基于领域本体和位置关系的信息检索模型

蒋宗礼,隋少鹏
(北京工业大学 计算机学院,北京 100124)

摘 要:向量空间模型是最常用的信息检索模型,它根据词频来计算文档之间的相关度,这种方法虽然能够满足用户的基本检索需求,但是对于检索要求较高的用户,其效果仍然不甚理想。文中在向量空间模型的基础上,首先通过领域本体和上层本体来计算特征词项之间的相似度,据此得出与查询词相关的词,在求词项频率和逆文档频率时考虑这些词,然后引入了词序相关度和词语相邻相关度这两个概念,把特征项的位置关系也考虑进来。实验结果表明,文中提出的模型相比原始向量空间模型,在准确率上有了较大的改善。这完全说明,与原始向量空间模型相比,文中提出的检索模型不仅考虑了与原有词项具有相似语义的词项,而且还考虑了词项顺序和词项相邻信息,从而更能符合用户的检索要求。

关键词:检索模型;向量空间模型;本体;相似度
中图分类号:TP31 **文献标识码**:A **文章编号**:1673-629X(2015)01-0006-05
doi:10.3969/j.issn.1673-629X.2015.01.002

Information Retrieval Model Based on Domain Ontology and Position Relationship

JIANG Zong-li, SUI Shao-peng
(College of Computer, Beijing University of Technology, Beijing 100124, China)

Abstract: Vector space model, which calculates the relatedness between documents through word frequency, is a frequently used information retrieval model. This method can meet the user's basic retrieval requirements, but for users with higher requirements, its effect is still not very ideal. In this paper, based on vector space model, first calculate the similarity, which can produce words related to the query word, between words through the use of domain ontology and upper ontology. So can take advantage of the related word when calculate TF and IDF. Then by introducing the concept of word order relatedness and word adjacent relatedness, can embody the position relationship. The experimental results show that this method can improve the precision considerably. This fully shows that, compared with the original vector space model, the retrieval model proposed not only considers the terms which have similar semantics with the original words, but also thinks about the word order information and word adjacent information, thus can meet users' retrieval requirements better.

Key words: retrieval model; vector space model; ontology; similarity

0 引 言

信息检索[1](Information Retrieval, IR),是指将信息按一定的方式组织和存储起来,并根据用户的需要找出有关信息的过程。信息检索模型[2]是指依照用户查询,对文档集合进行相关排序的一组前提假设和算法。IR 模型可形式地表示为一个四元组 $<D, Q, F, R(q_i, d_j)>$,其中 D 是一个文档集合,Q 是一个查询集合,F 是一个对文档和查询建模的框架,$R(q_i, d_j)$ 是一个排序函数,它给查询 q_i 和文档 d_j 之间的相关度赋予一个排序值。常用的信息检索模型有:布尔模型、向量空间模型[3]、概率模型[4]、基于排序的语言模型等[3]。

在计算机科学与信息科学领域,本体[5]是对共享概念模型的明确的形式化规范说明。本体就是一种特殊类型的术语集,具有结构化的特点,且更加适合于在计算机系统之中使用。领域本体[6]所建模的是某个特定领域,或者现实世界的一部分。领域本体所表达的是那些适合于该领域的术语的特殊含义。可以通过本体来得到词项之间的语义相似度[7],利用此相似度来改进排序算法。

在各种信息检索模型中,最为常用的是向量空间模型。但是向量空间模型有其固有的缺点,因此有许多文献对其进行了改进。例如,文献[8]通过引入语

收稿日期:2014-01-26 **修回日期**:2014-04-29 **网络出版时间**:2014-11-17
基金项目:教育部国家级教学团队建设项目(00700054J1901)
作者简介:蒋宗礼(1956-),男,博导,研究方向为网络信息搜索与处理;隋少鹏(1988-),男,硕士研究生,研究方向为搜索引擎、语义 Web。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20141117.2205.017.html

义增量的方式解决了特征词之间相互独立的问题;文献[9]通过同义关系、继承关系以及关联关系三个方面对权值扩展解决了特征词权值不能反映语义的问题;文献[10]根据 Wordnet 对查询中的特征词进行语义扩展,将与原特征词相似的特征词考虑进来。这些方法都在一定程度上改进了向量空间模型,但是它们都没有考虑特征词项之间的位置关系,从而使检索结果不够准确。

基于上述分析,文中在向量空间模型的基础上,提出了一种结合查询词项语义及位置关系的检索模型,该检索模型使用的算法改进了传统的基于词频统计的算法未考虑词项语义以及位置关系从而使检索结果不够准确的问题。实验结果表明,通过这些改进,文中提出的算法能获得较好的检索准确率。

1　经典向量空间模型

向量空间模型(VSM)是 60 年代末由 Salton 提出来的,它采用了“部分匹配”的检索策略,即仅出现部分索引词的文档也可以出现在检索结果中。但是作为一个检索模型,它有一些缺点,虽然提供了方便的计算框架,但是没有考虑词项语义以及位置关系,从而使检索结果不够准确。

模型中,文档和查询都被假设是一个 t 维向量空间中的向量(点),其中 t 是特征词项(词语、词干、短语等)的个数。一篇文档 D_i 表示为由特征词项组成的向量:$D_i=(d_{i1},d_{i2},\cdots,d_{ij},\cdots,d_{it})$。其中 d_{ij} 表示文档 D_i 的第 j 个特征词项的权值。一个包含 n 个文档的文档集,可以表示成一个词项权值的矩阵,其中每一行表示一篇文档,每一列表示相应文档在相关词项上的权值大小。如下所示:

	词项$_1$	词项$_2$	…	词项$_t$
文档$_1$	d_{11}	d_{12}	…	d_{1t}
文档$_2$	d_{21}	d_{22}	…	d_{2t}
⋮	⋮			⋮
文档$_n$	d_{n1}	d_{n2}	…	d_{nt}

查询项采用与文档相同的方式表示,即查询项 Q 表示为有 t 个权值的向量:$Q=(q_1,q_2,\cdots,q_j,\cdots,q_t)$,其中 q_j 是查询项 Q 中第 j 个词项的权值。

基于这种表示,可以通过计算文档和查询向量之间的距离来对相关文档进行排序。通常这种距离用相似度来度量,得分最高的文档被认为和查询具有最高的相似度。为此,已经有很多相似度度量方法先后被提出,其中应用范围最广的是余弦相似度度量方法。

$$\mathrm{Score}_{\mathrm{original}}(Q,D_i)=\frac{\sum_{j=1}^{t}q_jd_{ij}}{\sqrt{\sum_{j=1}^{t}d_{ij}^2\cdot\sum_{j=1}^{t}q_j^2}} \tag{1}$$

关于词项权值有很多不同的计算方法,其中最常使用的是 tf. idf 加权方法的变形。tf 表示文档中词项的频率,反映了一个词项在文档 D_i(或查询)中的重要性。但是为了减少文档的长度对重要性的影响,这个频率通常都是通过对词项在文档中的出现次数进行归一化后得到,例如

$$\mathrm{tf}_{ik}=\frac{\log(f_{ik})+1}{\sqrt{\sum_{k=1}^{t}[\log(f_{ik})+1]^2}} \tag{2}$$

反文档频率(idf)反映了词项在文档集的重要性。如果包含某个词项的文档越多,这个词项在文档之间就越没有区分性,也就对检索越没有用。这个权值的典型形式如下

$$\mathrm{idf}_k=\log\left(\frac{N}{n_k}\right) \tag{3}$$

两种权值的影响通过相乘结合起来就是所谓的 tf. idf 加权算法。因此,文档词项权值的形式为:

$$d_{ik}=\frac{(\log(f_{ik})+1)\cdot\log(N/n_k)}{\sqrt{\sum_{k=1}^{t}[(\log(f_{ik})+1.0)\cdot\log(N/n_k)]^2}} \tag{4}$$

考虑到余弦相似度的归一化已经融合到权值计算中,一个文档的分值可以通过将文档向量和查询向量求点积获得。

此算法的缺点是只能依据某词项出现次数的多少来决定某文档是否命中,而不能查询出包含某些与原词项相关的词项的文档。同时文档中词项出现的位置先后顺序以及邻近关系也没有在此算法中体现。因此文中在该算法的基础上将这两个因素加入进来,从而对其进行了改进。

2　向量空间模型的改进

2.1　语义相似度算法的改进

如前所述,语义相似度计算方法有多种,文献[11]以基于语义距离[12]的计算模型为基础,提出了一种改进的基于领域本体的概念语义相似度计算方法。相似度算法如下:

$$\mathrm{sim}_1(c_1,c_2)=\frac{\omega}{\mathrm{dist}(c_1,c_2)+\omega}\left(1-\frac{\mathrm{dep}(c_1)+\mathrm{dep}(c_2)}{\varphi|\mathrm{dep}(c_1)-\mathrm{dep}(c_2)|+1}\right) \tag{5}$$

其中:

$$\mathrm{dist}(c_1,c_2)=\frac{\lambda}{\mathrm{weight}(c_1,c_2)}-\lambda \tag{6}$$

$$\mathrm{weight}(c_1,c_2)=\begin{cases}1,\text{如果 } c_1,c_2 \text{ 为等价关系}\\ \alpha\times\mathrm{type}(c_1,c_2)+\beta\times\dfrac{\mathrm{num}(\mathrm{attr}(c_1)\cap\mathrm{attr}(c_2))}{\mathrm{num}(\mathrm{attr}(c_1))+\mathrm{num}(\mathrm{attr}(c_2))}+\\ \quad\gamma\left(\eta+(1-\eta)\times\theta^{\frac{2\mathrm{dgr}(*)}{\mathrm{dgr}(c_1)+\mathrm{dgr}(c_2)}}\right),\text{其他}\\ \quad(\text{其中 }\alpha+\beta+\gamma=1)\end{cases} \tag{7}$$

它确定了三个对相似度产生影响的量化因素：节点关系类型、属性重合度、密度。节点关系类型是指根据节点之间的关系来评分。等价关系评分较高，继承关系次之，其他关系最低。属性重合度是指两个概念拥有的共同属性越多，说明两节点的关系也就越密切，由它们构成的有向边的权重也就应该越大；拥有的不同属性越多，权重越小。密度是指在本体有向图中某一局部的节点密度越大，说明在该处对概念的细化也就越大，那么对应的有向边的权重也就越大。

虽然这种方法对特定领域的检索具有较好的效果，但是它只考虑了概念在领域本体中的若干特征，而未考虑概念在语料库中的特征—概率，从而带有一定的主观性，与实际情况不符。

可以通过将两个概念在 HowNet 中根据信息理论[13-14]得到的语义相似度融合到以上相似度中得到一种新的相似度算法。

根据信息理论的定义，两个概念之间的语义相似度是：

$$\mathrm{sim}_2(c_1,c_2)=\frac{2\times\log P(\mathrm{common}(c_1,c_2))}{\log P(\mathrm{description}(c_1,c_2))} \tag{8}$$

其中，$\mathrm{common}(c_1,c_2)$ 是 c_1 和 c_2 的共性；$\mathrm{description}(c_1,c_2)$ 是对 c_1 和 c_2 的描述。在信息理论中，一个陈述(Statement)中的信息量的大小是由它的概率的负对数($-\log P(\mathrm{Statement})$)来衡量的。其中 $P(\mathrm{Statement})$ 表示一个随机选择的对象(Object)属于 Statement 的概率。

考虑如图1所示的本体(它是 HowNet 的一个片段)，其中括号内的数字表示对应概念的概率。现在要求“丘陵”和“海岸”这两个概念的相似度，则 $\log(P(\mathrm{common}(丘陵,海岸)))=\log(P(地质建造))=\log(0.001\ 76)=-2.754$，$\log(P(\mathrm{description}(丘陵,海岸)))=\log(P(丘陵))+\log(P(海岸))=\log(0.000\ 018\ 9)+\log(0.000\ 021\ 6)=-9.389$，所以 $\mathrm{sim}(丘陵,海岸)=0.59$。

综合式(5)和(8)，可以用式(9)求得两个概念 c_1 和 c_2 的语义相似度。

$$\mathrm{sim}(c_1,c_2)=d\times\mathrm{sim}_1(c_1,c_2)+e\times\mathrm{sim}_2(c_1,c_2) \tag{9}$$

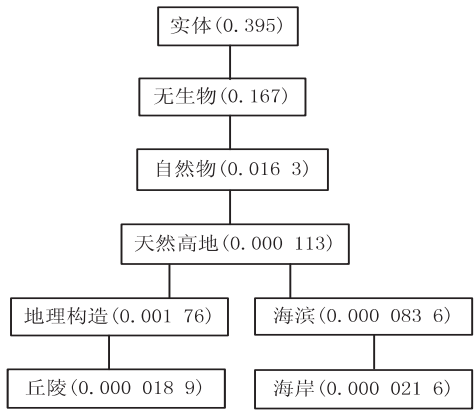

图1　HowNet 片段

这样既考虑了领域本体中概念的若干特征，又考虑了概念在真实语境中出现概率的大小。其中，$d+e=1$，通过调整它们的值，使 $\mathrm{sim}(c_1,c_2)$ 最接近专家给出的相似度。

2.2　未考虑位置关系时修改排序函数

对于查询 q 中的每个词项 p，首先根据上面的相似度计算公式得到与 p 相似的词的集合。然后在求 tf 和 idf 时，把它们也考虑进来。

具体做法是在求 tf 时，不仅计算与查询项一模一样的词在某篇文档中出现的频率，还要计算与查询词项相关的词项出现的频率。当然需要设定一个阈值，当两个词项的相似度大于这个阈值时，这两个词项才算相关。求 idf 时，采用相似的方法。然后将改进的 tf 和 idf 带入公式(4)，就可以得到某个词项的权值。当求出所有词项的权值后，就可以根据公式(1)来得到查询与文档的相关度。

这种方法从语义方面扩展了经典的向量空间模型，使得检索结果包含了跟查询词项相关的词项所在的文档，但是其仍未能考虑词项位置关系。

2.3　考虑词项位置关系时修改排序函数

词项位置关系是影响评分的重要因素，考虑下面两个文档：

A：数据库系统是软件研究领域的一个重要分支。

B：数据库管理系统是一种重要的系统软件。

如果给定查询“数据库系统”，根据公式(1)，将得到文档 B 的评分较高，因为在文档 B 中词项“系统”出现了两次，而在文档 A 中词项“系统”只出现了一次。但是很明显，用户希望查到的是文档 A，因为在文档 A 中词项“数据库”和词项“系统”是相邻的，因此文档 A 的评分应该较高。

词项位置关系对评分的影响包含两个方面：词序因素和词语邻近因素。

词序相关度是与词项顺序有关的，它把查询中的词项看作是有序的，如果某篇文档中包含这些词项且

顺序与查询中词项的顺序一致,那么这篇文档的相关度较高,否则较低。

文档 A、B 的词序相关度 $\mathrm{Score}_{\mathrm{order}}(A,B)$ 由下述公式求出:

$$\mathrm{Score}_{\mathrm{order}}(A,B)=\begin{cases}1-\dfrac{\mathrm{InvNum}(A,B)}{|\mathrm{Intst}(A,B)|-1}, & \text{当}|\mathrm{Intst}(A,B)|>1\\ 1, & \text{当}|\mathrm{Intst}(A,B)|=1\\ 0, & \text{当}|\mathrm{Intst}(A,B)|=0\end{cases}\tag{10}$$

其中,$\mathrm{Intst}(A,B)$ 表示在文档 A 和 B 中都出现且仅出现一次的特征项的集合;$P_1(A,B)$ 表示 $\mathrm{Intst}(A,B)$ 中的特征项在 A 中的位序构成的向量;$P_2(A,B)$ 表示按照 $P_1(A,B)$ 中的分量对应特征项在 B 中的次序排列生成的向量。

考虑有下面两个文档:

$A:W_1\ W_2\ W_3\ W_4\ W_5\ W_6\ W_7\ W_8$

$B:W_4W_5\ W_2\ W_9\ W_1\ W_{10}\ W_3W_8$

其中,$\mathrm{Intst}(A,B)=\{W_1,W_2,W_3,W_4,W_5,W_8\}$。

A 中特征项与序号的对应关系如表 1 所示。

表 1　A 中特征项与序号的对应关系

W_1	W_2	W_3	W_4	W_5	W_6	W_7	W_8
1	2	3	4	5	6	7	8

由表 1 可得 $P_1(A,B)=(1,2,3,4,5,8)$。B 中特征项与序号的对应关系如表 2 所示。

表 2　B 中特征项与序号的对应关系

W_4	W_5	W_2	W_9	W_1	W_{10}	W_3	W_8
4	5	2		1		2	7

由相同分量对应的特征项在 B 中的顺序得 $P_2(A,B)=(4,5,2,1,3,8)$。

$\mathrm{InvNum}(A,B)$ 表示 $P_2(A,B)$ 中各相邻分量的逆序数。求 InvNum 时,不仅要考虑文档中与查询中一模一样的词项的顺序,还要考虑与原词项相关的词项的顺序。本例中,由 4<5、5>2、2>1、1<3、3<8,得 $\mathrm{InvNum}(A,B)=2$。所以 $\mathrm{Score}_{\mathrm{order}}(A,B)=1-2/(6-1)=0.6$。

词语的位置关系并不仅仅体现在词语出现的先后顺序关系上,还有词语所在位置的距离关系。为了更好地体现这一关系,文中引入“词语相邻相关度”,来反映查询串和检索文档中的词语在相邻性关系上的相似程度。词语相邻相关度是指根据查询中具有相邻关系的两个词项在文档中是否相邻而对文档进行评分,不仅要考虑文档中与查询中的词项一模一样的词项是否相邻,而且将与查询中词项相关的词项也考虑在内,相邻且顺序不变评分较高,否则评分较低。文档 A 和 B 的词语相邻相关度 $\mathrm{Score}_{\mathrm{aja}}(A,B)$ 由公式(11)求出:

$$\mathrm{Score}_{\mathrm{aja}}(A,B)=\frac{|\mathrm{bigram}(A)\cap\mathrm{bigram}(B)|^2}{|\mathrm{bigram}(A)|\cdot|\mathrm{bigram}(B)|}\tag{11}$$

其中,$\mathrm{bigram}(D)$ 表示 D 中相邻词项两两组成的二元组集合。考虑下面的文档 A 和 B:

$A:w_1\ w_2\ w_3\ w_4\ w_5$

$B:w_6\ w_4\ w_5\ w_7\ w_1\ w_2\ w_8\ w_9$

$\mathrm{bigram}(A)=\{w_1w_2,w_2w_3,w_3w_4,w_4w_5\}$

$\mathrm{bigram}(B)=\{w_6w_4,w_4w_5,w_5w_7,w_7w_1,w_1w_2,w_2w_8,w_8w_9\}$

所以 $\mathrm{Score}_{\mathrm{aja}}(A,B)=2*2/(4*7)=1/7$

2.4　最终排序函数及算法

改进后的向量空间模型的排序函数如下:

$$\mathrm{Score}(q,d)=f\cdot\mathrm{Score}_{\mathrm{original}}(q,d)+g\cdot\mathrm{Score}_{\mathrm{order}}(q,d)+h\cdot\mathrm{Score}_{\mathrm{aja}}(q,d)\quad(f+g+h=1)\tag{12}$$

由于 $\mathrm{Score}_{\mathrm{order}}$ 和 $\mathrm{Score}_{\mathrm{aja}}$ 的计算非常消耗时间,所以在计算 $\mathrm{Score}_{\mathrm{original}}$ 的时候可以设置一个阈值 Y。只有当 $\mathrm{Score}_{\mathrm{original}}>Y$ 时,才有必要计算 $\mathrm{Score}_{\mathrm{order}}$ 和 $\mathrm{Score}_{\mathrm{aja}}$,否则将 $\mathrm{Score}_{\mathrm{order}}$ 和 $\mathrm{Score}_{\mathrm{aja}}$ 置为 0。

综合排序算法如下:

输入:查询 q,文档集 D;

输出:排序后的文档集 D'。

步骤 1:根据所需处理的文本集 $D(d_1,d_2,\cdots,d_n)$,获取特征项向量 $(F_1,F_2,\cdots,F_n)$。

步骤 2:对 D 中的每篇文档 d_i 进行分词,参照特征项向量获得分词后的向量 $(T_{i1},T_{i2},\cdots,T_{in})$。

步骤 3:结合领域本体和 HowNet,求特征项两两之间的相似度,将结果放在一个表中,表中每一行是一对特征项和它们的相似度。

步骤 4:根据阈值对表进行筛选,处理后的形式是一个哈希表,其键为某一特征项,其值为与这个特征项相似的特征项组成的集合。

步骤 5:对于文档集中的每篇文档,分别生成由其特征项二元组组成的集合。

步骤 6:依据原始公式计算 q 和 d_i 的分值 $\mathrm{Score}_{\mathrm{original}}(q,d_i)$,若其值小于阈值,则 $\mathrm{Score}(q,d_i)=\mathrm{Score}_{\mathrm{original}}(q,d_i)$,算法结束。

步骤 7:根据词序相关度公式计算 q 和 d_i 的词序相关度分值 $\mathrm{Score}_{\mathrm{order}}(q,d_i)$。

步骤 8:根据词语相邻相关度公式计算 q 和 d_i 的词语相邻相关度分值 $\mathrm{Score}_{\mathrm{aja}}(q,d_i)$。

步骤 9:求最终的分值 $\mathrm{Score}(q,d_i)$,并根据此值对文档集 D 进行排序,排序后的结果写入 D'。

与原始向量空间模型评分算法相比,新算法不仅保证了与原有词项具有相似语义的词项被考虑在内,

而且还考虑了词项顺序和词项相邻信息,从而更能符合用户的检索要求。

3 实 验

实验中使用的数据集是中科院自动化所的中文新闻语料库,它从凤凰、新浪、网易、腾讯等版面搜集文章,搜集时间在 2009 年 12 月—2010 年 3 月,分为 8 类:阅读、娱乐、历史、教育、社会、文化、军事、科技。由于文章数量太多,实验只使用科技类作为实验对象,它不仅可以作为查询的文档集,而且在求基于信息理论的相似度中的概率时也要用到。

实验所使用的是由领域专家构建的金融本体,它描述金融领域中的概念及其之间的关系,通过这个本体可以计算基于语义距离的相似度。所使用的上层本体是 HowNet,它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的常识知识库,可以针对里面的概念求基于信息理论的相似度。

实验用 Java 语言实现,经测工作正常。用 Jena 框架将本体中的相应概念转化成 Java 中的类,从而对这些概念进行操作。

求查询词项的相似词项集合时,需要设定阈值,只有相似度不小于这个阈值的词项才能作为求语义相关的 tf 及 idf 时的依据。将此值设为 0.71。

在求 $\text{Score}_{\text{original}}$ 时不是所有的文档都要进行下一步处理,只有此值大于阈值的文档才需要考虑词项位置关系,将此值设置为 0.6。公式中的各种参数可以做如下设置:$d=0.7$,$e=0.3$,$f=0.7$,$g=0.15$,$h=0.15$。

实验采用的评测指标为信息检索系统中常用的返回前 n 个结果的准确率 $P@n$。实验中 n 分别取 50,100,200。$P@n$ 的计算公式如下:

$$P@n = \frac{\text{前 } n \text{ 个结果中相关文档的个数}}{n} \times 100\% \tag{13}$$

数据集上的实验测评结果如表 1 所示。

表 1 数据集上的实验测评结果

准确率	检索模型		提高百分比/%
	向量空间模型	文中提出的检索模型	
$P@50$	0.402 5	0.467 9	6.54
$P@100$	0.367 1	0.431 1	6.40
$P@200$	0.345 1	0.403 7	5.86

从表中可以看出,相比原始向量空间模型,文中提出的检索模型在准确率上平均有了 6.26% 的提高,这表明文中提出的检索模型是有效可行的。

4 结束语

文中在向量空间模型的基础上,通过结合领域本体和上层本体来求两个词项的相似度,然后在对文档评分时将与查询词项相似的词也考虑在内。而且,还将词语位置关系加入到排序函数中,使词项间的位置关系成为影响评分的因素。

经过实验验证,文中的算法在准确率上取得了较理想的效果,更能符合用户的检索要求。但是在计算词项位置关系相关的评分时需要消耗较多时间,如何提高计算速度仍然是一个问题。此外在排序函数中有大量的参数设定,其值设定的不同会对结果产生影响,未来需要通过实验来确定最合适的参数。

参考文献:

[1] Gao J,Nie J Y,Wu G,et al. Dependence language model for information retrieval[C]//Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval. Sheffield:ACM,2004:170-177.

[2] Christopher D,Prabhakar R,Hinrich S. Introduction to information retrieval[M]. [s. l.]:Post & Telecom Press,2010.

[3] Vector space model[EB/OL]. [2013-11-25]. http://en.wikipedia.org/wiki/Vector_space_model.

[4] 范晨熙,黄理灿,李雪利. 基于 Lucene 的 BM25 模型的评分机制的研究[J]. 工业控制计算机,2013(3):78-79.

[5] Uschold M,Tate A. Putting ontologies to use[J]. The Knowledge Engineering Review,1998,13(1):1-3.

[6] Pascal H,Markus K,Sebastian R. Foundations of semantic Web technologies[M]. [s. l.]:Chapman & Hall/CRC,2009.

[7] Corley C,Mihalcea R. Measuring the semantic similarity of texts[C]//Proceedings of the ACL workshop on empirical modeling of semantic equivalence and entailment. [s. l.]:Association for Computational Linguistics,2005:13-18.

[8] 唐明伟,卞艺杰,陶飞飞. 基于语义向量空间模型的文档检索系统研究[J]. 情报杂志,2010,29(5):167-170.

[9] 韩美灵,杨 勇. 一种面向语义检索的向量空间模型改进方法[J]. 农业网络信息,2012(10):39-41.

[10] 纪兆辉. 一种基于本体语义的信息检索模型[J]. 计算机与数字工程,2010,38(11):118-121.

[11] 王 欢,孙瑞志. 基于领域本体和 Lucene 的语义检索系统研究[J]. 计算机应用,2010,30(6):1655-1657.

[12] 陈沈焰,吴军华. 基于本体的概念语义相似度计算及其应用[J]. 微电子学与计算机,2008,25(12):96-99.

[13] 董振东,董 强,郝长伶. 知网的理论发现[J]. 中文信息学报,2007,21(4):3-9.

[14] Bin You,Liu Xiaoran,Ning Li,et al. Using information content to evaluate semantic similarity on HowNet[C]//Proc of 2012 eighth international conference on computational intelligence and security. [s. l.]:[s. n.],2012.

基于领域本体和位置关系的信息检索模型

万方数据 WANFANG DATA 文献链接

作者： 蒋宗礼， 隋少鹏， JIANG Zong-li， SUI Shao-peng
作者单位： 北京工业大学 计算机学院,北京,100124
刊名： 计算机技术与发展 ISTIC
英文刊名： Computer Technology and Development
年，卷(期)： 2015(1)

引用本文格式：蒋宗礼.隋少鹏.JIANG Zong-li.SUI Shao-peng 基于领域本体和位置关系的信息检索模型[期刊论文]-计算机技术与发展 2015(1)