2007年 1月 第 1期
Vol. 28 No. 1 2007

一种利用 BC方法的关键词自动提取算法研究

张 敏, 耿焕同, 王煦法

(中国科学技术大学 计算机科学与技术系, 安徽 合肥 230027)
E-mail: zhangmin@ mail.ustc.edu.cn

摘 要: 通过分析几种常见关键词自动抽取方法的特点和不足,以 KeyGraph算法思想为基础,构建词语网络并利用网络节点中心度(Betweenness Centrality)理论,提出了一种新的自动抽取关键词算法.通过分析和比较,新算法提取的关键词更能体现文档内容,并且相对低频而意义重要的关键词也能被提取出.最后,通过与 TF和 TFIDF算法的比较和分析,获得了令人满意的结果.

关 键 词: 关键词抽取;中心度;词语网络;社会网络
中图分类号: TP18 文献标识码: A 文 章 编 号: 1000-1220(2007) 01-0189-04

Automatic Keyword Extraction Algorithm Research Using BC Method

ZHANG Min, GENG Huan-tong, WANG Xu-fa

(*Department of Computer Science and Technology, University of Science and Technology of China, Hefei* 230037, *China*)

Abstract In this paper, the characteristics and disadvantages of several common automatic keyword extraction methods are introduced firstly. Then based on the ideas of KeyGraph algorithm, a new automatic keyword extraction method is proposed. After analysing and comparisons, the advantages of the new algorithm are revealed the extracted words can represent more information of the document content, and the low frequent but important terms can also be extracted. Compared with the classic TF and TFIDF algorithms, a satisfying result has been gotten from the experiment and analysis.

Key words keyword extraction; betweenness centrality; term network; social network

1 引 言

随着 WWW的普及与发展,网上信息量成指数剧增, Internet已成为信息发布的第四媒体,因此文本信息处理的需求也就更为迫切.文档常由一个词语集合表示,且对文档的检索、比较等操作都是在此基础上进行的,而这组词语被称为文档的关键词.文档的关键词是信息检索技术和文档摘要生成的基础.

语言学专家可以手工从文档中提取出令人满意的关键词,但对海量的文档信息而言,依靠人类专家手工提取关键词显然是不可行的,为此很多自动提取关键词方法被提出: 1)统计方法: 统计文档中每个词语出现的频率,将频率高于某一阈值的词语作为关键词[1],该方法虽简单快速,但由于一些高频词语的重要性低而一些相对低频词语的重要性却很高,因此,[2,3]中提出了一些改进方法; 2)基于结构的方法[5]: 根据新闻文章中主要信息集中于固定位置的特征,在文章中的相应位置和标题处进行关键词抽取,但该方法在处理其他类别(如科技)文章时往往并不有效; 3)基于自然语言理解: 其能从文档中提取出最高正确率的关键词,但该领域的研究还未达到解决通用问题的地步.例如,背景知识库是必备的,该知识库要包括句法、语义等知识[4]; 4) KeyGraph算法[5]: 该算法将高频词语及其相互关系映射为图,利用该图计算文档中每个词语的 Key值进行出关键词抽取,该算法不仅考虑了文档中的高频词集,还可以抽取出重要而相对低频的词语.但该算法需要设定的参数过多,如顶点数、边数等,因而常造成边界上的取舍问题,影响算法的稳定性和精度; 5) KeyWorld算法[6]: 该算法将文档中的词语按给定规则构造为一个网络,通过验证网络的小世界特征[7],提取出对网络平均路径长度有剧烈影响的节点即关键词,但其未能解释关键词与平均路径长度变化量之间的关系且网络连通性常难以确保.

针对上述算法的特点和不足,本文提出的关键词提取算法力求达到以下目标: 1)能够体现作者主要思想的重要词语,而不仅仅是文档中的高频词语; 2)仅仅利用文档的文本信息,而不需要其他的背景知识; 3)该算法可以对提取出的关键词给出合理的解释,而不仅仅是实验验证.为此,本文基于社会网络理论,将预处理后的文档映射为一个词语网络,通过计算网络中节点的中心度 Betweenness Centrality(简称 BC)进行关键词提取.

2 算法的思想与流程

[5]中提出文档 D的主要内容可以由词语网络 G来表

收稿日期: 2005-10-10 收修改稿日期: 2006-02-23 作者简介: 张 敏,男, 1981年生,博士研究生,主要研究方向为数据挖掘,多目标优化的进化算法;耿焕同,男, 1973年生,博士研究生,主要研究方向为人工智能,知识发现;王煦法,男, 1948年生,教授,博士生导师, CCF高级会员,主要研究方向为智能信息处理.

示,即 G表示了文档 D的主题,同时文档的主题又是由一系列子主题组成,这些子主题对应的就是 G的连通子图,而连通子图中的中心高频词语和连接两个子图的相对低频词语就是对 G具有关键作用的词语即关键词.基于该思想,在图 1中子图的中心词语 b、c和 k,及子图间起连接作用的词语 f都将作为关键词被提取出来.

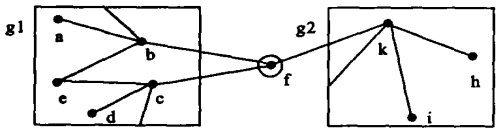

图 1 词语网络图示例

Fig. 1 Example of term network

社会网络理论将人及其相互关系映射为网络结构中的节点与边,用于研究社会中人与人之间的各种关系和人在网络中的作用及网络的发展趋势.随着社会网络理论研究的不断发展,一系列计算网络中节点重要度的方法被提出[12],其中以研究网络节点的中心度为主.

在社会网路中,节点代表社会中的个体,其相互关系用边来表示,网络中连接紧密的子图表示个体特征相似的群体或该群体有着相同的偏好,而整个社会网络也就一系列这样的群体所构成的,因此,网络中子图的中心节点和连接若干个子图的节点就能反映出社会网络的主要特征,将图 1视为社会网络,则节点 b、c、k、f就体现了该网络的主要特征,因此只要找到一种合适的方法将文档信息映射为社会网络,利用节点中心度方法就可以提取出体现文档特征的词语.

设网络的拓扑结构为图 $G=\{V,E\}$,g_{ij}表示从节点 i到节点 j的最短路径的个数,$g_{ij}(k)$表示从 i到 j的最短路径中通过 k的个数,则对一个顶点 v的 BC值可定义为:

$$BC(v)=\sum_{u\neq v\neq w}\frac{g_{uw}(v)}{g_{uw}},u\neq v\neq w \qquad (1)$$

对 $BC(v)$规一化后得到

$$BC(v)=\frac{1}{(|V|-1)(|V|-2)}\sum_{u\neq v\neq w}\frac{g_{uw}(v)}{g_{uw}},u\neq v\neq w \qquad (2)$$

同时按照该定义也可计算出边的 BC值,在此不再赘述.

从上面的描述中可以发现,在群体中处于中心位置的个体或在多个群体之间起连接作用的个体具有很高的 BC值,而这些个体也代表了整个社会网络的主要特征.在此,本文将社会网络中的 BC方法应用到文档的关键词提取中,从而可以提取出反映文档内容信息的词语,同时也考虑了相对低频的词语.

从文档 D来看,子图的中心节点处于子图内最短路径的次数往往很多,则这些中心节点的 BC值就相当高.以图 1为例,节点 b、c在子图 $g1$中处于中心位置,从而 a、e、d节点对之间的最短路径都会通过 b或 c,b和 c的 BC值就远远高于 $g1$中其它节点,最终 b、c和 k作为关键词被提取出来.

另一方面,连接两个子图的节点常位于两个子图之间对应顶点之间的最短路径上,这样的节点的 BC值也很高.例如在图 1中,节点 f将子图 $g1$和 $g2$连接起来,则 $g1$内节点与 $g2$内节点之间的最短路径都经过 f,从而将 f也作为关键词提取出来.另外在图 1中,若连接 $g1$和 $g2$的词语是多个时,利用 BC值仍然可以将其重要连接作用的词语提取出来.

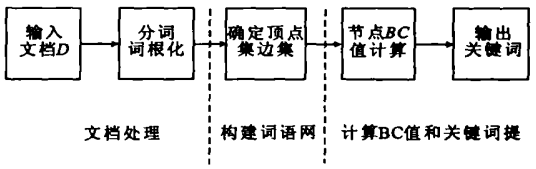

图 2 算法流程图

Fig. 2 Flow diagram of algorithm

从上述两个方面的分析可以看出,选择顶点的 BC值计算方法恰好满足了算法的思想,达到了既定目标.基于上述思想和方法,本文提出的算法的流程图如图 2所示,并将在第 3节中加以详细描述.

3 关键词提取主要算法描述

3.1 文档预处理

文档(记为 D)由句子组成,同时每个句子又由若干个词语组成.在对文档 D进行预处理之前,需要建立一个"停用词表":其中的词语多为冠词和介词且语义内容少,如"才能"、"也"、"对于"、"and"等词语[4],文档中的这些词语在分词时将被过滤.对英文文档,在分词处理前要进行词根化处理[8],如"run"、"running"、"runs"都被词根化为"run".对中文文档,需要建立字典来进行分词,同时还需要根据词语的出现频率来识别短语,如人物名称在文档中出现多次后,分词处理系统可自动抽取出该短语.

经过对文档 D的预处理后,用 D_{terms}表示从 D中抽取的词语集合.

3.2 构建词语网络

将文档 D映射为一个图 G,根据图论中图的定义 $G=\{V,E\}$,则需要对 G中的顶点集和边集给出相应的定义.文档 D对应的图 G中,顶点表示词语,边表示两个词语的关联程度即两个词语在 D中的同现(在同一个句子中出现)频率.

◦ 顶点集 对 D_{terms}中的词语按在 D中的出现频率进行统计,选择词频高于指定阈值的词语作为 G中的顶点集 V,其依据是文档中的高频词语可以反映作者的主要思想,同时将频率过低的词语去除可以降低计算复杂度.

◦ 边集 首先给出顶点集 V中的两个词语 w_i和 w_j的关联度定义,如下

$$assoc(w_i,w_j)=\sum_{s\in D}\min(|w_i|_s,|w_j|_s) \qquad (3)$$

其中,$|x|_s$表示词语 x在句子 s中出现的次数.该公式的建立是来源于 w_i只与最邻近的 w_j相关的假设.在关联度大于零的两个词语顶点间都加入一条无向边,边集用 E表示.

在确定 G中的边集之后,对边赋予权重 $weight$,表示在图 G中从顶点 v_i到达顶点 v_j的代价,与社会网络中个体的相互关系相似,定义如下:

$$weight(v_i, v_j) = f(assoc(w_i, w_j)) \quad (4)$$

其中,v_i和v_j对应的词语分别为w_i和w_j,按照关联度、权重的定义和社会网络个体的相关性定义,$f(x)$在$x>0$时为单调减函数,实验中定义的权重函数为:

$$weight(v_i, v_j) = 1/assoc(w_i, w_j) \quad (5)$$

上边的步骤成功地将文档D映射为图G,同时G也反映了D中的主要信息和作者的主要思想.下面将利用图G来提取D中的关键词语.

3.3 计算词语的*BC*值和提取关键词

从第2节中的分析和构建词语网络的方法中,可以发现图G表示了文档D的主题,而其子主题则蕴含在G中,通过计算图G节点BC值就可以对D进行关键词提取.按节点的BC值排序,选出指定数目的词语即文档D中的关键词语.整个算法的主要形式化描述如下:

Step1. D_{terms} = Split_Document(D)
/对文档D进行分词处理,获得词语列表

Step2. V = Select_Vertex(D_{terms})
/从词语列表中选择词语作为网络顶点

Step3. E = Create_Edge(V,D)
/根据顶点集和文档构造边集

Step4. For each v in V do
/计算网络中每个节点的BC值

$$BC(v) = \frac{1}{(|V|-1)(|V|-2)} \sum_{u \neq v \neq w} \frac{g_{uw}(v)}{g_{uw}}$$

Step5. *Keywords* = Select_KW($BC(V)$)
/根据节点的BC值选择关键词

设D中的词语数为n,$|V| = m$,$|E| = l$,$|S(D)|$表示D中的句子数,则整个算法的时间复杂度为:

$O(n) + O(n) + O(m*(m-1)*|S(D)|) + O(ml + m^2 \log m) + O(m \log m)$[10],

每一项分别依次对应算法中的每一步骤.通常m为指定提取关键词数的常数一般在10~20倍之间,因此$m << n$且$|S(D)| < n$,则整个算法的时间复杂度为$O(n)$.

4 实验设计与分析

为了检验本文提出的关键字提取算法,我们从"人民网"(http://www.people.com.cn/)选择了一些文档进行实验,同时与TFIDF算法的结果进行对比.图3中显示的是对文档"温家宝总理在庆祝香港回归祖国六周年酒会上的讲话"(http://www.people.com.cn/GB/paper464/9567/883861.html)进行相关处理后的词语网络的可视化显示(词语的阈值为1),从中看到词语"香港"与"发展"是文章中的中心词语,与文章中的很多词语都有关联.利用本文的算法计算网络中顶点的BC值,从而得到了表1中所列的关键词.同时,我们从"人民日报"中提取了关于香港回归主题的一个月文档集,以此利用TFIDF提取实验文档的关键词和利用TF

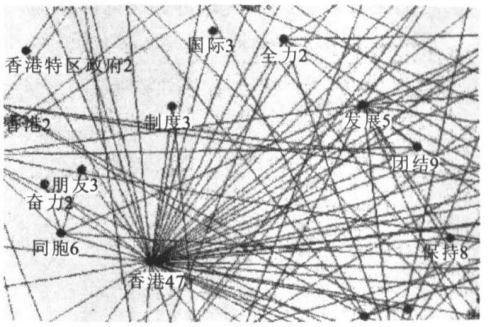

图3 词语网络图的可视化

Fig. 3 Visualization of the term network

方法提取关键词,分别如表2和表3中所示.

表1 利用BC值提取的关键词

Table 1 Keywords extracted by the BC values

关键词	$BC(v)$	频率
香港	0.849406	47
发展	0.065069	13
政府	0.055426	4
一国两制	0.049484	5
方针	0.044808	4
中央政府	0.003604	5
社会	0.002922	6
经济	0.002533	6
稳定	0.001559	7
同胞	0.000584	6

对比表1、表2和表3,可以发现TFIDF提取出的关键词主要体现了文档的标题信息,而难以反映出文档的内容信息,同时TF算法未能考虑到一些相对低频词语,而本文算法提取的关键词却体现了文档的主要内容,达到了既定目标,体现了算法的优越性.

表2 利用TFIDF提取的关键词

Table 2 Keywords extracted by TFIDF

关键词	权重	频率
香港	0.14434	47
温家宝	0.04752	2
回归祖国	0.04189	1
酒会	0.03919	2
一国两制	0.03576	5
中央政府	0.03311	5
回归祖国六周年	0.03089	3
同胞	0.03054	6
总理	0.03046	2
发展	0.02910	13

为了进一步验证算法的有效性,我们从"中文自然语言处理开放平台"(http://www.nlp.org.cn/)获取了10个类别共500篇的科技文章进行实验,这些文章的关键词作者已在文章中给出,使用TF、TFIDF和自己的算法进行对比实验,

表 3 利用 TF提取的关键词

Table 3 Keywords extracted by IF

关键词	频率
香港	47
发展	13
明天	8
美好	8
稳定	7
同胞	6
经济	6
社会	6
中央政府	5
祖国	5

利用精确率、召回率和综合指标 F_U[11]来评测自动提取的结果,其定义如下:

$$\text{精确率} = \frac{\text{提取正确的关键词数}}{\text{提取的关键词数}} \quad (6)$$

$$\text{召回率} = \frac{\text{提取正确的关键词数}}{\text{文档中的关键词数}} \quad (7)$$

$$F_U = \frac{2\times \text{查全率}\times \text{查准率}}{\text{查全率}\times \text{查准率}} \quad (8)$$

在实验中,词语的阈值设置为 4,同时每种方法提取的关键词数设置为 10,实验结果如表 4中所示.

表 4 各种方法的实验结果

Table 4 Experimental results of the three methods

方法/指标 实验数据	精确率			召回率			综合指标		
	TF	*TFIDF*	*BC*	*TF*	*TFIDF*	*BC*	*TF*	*TFIDF*	*BC*
Space (50)	0.352	0.348	0.366	0.702	0.685	0.711	0.458	0.441	0.473
Energy (50)	0.328	0.330	0.320	0.631	0.643	0.622	0.423	0.426	0.419
Electronics (50)	0.342	0.336	0.348	0.661	0.655	0.672	0.442	0.438	0.452
Computer (50)	0.320	0.360	0.376	0.620	0.742	0.786	0.421	0.504	0.508
Mine (50)	0.332	0.328	0.340	0.643	0.631	0.655	0.428	0.422	0.435
Transport (50)	0.304	0.312	0.318	0.612	0.624	0.631	0.393	0.407	0.416
Environment (50)	0.350	0.342	0.348	0.696	0.688	0.695	0.451	0.442	0.447
Agriculture (50)	0.346	0.348	0.352	0.682	0.691	0.703	0.449	0.453	0.461
Economy (50)	0.312	0.308	0.320	0.615	0.605	0.629	0.405	0.399	0.412
Law (50)	0.330	0.336	0.342	0.641	0.653	0.661	0.432	0.441	0.448

表 4中的数据都是对 50篇文档进行计算后的均值,从实验数据中可以看出利用 *BC*方法进行关键词提取的有效性.例如,在 Computer类别中的文章中,文档"一种理性 Agent的 BDI模型"(软件学报 Vol. 10 No. 12 1999)中 *BC*算法不仅提取出高频关键词: 信念(63)[1],愿望(68),意图(54),agent(38),还提取出相对低频的关键词: 理性(8),思维状态(8),而使用 TF TFIDF算法却仅能提取高频关键词,从而验证本文算法思想的正确性.

5 结论与展望

利用网络中计算顶点 *BC*值的方法来提取关键词,本文对其合理性和有效性进行了详细阐述和实验验证.同时,将文档信息映射为词语网络,提取出的关键词更能体现了文档特征,而不仅仅是文档中的高频词语.

由于词语网络表示了文档的重要信息: 词语及其相关关系,因此,文档的中心思想也可从词语网络中进行提取,从而有助于文档的自动摘要生成.另一方面,对词语网络中边的 BC值进行计算,从而对词语网络进行划分获得文档的子主题,从而有利于对文档的分类研究,这些都是进一步的研究工作.

References

[1] Luhn H P. A Statistical approach to the mechanized encoding and searching of literary information[J]. IBM J. Research and Development, 1957, 1(4): 309–317.

[2] Salton G, Yang C S. On the specification of term values in automatic indexing[J]. Documentation, 1973, 29(4): 351–372.

[3] Cohen J D. Highlights: language- and domain-independent automatic indexing terms for abstracting[J]. Journal of American Society for Information Science, 1995, 46(3): 162–174.

[4] Swaminathan K. Tau a domain-independent approach to information extraction from natural language documents [Z]. DARPA Workshop on Document Management, Palo Alto, 1993.

[5] Ohsawa Y, Benson N E, Yachida M. KeyGraph: automatic indexing by co-occurrence graph based on building construction metaphor[Z]. Research and Technology Advances in Digital Libraries, 1998 12–18.

[6] Matsuo Y, Ohsawa Y, Ishizuka M. KeyWorld extracting keywords from a document as a small world[C]. Discovery Science, 4th International Conference, 2001: 271–281.

[7] Watts D J, Strogatz S H. Collective dynamics of Small-World networks[J]. Nature 393, 1998 440–442.

[8] Porter M F. An algorithm for suffix stripping[J]. Program, 1980, 14(3): 130–137.

[9] Freeman L C. A set of measures of centrality based on betweenness[J]. Sociometry, 1977, 40: 35–41.

[10] Brandes U. A faster algorithm for betweenness centrality[J]. Journal of Mathematical Sociology, 2001, 25(2): 163–177.

[11] Li J S, Wang H F, Yu S W, et al. Research on maximum entropy model for keyword indexing[J]. Chinese Journal of Computers, Sep. 2004, 27(9): 1192–1197.

[12] Latora V, Marchiori M. A measure of centrality based on the network efficiency[Z]. Cond-Mat/0402050, 2004.

附中文参考文献:

[11] 李素建, 王厚峰, 俞士汶,等. 关键词自动标引的最大熵模型应用研究[J]. 计算机学报, 2004, 9, 27(9): 1192–1197.

[1] 括号内表示该词语的频率.