基于 TFIDF 的专业领域词汇获取的研究*

刘桐菊 于浩 杨沐昀

（哈尔滨工业大学计算机系，哈尔滨 15001）

E-mail: ltj,yu,ymy@mtlab.hit.edu.cn

摘要： 针对目前专业词典的建设中如何获取专业词汇这一问题，本文采用了 TFIDF 算法求加权词频，按词频高低确定专业领域词汇的方法，并针对这一思路给出了在体育领域内研究的样例，同时对这一设想作了预期评估和展望。

关键词:TFIDF; 词频; 前景语料; 背景语料

引言

词典是乔姆斯基模型中必不可少的一部分，是我们进行自然语言处理的一块基石，因此它的建设对自然语言处理有重大的意义。随着机器翻译的实用化进程，专业词典的建设也提上日程。为了有效的获取专业领域的词汇作为专业词典的收录词条，本文题提出了基于 TFIDF 求加权词频来获取专业领域的汉语基本词条的方法，并在体育领域做了相应的实验证明此方法的有效性。我们选用的是体育领域的文章作为前景语料，首先将生语料进行加工，用分词程序分词，然后生成一元组和多元组，求得 TF 值；把背景语料分词，生成一元组和多元组，求出 IDF 值，然后以前景语料为基础求每个元组的加权词频并排序，默认词频高的为体育领域的专有词汇。

1 算法的选择与介绍

我们很容易可以理解体育领域内的专业词汇应该是体育领域文章中的高频词[6]，所以，首先应该对生语料进行分词，然后生成一元组和多元组，求这些元组的频率即可。可是，这样做有一个潜在的问题，就是在各个领域内都是高频词的词（如的，了，个等等）将淹没体育领域内的词汇（实验证明结论就是如此，参考表 1），很自然想到应该求加权词频，让他们按照不同的幅度增长，在相对意义上是此消彼长的，从而使体育领域内的词汇频率增高。在这里我们选择了 TFIDF 算法[2][3][5]来求加权词频。

TFIDF 算法介绍

$F(w)=f(w)*N/n$

F(w)——用 TFIDF 方法求得的加权词频；

f(w)——体育领域内的词频；

N——背景文本数；（远远大于体育领域语料的文本数）；

n——该词在背景文本中出现的文本数；

* 本文研究受到国家 863 计划资助(项目编号：2001AA114101)。

令 TF=f(w),IDF=N/n

TFIDF 算法就是求词的加权词频，这个权值（IDF）等于背景语料文本数 N 和该词在背景语料中出现的文本数 n 的比值，也就是说，当这个词在背景语料许多文本中出现时（即 n 很大），我们就可以认为这个词不是体育领域内的专有词汇，应该让它的权值很小，此时的 IDF 值恰好因为 n 很大而变得很小，经过运算后词频增大的幅度比较小，F(w)就相对来说要小；相反，当某个词为体育领域内的专有词汇时，它在整个背景语料中出现的文本数会很低，那么，它将获得一个很大的权值，词频会大幅度提高，即 F(w)值很大，这样，按 F(w)值排序后就有可能使体育领域专有词汇排在前面，从而可以提取出来。例如：在大背景语料的实验里，“的”字，它的词频 TF 为 109, 5401，IDF 为 10，这样，它的词频 F(w)就变为 1095, 4010。“角球”一词，它的 TF 值为 4247，明显低于“的”的 TF 值，但它的 IDF 为 21,0780，最终它的 F(w)为 8,9518,2660，远远大于“的”的 F（w)。显然，按加权词频 F(w)排序，体育领域词汇会被排在前面可以被提取出来。基于这种思路，我们进行了一些试验，证明采用 TF/IDF 算法求加权词频方法提取专业词汇的有效性。

2.体育领域词汇获取的实验

2.1 语料的来源

体育网页上常用的汉语体育词汇，选用的前景语料为新浪网的体育信息，共计 5, 1985 个文本；背景语料也选自新浪网，小背景语料为新闻和军事领域，共计 11, 2632 个文本，大背景语料为健康、汽车、文化、科技、金融、住房、生活、娱乐、新闻和军事领域，共计 42, 1560 个文本。

2.2 词汇获取流程框图

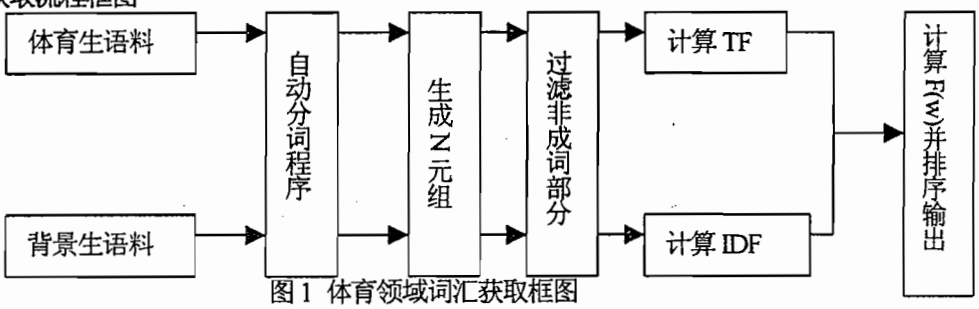

图 1 体育领域词汇获取框图

2.3 实验及数据

试验一：验证 TF/IDF 有效性实验

提取过程如（图 1）所示，从中我们共得到 9, 8494 个一元组，以一元组为例，列举出部分结果来证明 TFIDF 算法的效果：

表 1　　绝对词频和 TF/IDF 加权词频对比

次序	绝对词频高的词	加权词频高的词（小背景语料）
1	的	赛季
2	在	任意球
3	队	客场
4	了	点球
5	一	助攻

6	是	国脚
7	他	上半场
8	中	转会
9	不	射门
10	这	主队
11	比赛	主教练
12	和	下半场
13	个	比分
14	我	球员
15	也	后卫
16	有	篮板球
17	中国	传球
18	上	投篮
19	分	足协
20	第	男篮

我们来分析一下上面的数据：

我们选取的是高频的前 20 个词，在第一列里，列出的是绝对词频高的词（加权前），我们可以看出这些词汇多数是常用词汇，仅有“比赛”一词勉强可以认为是体育领域的专业词汇，但它在其他领域也常出现，如：歌咏比赛、书画比赛等等，总体上来讲，将绝对高频词作为体育领域的专业词汇效果不好；第二列列出的是加权后的高频词，很明显，这 20 个词都是体育领域的专业词汇，那么加权后的高频词作为体育领域的专业词汇的效果就会相当不错。这组数据足以证明将 TFIDF 算法应用于专业领域词汇获取方面的有效性。

试验二：验证背景语料规模对结果的影响

不难看出采用 TFIDF 算法加权后效果非常明显，那么背景语料大小的选择对结果会产生什么影响呢？我们针对这一问题进行了另外一个实验，分别对小背景语料和大背景语料两种情况进行比较，在得到加权词频后，对两组数据各取前 5000 个高频词，做差集来验证哪个效果更佳（我们从直观上来判断，认为哪个差集中包含的体育词汇多哪个就好）。做差集后得到 1990 个词，其中小背景语料独有的体育词汇为 2 个，而大背景语料独有的体育词汇为 95 个，不言而喻，选取大背景语料的效果更佳。下面列出部分结果（以词频为序）。

表 2　小背景语料和大背景语料的对比

次序	小背景语料独有的词	大背景语料独有的词
1	佐料	角球
2	老气横秋	主罚
3	悠着	下半时
4	看轻	上半时
5	看官	篮板
6	复选	停赛
7	出局	头球
8	脱兔	混合泳

9	快议通	罚球
10	屁滚尿流	客队
11	绵软	界外球
12	割爱	种子队
13	独孤	短传
14	丹参	左后卫
15	自视甚高	海牛
16	曾经沧海	加时赛
17	下三滥	男双
18	票房价值	球风
19	马后炮	升班
20	看低	棋坛

我们来分析一下上面的数据：

在第一列小背景语料独有的词汇里“出局”、“复选”可以认为是体育领域的词汇，而在第二列的大背景语料独有的词汇中，基本上都可以认为是体育领域的专业词汇，由此证明，背景语料选的大效果会更好一些。

实验三： 多元组的获取及过滤

词是不断涌现的，不同词的组合被赋予新义，对于这些新词的获取，仅限于一元组是不行的，它获取的是收录到通用词典中的词汇。新词获取，我们就可以借助多元组的获取来实现，为此，我们进行了该实验。

流程和一元组获取方法基本一致，改动的地方是过滤，计算出 TFIDF 值后还要加一些 stopwords 进行过滤。Stopwords 集合过滤是很重要的一步，因为有些元组进常在一起搭配出现，但他们不一定是新词，例如：“被、了、但是、因为、你、他、啊”等词汇，经常和一些词搭配出现，但他们是不会构成新词的，这些词的存在无疑会淹没体育领域的词汇，会给提取带来极大的麻烦，严重影响结果。根据实验获得的数据，最终确定的 stopwords 集合包含词典中的词性为 nd(地名)、p（介词）、nm（人名）、e（感叹词）、o（拟声词）、m（数词）、nx（姓）、c（连词）、i（成语）、l（俗语）、t（时间）、y（语气助词）、ny（外国译名）、d（副词）的一些词，其中，兼词（多个词性的词）没有加入。

在得到的结果中，选取出词频高的前 5000 个词，然后用 stopwords 集合进行过滤，其正确率为 91.6%，所以，在这里选用的 stopwords 集合是安全的。

3.结论及后续研究展望

从体育领域词汇获取这个实验来看，通过给定前景语料和背景语料，采用 TFIDF 算法求加权词频，默认词频高的词为体育领域专有词汇的方法是行之有效的，对于背景语料的大小，实验证明选取大背景语料获得的效果更好。在以后的研究中，我有以下的设想：

1) 本实验未曾考虑词之间的互信息，在今后的研究中，应该把词之间的关联强度加进去，修改 F(w) 的计算公式。
2) 在生成多元组时，应该加入语言知识加以引导，剔除掉不能成词的组合，可以在很大程度上提高准确率。
3) 对于 N 元组的生成，是借助了分词程序，这样就会引入一些错误蔓延，在今后的研究中，还可

以试图在这方面下功夫，不用分词程序，直接将生语料按字生成1到8元组（假设的），然后通过有效字符串的识别来获取词[1][4]，接下来再按照上述方法进行专业词汇的获取。

4) 可以通过反复试验来确定一个更有效的stopwords集合，把过滤加到前端，不但可以大大缩减问题的规模，还可以提高准确率。

参考文献:

[1] Robertson, Sparck-Jones. Relevance weighting of search terms (context), 1976

[2] Salton, Buckley . Term-weighting approaches in automatic text retrieval (context), 1998

[3] The automatic creation of literature abstract. IBM Journal of Research and Development, 2:159-165.

[4] Salton, G. Automatic Text Processing. Reading, MA :Addison-Wesley, 1998

[5] Text Retrieval Conference (TREC), Gaithersburg, Maryland, 1994, 109~126.

[6] Salton, G., Singhal, A., Mitra, M., and Buckley, C., Automatic text structuring and summary. Info. Proc. and Management, 1997, 33(2):193 - 207.

致谢 哈尔滨工业大学计算机系机器翻译研究室的于传武、张博同学对本文的完成做了很多的工作，在此一并表示感谢！

作者简介: 刘桐菊（1978—），女，吉林省德惠人，硕士生，主要研究领域为专业域词汇的获取技术。于浩（1971—），男，博士，副教授，主要研究领域是自然语言理解和信息处理；杨沐昀（1971—），男，博士生，主要研究领域是翻译知识处理。

The Research of Term Extraction in Professional Field*

Liu Tongju, Yu Hao, Yang Muyun

(Harbin institute of technology, Harbin 15001, China)

E-mail: ltj,yu,ymy@mtlab.hit.edu.cn

Abstract: Getting vocabulary is essential to building professional dictionary. To solve the problem, this paper puts forward a kind of term extraction method in sport field with TFIDF algorithm .The experiment's results prove that the method is effective. In addition a preliminary evaluation and future improvements are discussed.

Key words: TFIDF;word frequency; foreground corpus; background corpus

* Supported by, the National Natural Science Foundation of China under Grant No.(!!!!!!!!)