扩展的文本概念权重计算方法

文献[4]提出了基于概念查询扩展方法的基本思想:在用户初始查询的基础上抽取概念来建立用户查询空间,以保证加入的扩展词不再局限于相似度高或者同时出现频率高的词;基于精确性的考虑,对扩展词进行分组查询扩展并对查询结果整合排序以提高查准率;考虑到自动查询扩展不一定完全满足用户查询需求,提出了概念图的思想并加入手动查询扩展方法。文献[5]提出了基于体现层次关系的特征项树型存储技术的查询扩展、文献[6]提出了基于语义关系查询扩展的文档重构方法,其基本思想都是根据每个用户查询的扩展词表,对文本的表达方式进行重新组织和替换,将文本中表示相同信息概念的单元聚集起来,统一用查询中的信息概念来表示,然后再进行检索。设查询词 q 有 m 个扩展词,在文本 t 中:q 的词频为 tf、第 i 个

扩展词的词频为 tf_i；文本集合中的文本数为 N、含有查询词 q 的文本数为 n_q。q 及所有扩展词合成的查询概念在 t 中的权重 $w_t(q)$：

$$wt(q)=(tf+\sum_{i=1}^{m} tf_i)\times\log(\frac{N}{n_q}+1)$$

文献[5-6]将查询词的扩展词(查询词的同义词、上下义词)在文本中视为与查询词具有同等的权重，片面地扩大了部分扩展词相对于查询的权重。为此，本文提出一种基于检索概念扩展(Retrieval Concept Expansion，RCE)的文本概念权重计算方法。

4.1 基于 RCE 的文本概念权重概述

基于概念树合并的文本概念权重计算方法流程如图 1 所示。

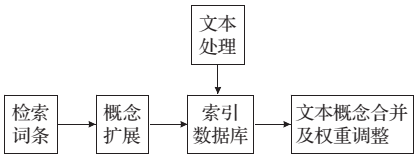

图 1 基于 RCE 的文本概念权重计算方法

基本思想：文本中查询词的每一个扩展词对查询的贡献在同等条件(词频、倒文本频率等)下应小于等于文本中的查询词，将扩展词的权重映射到查询词时：应乘上一个小于等于 1 的映射系数，再与查询词的权重进行相加。在此，将扩展词与查询词的语义距离作为其映射系数。

检索词条：用户原始查询语句中提取出来的检索词(查询词)。

概念扩展：查找检索词所在的领域概念树，在对应概念树中遍历所有的父概念和子概念节点，并标记存储。

文本处理：文本存储在索引数据库之前，对文本的分类、分词等便于存储用于索引之前的一切处理。

索引数据库：(1)存储文本可提供索引的有关信息；(2)接收检索扩展概念的有关数据，并与文本的相关数据比较计算。

文本概念合并及权重调整：(1)概念合并：在文本中将检索词及父子概念重新合并为文本的检索概念；(2)概念权重调整：将文本中检索词的父子概念权重合并到文本中的检索概念，删除其父子概念，降低文本向量化后的维数，降低计算复杂度。

4.2 基于 RCE 的文本概念权重计算

检索语句 q 包括检索概念 $q_1,q_2,\cdots,q_m$，根据概念树可展开为：$(q_1,q_1^F,q_1^1,q_1^2,\cdots,q_1^{S_1})$，$(q_2,q_2^F,q_2^1,q_2^2,\cdots,q_2^{S_2})$，$\cdots$，$(q_m,q_m^F,q_m^1,q_m^2,\cdots,q_m^{S_m})$，其中 q_i^F 为 q_i 的父概念，设 q_i 有 S_i 个子概念，q_i^j 为 q_i 的第 j 个子概念。

设 q_i 与文本 t 中的第 k 个词条 ct_k 对应，则 $(q_i,q_i^F,q_i^1,q_i^2,\cdots,q_i^{S_i})$ 在文本 t 中映射为 $(ct_k,ct_k^F,ct_k^1,ct_k^2,\cdots,ct_k^{S_i})$，合并权重记为 $w_{c_{t_k}}(t)$，并删除 ct_k 在文本 t 中的父子概念，$(ct_k,ct_k^F,ct_k^1,ct_k^2,\cdots,ct_k^{S_i})$ 在 t 中缺省的项，对应权重为 0，若 ct_k 缺省，则补上；t 中不归属于任何 q_i 及父子概念映射的词条 ct_j 不参与合并，权重不变，记为 $w_{ct_j}(t)$(之前已由 *TF-IDF* 计算出)。设文本 t 经概念词条合并调整后有 $u(u\leqslant m)$ 个合并概念，v 个非合并词条。

RCE 具体算法如下：

算法：RCE(改进的文本概念权重计算)//解决文本概念的权重计算//

输入：检索(查询)语句 q，用 TF-IDF 计算出词条权重的任一文本 t；

输出：t 中各概念(词条)的权重。

过程：

(1)将 q 分词(去除停用词后)为 m 个检索概念 $q_1,q_2,\cdots,q_m$；

(2)在概念树中遍历 q_i 的父概念 q_i^F、子概念 $q_i^1,q_i^2,\cdots,q_i^{S_i}$，$(i=1,2,\cdots,m)$；

(3)映射：$(q_i,q_i^F,q_i^1,q_i^2,\cdots,q_i^{S_i})\rightarrow t:(ct_k,ct_k^F,ct_k^1,ct_k^2,\cdots,ct_k^{S_i})$，权重合并为 $w_{c_{t_k}}(t)$；$ct_k,ct_k^F,ct_k^1,ct_k^2,\cdots,ct_k^{S_i}$ 在 t 中实际不存在的项，对应权重为 0，若 ct_k 在 t 中不存在，则补上，权重为 0；

(4)输出 $w_{c_{t_k}}(t)$：

$$w_{c_{t_k}}(t)=w_{ct_k}(t)+D(ct_k,ct_k^F)w_{ct_k^F}(t)+\sum_{l=1}^{S_i} D(ct_k,ct_k^l)w_{ct_k^l}(t)$$

$w_A(q)$ 为 A 的布尔权重，$D(X,Y)$ 为 X、Y 之间的语义距离；

(5)删除 $ct_k^F,ct_k^1,ct_k^2,\cdots,ct_k^{S_i}$；

(6)重复(2)~(5)，直到 m 个检索概念对应 t 中的 $u(u\leqslant m)$ 合并概念权重全部输出；

(7)输出 t 中未经合并的 v 个词条的权重：原权重不变输出。

5 基于概念树扩展的检索模型

5.1 检索概念权重合并增大的检索模型

检索概念权重经 MI 计算后对应的检索文本概念(词条)权重需按 RCE 计算，对应构建了检索概念权重合并增大的检索模型(Retrieval Concept Weight Merger Increase Retrieval Model，RCWMIRM)，模型中用户查询语句 q 与文本 t 的相似度 $S_{MI}(t,q)$ 为：

$$S_{MI}(t,q)=\frac{\sum_{i=1}^{m}(w_{c_i}(t)*w_{q_i}(q))}{\sqrt{(\sum_{j=1}^{v} w_{ct_j}^2(t)+\sum_{k=1}^{u} w_{c_k}^2(t))\sum_{i=1}^{m} w_{q_i}^2(q)}}$$

5.2 检索概念权重展开减小的检索模型

检索概念权重展开降低的检索模型(Retrieval Concept Weight Expansion Decrease Retrieval Model，RCWEDRM)只需将关键词模型的查询关键词权重按 ED 方法计算即可。

6 实验数据及结论

6.1 实验方案设计

6.1.1 实验安排及素材

对本文构建的检索概念权重合并增大的检索模型(RCWMIRM)、检索概念权重展开减小的检索模型(RCWEDRM)进行实验，并与文献[6]提出的检索方法(实验中简称“直接合并检索模型”：Directness Merger Retrieval Model，DMRM)、关键词检索模型(KRM)进行比较。

本次实验数据来源于网络收集整理的网页所得的文本，体育类：400 篇。

6.1.2 实验参数及性能评估

文本分词：采用中科院计算所研发的《中科院计算所汉语词法分析系统》。

概念(词汇)间语义相似度 $D(X,Y)$：采用中国科学院计算

技术研究所刘群等研发的《基于《知网》的词汇语义相似度计算》,其中的参数(参数含义见文献[7])为 α=1.60、β1=0.50、β2=0.20、β3=0.17、β4=0.13、γ=0.20、δ=0.20。

实验中将准确率和召回率视为同等重要,选择 F-Measure 法对检索模型进行综合评估,F 值越大,检索性能越好。

$$\text{检准率(准确率,}P\text{)}=\frac{\text{检索到的相关文本数}}{\text{检索返回的全部文本数}}$$

$$\text{检全率(召回率,}R\text{)}=\frac{\text{检索到的相关文本数}}{\text{文本集合中相关的全部文本数}}$$

$$F=\frac{2*\text{准确率}*\text{召回率}}{\text{准确率}+\text{召回率}}$$

6.2 实验性能比较及结论

实验采用"亚运会的球类比赛"作为查询语句,文本集合中与查询主题相关的文本有 152 篇,将"运动会"作为"亚运会"的父概念、"篮球"等 15 类球类作为"球类"的子概念进行实验。实验结果如表 1 所示。

表 1 KRM、DMRM、RCWEDRM 与 RCWMIRM 检索性能比较

	检索模型		KRM	DMRM	RCWEDRM	RCWMIRM
		返回相关文档数	113	118	136	127
		返回文档数	215	239	297	273
	0.08	*P*	0.526	0.494	0.458	0.465
		R	0.743	0.776	0.895	0.836
		F	**0.616**	**0.604**	**0.606**	**0.598**
		返回相关文档数	73	98	121	113
		返回文档数	159	198	264	223
λ	0.10	*P*	0.460	0.495	0.458	0.507
		R	0.480	0.645	0.796	0.743
		F	**0.470**	**0.560**	**0.581**	**0.603**
		返回相关文档数	41	70	95	97
		返回文档数	87	136	168	161
	0.12	*P*	0.471	0.515	0.568	0.602
		R	0.270	0.461	0.625	0.638
		F	**0.343**	**0.487**	**0.595**	**0.619**

性能比较如表 2。

表 2 KRM、DMRM、RCWEDRM 与 RCWMIRM 的平均 *F* 值

检索模型	KRM	DMRM	RCWEDRM	RCWMIRM
平均 *F* 值	0.476	0.550	0.594	0.607

实验结论:检索中文文本,用 RCWMIRM 和 RCWEDRM 来提高检索系统的 F 值,改进其性能是有效可行的。

7 结束语

本文对概念树扩展用于文本检索进行了探讨,将提出的权重计算方法用于构建文本检索模型,实验验证了所构造的检索模型综合性能优于关键词检索模型。研究中存在的不足:假设了除根节点以外的所有概念只有一个父概念,要求任一概念只能属于一个类;引用的词语相似度计算方法覆盖面很有限,部分相关的词语,还不能计算其间的相似度。这些都是需要进一步研究解决的问题。

参考文献:

[1] 李振东,费翔林.基于概念的信息检索模型研究[J].南京大学学报:自然科学版,2002,38(1):99-109.

[2] 邱树雄,李志蜀,王娣,语义网络及其 Web 信息检索机制研究[J].计算机工程,2004,30(23):118-120.

[3] Cooper W S.Getting beyond Boole [J].Information Processing and Management,1988,24:225-243.

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

[5] 王丽君,高迎,王锡钢.中文检索系统中查询的扩展[J].小型微型计算机系统,2002,23(7):894-896.

[6] 张敏,宋睿华,马少平.基于语义关系查询扩展的文档重构方法[J].计算机学报,2004,10:1395-1400.

[7] 刘群,李素建.基于《知网》的词汇语义相似度计算[C]//第三届中文词汇语义学研讨会论文集,中国台北中央研究院,2002-05.

(上接 131 页)

览的兴趣和目的,为网站设计者提供更科学的关于用户访问行为的信息。

参考文献:

[1] 陈莉,焦李成.Internet/Web 数据挖掘现状及最新进展[J].西安电子科技大学学报:自然科学版,2001,28(1).

[2] 刑东山,沈钧毅,宋擒豹.从 Web 日志中挖掘用户浏览偏爱路径[J].计算机学报,2003,26(11):1518-1523.

[3] 缪勇.匿名用户兴趣路径挖掘研究与实现[D].南京:南京理工大学计算机应用技术系,2006.

[4] Chen M S,Park J S,Yu P S.Data mining for path traversal patterns in a Web environment [C]//Proceedings of the 16th International Conference on Distributed Computing Systems,Hong Kong,1996:385-392.

[5] Mobasher B,Srivastava J.Data preparation for mining world wide web browsing patterns[J].Knowledge and Information System,1999,1(1):5-32.

[6] Jin X,Zhou Y,Mobasher B.Task-oriented Web user modeling for recommendation [C]//Proceedings of the 10th International Conference on User Modeling,Edinburgh,Scotland,July 2005.

(上接 143 页)

[11] Saggion H,Lapalme G.Concept identification and presentation in the context of technical text summarization[C]//Proc of the Workshop on Automatic Summarization.New Brunswick,New Jersey:Association for Computing Linguistics,2000:1-10.

[12] Morris A H,Kaspcr G M,Adams D A.The effects and limitations of automated text condensing on reading comprehension performance[J].Information Systems Research,1992,3(1).

[13] 傅间莲,陈群秀.一种新的自动文摘系统评价方法[J].计算机工程与应用,2006,42(18):176-177.

[14] 李绍山.语言研究中的统计学[M].西安:西安交通大学出版社,2001:165.