知识抽取中的停用词处理技术

化柏林

（中国科学技术信息研究所 北京 100038）

【摘要】 在知识抽取的分词过程中，需要提前把停用词标引出来。停用词处理的关键在于停用词的认定、停用词表的获取与组织技术以及停用词匹配技术。停用词的识别常常需要停用词表，识别过程中需要判断假停用词以降低噪声。实验表明，对停用词进行单独处理可以大大加快词语切分速度以及后续的句法分析归约速度。

【关键词】 知识抽取 停用词 中文分词 自然语言处理 文本信息分析 【分类号】 TP391 G356

Stop – word Processing Technique in Knowledge Extraction

Hua Bolin

(*Institute of Scientific and Technical Information of China*, *Beijing 100038*, *China*)

【**Abstract**】 It is indispensable to index stop – word before word segmentation in knowledge extraction. The key technique of processing stop – word is how to select stop – word, acquire and organize stop – word lists, and match stop – word. To recognize stop – word, constructing stop – word list is necessary. In processing stop – word, recognizing false stop – word can decrease noise. According to experiment, processing stop – word can not only save segment time, but also improve following syntactic analysis efficiency.

【**Keywords**】 Knowledge extraction Stop – word Chinese segmentation Natural language processing Text information analysis

1 引 言

知识抽取是把蕴含于文本文献中的知识经识别、理解、筛选、格式化，把文献的每个知识点（包括常识知识和专家知识）抽取出来，再以格式化的形式存入知识库。本研究为实现把用自然语言描述的句子通过知识表示转换成计算机可理解的形式，专门对停用词处理技术进行了研究。停用词处理是知识抽取过程中向量分词的一个步骤，它的单独处理会加快切分速度以及提高后续句法分析的速度与质量。

2 停用词的认定与选取

2.1 停用词的意义阐释

在基于词的检索系统中，停用词是指出现频率太高、没有太大检索意义的词，如“的、是、太、of、the”等；在基于支持向量机的自动分类中，停用词指没有实意的虚词和类别色彩不强的中性词[1]；在自动问答系统中，停用词因其问题的不同而动态变化[2]；在机器翻译、知识抽取中几乎没有真正的停用词，只是把频率太高的虚词作为临时的停用词特殊处理，切分完后仍然需要标记。

2.2 停用词的选取范围

周钦强等人在文本分类中认为，停用词主要包括英文字符、数字、数学字符、标点符号以及使用频率特高的单汉字等[1]。梁南元在 CDWS 系统中选定一些非自然切分标志的字或词为停用词，如只能充当词首字和词尾字的字、拟声字、复音单纯字等[3]。罗杰等人认为，除数字等切分标记外，停用词还包括数词、量词、代词、方位词、拟声词、叹词等，没有实际意义的动词，如“可能”等，以及一些太过于常用的名词，如“操作”等[4]。其实，大多数英文字符不应该作为停用词处理，如“XML”、“IC 卡”等，因为这些词从某种程度上反映文章的内容，可以根据 ASCII 码或机内码单独处理，而且要把整个单词一起处理。混在中文中的英文字符与中文一样，没有空格自然切分，所以处理起来与中文一样，只不过匹配所用的词典不同。

2.3 停用词的数量确定

停用词的数量非常少，往往就几十个。美国 Bell 实

收稿日期：2007 – 05 – 11
收修改稿日期：2007 – 06 – 19

验室的 Tin Kam Ho 论述到,在典型的英文文章中,停用词的使用数量占到一半以上,而这些停用词的数量却不足 150 个[5]。文献[6]列出了搜索引擎针对英文的停用词列表,其数量达到 658 个,其中包括 a,b,c,re,co 等很多并不构成单词的字母组合。顾益军等人提出了一种新的停用词选取方法,用词条在语料库中各个句子内发生的概率和包含该词条的句子在语料库中的概率计算它们的联合熵,依据联合熵选取停用词[7],他在实验中选取了前 25 个词,但对于为什么选取前 25 个词,而第 26 个词就不是停用词却没有说明,这种阈值的确定是很难的。

3 停用词的获取

3.1 停用词的获取方法

停用词表有通用停用词表与专用停用词表之分,其来源有人工构造与基于统计的自动学习两种方式。基于统计的自动学习方法是从语料中统计出高频停用词,自动构建停用词表并由人工进行核对,或者从初步的向量分词结果中得到停用词,然后分词过程中不断地更新频率并根据切分结果进行验证。美国德克萨斯大学达拉斯分校(University of Texas at Dallas)的 Feng Zou 等人提出一种基于统计与信息论模型的中文停用词抽取方法[8]。约旦阿莫克大学(Yarmouk University)的 Al - Shalabi,R. 等人提出一种基于有限状态自动机的阿拉伯语停用词过滤算法[9]。瑞士纳沙泰尔大学(Université de Neuchâtel)的 Jacques Savoy 设计了针对葡萄牙语通用停用词词表的获取程序[10]。

3.2 本系统中的停用词获取

在本文所述的知识抽取系统中,第一趟扫描并没有停用词表,只是用关键词构成词库进行分词。把第一趟扫描的结果中所有未能根据词典分出来的字符串全部提出来,按频率降序排列,频率较高的作为停用词。然后利用这些停用词对未能分出来的字符串进行二次切分,这次切分主要是处理未登录词的情况。如由“并提出 知识/ 博客/ 在 图书馆联盟/ 中的重要 应用/ 。”变成了“并 提出 知识/ 博客/ 在 图书馆联盟/ 中 的 重要 应用/ 。”,其中“并”、“中”、“的”是停用词,而“提出”、“重要”就是未登录词,因为实验中向量分词最初使用的词典是由图书情报学核心期刊论文中的 43 980 个关键词构成的,这些关键词中不会包含“提出”、“重要”等词语,因此,可以把这样的词追加到词库里。停用词的选取是一件很难的事情,一般是根据出现的频率来选取。

从 1989 - 2005 年图书情报学中文核心期刊的 42 989 篇论文的摘要中(其中 1996 年以前的很多论文没有摘要)经过分词提取,得到高频停用词如表 1 所示。

表 1 实验中自动提取的停用词表

第一趟统计结果		第二趟统计结果	
的	48 456	的	81 352
、	15 756	,	54 806
和	12 490	了	27 624
,	11 917	、	21 502
了	6 223	和	16 241
与	5 936	。	11 900
在	3 417	对	10 459
及	2 990	在	9 550
。	2 921	与	7 809
是	2 428	是	7 187
对	2 266	中	6 659
中	1 536	并	6 270
为	1 387	从	4 790
从	1 223	及	4 201
等	1 106	为	4 085
;	817	等	3 832
上	433	一	3 653
”	417	上	3 203
以	416	以	2 773
“	393	“	2 660
下	391	”	2 564
(	367	;	2 537
:	349	个	2 478
其	320	种	2 445
于	303	其	2 407
合计	124 258	合计	302 987

上述论文摘要中共出现单字 2 440 个,计 426 431 次;非停用词非关键词共出现 5 321 个,计 146 377 次。从第一趟和第二趟结果中分别选取前 25 个词,除标点符号外,基本上是相同的,但顺序与频率有一些变化。第一个词频率增加了 1 倍,而最后一个词频率却增加了 7 倍,说明第二趟结果的停用词分布相对更集中一些。

4 停用词表的组织方式

4.1 停用词表的关系型组织方式

保加利亚瓦尔纳医科大学的 D. T. Tomov 把停用词分为真正的停用词(True Full - stop Words)与半停用词(Semi - stop Words)[11]。笔者在知识抽取项目过程中,把停用词表按其复杂程度分为简易停用词表和复杂停用词表。简易停用词表就是把符合停用词条件的字符组织起来,按照使用频率、字母顺序进行排序,或者按照停用词类型进行分组。复杂停用词表包含真停用词与假停用词。假停用词指含有停用字的非停用词,有些停用词(为行文方便,以下称停用字)也可能包含在某个非停用词里。如含有“的”的“的确”、“有的放矢”,含有“了”的

“知了”、“了解”、“了如指掌”等就不是停用词。为了处理这些情况,可以在建立停用词库时,把含有停用字的非停用词都收录到一个表里,并与停用字建立关联关系,如表 2 所示。

表 2 复杂停用词表结构示意

停用字	词长	位置	词项
了	2	1	了解
	2	2	知了
	2	2	明了
	4	1	了如指掌
			…
的	2	1	的确
	2	2	目的
	4	2	有的放矢
			…

4.2 停用词表的其它组织方式

表 2 所示的关系型数据索引速度比较慢,为了提高索引速度,一般采用 Hash 索引。孙茂松等人对整词二分法、Trie 索引树及逐字二分法 3 种常用的分词词典机制作了比较[12]。但是停用字在假停用词中并不总是处在起始位置,因此,停用词表不适合采用首字散列、后缀数组等索引方式。

5 真假停用词的识别

5.1 真假停用词的识别流程

大多数应用系统通常先进行停用词处理,后进行常规分词。停用词的匹配与识别主要体现在真停用词还是假停用词的判别上。分析时遇到停用字,首先判断是不是含停用字的非停用词,如果是就不再切分;如果不是就把它切分出来。停用字可能出现在词首、词尾与词中,例如,“的”在“的确”中出现在词首,在“有的放矢”中出现在词中,在“目的”中出现在词尾。因此,从待分析串中靠近停用字取前 n 个或后 n 个字符,然后去通用词典里匹配,这种方法要花费很多时间。在待分析串中遇到停用字,去停用字的非停用词表(见表 2)里查找,然后根据词长和在词中的位置去待分析串中取词长个字符,判断是否为含停用字的非停用词,这样的匹配算法简化了很多。向后向前取字符的长度可以根据表 2 中的位置字段,如果位置是 1 就向后取;如果位置大于 1 就从前面取,也就是从停用字在句子中的位置(iSenPos)减去停用字在词中的位置(iWordPos)开始取,取词长个字符进行比较即可。这种算法非常简单,只需要取 n(n 为含“了”的非停用词个数)个词然后各匹配一次就可以了。而向量分词中却要执行 iMaxVector - i 次查找(iMaxVector 代表设定的最大向量长度,i 代表停用字在按最大向量长度所取待分析串中的位置),而每一次查找要耗费大量的时间。停用词识别切分算法如例 1 所示。

例 1:停用词识别切分处理算法

```
//sNotStop 为待分析句子中含停用字的词串,iSenPos 为停用字在句子中的位置
// iWordPos 为停用字在非停用词中的位置,iLength 为含停用字的非停用词词长
sNotStop = Mid(sSentence,iFind - iInPos + 1,iLength)
//iStopinNot 为停用词表中含停用字的非停用词个数
For k = 1 To iStopinNot
  //如果找到,说明是假停用字
  If NotStop = sStopinNot(k) Then
  sSegment = Left(sSentence,iSenPos - iWordPos) + " " +
             sNotStop + " " + Mid(sSentence,iSenPos -
             iWordPos + iLength + 1)
  bFind = True
  exit for
  end if
Next
//如果都不符合,说明是真停用字
If bFind = False Then
  sSegment = Left(sSentence,iSenPos - 1) + " " + sStopWord
    + " " + Mid(sSentence,iSenPos + 1)
End If
```

5.2 真假停用词的识别方法

假停用词的判断有两种方式,一种是从待分析串中停用字所在位置分别向前或向后取 n 个字符,然后去假停用词词表中匹配,笔者把这种方法称为盲目判别法,因为对每个停用字都要判断 $\sum_{i=1}^{n} i$ 次,其中 n 为假停用词词表的最大词长。另一种方式是从假停用词词表中取含有停用字的假停用词去待分析串中匹配,把这种方法称为关联判别法,是一种有针对性的判别。这两种方式的不同在于:前者是从待分析串到词表的匹配,后者是词表到待分析串的匹配;在首先确定了停用字的情况下,后者比前者有着明显的匹配速度优势。

5.3 真假停用词识别的歧义处理

上述停用词的处理依然存在不足,因为这种方法只考虑停用词本身,而不考虑停用词所在的语境,特别是上下文。如“对于阈值的确定”,利用上述方法会错误地切分成“对于 阈值 的确 定”。在进行词性标记时发现有错误,然后就需要重新分词。在分词过程中尚不能判断切分结果是错误的,因为分词算法能确定分词结果中的“定”不是停用词,也不是未登录词,但它在行文中是可以单独成词的,如“他定能成功”就可以切分成“他 定 能 成

功”,而不能切分成“他 定能 成功”,因为“定能”是个稳定的组合,但不是个词。实验中词的界定必须是能明确词性的,而例子中的“定能”是两个词,分别是副词与情态动词。无论用隐马尔科夫模型[13],还是互信息[14],这种基于转换的错误驱动学习算法[15],都难以识别这种切分错误的存在。

6 结 语

6.1 停用词处理的结果标记方式

(1)直接把停用词滤掉,以空格来代替。用空格代替的优点是保证每个索引词在句子中的位置不发生变化;缺点是只适用于那些基于词的应用系统。

(2)保留停用词在句子中的位置,与其它词采用不同的标记符标记。例如,根据词典分出来的词采用斜杠加空格标记,而停用词只用空格标记,不加斜杠。

(3)采取与其它词一样的处理方式,采用斜杠加空格进行标记。

在需要进行句法分析处理的系统中,常用第 3 种结果标记方式,因为“了”、“与”、“的”等停用词更能反映句子的结构与时态等句法信息。

6.2 停用词处理的关键

停用词处理的关键在于真假停用词表的构造以及停用词的识别。同一般的向量分词处理一样,主要有 3 个过程:待切分串的截取技术、停用词表的获取与组织技术以及停用词的匹配技术。待切分串的截取技术比较简单,一个简单的字符串查找函数就满足需求。停用词表的构造需要经验语言学与统计语言学知识的支撑,针对不同的应用系统还要有不同的处理,根据处理结果对停用词表的更新也是必须的。停用词的匹配技术中主要涉及去掉假停用词。整个停用词的处理过程相对来讲并不困难。

6.3 停用词处理的难点

错误的识别是很困难的,需要有丰富的语言学知识库才能识别出错误,错误的修正要比识别容易得多。既然能识别出错误,就可以利用识别错误的依据,确定错误的类型,根据错误的类型,选择相应的修正方案。完全通用的算法其准确度是有限的,在处理过程中,针对具体的语言现象设定特定的算法,算法中考虑的语言现象越充分,处理的准确度就会越高。

参考文献:

[1] 周钦强,孙炳达,王义. 文本自动分类系统文本预处理方法的研究[J]. 计算机应用研究,2005(02):85 - 86.

[2] 熊文新,宋柔. 信息检索用户查询语句的停用词过滤[J]. 计算机工程,2007,33(06):195 - 197.

[3] 梁南元. 书面汉语的自动分词与一个自动分词系统—CDWS[J]. 北京航空学院学报,1984(4):97 - 104.

[4] 罗杰,陈力,夏德麟,等. 基于新的关键词提取方法的快速文本分类系统[J]. 计算机应用研究,2006,4:32 - 34.

[5] Ho T K. Stop Word Location and Identification for Adaptive Text Recognition[J]. International Journal on Document Analysis and Recognition,2000,3(1):16 - 26.

[6] Stop Word List—Words Filtered out by Search Engine Spiders[EB/OL]. [2007 - 06 - 14]. http://www. seo - innovation. com/support - files/stopwordlist. pdf.

[7] 顾益军,樊孝忠,王建华,等. 中文停用词表的自动选取[J]. 北京理工大学学报,2005,25(04):337 - 340.

[8] Zou F, Wang F L, Deng X T, et al. Stop Word List Construction and Application in Chinese Language Processing[J]. WSEAS Transactions on Information Science and Applications,2006,3(6):1036 - 1044.

[9] Al Shalabi R, Kanaan G, Jaam J M, et al. Stop - word Removal Algorithm for Arabic language[C]. Information and Communication Technologies: From Theory to Applications,2004. Proceedings. 2004 International Conference on.

[10] Savoy J. Data Fusion for Effective European Monolingual Information Retrieval[C]. Workshop of the Cross - Language Evaluation Forum (CLEF 2004),2005:233 - 244.

[11] Tomov D T. Research Brief: Some Critical Remarks on the Stop Word Lists of ISI Publications[J]. The Journal of Documentation,2001,57(6):798 - 808.

[12] 孙茂松,左正平,黄昌宁. 汉语自动分词词典机制的实验研究[J]. 中文信息学报,2000,14(1):1 - 6.

[13] 刘颖. 用隐马尔柯夫模型对汉语进行切分和标注排歧[J]. 计算机工程与设计,2001,22(4):58 - 62.

[14] 刘开瑛. 中文文本自动分词和标注[M]. 北京:商务印书馆,2000.

[15] Bril E. A Simple Rule - based Part - of - speech Tagger[C]. In: Proceedings of the Third Conference on Applied natural Language Processing. ACL. Trento, Italy. 1992:152 - 155.

(作者 E - mail:huabolin@ istic. ac. cn)

更 正

2007 年第 7 期目录中,《网络环境下多媒体关联规则数据挖掘方法研究》一文作者“羊牧”被误排为“关牧”,特此更正。编辑部向作者及广大读者表示歉意。

《现代图书情报技术》编辑部