基于互信息的中文术语抽取系统

张 锋[1], 许 云[1], 侯 艳[2], 樊孝忠[1]
(1.北京理工大学 计算机科学与工程系, 北京 100081; 2.广东工业大学 计算中心, 广东 广州 510520)

摘 要: 介绍了一个中文术语自动抽取系统,该系统首先基于互信息计算字串的内部结合强度,从而得到术语候选集;接着从术语候选集中去除基本词,并利用普通词语搭配前缀、后缀信息进一步过滤;最后对术语候选进行词法分析,利用术语的词性构成规则进行判别,得到最终的术语抽取结果。实验结果表明,术语抽取正确率为72.19%,召回率为77.98%,F测量为74.97%。

关键词: 术语抽取; 互信息; 语料
中图法分类号: TP391 **文献标识码**: A **文章编号**: 1001-3695(2005)05-0072-02

Chinese Term Extraction System Based on Mutual Information

ZHANG Feng[1], XU Yun[1], HOU Yan[2], FAN Xiao-zhong[1]
(1. *Dept of Computer Science & Engineering, Beijing University of Technology, Beijing* 100081, *China*; 2. *Computing Center, Guangdong University of Technology, Guangzhou Guangdong* 510520, *China*)

Abstract: An automatic Chinese term extraction system based on mutual information is presented. Firstly, the system gets term candidates by calculating internal associative strength of characters string using mutual information. Then, term selection from term candidates is done using basic word dictionary, common collocation's suffix, prefix bank and some term POS composing rules. Experiment shows the precision is 72.19%, recall is 77.98% and F-measure is 74.97%.
Key words: Term Extraction; Mutual Information; Corpus

术语的自动抽取是自然语言处理的一个重要问题。目标是在文本集中抽取一定意义的词语搭配[1]。术语抽取可以应用在机器翻译、自动索引、信息检索、信息抽取、构建词汇知识库等领域。目前,国内外进行词语搭配抽取的研究方法主要是基于统计。作为一种特殊的词语搭配,术语的抽取过程一般有两个步骤:①进行术语候选抽取(Term Candidate Extraction);②在候选集中进行术语选择(Term Selection)。通常基于统计计算字串的内部结合强度来决定是不是候选术语。常用的方法有频率、互信息、Dice公式等。其中互信息方法在两字新词抽取方面结果较好,它的F测量为57.82%[2]。术语选择的方法有依据频次排序选择法,即根据候选术语在语料中出现的频次从多到少排序,按顺序选择一定数目的候选术语作为术语选择的结果[1];另外就是利用术语的词法、句法信息和语义信息等进行术语选择[3,4]。我们设计了一个中文术语自动抽取系统。系统利用互信息计算字串的内部结合强度,从而得到术语候选集。针对中文的特点,我们建立了普通词语搭配前缀、后缀信息库,总结了术语的词性构成规则进行术语选择。

1 系统结构

系统的应用对象是专业领域的文本,而原始领域文本中汉字是以"字"为单位的,直接对原始文本逐字分析来进行术语提取,不但速度慢,而且没有语言知识的引导,容易出现术语边界判断不准等错误。我们对原始领域文本先进行粗切分,即按常见的切分标志进行切分,得到字串集合。常见的切分标志为标点符号、中西文数字以及常见的助词(的、了、着等)。

在领域文本粗切分后,系统采用互信息计算待识别字串的内部结合强度,进行术语候选的抽取。然后利用普通词语搭配前缀、后缀信息库和术语的词性构成规则进行术语选择,得到最终术语抽取结果。系统结构如图1所示。

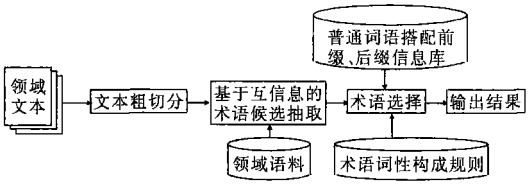

图1 基于互信息的中文术语抽取系统结构

2 基于互信息的字串内部结合强度计算

记待识别字串为 $c = c_1 c_2 \ldots c_n$, c的两个最长子串记为 $a = c_1 c_2 \ldots c_{n-1}$, $b = c_2 c_3 \ldots c_n$。例如,字串"自然语言处理",c=自然语言处理; a=自然语言处; b=然语言处理;记 $f(c)$为字串 $c_1 c_2 \ldots c_n$在语料中的共现频率; $p(c)$为字串 $c_1 c_2 \ldots c_n$在语料中的共现概率。根据最大似然估计,在语料规模足够大的情况下,可以认为 $p(c)$等于 $f(c)$。其中 n为字串的长度,要求 $n > 1$。

在信息论中两个事件AB的互信息计算如下:

$$MI = \log_2 \frac{p(AB)}{p(A)\,p(B)} \tag{1}$$

那么对于字串 $c = c_1 c_2 \ldots c_n$,可以定义其互信息[5]为

收稿日期: 2004-05-31; 修返日期: 2004-07-15

$$MI = \log_2 \frac{p(c)}{p(a)\,p(b)} = \log_2 \frac{f(c)}{f(a)\,f(b)} \quad (2)$$

如果字串 c结合十分紧密,那么 $f(c)$就与 $f(a)$或 $f(b)$相差不大,依据式 (2)计算的字串互信息就比较大;反之, $f(a)$和 $f(b)$就会远远大于 $f(c)$,这样计算出来的互信息就比较小。因此,互信息可以用来表示一个字串的内部结合强度。

基于统计的思想认为,一个词语搭配如果在语料中出现,那么它肯定不止出现一次。因此运用上述公式分析字串的内部结合强度时,一般只对在语料中出现次数大于两次的术语进行考察。

3　术语候选抽取及术语选择

对原始领域文本经过粗切分后得到的每一个字串,系统以字为单位顺序扫描,通过基于大规模领域语料计算其子串的内部结合强度 (互信息),把内部结合强度超过预先设定阈值的子串抽取出来,完成术语候选的抽取。

术语选择就是从术语候选集中选出正确的术语。对术语候选集观察分析后,我们发现术语候选一般包括基本词、人名、地名、机构名、普通词语搭配、正确的术语和无意义的字串组合。基本词,就是分词词表中已经有的词,这其中也有部分术语,如"机器翻译"、"正态分布"等,因为可以被分词系统正确切分,我们直接从候选中去除;对人名、地名和机构名等未登录词的识别,国内外的研究很多,而且已经达到不错的识别效果,因此,本系统也不作为处理对象。那么在所剩的术语候选中,术语干扰项就剩下普通词语搭配和无意义的字串组合了。

3.1　普通词语搭配前缀、后缀库的建立

普通词语搭配的内部结合强度也很高,例如"研究工作"、"理论体系"、"非常重要"、"十分高兴",因此第一步是从术语候选集中去除普通词语搭配。对大量的术语候选进行分析,我们发现普通词语搭配一般都有一些使用频率极高的前缀和后缀。因此,利用这些前缀和后缀把普通词语搭配从术语候选中去除,是一个合理的思路。在基于统计的基础上,我们用人工方式建立了普通词语搭配前缀、后缀库。表 1给出了普通词语搭配前缀、后缀的一些例子。

表 1　普通词语搭配前缀、后缀示例

序号	前缀	后缀	序号	前缀	后缀
1	基于	问题	4	有关	错误
2	可以	方法	5	可能	进行
3	处理	结果	6	进行	困难
…	…	…	…	…	…

3.2　术语词性构成规则

术语中的相当一部分是基本名词短语,但也有例外,如术语"自然语言理解"、"中文信息处理",结尾词为动词,并不符合基本名词短语的定义[6]。结合基本名词短语的词性构成信息,在对大量术语分析的基础上,我们总结了术语词性构成规则,如表 2所示。

表 2　术语词性构成规则

编号	规则描述
1	术语中至少含有一个动词、名词或名词性成分 (n, vn, an, Ng)
2	术语最后一个词为动词、名词或名词性成分 (v, n, vn, an, Ng)
3	术语第一个词不为介词 (p)、量词 (q)
4	术语中没有连词 (c)、代词 (r)、语气词 (y)

针对每一个术语候选的词性序列应用这些规则,得到系统术语抽取的最终结果。

4　实验结果及分析

4.1　术语抽取评价指标

本文采用的术语抽取评价指标如下:

(1)术语抽取准确率

$$p = \frac{\text{系统抽取的正确的术语数}}{\text{系统抽取的术语总数}} \times 100\%$$

(2)术语抽取召回率

$$r = \frac{\text{系统抽取的正确的术语数}}{\text{文本集中包含的术语总数}} \times 100\%$$

(3) F测量

$$F\text{-}Measure = \frac{2pr}{p+r}$$

4.2　实验结果

我们以《计算语言学概论》[7]的电子书稿约 21万字作为领域文本和统计语料进行了术语抽取实验,其中对术语候选的词法分析采用中科院计算所的 ICTCLAS词法分析系统。通过人工方式获得该语料中术语集合 (共 486个)作为术语抽取结果评价的依据。需要说明的是,分词词表中已经存在的和语料中出现次数少于三次的术语没有列入该术语集合。实验结果如表 3所示,其中最佳阈值是以 $F\text{-}Measure$为指标确定的。

表 3　实验结果表

最佳阈值	术语总数	抽取数目	正确数目	准确率 (%)	召回率 (%)	F-Measure(%)
13.4	486	525	379	72.19	77.98	74.97

系统抽取出的一些术语及其在语料中的出现频次如表 4所示。

表 4　系统抽取出的一些术语及其出现频次

术语	频次	术语	频次	术语	频次
自然语言	200	机器翻译系统	52	产生式规则	34
句法分析	145	信息提取系统	51	概率上下文无关文法	34
计算语言学	97	目标语言	48	句法分析技术	33
语言知识	82	语义知识	47	输入缓冲区	31
信息提取	80	部分分析	42	最大匹配	24
信息检索	65	句法结构	38	链接表达式	23
特征结构	63	信息检索系统	38	部分分析技术	22
上下文无关文法	60	自然语言处理	38	状态转移网络	22
状态转移	57	决策树	35	合一运算	20

4.3　分析

没有召回的术语可以分为两类:一类是字串内部结合强度低于最佳阈值的术语,主要为两字术语,如"论元"、"切词"、"施事"、"受事"、"树库"、"文法"、"语块"、"语料"、"组块"等和含有一个基本词的术语,如"语料库"、"语言学"等。两字词的最长子串是一个字,很显然字出现的频次比较高,导致用式 (2)计算的字串内部结合强度很低;含有一个基本词的术语,由于其最长子串——基本词的出现频次较高,同样导致字串内部结合强度很低。目前这种情况的术语还没有好的方式召回。另一类是结合强度虽然高于阈值,但由于术语选择阶段的分词错误,应用术语词性构成规则判别时被从术语候选中去除的术语。例如,"句 /q 法规 /n 则 /c"、"名 /q 词性 /n 成分 /n"、"自 /p 动机 /n"等,由于第一个词为量词 (q)或介词 (p),应用术语的词性构成规则,系统认为不是术语。 (下转第 77页)

表 2　实验样本中 10家企业的部分评估指标

样本代号	贷款回收率	流动比率	速动比率	净产值	贷款偿还率	…
1	14. 85	11. 46	11. 20	8750	36. 25	…
2	25. 10	5. 7	5. 8	11611	3. 9	…
3	1. 7	0. 54	0. 57	4174	1. 4	…
4	1. 6	1. 1	0. 92	11893	15. 37	…
5	10	6. 33	5. 44	6547	- 7. 7	…
6	20. 07	7. 7	8. 3	9690	- 5. 1	…
7	- 5. 5	- 6. 9	5. 7	2742	5. 39	…
8	31. 15	7. 55	7. 6	17998	- 16. 45	…
9	8. 59	3. 2	2. 95	7197	- 35. 49	…
10	0	- 4. 62	- 4. 48	2289	1. 4	…

表 3　归一化后的效用函数值

样本代码	*Y*1	*Y*2	*Y*3	*Y*4	*Y*5	…
1	0. 818	1	1	0. 257	0. 998	…
2	0. 98	0. 99	1	0. 41	- 0. 067	…
3	- 0. 301	- 0. 674	0. 99	- 0. 135	0. 331	…
4	- 0. 312	0. 817	1	0. 491	0. 836	…
5	0. 545	1	1	0. 16	- 0. 875	…
6	0. 93	1	1	0. 332	- 0. 788	…
7	- 0. 784	0. 998	1	- 0. 255	0. 1	…
8	0. 995	1	1	0. 786	- 0. 869	…
9	0. 426	0. 983	1	0. 127	0. 998	…
10	- 0. 462	0. 97	- 0. 97	- 0. 292	0. 33	…

表 4　遗传算法快速 k均值聚类后的分类信息

遗传算法样本	1	2	3	4	5	6	7	8	9	10
k均值聚类类别	1	3	4	5	5	2	7	1	6	8
遗传算法快速样本	1	2	3	4	5	6	7	8	9	10
k均值聚类类别	1	3	4	5	5	2	7	1	6	8

4　结论

由于企业资信评估指标体系包括定性指标和定量指标,在研究中,我们利用模糊数学隶属度关系,将定性指标定量表示,将在不同类型、不同量纲的原始评估值转换到[-1,1]区间,再利用遗传算法的快速 k均值算法聚类。由计算结果表明,常规聚类方法不能有效地处理局部极值问题,因此当初始聚类中心在整个样本空间不平衡时,很难将这种不平衡纠正,从而导致聚类结果对初始聚类中心的选取有着很大的敏感性。而基于 GA的 k均值聚类方法因具有很好的处理局部极值能力,对初始聚类中心的选取以及样本的输入次序没有任何要求。另一方面,从它们各自的收敛速度上看,基于 GA的 k均值聚类方法的收敛速度较慢,但是这显然比用常规方法对不同初始聚类中心进行聚类来获取全局最优解要有效得多。FGKA和 GKA可以获得全局最优解,且 FGKA算法时间性能明显比 GKA算法快。实验证明,FGKA中的目标函数,避免了消除非法个体的损耗,简化了变异算法,改善了 GKA算法的性能,提高了算法的收敛速度。

参考文献:

[1] 牟锐,张洪伟,刘向锋. SCM和 ERP结合下的供应商评估与选择决策模型[J]. 计算机应用研究, 2004, 21(4): 23-25, 57.

[2] 谢澍,张洪伟,魏筱毛,等. ERP与供应链结合的采购管理研究[J]. 计算机应用研究, 2002, 19(9): 91-93.

[3] Y Lu, S Lu, F Fotouhi, *et al*. Fast Genetic K-means Algorithm and Its Application in Gene Expression Data Analysis[R]. Technical Report TR-DB-06-2003.

[4] K Krishna, M Murty. Genetic K-means Algorithm. IEEE Transactions on Systems, Man and Cybernetics [J]. Cybernetics, 1999, 29(3): 433-439.

[5] 李敏强,等. 遗传算法的基本理论与应用[M]. 北京:科学出版社, 2003.

[6] 朱剑英. 智能系统非线性数学方法[M]. 武汉:华中科技大学出版社, 2001.

[7] 邢文训,等. 现代优化计算方法[M]. 北京. 清华大学出版社, 1999.

[8] 中国信用在线[EB/OL]. http://www.cn-co.com, 2003-10.

[9] 张文修,梁怡. 遗传算法的数学基础[M]. 西安:西安交通大学出版社, 2003.

作者简介:

朱丽(1980-),女,甘肃人,主要研究方向为智能信息系统数据库与计算机网络;张洪伟(1955-),男,四川人,教授,西德博士,西德博士后,主要研究方向为智能信息系统数据库与计算机网络;谭辉(1978-),男,江西人,主要研究方向为智能信息系统数据库与计算机网络。

(上接第 73页)对于这类情况,可以通过提高分词系统的准确率加以解决。错误抽取的术语一部分为普通词语搭配,如"计算公式"、"基本过程"、"一个名词组块"等,不能通过前缀、后缀信息去除,又不符合术语的词性构成规则。另一部分是无意义的字串组合,如"空格分开"、"分析参见"等。同样两种方法均不能去除,这也是导致准确率不太高的主要原因。

5　结论及未来工作

本文介绍了基于互信息的中文术语抽取系统,我们的主要工作是:基于互信息进行中文术语候选的抽取;统计建立了普通词语搭配前缀、后缀信息库;总结了术语词性构成规则;利用前缀、后缀信息库和术语词性构成规则进行术语选择。实验取得了较好的效果。未来的工作包括:研究对两字术语和含一个基本词的术语的抽取方法;通过对语料文本集合建立索引以提高系统的术语抽取速度等。

参考文献:

[1] Patrick Pantel, Dekang Lin. A Statistical Corpus-based Term Extractor[C]. Ottawa, Canada: Lecture Notes in Artificial Intelligence, 2001. 36-46.

[2] Shengfen Luo, Maosong Sun. Two-Character Chinese Word Extraction Based on Hybrid of Internal and Contextual Measures[C]. Sapporo, Japan: Proceedings of the 2nd SIGHAN Work Shop on Chinese Language Processing, 2003. 24-30.

[3] Munpyo Hong, Sisay Fissaha, Johann Haller. Hybrid Filtering for Extraction of Term Candidates from German Technical Texts[C]. Nancy: Proceedings of Terminology & Artificial Intelligence, 2001.

[4] Diana Maynard, Sophia Ananiadou. Terminological Acquaintance: The Importance of Contextual Information in Terminology [C]. Patras, Greece: Proceedings of NLP 2000 Workshop on Computational Terminology for Medical and Biological Applications, 2000. 19-28.

[5] Thian-Huat Ong, Hsinchun Chen. Updateable PAT-Tree Approach to Chinese Key Phrase Extraction Using Mutual Information: A Linguistic Foundation for Knowledge Management[C]. Taipei, Taiwan: Proceedings of the 2nd Asian Digital Library Conference, 1999. 63-84.

[6] 赵军,黄昌宁. 基于转换的汉语基本名词短语识别模型[J]. 中文信息学报, 1998, 13(2): 1-8.

[7] 俞士汶,等. 计算语言学概论[M]. 北京:商务印书馆, 2003.

作者简介:

张锋(1978-),男,博士研究生,研究方向为自然语言处理;许云(1976-),男,博士研究生,研究方向为自然语言处理;侯艳(1977-),女,硕士研究生,研究方向为信息检索;樊孝忠(1948-),男,博士生导师,研究方向为自然语言处理、数字化网络教学。