●仲云云　侯汉清　杜慧平

电子政务主题词表自动构建研究*

摘　要　电子政务主题词表是电子政务信息组织和检索的重要语义工具。传统手工编制叙词表的方法已不再适用于网络环境。电子政务词表的自动构建技术主要有基于 N-gram 方法的词汇收集和选择词间关系的自动识别。要想编制一部性能优越且容易应用的词表,应将计算机自动构建与传统方式编制结合起来,取长补短。表 7。参考文献 10。

关键词　电子政务　主题词表　词表编制　自动构建　N-gram
分类号　G254.24

ABSTRACT　Thesauri for e-government are important semantic tools for the organization and retrieval of e-government information. Traditional manual thesaurus compilation methods are no longer suitable to the networked environment. Among the technologies for the automatic construction of e-government thesauri, there are the N-gram-based vocabulary-collecting technology and the automatic recognition for the selection of word relationships. To compile a good and easy-to-use thesaurus, we should use both computer-based automatic methods and traditional manual methods. 7 tabs. 10 refs.

KEY WORDS　e-Government.　Thesaurus.　Compilation of thesaurus.　Automatic construction.　N-gram.
CLASS NUMBER　G254.24

目前国内外所研究的自动构建词表的方法包括“从 WordNet 转化”[1]、“概念空间”[2]、“整合既有词表”等。但这些方法基本上都是识别词与词之间的相关关系,即所编制的词表只能称为关联词表[3]。这对于编制一部比较正规的叙词表是不够的,必须要进一步识别其他词间关系。本文将尝试用计算机来自动识别等同、等级和相关关系,从而自动构建一部电子政务主题词表。

1　电子政务词表的自动构建技术

1.1　基于 N-gram 方法的词汇收集和选择

本文所建词表的词汇来源于现有词表及电子政务网页。所用的词表是《综合电子政务主题词表》和《中国分类主题词表》。所用的网站包括江苏共青团网、江苏共青团电子政务网、中国共青团网。网站中的网页数据主要是通过计算机编写的机器人下载程序完成对网站的下载,共下载 12000 余篇网页文献,下载完成后,将 HTML 文件进行预处理,转化为文本文件。

N-gram 分词方法的思想是:一个单词的出现与其上下文环境中出现的单词序列密切相关,第 n 个词的出现只与前面 n－1 个词相关,而与其他任何词都不相关。即是指将长度为 n(个字符)的窗口从文本的第一个字符处开始,自左向右连续移动,每次移动的步长为 1 个字符,窗口中出现的 n 个字符即为 N-gram。例如,“叙词表的自动构建”可以生成如下字符串:

n＝1:叙,词,表,的,自,动,构,建

n＝2:叙词,词表,表的,的自,自动,动构,构建

n＝3:叙词表,词表的,表的自,的自动,自动构,动构建

……

n＝8:叙词表的自动构建

鉴于中文关键词一般不超过 15 个汉字,设定最大抽取长度为 15,利用 GF/GL 权重值计算和关键词筛选算法来选择关键词。

GF/GL 权重法:词汇的重要性与其长度和在文献中的出现频率呈正相关,关键词在一篇文献中至少会出现两次。另外考虑到文本长度的影响,用文献长度对公式进行规范,具体定义如下[4]:

$$GF/GL = \frac{\log_2(freq) \times \log_2(len+1)}{\log_2(Tlen) \times \log_2(\max(len)+1)}$$

注:freq 表示字符串在一篇文献中出现的频率;len 表示字符串的长度;Tlen 表示该文献的长度。

关键词筛选算法:即是对分词后、通过 GF/GL 计算出权值的词汇进行筛选,根据选出的结果,子串和父串进行比较,设定它们之间长度差异最大值为 k(k≥4)。表 1 为关键词“共青团”筛选样例:

* 本文为国家社会科学基金项目资助项目“基于知识组织系统的电子政务信息资源管理”(05BTQ021)的成果之一。

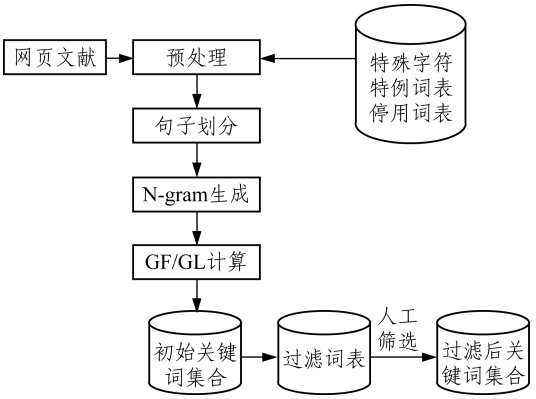

图 1　N-Gram-Keyword 方法流程图

表 1　关键词筛选样例 *

ngrams	len	freq	GF/GL	kd
共青	2	3	0.114	Y
青团	2	3	0.114	Y
共青团	3	3	0.144	x

以上算法,容易实现并且收词全面,但运算量大、效率较低、且容易生出很多噪声词,例如:个代表、被感动、力争、市鼓楼区、星星火炬、不公平、小张等等,由此构建一部过滤词表,将这些噪声词纳入到过滤词表中。根据此方法,从 12000 篇文章中,初步选出关键词 19445 个。经规则结合人工过滤,最后选定关键词 10824 个,这些词汇将作为构建词表的主要词汇来源。

1.2　词间关系的自动识别

叙词表是由概念集合以及概念之间的关系组成,在信息检索领域的概念关联包括等同、等级和相关三种关系。

词表的自动构建是指通过计算机技术,由程序根据一定的规则来自动收集词汇构建词表。自动构建的词表按词间关系的显示可以分为两种:一是只显示等同和相关关系,将等级关系作为相关关系来显示;二是将等同、等级和相关关系全部显示出来。前者的词间关系罗列比较粗泛,近似于关联词表,构建起来比较简单,但难以实现概念检索,不易扩检和缩检,影响检全率和检准率;后者的词间关系细致、全面,类似于传统的叙词表,构建起来较为复杂,但可以实现概念检索,容易进行扩检和缩检,检索效率较高。综合分析,本文选择后者进行构建。其核心内容就是通过计算机来完成对词间关系包括等同、等级和相关关系的自动识别,从而自动生成一部主题词表。

1.2.1　基于模式匹配或同义词典的等同关系的识别

目前,国内识别汉语同义词的主要方法有:①基于单汉字的字面相似度算法;②基于词素的字面相似度算法;③基于《同义词词林》等的语义相似度算法;④基于词汇共现分析的算法;⑤基于模式匹配的识别方法。本文采用"模式匹配 + 同义词典"两种方法相结合进行同义词识别。

(1)基于 WEB 网页的模式匹配方法。根据网页自身的特点,将同义词提取模式分为如下几种[5]:

① < Prefix > "词汇"简称/也称/又称/俗称/以下简称 < Postfix > 左括号 + 同义词 + 右括号。例如:"中国共产主义青年团中央委员会"(以下简称团中央)。

② < Prefix > "词汇"简称/也称/又称/俗称/以下简称/是……的简称 < Postfix > 同义词。例如:计算机俗称电脑。

③ < Prefix > "中文词汇" + 左括号英文同义词 < Postfix > 右括号。例如:什么是"非典"(SARS)。

④多层同义。如:"采样"又称取样或抽样。

表 2　模式匹配结果分析

总网页数	抽出的同义词对	正确的同义词对	正确率(%)	执行效率(%)
13859	512	386	75.4	2.8

因为网页本身行文格式的不规范,真正符合上述模式的格式较少,很难用计算机将网页中的同义词全部自动提取出来,其执行效率要远远低于同义词典的模式匹配。

(2)同义词典匹配。同义词典共收录同义词 3587 对,词对来源于两个部分,一个是综合电子政务主题词表的同义词[6],另一个来源于《中分表》中的部分同义词。将关键词与同义词典中的同义词对进行匹配比较来识别同义词和同义词组。

表 3　模式匹配及同义词典方法抽出同义词对的比较

总同义词对	模式匹配生成的词对数及百分比	同义词典生成的词对数及百分比
1341	386(29%)	955(71%)

* kd 表示关键词是否保留,Y 表示过滤掉的字符串,x 表示保留的字符串,即关键词。

1.2.2　基于字面相似度算法的等级关系的识别

词素是构成词的单位,在意义上不能再分解。具有相同词素的词和词组之间,必然绝大多数在意义上有某种联系,存在着聚类现象——字面成族现象。据统计,在《综合电子政务主题词表》中,字面成族的占70%以上[7],其他则为概念成族或按政务工作成族。通常,汉语语词的中心在于词的后部。因而本文结合字面相似度,根据后方一致原则进行词的入族处理并进行上下位类的划分。即:根据字面相似度的结果,如果两个词包含相同的词素,且相同的词素位于词的后方,那么包含字数少的词作为包含字数多的词的上位词;反之,作为下位词处理。根据这个规则,"青年"和"知识青年"就可以归为一族,其中,"青年"是"知识青年"的上位词,而"知识青年"是"青年"的下位词。

本文利用的字面相似度算法考虑三个因素:匹配度、匹配序、重心后移规律。

计算公式如下:

$$Sim = 60 \times \left(\frac{xsword}{ctrlword} + \frac{xsword}{keyword}\right)\Big/2 + 40 \times dp \times \left(\sum \frac{c - xsword(i)}{\sum ctrlword(i)} + \sum \frac{k - xsword(i)}{\sum keyword(i)}\right)\Big/2$$

根据以上算法,"计算机"和"微型计算机"的字面相似度为:

$$sim = 60 \times \left[\frac{3}{3} + \frac{3}{5}\right]\Big/2 + 40 \times \frac{3}{5} \times \left[\frac{6}{6} + \frac{4}{5}\right]\Big/2 = 69.6(\%)$$

按照后方一致的原则,以上两个词拥有相同的词素"计算机"且"计算机"位于词的后方,那么"计算机"和"微型计算机"可以归为一族,且"计算机"是"微型计算机"的上位词。

1.2.3　基于词聚类算法的等级关系的识别

按字面成族原理识别等级关系词汇,无法识别无字面相似特征的等级关系词。现试用一种基于相似度的词聚类算法,把表达不同主题范畴的词汇分别聚集成族,待进一步识别词族内的等级关系。

叙词表的词族包含表达同一主题的词汇,这些词汇在语义上是相似的。为了对同族词汇实施聚类,首先要计算词汇之间的语义相似度。

语义相似的两个词在特定的上下文中可以互相替代[8],而未知词汇的涵义常常能从它的上下文推导得出。这样,词汇 T_i 的语义可以用其在语料库中经常同现的词汇来表达,如果目标词汇 T_i 和 T_j 的同现词汇有很大重叠,那么它们在语义上很相似。

首先以同一篇网页文本为同现窗口,利用 Dice 测度算法计算词汇之间的关联度,用与词 T 最相关的前 K 个词汇代表其特征,构成词 T 的特征向量 $T(<T_1,W_1>,<T_2,W_2>,\ldots<T_K,W_K>)$,其中,$T_i$ 表示相关词汇,W_i 表示相关词汇 T_i 与 T 的关联度值。那么,两个词 T_i 与 T_j 间的语义相似度 $Sim(T_i,T_j)$ 就可以借助于向量之间的某种相似性函数来度量。本文采用了向量空间模型中常用的余弦相似度算法来计算两词汇之间的语义相似度。词汇向量之间的夹角越小,它们的语义相似度越大。计算公式如下:

$$Sim(T_i,T_j) = \frac{\sum_{k=1}^{m} t_{ki}t_{kj}}{\sum_{k=1}^{m} t_{ki}^2 \sum_{k=1}^{m} t_{kj}^2}$$

聚类算法过程如下所示:

步骤 1:初始化相似度矩阵。把词表中的每个词作为一个单独的簇,通过余弦相似度系数计算簇与簇之间的距离,生成距离矩阵。

步骤 2:找出最相似的两个簇。

步骤 3:合并最相似的两个簇,生成一个新词簇。

步骤 4:计算刚合并的簇与剩余其他簇之间的相似度,更新距离矩阵。

步骤 5:判断最大相似度是否小于阈值,是则结束程序,否则转步骤 2。

计算簇与簇之间相似度时,可以采用等级聚类算法中单连通、全连通和平均连通算法中对簇间距离的定义[9]。

本文主要利用字面相似度算法来识别等级关系,词聚类算法作为等级关系识别的补充和参考。

1.2.4　基于相关度算法的相关关系的判断

相关关系的挖掘一般都是采用计算语言学与统计学的知识来实现。计算两种信息之间相关度的算法有很多,主要包括:互信息、Dice 测度、Jaccard 系数、开方统计、极大似然比等方法[10]。本文采用 Dice 测度来计算词与词之间的关联。原因是,公式中各测度因素设置较为合理,可以有效克服"零概率事件"和低频现象。Dice 测度公式如下:

$$D(S_1,S_2) = \frac{P(AB)}{\frac{1}{2}[P(A) + P(B)]}$$

用 Dice 测度来构建关联词表的步骤包括:

将用 N-gram 算法筛选出的关键词按字顺排序→生成关键词表→按字顺计算各个词的关联度→确定同现窗口(同一篇网页)→统计两个词(A 和 B)分别

出现的频次(即统计分别包括 A 词和 B 词的文献数)→统计两个词共现的频次(即同时包括 A 词和 B 词的文献数)→利用算法计算关联度→选出关联度排在前 10 位的词汇作为该词的关联词。下面以“计算机”为例来说明其关联词的生成情况。

设“计算机”为 ka,被比较词为 kb,二者的关联度为 rp,经过和文件库中所有符合条件的关键词(即凡是和“计算机”在同一篇网页中出现的关键词)进行比较计算,最后生成和“计算机”最关联的 10 个词及其结果见表 4:

表 4　关联度生成结果显示

关键词对(ka – kb)	Ka 出现频次	Kb 出现频次	Ka, Kb 共现频次	rp	排序
计算机 – 电脑	113	121	22	0.188	1
计算机 – 英语	113	91	16	0.157	2
计算机 – 软件	113	59	13	0.151	3
计算机 – 网络	113	367	28	0.117	4
计算机 – 电子	113	95	12	0.115	5
计算机 – 网站	113	178	15	0.103	6
计算机 – 专业	113	431	27	0.099	7
计算机 – 上网	113	92	10	0.098	8
计算机 – 互联网	113	79	9	0.094	9
计算机 – 用户	113	64	8	0.091	10

以上,生成的关联结果基本是合理的,“计算机”和“电脑”本身是同义词,关联度也是最高的,而“网络”、“软件”等作为“计算机”的相关词被抽出,也是合理的,但是“专业”和“英语”也作为相关词被抽出,就不是很确切。分析其原因,主要是以一篇文献作为共现窗口过于宽泛。可考虑作如下改进:适当调整同现窗口,将同现窗口缩小到一个段落或题名或标引词中来判断。

1.3　实例分析

以“团员青年”一词为例,说明生成等级和相关关系的过程:根据上述字面相似度以及 Dice 测度的结果,“团员青年”可以生成很多词对,为避免有太多的冗余数据,选择关联词表中相关度排在前 5 位的词对,结果如下:

字面相似度:

团员青年——社区团员青年——0.746999979019165
团员青年——青年——0.620000004768372
团员青年——城乡青年——0.579999983310699
团员青年——城镇青年——0.579999983310699
团员青年——出国青年——0.579999983310699

Dice 测度:

团员青年——共青团——0.113597244024277
团员青年——团组织——0.106707319617271
团员青年——团委——0.104830421507359
团员青年——非典——0.0995850637555122
团员青年——团支部——0.0976863726973534

以上提到根据字面相似度中后方一致的原理来生成等级关系,根据上述结果,很显然,青年是团员青年的上位词,即作为团员青年的属项;社区团员青年是团员青年的下位词,即作为团员青年的分项;那么剩下的词作为团员青年的相关词出现,即作为团员青年的参项。计算机可以自动生成如下词间关系:

团员青年
　F　社区团员青年
　S　青年
　C　城镇青年
　C　出国青年
　C　城乡青年
　C　共青团
　C　团组织
　C　团委
　C　非典
　C　团支部

以“共青团”为例,来说明等同关系的自动生成过程:本文词汇的等同关系的确立来源于两个方面,

其一是网页中模式匹配的方式得出的同义词;其二是来源于自动构建的共青团电子政务同义词典。中国共青团网站“http://www. ccyl. org. cn”出现这样一句话:

> 中国共产主义青年团(简称共青团)是中国共产党领导的先进青年的群众组织,是广大青年在实践中学习中国特色社会主义和共产主义的学校,是中国共产党的助手和后备军。

根据上述模式,即可抽出“中国共产主义青年团”和“共青团”是同义词对。另外,在共青团电子政务同义词典中,出现“共产主义青年团 Y 共青团”。由此,将两种模式得出的结果进行合并,生成“共青团”一词的等同关系:

共青团
　D　共产主义青年团
　D　中国共产主义青年团

2　自动构建电子政务词表的性能分析

2.1　自动构建词表的性能分析

经统计,自动生成的《共青团电子政务词表》各项指标如表 5 和表 6 所示:

表 5　《共青团电子政务词表》词汇性能参数

总词量	非正式主题词数	属项词数	分项词数	参项词数	无关联词数
9440	1341	3696	2797	50076	145

表 6　《共青团电子政务词表》词长统计

总词数	1 字词数	2 字词数	3 字词数	4 字词数	5 字词数	6 字词数	7 字至 15 字之间词数
9440	12	4048	1056	2882	474	587	381

$$\text{入口率} = \frac{\text{非正式主题词数}}{\text{正式主题词数}} = \frac{1341}{9440-1341} = 0.17$$

$$\text{参照度} = \frac{\text{分项词数} + \text{属项词数} + \text{参项词数}}{\text{正式主题词总数}} = \frac{3696+2797+50076}{8099} = 6.98$$

$$\text{属分参照度} = \frac{3696+2797}{8099} = 0.80$$

$$\text{参项参照度} = \frac{50076}{8099} = 6.18$$

$$\text{关联比} = \frac{\text{正式主题词总数} - \text{无关联词数}}{\text{正式主题词总数}} = \frac{8099-145}{8099} = 0.98$$

$$\text{先组度} \approx \frac{\text{三字以上词数}}{\text{总词数}} \approx \frac{5380}{9440} \approx 0.57$$

注:因为先组度比较难以定义,这里取大于 3 字以上词长数作为先组词来对先组度大概进行估算。

根据以上的统计结果及比较分析,可以总结出计算机自动构建的《共青团电子政务主题词表》有如下特点:①《共青团电子政务主题词表》收录的词量偏多,因为绝大部分的主题词是通过计算机自动从有关共青团网站中筛选的,但所选词汇并不局限于共青团类,很多和教育、科技类比较相关的词都被选进来了,所以收录词汇的范围比较宽泛。②《共青团电子政务主题词表》的性能总体较好:其参照度、关联比、先组度较高。但同时也存在一定问题:入口率偏低、生成的词间关系不够准确,时有冗余甚至错误的词间关系生成且参项词过多。

2.2　自动构建词表方式与传统编表方式的比较

表 7　计算机自动构建词表与手工编表的比较

编表方式	效率	成本	更新	应用	词汇冗余	严谨性	准确性
传统编制	低	高	较难	较复杂	高	好	好
自动构建	高	低	较容易	较容易	低	较差	较差

从表7可以看出,计算机自动构建的叙词表在效率、成本、更新、应用方面均优于传统方式编制的词表,但准确性及严谨性则不如。所以,如果想编制一部性能优越且容易应用的词表,可考虑将两种方式结合起来,取长补短,在计算机自动构建的基础上适当加以人工干预,即参照已有的一些词表,对生成的词间关系进行挑选、调整,这样编制出来的词表一定具有比较好的性能和应用的价值。

参考文献

1 张俐等. 中文 Wordnet 的研究及实现. 东北大学学报,2003(4)

2 Hsinchun Chen. A Concept Space approach to Addressing the Vocabulary Problem in Scientific Information Retrieval: An Experiment on the Worm Community System. Journal of the American Society for Information Science,1997,48(1)

3 Yuen-Hsien Tseng. Automatic Thesaurus Generation for Chinese Documents. Journal of the American Society for Information Science and Technology,2002(11)

4 张雪英. 基于粗糙集理论的文本自动分类研究. 南京理工大学博士学位论文,2005

5 陆勇. 面向信息检索的汉语同义词自动识别. 南京农业大学硕士学位论文,2005

6 赵新力等. 综合电子政务主题词表(试用本)字顺表. 北京:科学技术文献出版社,2005

7 倪静. 电子政务主题词表的编制及应用研究. 中国科技信息技术研究所硕士学位论文,2003

8 Word clustering. http://www. ilc. cnr. it/EAGLES96/rep2/node37. html

9 Manning C D, Schutze H 著;苑春法等译. 统计自然语言处理基础. 北京:电子工业出版社,2005

10 夏祖奇. 基于关联概念空间的自动标引与自动分类研究. 南京农业大学硕士学位论文,2004

仲云云 南京农业大学信息科技学院硕士研究生。通讯地址:南京。邮编210095。

侯汉清 南京农业大学信息科技学院教授,博士生导师。通讯地址同上。

杜慧平 上海师范大学图书馆。通讯地址:上海。邮编200234。

(收稿日期:2007-07-16)

(上接第93页)

5 图林丫枝. 用户永远都是正确的——你图林旅途中必须正视的一座高山. [2007-01-19]. http://blog. sina. com. cn/u/53acbee9010007aw

6 耄耋少年. 不加永远可能更好. [2007-01-24]. http://blog. sina. com. cn/u/4bd4c87b010006u7

7 思考的乐趣——雨禅的博客. "用户永远正确"定理成立的条件. [2007-01-19]. http://rainzen. bokee. com/6053373. html;思考的乐趣——雨禅的博客. 用户正确是个伪问题. [2007-01-25]. http://rainzen. bokee. com/6066222. html;蓝天白云. 用户并非永远正确. [2006-06-20]. http://hi. baidu. com/blueyye/blog/item/29b1d52a3fe12c2dd52af165. html;蓝天白云. 用户命题正说与趣解. [2007-01-28]. http://hi. baidu. com/blueyye/blog/item/6e670508af88b530e924889e. html

8 程焕文. 新世纪中国大学图书馆发展之我见(之五)——中山大学图书馆馆长程焕文访谈录. 大学图书馆学报,2001(6);程焕文,王蕾. 21世纪高校图书馆管理的新理念. 大学图书馆学报,2003(2)

9 司步林. 图书馆的学术地位刍议——兼与程焕文先生商榷. 大学图书馆学报,2002(3)

10 梁启超. 中国近三百年学术史. 北京: 东方出版社,1996,3: 14

程焕文 中山大学资讯管理系教授,博士生导师。通讯地址:广州市。邮编510275。

(收稿日期:2007-10-25)