基于概念空间的文本语义索引*)

Text Semantic Indexing Based on Concept Space

李 源[1] 郑 毅[2] 何 清[2] 史忠植[2]

(中国科技大学研究生院计算机学部 北京 100039)[1]

(中科院计算技术研究所智能信息处理开放实验室 北京 100080)[2]

Abstract A new method of using cluster analysis to establish a text semantic indexing based on concept space is presented in order to get the needed information promptly from a great number of text documents. The keywords with high semantic proximity are direct clustered, an indexing based on the semantic proximity between keywords is set up on text concept space, and a strategy of indexing expand is presented. Therefore corresponding text files can be obtained after users input keywords. According to the results of experimentation, it is clear that the indexing efficiency and accuracy have been improved using this method.

Keywords Text indexing, Concept space, Cluster analysis, Direct clustering, Semantic proximity

1 引言

据统计,在现今的联机存储信息中,80%以上的信息以文本的形式存在。信息的多元化、复杂化,致使信息的自动索引成为急需解决的问题。本文研究的内容是建立一个基于概念空间的文本语义索引。目前的文本索引都是建立在文本空间,或关键词空间上的,而建立在概念空间上的索引具有条理清晰、人机界面友好、符合通常检索习惯等许多优势,这也是文本语义索引发展的方向。另外,在建立文本索引的过程中,国内外大多使用 Hopfield 神经网络联想的方法[5,6,8],本文首次使用直接聚类法代替了 Hopfield 神经网络联想功能,这样使得索引具有很好的可扩展性。基于语义关联度的文本索引可以广泛应用于 Internet 搜索引擎、数字图书馆、电子商务等众多领域中。建立文本索引的过程主要有以下几部分:

1)对文档分类,建立文档的概念空间,在概念空间的层次上组织文档并确定文档中出现的关键词。

2)分析概念空间中的文档,计算关键词的语义关联矩阵,标识出任意两个单词的语义关联度。

3)对关键词根据语义相近的原则聚类,建立分层索引。采用直接聚类法对关键词聚类,返回同类关键词共现率较高的文档。

为了对文本索引的准确性以及算法的可行性进行验证,选取 1061 篇足球类文档,对文档切词、分词后得到 876 个关键词,在此基础上建立起一个完整的基于概念空间的文本语义索引,并进行了检索试验。实验证明,这种索引能够大大提高用户的检索效率和准确度。

2 文档概念空间

文档的概念空间是指对文档自然聚类后所得的一种概念的层次结构,简单地说是一个树型的结构。从一个根节点依次地打开,逐层得到最终的叶子节点,形成一个空间的结构。它的叶子节点是一篇篇文档,各层的根节点是一个个从文档中提取出的概念,也就是一个个能概括其子节点共同特征的关键词。概念空间与目录的性质有些相同,但它是机器通过学习,由下而上自动生成,而不是由人工逐层分类、自上而下构成的。生成概念空间的方法如下:

1)采用传统的向量空间模型的方法来实现文本的数学抽象。

*)本文得到国家自然科学基金资助(课题号 60073019、69803010)。李 源 研究生,主要研究方向:人工智能。

验。Zeillinger 使用光子作为他的研究对象,并使光子在他们实验室的一边移动到另一边,但距离仅为一米。然而三年多以后,Zeillinger 等已经工作在第二步了,他们把光子传送了超过一公里之遥。

在 Zeillinger 等人的突破之后,Cirac 和 Zoller 提出,远程移动可以变成量子互联网的一个基础。2000 年 3 月,麻省理工学院的 Seth Lloyd 和 Selim Shahriar 以及马萨诸塞州林肯市的美国空军实验室的 Phillip Hemmer 提议,在光纤上传送缠绕的光子到包含冻原子的节点上,这些原子将吸收光子,因此将保存缠绕。这个缠绕然后可用于误差校正、远程移动以及许多其它有价值的应用。一些研究小组正在对这个思想开展研究,其中包括加州理工学院的 Jeff Kimble 和在洛杉矶的加州大学的 Eli Yablonovitch。他们希望,在 10 年内有包含三个节点的量子计算机网络投入运行。因此,我们作如下展望:

10 年内,肯定将会有真正意义下的量子计算机问世,人们将不会像现在这样对于量子计算机感到陌生。

然而,由于对于量子计算机的使用要求全新的程序设计技术,如前边所说的量子算法,误差校正算法,特别是真正用于解决确定问题的算法,因此它的普及应用将需要更长得多的时间。

量子互联网和分布式量子计算机系统,是使量子计算机系统走向普及达到当前通常计算机发展水平的必经之路。

参 考 文 献

1 Knuth D E. The Art. of Computer Programming. Addison-Wesley. 1998,2

2 Mullins J. The topsy turvy world of quantum computing. IEEE Spectrum,2001(2):42~49

2)采用自组织特征映射 SOM 算法(一种无监督的自学习过程)进行文本聚类,将每一篇的文本归入所属的类别,并完成类别概念的抽取,而每篇文档都作为最底层的叶子节点。

3)在生成了概念空间的底层后,我们可以通过对中层的类别再实行概念归并,形成概念空间。归并将不涉及到文本,而是对 SOM 生成的各个类别的模板向量直接进行聚类操作。

例如:从一批体育类文档中,先通过 SOM 算法提取出底层概念,如国安、申花、罗马等,再完成概念的归并,比如将国安、申花归并为甲 A 联赛这一中层概念,再对中层概念归并,最终形成以下的概念空间,如图 1 所示。

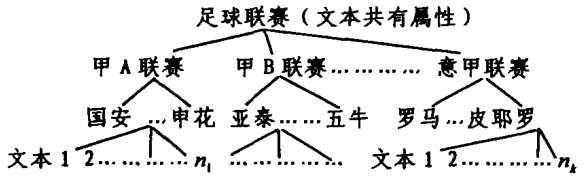

图 1 文档概念空间

与普通的在文档空间上直接建立索引相比,在概念空间上建立索引具有许多优势。由于概念空间已对文本做了大致的分类,因此可以避免聚类过程中的噪音干扰,在此基础上建立的索引其准确性就可以得到保证。更重要的是,可以在概念空间的层次结构上,分层次地建立起索引,分层次的索引具有很多优点:

1)分层次的索引在建立时将大批文档分批处理,生成索引,因此减少了总的计算量,当文档数量、结构发生变化,概念空间的部分结构随之变化时,只需对有变化的那类文档更新索引,而无须对其它索引重新计算。这样使索引具有很好的扩展性,能实时地反映文档的变化,具有这个特点的索引特别适用于文档更新极快的 Internet 搜索引擎。

2)分层次的索引可提供友好的用户界面,由于索引是有组织有层次,用户可以根据自己所需的查全率要求,控制检索面向文档的范围,例如,在上面的概念空间中,用户可以很容易地查到与关键词相关的甲 A 类文档或意甲类的文档。

3 语义关联权计算与关联矩阵

要建立索引,首先要分析文档,使用簇分析技术建立一个关键词的语义关联权关联矩阵[1,2]。建立关键词的语义关联权关联矩阵是为了标识任意两个单词一起出现的可能概率(Co-occurrence probability),它可以用一个反映关键词之间的语义关联度的模糊自反矩阵来表示。只有该矩阵能准确地反映出在这批文档中词与词之间的语义关联度,聚类的结果才能与实际相符,索引才能准确。簇分析的主要过程简述如下:

首先计算文档关联权 d_{ij}(单词 j 和文档 i 的关联权)。为了筛选出代表大多数文档共性的索引,则

$$d_{ij}=tf_{ij}\log df_j \tag{1}$$

如果为了筛选出那些代表各个文档独特之处的索引,则

$$d_{ij}=tf_{ij}\log(\frac{N}{df_i}w_j) \tag{2}$$

式中:tf_{ij}为单词频率(索引单词 j 在文档 i 中出现的次数),df_j 为文档频率(文档空间中包含索引单词 j 的文档的数目);N 为文档空间中的文档总数;W_j 为索引中的单词数。

接着将原始数据(单词频率和文档关联权)转换为关联矩阵。由于单词 T_j 到 T_k 的关联权不等于从单词 T_k 到 T_j 的关联权,所以采用下列非对称簇分析函数计算单词的关联权,即

$$clusterWeight(T_j,T_k)=\sum_{i=1}^{n}d_{ij,k}/\sum_{i=1}^{n}d_{ij}$$

$$clusterWeight(T_k,T_j)=\sum_{i=1}^{n}d_{ij,k}/\sum_{i=1}^{n}d_{ik} \tag{3}$$

以上两式表示从单词 T_j 到 T_k 的关联权和从单词 T_k 到 T_j 的关联权。在文档 i 中同时出现的索引到 T_j、T_k 和文档 i 的关联权值为

$$d_{ij,k}=tf_{ij,k}\log(\frac{N}{df_{j,k}}W_j) \tag{4}$$

式中:$tf_{ij,k}$为索引 j 或 k 在指定文档 i 中出现次数最少的数目,若在文档 i 中索引 j 出现了 3 次,而索引 k 出现了 5 次,则 $tf_{ij,k}=\min(3,5)=3$;$df_{j,k}$为在 N 个文档组成的文档空间中同时包含索引 j、k 的文档数;W_j 为索引单词中的单词数,本文取 1。

通过以上的计算,得到的矩阵就是关键词的语义相关权关联矩阵。

4 直接聚类法

经过簇分析处理后,原始文档和关键词被处理为具有关联权的矩阵。建立索引的过程的实质和聚类方法的实质相一致,采用对权关联矩阵直接聚类的方法[3]。在关联权的矩阵中对关键词聚类,找到联系最紧密关键词,形成多个关键词类。

以下举一个具体的例子说明聚类的过程:

如图 2 所示 $S=(S_{ij})_{n\times n}$为一个文档和关键词经簇分析处理后得到的关键词权关联矩阵,S_{ij}表示从单词 i 到 j 的关联权,关联权值越大,表明从单词 i 到 j 的关系越紧密,S 具有以下特点:

$S_{ii}=1;i=1,2,\cdots,n;S_{ij}\leqslant 1;i,j=1,2,\cdots,n$。

S 是一个模糊自反矩阵,可直接聚类,步骤如下:

1)对 S 取一个域值 $\lambda=0.81$ $(0\leqslant\lambda\leqslant 1)$,对矩阵进行等价划分(大于等于 λ 者转化为 1,小于 λ 者转化为 0),得到矩阵 S_1。

2)将 S_1 对应为一个有向图 G_1 如图 3 所示。规则如下:S_1 的行向量 S_i 对应顶点 x_i,若 $S_{ij}=1$,则从 x_i 到 x_j 有一条矢线。若 $S_{ij}=0$,则没有。

3)对有向图 G_1 分析,如果出现了回路,那么将回路各节点合并成类,若两回路有公共节点,则将两回路合并,由于 x_i 对应的就是关键词 I,这样就完成了在域值 $\lambda=0.81$ 的情况下的对关键词的直接聚类。聚类结果如下:$\{X_1,X_2,X_3,X_4\}$,$\{X_5\}$,$\{X_6,X_7\}$。

$$S=\begin{bmatrix}1 & 1 & 0.3 & 0.4 & 0.5 & 0 & 0.1\\ 0.8 & 1 & 0.9 & 0.1 & 0 & 0.2 & 0.8\\ 1 & 0.8 & 1 & 0.9 & 1 & 0.6 & 0.2\\ 0.8 & 1 & 0.7 & 1 & 0 & 0.1 & 0\\ 0.1 & 0.2 & 0.6 & 0.3 & 1 & 0.4 & 0.6\\ 0.5 & 0 & 0.4 & 0.4 & 1 & 1 & 1\\ 0.6 & 0.2 & 0 & 0.1 & 0.2 & 0.8 & 1\end{bmatrix}$$

$$S_1=\begin{bmatrix}1 & 1 & 0 & 0 & 0 & 0 & 0\\ 1 & 1 & 1 & 0 & 0 & 0 & 1\\ 1 & 1 & 1 & 1 & 1 & 0 & 0\\ 1 & 1 & 0 & 1 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 0 & 1 & 1 & 1\\ 0 & 0 & 0 & 0 & 0 & 1 & 1\end{bmatrix}$$

图 2 关键词权关联矩阵及其等价划分矩阵

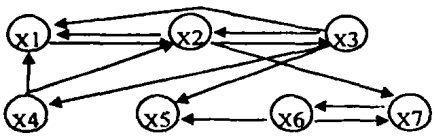

图 3　矩阵对应有向图

目前国内外对权关联矩阵主要采用 Hopfield 神经网络算法进行联想记忆，这种方法的一个缺点就是网络不具有扩展性，如增加一个节点，则整个计算结果都要重新计算，无法利用原有的计算结果。而采用直接聚类法，在增加节点后，只需计算是否有新增的回路，并与已有的回路查找是否有公共节点，再做相应的归类处理，而原有的分类结果基本不用改变，使整个算法具有了可扩展性。

5　试验结果及分析

为了对文本索引的准确性以及算法的可行性进行验证，进行了以下试验。在概念空间上选取足球子类，其中包含有甲A联赛、意甲联赛、世界杯等多个子类，叶子节点共有 1061 篇相关文档，这其中又包含有 876 个关键词，包括了大量的球队、球员名称，地名等经常被查询的词汇，使用簇分析的方法对这批文档和关键词处理，得到关键词权关联矩阵：

$S=(S_{ij})_{876\times876}(0\leqslant i,j\leqslant876)$

由于 S_{ij} 值普遍较小，经反复试验，最后选定域值 $\lambda=0.25$，直接聚类后，得到以下结果：

关键词共 876 个，每个节点代表一个关键词，节点的分布大致如下：

类中含节点个数	类别数	说明
20<N	1	该类中有关键词 180 个，是最大的一个类
10<N≤20	1	
5<N≤10	22	
2<N≤5	57	
N=2	102	
N=1	88	这一部分词与其他词没有构成类

由于存在一个尚未分开的大的类别，因此将此类别的节点抽取出来，构成一个子矩阵：

$S_{21}=(S_{ij})_{180\times180}(0\leqslant i,j\leqslant180)$

由于域值的选取会直接影响到直接聚类的结果，当域值 λ 越小，聚出的类的数目也越少，例如当 $\lambda=0$ 时，类的数目为 1。因此，对于 S_{21}，要选取更大的域值，再次直接聚类，试验中取 $\lambda=0.38$，对这 180 个关键词再聚类，结果如下：

类中含节点个数	类别数	说明
10<N	1	
5<N≤10	8	
2<N≤5	16	
N=2	18	
N=1	16	这一部分词与其他词没有构成类

将此结果与第一次聚类的结果和在一起就建立起基于关键词语义关联度的索引。结果如下：

Cluster1：北京、北京国安、大连实德、山东鲁能…

Cluster2：英超、阿森纳、利物浦、福勒、曼联、福格森

Cluster3：西班牙、意大利、世界杯、德国

Cluster4：罗马、AC 米兰、犹文图斯、佩鲁贾、中田英寿

…

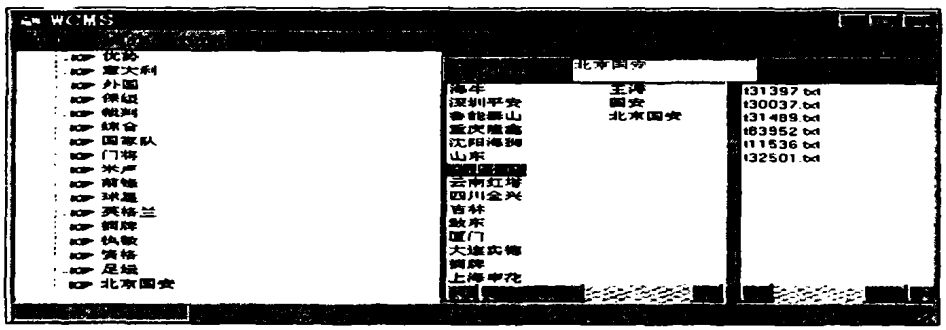

图 4　试验结果

如图 4 所示，中间的一列词为与“北京国安”聚为一类的词，右侧为相关文档，由此可见，绝大部分类中关键词的语义相关度非常高，符合实际情况，在对一些原文档进行了对照后发现，对照的结果很令人满意，索引的准确性得到了验证。

同时还对未被归入任何类的近 100 个关键词进行了研究，通过对文档的仔细观察，发现绝大多数未被归类的词仅在几篇文档中出现过很少的几次，未能与其他关键词构成联想。这些词都不是检索时的常用关键词，因此，索引的查全率也比较理想。

参 考 文 献

1　Salton G. Automatic Text Processing. Addison-Wesley Publishing Company, Inc., Reading, MA, 1989

2　Salton G, Allan J, Buckley C. Automatic structuring and retreval of large text files. CACM, 1994, 37(2): 97～108

3　张乃尧，阎平凡. 神经网络与模糊控制[M]. 北京：清华大学出版社，1998

4　Chen H, Hsu P, Orwig R, Hoopes L, Nunamaker J F. Automatic concept classification of text from electronic meetings. Communications of the ACM, 1994, 37(10): 56～73

5　Chen H, Lynch K J. Automatic construction of networks of concepts characterizing document databases. IEEE Transactions on Systems, Man and Cybernetics, 1992, 22(5): 885～902

6　Chen H, et al. A Concept Space Approach to Addressing the Vocabulary Problem in Scientific Information Retrieval: An Experiment on the Worm Community System. J. American Soc. Information Science, 1997, 48(Jan.): 17～31

7　何清，等. 基于因素空间和模糊聚类的概念形成方法. 系统工程理论与实践，1999, 19(8): 99～104

8　刁倩，等. 基于神经网络的中文的信息概念联想构造算法. 情报学报，2000, 19(4): 170～175