科技文献关键词冗余解决方案研究

邢美凤[1,2,3]

[1]（中国科学院国家科学图书馆 北京 100190）

[2]（中国科学院研究生院 北京 100049）

[3]（晋中学院图书馆 晋中 030600）

摘要：

由于科技文献中作者选用关键词不规范，经常会造成同一研究主题下关键词的冗余。针对这一问题，本文提出了一种改进的基于相似度计算的科技文献关键词选取算法。先利用 n-gram 算法提取领域词库，再综合利用领域词库和常识词库，对最初选的关键词重新切分，进行给定关键词之间的语义对比。语义相似度大于一定阈值的关键词认为是表达同一意义的同义词，将同义词在文献库中合并，解决了关键词冗余问题。实验结果表明该方法的有效性。

关键词：

科技文献关键词；冗余；解决方案；语义相似度；特征降维

Study on solution to redundancy of Scientific literature Keywords

Xing Meifeng[1,2,3]

[1] (National Science Library, Chinese Academy of Sciences, Beijing 100190, China)

[2] (Graduate University of Chinese Academy of Sciences, Bejing 100049, China)

[3] (JinZhong University Library, Jinzhong 030600, China)

Abstraction：

Irregular keywords that authors gave are often made in some scientific literature, which cause high redundancy in the same research topic. To address the issue, this paper proposed an improved keywords selection algorithm based on similarity calculation. Re-segmented keywords using field dictionary and common-sense knowledge database thesaurus. When the total semantic similarity is greater than a given threshold, the two compared keywords is considered to express the same meaning. Merge and keep only one of them in Library. Achieve the purpose of the dimension reduction. Experimental results show the effectiveness of the method.

Keywords：

Scientific literature keywords; Redundancy; Solution; Semantic similarity; Feature Reduction

引言

利用作者给定关键词进行科学研究的过程中，由于关键词数量庞大，经常要截取词频较大的一部分进行分析。这种方法有一定的科学依据，但由于作者给定关键词不规范，同一意义在关键词中会以多个形式出现，以词频的方式选取关键词会丢失大量有用的信息。如果在利用词频处理之前，先进行关键词之间的合并或修正，将关键词无损地压缩，就可以最大程度地保留所要分析的信息内容。

1 研究背景

本文的方法归结为语义特征降维。将作者给定的同一主题关键词组成最初的高维特征向量，选用一种合适的方式识别和合并同义词，实现关键词向量空间的有效降维。

目前，基于语义的特征降维主要利用一些常识知识库如 wordnet，《知网》，《同义词词林》等进行。chua[1] 和 li[2]利用 wordnet 中提供的同义关系，上下位关系，部分、整体关系等计算词与词之间的语义相似度进行特征降维。也有一些研究人员[3][4]将语义知识词典《知网》[5]应用于中文文本表示降维研究。还有一些研究人员利用同义词词林进行同义词合并[6]，实现特征降维。以上的研究对象主要是新闻文稿和社会科学领域的文档，利用常识知识库可以有效完成文本特征降维任务。但是，自然科学领域的科技文献会涉及到大量的领域词汇，这些词汇大都没有被常识知识库收录，仅利用这些常识知识库对科技文献进行相关的特征降维研究，效果并不是很好。

本文提出一种利用 n-gram 算法获取领域词库，结合《知网》进行相似度计算，达到对科技文献关键词进行特征降维的目的。

2 问题及解决方法

2.1 科技文献关键词的特点及问题

(1) 相同意义的词汇写法多样，如算法和方法，原理和定理，选择、选取和筛选，量和值等，这些词都可以利用常识性词典如《知网》或 wordnet 通过语义计算后达到无损特征降维。

(2) 对于科技文献，很多专业词汇是几个一般词汇的组合，组合在一起表达一个完整的意义。这些专业词汇是某个领域的专用词，没有必要再进行切分，但这些词在常识性知识库中并没有收录，所以直接利用常识性词典进行语义计算不能完全解决实际问题。

(3) 有些领域词汇没有规范写法，一个意义的表示方法可能有好几种，这几种都有可能会用在关键词中，如“因特网”、“互联网”和“internet”等。也有一些是简写和正式写法混用，如“自然语言处理”和“NLP”、“词频”和“TF”、“倒排文档频”和“IDF”等。这些关键词出现的频率几乎不相上下，在利用词频的方法进行特征选择时，两者都有可能被选入，错认为是两个独立的特征项。造成关键词向量的冗余。这些词合并后都可以无损地进行特征降维。但这些词在常识性词典中没有收录，直接利用常识性词典不能完全解决相似问题。

2.2 上述问题的解决办法及本文所提出的特征降维方法

为解决关键词标引中出现的问题，国家标准 GB/T 7713.1-2006[7]（学位论文编写规则）中要求每篇论文应“选取 3～8 个词作为关键词”，“尽量用词表提供的规范词”。许多期刊要求按照标准进行关键词的标引。但随着信息技术的飞速发展，词表不能及时跟上飞跃发展的科学技术，难以满足各专业的要求。为此，一些研究人员提出了提高关键词标引质量的方法：马开俊[8]提出“有控制的关键词标引”，谭慧华[9]提出“论文作者标引与专业标引人员标引相结合的方法”；郭淑敏[10]提出利用辅助词表来提高关键词标引的质量。赵宗蔚[11]提出“采用自然语言与人工语言结合的后控制词表来提高期刊论文关键词标引质量等。

针对由于关键词标引不当造成的冗余问题，本文提出一种新的关键词冗余解决办法。主要思想为：根据关键词冗余问题的特点，对已有的语义特征降维方法进行改进。在分析词与词之间的语义相似度时加入了提取领域词的过程；先进行简写和正式写法的合并，然后利用 N-gram 过滤算法构建领域词库。在计算词与词之间的语义相似度时，领域词作为整体来考虑，最后将计算所得的同义词加到同义词词典中，并且将这些同义词在向量空间中合并，实现关键词向量的特征降维。

利用 java 语言实现所提出的方法。实验部分下载 CNKI 某一领域文献关键词。利用本文所提出的方法进行领域词库的构建。对原有关键词进行分词，并区分为领域词和非领域词。对这两种词分别进行计算。关键词之间的语义相似度是几个部分相似度的加权结果。计算结果大于一定阈值，则认为是表达同一意义的同义词。将这些同义词分别放到同义词词典中，并且在原有的向量空间中合并，以达到利用特征降维方法消除冗余关键词的目标。

3 主要算法具体完成过程

通过在同义词词典中加入领域术语简写和正式写法的同义词项，实现对这些意义相同的领域术语简写和正式写法的合并。针对领域术语的特点，选择具体的 n-gram 候选词条的长度范围，按照过滤效果为不同长度的候选词设置不同的阈值，大于阈值的词被选入领域词库。根据本文建立的领域词库将关键词分为不同的组，同一个组的关键词之间进行对比，组与组之间不再进行对比。在组内进行对比过程中，首先利用领域词库将关键词切分为三个部分：领域前缀、领域词和领域后缀，然后分别计算两两关键词之间领域前缀部分和领域后缀部分的相似度，综合得出关键词之间的相似度。给定一个阈值，将大于阈值的这些词视为同义词，选取其中一个作为标准词，其余放入同义词典，达到特征降维的目的。

3.1 同义词词典生成

同义词词典最初存放的是领域术语的简写和正式写法的同义词对照表。计算后所得的同义词也存入该词典中，为后续研究使用。同义词词典格式为每个同义词集合各占一行，最后一个为标准写法。同义词之间用固定分隔号分隔。本文使用的同义词合并算法流程如图 1 所示：

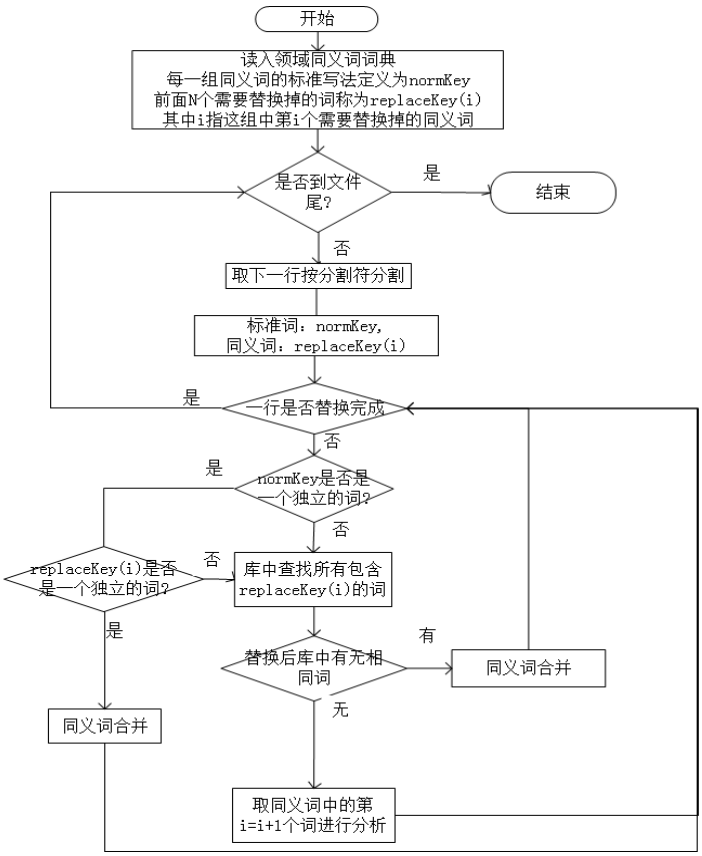

图 1:同义词合并算法流程

3.2 领域词库的确定

n-gram 是应用很广的统计语言模型，在语音识别[12]、机器翻译[13]、手写识别[14]、拼音输入[15]、信息检索[16]、分词和词性标注[17]等许多应用领域都取得了较大的成功。本文利用 n-gram 算法确定领域词库。将关键词切分为 2-gram, 3-gram,…，n-gram，具体最大值 n 的确定视领域术语的特点而定。利用过滤算法考察这些 n- gram 成为领域词汇的可能性，最终确定领域词库。领域词库生成算法如下所示：

输入：作者所给的关键词列表 AK

输出：领域词库

中间参数:n-gram 片断中最大的词长为 n

对于每个 n-gram 列表，粗过滤词频值为α_n

利用 n-gram 列表进一步过滤子串的参数为β_n

① 对作者所给的关键词列表 AK 进行 n-gram 切分；

② 将长度相同的 n-gram 切分词组成一个列表，把这个列表中频率值小于α_n的词过滤掉，形成 n-1 个 n-gram 列表（分别为 2-gram,3-gram,…,n-gram 列表）；

③ 对于每一个列表，取每个候选词的词频θ；

④ 分别取候选词子串中的每个词及词频θ'；

⑤ 如果$\theta' - \theta < \beta_n$时，在子串中过滤掉该词，如图 2 所示，当每一个候选词和候选子串中对应的词的比较阈值β_n都定义为 3 时，“支持向量机”，“支持向量”和“向量”被作为领域词，其它的子串被过滤掉；

⑥ 将 n-gram 列表中没有执行过滤操作的词选作领域词汇。

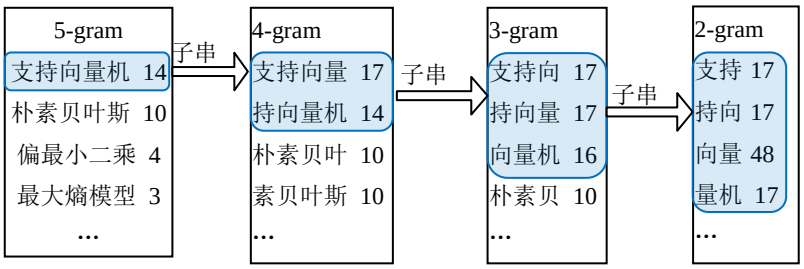

图 2：n-gram 词串以及每个词对应子串

3.3 语义相似度计算方法

利用领域词库对关键词进行切词；利用一般分词程序对上一步的非领域词部分进行重新切分；切词后，将词表示为三个部分，分别为领域前缀、领域词和领域后缀。领域前缀指其它词位于领域词的左边，领域后缀指其它词位于领域词的右边。

如：“树状贝叶斯方法”和“树形贝叶斯理论”二个关键词分别切分为：

树状贝叶斯方法= 树状 + 贝叶斯 + 方法

树形贝叶斯理论= 树形 + 贝叶斯 + 理论

作者所给关键词= 领域前缀 + 领域词 + 领域后缀

利用知网计算非领域词的语义相似度：设关键词 W_1 经过切分后为三个部分 W_{1l} 、W_{1f} 、W_{1r} ，关键词 W_2 经过切分后为三个部分 W_{2l} W_{2f}、W_{2r}，其中 W_{1f} 和 W_{2f} 分别对应领域词。比较时，领域词和领域词相比较，领域前缀和领域前缀相比较，领域后缀和领域后缀相比较。前缀和后缀的比较利用常识库的相似度计算方法进行。只有当领域词相同时，才进行比较；领域词不相同时，认为是相似度很小的词，没有必要再进行前后缀的比较。本文提出用来计算两个关键词的语义相似度公式为：

$$S(W_1,W_2)=\begin{cases}\alpha\ S(W_{1l},W_{2l})+\gamma\ S(W_{1r},W_{2r}) & \text{当}W_{1f}=W_{2f}\text{时 其中}\alpha,\gamma\text{ 为引入的调节参数，}\alpha+\gamma=1\\ 0 & \text{当}W_{1f}\neq W_{2f}\text{ 时}\end{cases}$$

3.4 特征降维的方法

所给定的关键词没有必要一一对比，只有当分词后领域词一致时，这两个关键词才有可比性。计算每个领域词包含前后缀的相似矩阵，然后利用所给的相似性公式计算词与词之间的语义相似度，将语义相似度大于一定阈值的关键词视为同义词，将这些同义词存放到同义词词典中，以备以后计算时用。将这些同义词合并，完成特征降维过程。

特征降维的具体实现算法：

输入：作者所给的关键词列表 AK

输出：降维以后的关键词列表

中间辅助变量：每个领域词对应一个相似矩阵，这个矩阵中存放所有与本领域词相关的前后缀的语义相似度值。在关键词列表 AK 中为每个关键词设删除标志位。

① 利用相似矩阵，分别计算两个关键词对应的前后缀的相似度，再利用本文 3.3 所提出的语义相似度计算方法加权计算整个相似度，最终所有关键词之间的语义相似度存放到 N×(N-1)维的二维数组中，其中每一行数据存放某个关键词与其它 N-1 个关键词的语义相似度；

② 给定一个循环变量 i，初始赋 i=1；

③ 判断 i 值是否大于 N，如果是，执行第⑧步操作，否则取第 i 个关键词的删除标志位以及在二维数组中的第 i 行数据；

④ 首先判断这个关键词删除标志位是否为 1，如果是，说明这个词已被认定为其它词的同义词，没有必要进一步对比，取下一个关键词 i=i+1，返回第③步，否则往下执行；

⑤ 在对应的 N-1 维数组中取语义相似度大于阈值δ 的所有关键词 Key_j，执行以下操作；

⑥ 关键词 Key_j 存放到同义词词典中，并定义为 Key_i 的同义词；

⑦ 在关键词列表 AK 中设置关键词 Key_j 的删除标志位为 1，取下一个关键词 i=i+1，返回第③步；

⑧ 将 AK 列表中没有设为删除标志的关键词放到降维以后的关键词列表中，实现关键词的降维操作。

4 实验及对比分析

4.1 实验方法

本文实验在 32 位的 win7 系统平台下进行，利用 java 语言实现关键词向量的冗余处理。利用 lucene 建立索引和进行 n-gram 统计[18][19]，利用 imdict-chinese-analyzer[20]进行非领域词分词，参照[21]设计非领域词的相似性对比计算，领域词之间相互独立，不再进行对比。

实验数据从 CNKI 选取 1113 篇主题词为“文本分类”的研究论文，从题录信息中提取作者给定的关键词 1731 个。以下是作者给定的关键词的一部分，选中部分是与“K 近邻”有关的关键词，前面的数字表示共有多少篇文章用到了这个关键词，如图 3 所示，可以看出作者给定的关键词的冗余程度之大。

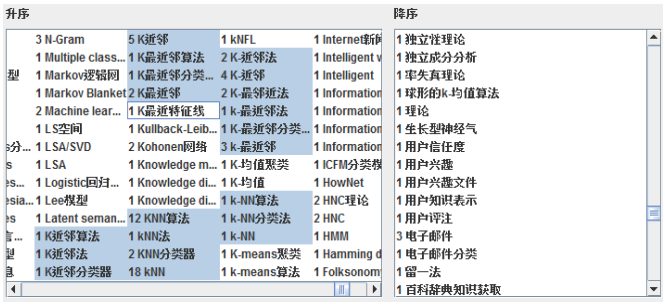

图 3：作者给定的部分关键词

通过查看作者给定的关键词，将简写的领域同义词手工提取出来，生成同义词词典，图 4 所示是本文最初生成的领域同义词词典的一部分，同义词之间以分号分隔。利用本文提出的方法进行领域同义词合并，上面选中部分可以合并为：K 近邻法、K 近邻算法、K 近邻分类器和 K 近邻四个关键词，进一步的合并需要后续的相似性计算。

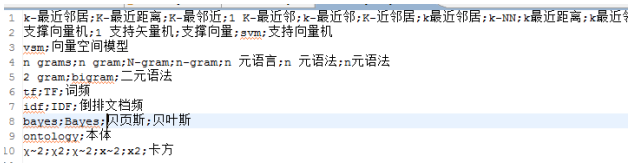

图 4：领域同义词词典

利用 n-gram 算法提取领域词库共 129 个领域词。包括支持向量机、散度、本体、朴素贝叶

斯、模糊、潜在语义、特征、K 近邻、矢量、神经网络、贝叶斯、降维、隶属度和领域等。同时需要根据实际情况对这些计算得来的领域词汇进行细微调整。

利用以上生成的领域词库以及知网知识库，按照本文 3.3 所给出的语义相似度计算公式计算相似度，利用本文 3.4 提取的流程识别同义词集合，并进行特征提取。对给定的 1731 个关键词进行降维后，最终关键词为 1491 个，压缩比为 13.9%。

4.2 对比分析

在本文实验分析的关键词中选出 6 个关键词，从分词方法和矩阵存储量两个方面进行对比分析。从表 1 中可以看出不加入领域词进行分词时，每个关键词会分成粒度非常小的几个普通词。本文利用 n-gram 算法提取领域词库，根据领域词库进行分词，领域词作为一个整体被提取出来，提高了分词的准确性。从表 2 中可以看出，利用一般的方法进行相似性对比时，需要将所有分词后的结果一一对比，而领域词本身相互独立，没有必要进一步对比，本文中只在领域词相同时才会进行对比，用来对比的相似矩阵的辅助矩阵内存占用量大大减小。

序号	词	一般分词	本文分词	一般相似计算	本文相似计算
1	贝叶斯方法	贝/叶/斯/方法	贝叶斯/方法	所有的词一一进行对比，并且分词颗粒太细，对比速度很慢。	领域词相同时才进行对比。本例中序号 1-4 之间关键词对比；序号 5-6 之间关键词对比。
2	贝叶斯算法	贝/叶/斯/算法	贝叶斯/算法		
3	贝叶斯理论	贝/叶/斯/理论	贝叶斯/理论		
4	贝叶斯模型	贝/叶/斯/模型	贝叶斯/模型		
5	隶属度矩阵	隶属/度/矩阵	隶属度/矩阵		
6	隶属度向量	隶属/度/向量	隶属度/向量		

表 1 分词方法对比

	贝	叶	斯	方法	算法	理论	模型	向量	矩阵
贝	xx	xx	xx	xx	xx	xx	xx	xx	xx
叶	xx	xx	xx	xx	xx	xx	xx	xx	xx
斯	xx	xx	xx	xx	xx	xx	xx	xx	xx
方法	xx	xx	xx	xx	1	0．44	0.17	xx	xx
算法	xx	xx	xx	xx	xx	0.44	0.15	xx	xx
理论	xx	xx	xx	xx	xx	xx	0.17	xx	xx
模型	xx	xx	xx	xx	xx	xx	xx	xx	xx
向量	xx	xx	xx	xx	xx	xx	xx	xx	0．88
矩阵	xx	xx	xx	xx	xx	xx	xx	xx	xx

表 2 相似矩阵存储量对比

（说明：整个矩阵表示一般方法占用的相似矩阵存储量，xx∈[0，1]，颜色加深部分是本文使用的矩阵存储量）

5 结论

本文的创新之处在于：针对关键词冗余问题，在分析词与词之间的语义相似度时加入了利用 n-gram 算法提取领域词的过程；分词时领域词作为一个整体来对待，分词后每个关键词分为领域词部分和非领域词部分，如果领域词不相同，两个词不再对比；如果领域词相同，再进行非领域词部分语义相似度计算。利用这一改进的方法进行关键词的特征降维处理，有效地解决了关键词的冗余问题。并且和没有加入领域词库的相似性计算方法相比较，减少了内存的使用量以及词与词之间对比计算的次数。不足之处在于：n-gram 统计生成领域词汇时会出现一些垃圾词汇，需要手动去除；相似度计算公式还需要改进；程度还需要进一步优化，进一步改进词库存放格式和相似性对比的过程。

参考文献

[1] Chua, S, Kulathuramaiyer, N. Semantic Feature Selection Using WordNet[C]. In. Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence. Beijing: IEEE Computer Society, 2004: 166-172.

[2] Li Xiaobin, Stan Szpakowicz, Stan, Matwin. A WordNet-based algorithm for Word Sense Disambiguation[C]. in proceedings of the IJCAI-95, Montreal, Canada,August. 1995:1368-1374.

[3] 熊忠阳,付玲玲,张玉芳,文本分类中基于概念映射的二次特征降维方法 [EB/OL]. [2011-3-10]. http://www.cnki.net/kcms/detail/11.2127.TP.20110223.1435.007.html?uid=WEEvREcwSlJHSldRa3JPV0dvSFpWamplRWN1SW9vVW91ZlRaY0xYV2cxZFMzVVkzTkpOemo1cXN6ckVhNGx3PQ==

[4] 唐歆瑜,乐文忠,李志成. 基于知网语义相似度计算的特征降维方法研究[J] 科学技术与工程 2006, 6(21):3442-3446.

[5] 董振东,董强.知网 [DB/OL].[2011-2-10]. http: //www.keenage.com.

[6] 吕震宇,林永民,赵爽,朱卫东. 基于同义词词林的文本特征选择与加权研究[J]. 情报杂志 2008,27(5):130-132.

[7] 中华人民共和国国家标准. GB/T 7713.1-2006 学位论文编写规则 [S] 2006

[8] 马开俊 数字化建设中文献信息主题标引方式管见[J]. 情报资料工作，2004 年年刊, 355-356

[9] 谭慧华 CAJ- CD 关键词标引质量探析[J]. 情报杂志，2003,（3）: 79-80

[10]郭淑敏 医学期刊编辑中的关键词标引[J]. 中华医学科研管理杂志，2006,19（3）: 178-179

[11]赵宗蔚 提高期刊论文关键词索引质量—自然语言与人工语言的结合.[J] 图书馆论坛 2005 ,25(1): 119-121

[12]F. Jelinek, Continuous speech recognition by statistical methods[J]. Proceedings of the IEEE. 1976, 64(4): 532-556

[13]Yuqing Gao, Bowen Zhou, Zijian Diao, Jeffrey Sorensen, Michael Picheny. MARS: A Statistical Semantic Parsing and Generation-Based Multilingual Automatic tRanslation System[J]. Machine Translation, 2002, 21(2) :185-212.

[14]A. L. Koerich, R. Sabourin , C. Y. Suen. Large vocabulary off-line handwriting recognition: A survey. [J] Pattern Analysis & Applications . 2003 6(2), 97-121,

[15]Zheng Chen . Kai Fu Lee . A new statistical approach to Chinese pinyin input[A]. ACL-2000. In：The 38th Annual M eeting of the Association for Computational Linguistics[C]. Hong Kong，3-6 October，2000.

[16]Ponte J M, Croft W B . A Language Modeling Approach to Information Retrieval. [C]. In Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, USA. 1998: 275-281

[17]刘群，张华平，俞鸿魁，程学旗. 基于层叠隐马模型的汉语词法分析[J]. 计算机研究与发展 2004, 41(8):1421-1429

[18]Lucene: [EB/OL][2011-3-20]. http://lucene.apache.org.

[19]Kumar N, and Srinathan K. Automatic keyphrase extraction from scientific documents using N-gram filtration technique[C]. In: Proceedings of the 2008 ACM Symposium on Document Engineering, Sao Paulo, Brazil. 2008:199-208.

[20]ICTCLAS [EB/OL].[2011-5-1]. http://ictclas.org/ictclas_files.html.
[21]刘　群,李素建. 基于知网的词汇语义相似度计算[A].第三届汉语 词汇语义学研讨会[C].台北: 2002.

（作者 Email:xingmeifeng@mail.las.ac.cn）