基于 VSM 的中文信息检索

Chinese Text Information Retrieval Based on VSM

徐云青　徐义峰　（衢州学院 浙江 衢州 324000）
李舟军　（北京航空航天大学，计算机科学与工程学院 北京 100083）

摘要：本文介绍了在向量空间模型下，使用 Cosine－Measure 和 OKAPI－Measure 两种不同的相似度评测方法，来评测查询与文本之间的相关性。通过针对 10 字以内的短查询的实验分析，作者发现在相同召回率的情况下，使用 OKAPI 法来计算相似度得到的检索结果，准确率要比 Cosine 法的高。

关键词：VSM　中文信息检索　Cosine－Measure　OKAPI－Measure

1　引言

对于西文的检索，技术已经发展到了一种相当高的阶段，但是，对于中文检索，由于种种原因，现在还不能达到如西文那样让人满意的程度。为什么这么多年的研究对中文检索的效果还是不大呢？在文献[1]中对中文检索研究进行分析，总结出了中文信息检索中存在的一些问题，其中主要问题之一就是关键字（词）的索引问题（Keyword Indexing）。这是检索的关键技术问题。在传统的索引方法中，人们通常采用 Inverted Files、Signature Files 方法，然后研究者们又对中文使用了 Pat－Tree 方法进行索引，对这种方法台湾学者研究的比较多，使用这一结构，台湾中央研究院资讯科学研究所开发了 Csmart 系统；这些方法都是完成对文本进行匹配查找；后来又使用自然语言学处理的方法，把向量空间模型（Vector Space Model）引入信息检索里面来。这种模型实际上是对 Inverted Files 的一种模仿，但这种模型又比 Inverted Index 灵活，可以单独计算每一个索引项（字、词）的权重，克服了传统的词索引所不能解决的问题[2]。本文主要采用向量空间模型的索引技术，使用两种不同的相似度评测方法来对 10 字以内的短查询进行实验。本文后面的部分将主要介绍以下内容：向量空间模型索引、两种不同的相似度评测算法、实验结果的比较分析。

2　向量空间模型（VSM）

向量空间模型（Vector Space Model）由 Salton 等人于 60 年代末提出，主要思想是通过使用空间的相似性可以用来解决语义上的相似性。这种方法在原理上很简单，就是把文档和查询在高维空间中用向量表示出来，即把文档集合中的每篇文档都形式化为高维空间中的一个向量，同样把每个查询都形式化为高维空间中的一个向量，每一个维对应着文档集合中的一个词。从理论上讲，和查询向量越接近，向量间的夹角越小的文档向量所代表的文档就和查询的相似度越高，越接近查询要求。

2.1　文本特征的表示

文本特征是指关于文本元数据信息，如作者、机构、内容等[5]。对于如作者、机构等信息则是易于获取和表示的，这里作者主要讨论对于文本内容的表示。对于一篇文本，我们可以认为它是由字、词、词组或短语组成。这些字、词、词组或短语，我们称之为文本的特征项。因而，我们可以将文本抽象为：$V(d_i)=((t_1,w_1(d_i)),\cdots\cdots(t_i,w_i(d_i))\cdots\cdots,(t_n,w_n(d_i)))$，其中，$t_i$ 是特征项（term），可以是字、词、短语及其他语言单位，$w_i(d_i)$ 是 t_i 在文档 d_i 中的权值。

对于一个训练文本集合，我们就可以得到如图 1 所示的一个向量空间：

其中，d_i 表示文档，t_i 表示权值，$w_{ij}(d_i)$ 主要是通过统计特征项 t_j 在文档 d_i 中出现的次数来表示，W 通常是一个稀疏矩阵，然后根据所采用的相似度评估算法再进行相应的处理。

2.2　文本特征的选取

文本特征项的选取，有多种选择，可以选择字、词（或汉字串，n－Gram）作为特征项。一般情况下，人们普篇认为以单字（1－Gram）作为特征项对文档内容的代表性不如以词语（n－Gram）作为特征项对文档内容的代表性强。因此，在大多数情况下，词语会被选作特征项。现在对词语的获取方法很多，有人通过使用互信息、信息熵等手段对文档中的词进行抽取，这样就可以抽出对该文档内容最具表现力的若干条词语作为该文档的特征项。鉴于这种方法现在还不能达到人们所期望的程度，特别是对中文的特征项抽取，现在的准确率也只是在 70% 左右。获取文本特征项当前使用较多的方法还是用分词的方法来进行，然后，再利用一些过滤手段对特征项进行筛选，最终得到一组具有比较强的代表性的特征项。

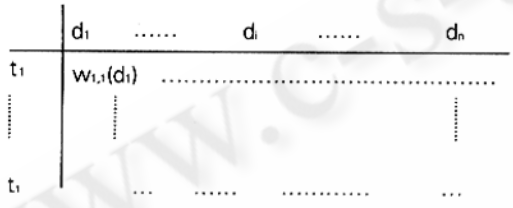

图 1

本文采用分词的方法，先对训练样本进行分词处理，然后进行过滤掉一些停用字和长度大于 5 的词（汉字串），总共得到 35982 个特征项，以此作为特征项集。

3　Cosine－Measure 与 OKAPI－Measure

在信息检索中，通常要计算查询向量与文档向量的相似度来评估查询与文档的相关性，从而根据相似度的高低来决定检索结果。本文采用两种方法计算相似度来对查询题文与文档之间的相关性进行评估。

3.1 Cosine Measure

在 2.1 节中对训练文本集进行了处理得到训练文本集向量空间称之为“特征项－文档”（Term－Document）矩阵空间 Wtxd，在使用 Cosine－Measure 之前要用 TF－IDF 形式对原始词频矩阵进行处理：设原始词频矩阵（在向量空间模型中，最初构造的词频矩阵即为原始词频矩阵。）为 W_{txd}，$w_{i,j}$ 为其非零元素，为了更好地表示词 t_i 在 d_j 中的重要性，则用下式对其进行标准化处理：

$$w_{i,j}=\frac{local(i,j)*Global(i)}{\sqrt{\sum_{i=1}^{t}(Local(i,j)*Global(i))^2}} \quad (1)$$

其中，$Local(i,j)=\log_2(1+tf_{i,j})$，$tf_{i,j}$ 表示特征项 i 在文本 j 中的出现频率；

$Global(i)=\log_2((n/df_i)+1)$，n 表示训练集文本总数，$df_i$ 表示出现特征项 i 的文本数。同样，我们对查询也要进行相同的处理。我们把查询看成是一个文档，那么也可以得到一个查询向量 Q，用与训练集相同的方法进行标准化后得到最终的查询向量 q. 处理完后，使用 Cosine－Measure 进行相似度计算，整个检索的步骤大致如下：

Step1：处理训练集文档，求得训练集矩阵空间；

Step2：当新的查询到来，对其进行分词处理，求得特征向量 并对其进行标准化处理；

Step3：利用如下公式计算查询向量与训练文档的相似度：

$$Cos(q,d)=\sum_{t\in q\wedge t\in d}(w_{q,t}\times w_{d,t})/\sqrt{(\sum_{t\in q}w_{q,t}^2)\times(\sum_{t\in d}w_{d,t}^2)} \quad (2)$$

其中，$w_{x,t}$ 表示 t 在 x 中的权重，q 表示查询向量，d 表示训练样本向量

Step4：对相似度按降序排序，将相似度排序靠前的文档作为检索结果输出。

3.2 OKAPi－Measure

当我们使用 OKAPI[4] 方法来评估查询与文档的相似度时，则不需对原始词频矩阵做标准化处理，只需使用下式进行相似度的计算：

$$OKA(q,d)=\sum_{t\in q\wedge t\in d}\left[\left(f_{q,t}/l_q\right)\cdot\log\left(\frac{N-f_t}{f_t}\right)\left(\frac{f_{d,t}}{f_{d,t}+\sqrt{f_d}/av\sqrt{f_d}}\right)\right] \quad (3)$$

其中，$f_{x,t}$ 表示 t 在 x 中的频率，f_t 是训练集包含 t 的文档总数，f_d 表示 d 中包含的特征项数，l_q 表示查询的长度，N 表示训练集的文档数。

其整个检索步骤与 3.1 节中的基本相同，只是计算相似度算法有所不同，在这里用（3）式进行计算，然后再将检索结果输出。

4　实验分析

本文的实验是从网上下载网页经人工分析、归类

后作为训练语料。整理后得到 11 个相关主题文档，共 550 篇作为训练集进行实验，每个文档大约 2000 字左右。本文在实验时针对各个主题设计了查询，每个查询的汉字数不超过 10 个，经分词后部分查询如下：

(1) 音乐/ 的/ 教育/ 和/ 欣赏/

(2) 思想/ 政治/ 哲学/ 的/ 探讨/

(3) 生长/ 发育/ 卫生/

(4) 军事/ 兵法/ 战/ 策/

(5) 数学/ 命题/ 与/ 物理/ 理论/

(6) 动/ 、/ 植物/ 的/ 生化/ 反应/

(7) 天体/ 物理/ 及/ 地理学/

(8) 海洋生物/ 生长/ 与/ 繁殖/

(9) 果树/ 的/ 栽种/培育/

.....

经过使用第 3 节所描述的相似度评测算法得到结果如图 2 和表 1 所示。通过对上表中的数据进行比较，我们不难看出：

① 召回率(即实际检索出的相关信息文档数与信息库中的相关信息文档数之比)与准确率大致呈现一种朝相反方向发展的趋势，召回率高时准确率低，准确率高时召回率低。

② 总体上来看，在相同召回率的情况下，OKAPI 法要比 Cosine 法的准确率要高一些，这说明在(10 字以内)短查询中用 OKAPI 相似度评测法要比 Cosine 法总体上要好些。

③ 由于本实验选用了处理比较好的平衡语料库的样本作为训练样本，特征相对而言还是比较明显，实验的效果还是很明显的。

表 1　Cosine与OKAPI相似度评测在相同召回率情况下的准确率比较

Recall	Precision										Average Precision
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	
Cosine	0.8547	0.7784	0.7425	0.7045	0.6746	0.6301	0.5865	0.5491	0.4935	0.4260	0.5855
OKA	0.9491	0.8861	0.8359	0.770	0.7233	0.5960	0.5658	0.5721	0.4516	0.4155	0.6250

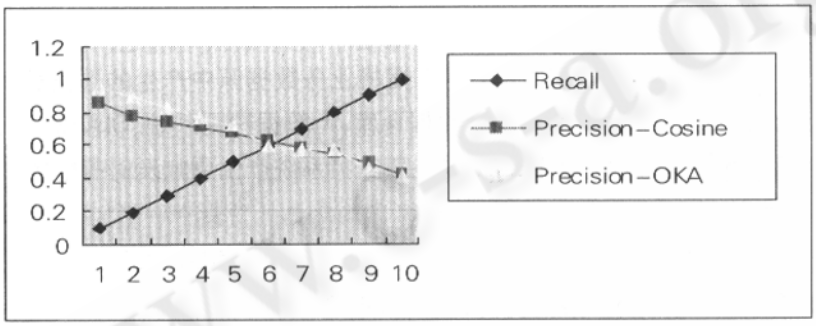

图 2　准确率与查全率变化图

5　结束语

本文采用空间向量模型作为索引模型，使用 Cosine 和 OKAPI 两种不同的相似度计算方法，来实现对中文短查询进行检索实验。通过对比发现，针对短查询检索，OKAPI 相似度评测法要比 Cosine 法的效果要好。本文主要是对短查询进行了分析比较，在以后的工作中，将对各种不同长度的查询以及查询算法进行分析比较。

参考文献

1 Lee - Feng Chien, Hsiao - Tieh Pu Important Issues on Chiese Information Retrieval Computational Linguistics and Chinese Language Processing vol. 1, no. 1, August 1996, pp205 - 221.

2 T. A. Letsche and Michael W. Berry Large - Scale Information Retrieval with Latent Semantic Indexing Information Science Applications , Volume 100, Number 1, August 1997, pp. 105 - 137.

3 C. D. Manning H. Schutze Foundations of Statistical Natural Language Processing The MIT Press Cambridge, Massachusetts London, England, 1999.

4 Ross Wilkinson, Justin Zobel, Ron Sacks - Davis Similarity Measures for Short Queries . SIAM Review, 37(4):573 - 595, 1995.

5 史忠植著，知识发现[M]，清华大学出版社，2002.