第 33 卷第 1 期　　河 北 科 技 大 学 学 报　　Vol. 33, No. 1
2012 年 2 月　　Journal of Hebei University of Science and Technology　　Feb. 2012

文章编号：1008-1542(2012)01-0065-04

后缀树算法在舆情聚类中的应用

彭　静[1]，翟　英[2]，冯　爽[3]

（1. 河北科技大学信息科学与工程学院，河北石家庄　050018；2. 河北经贸大学信息技术学院，河北石家庄　050061；3. 河北科技大学教务处，河北石家庄　050018）

摘　要：针对网络舆情分析的需求背景，研究了通过后缀树算法发现文本文档之间的公共短语串，按公共短语串实现文档聚类。网页文档的标题和摘要能代表文档的主要思想，应用后缀树算法实现对标题和摘要自动聚类，从而实现舆情信息自动聚类。

关键词：网络舆情；后缀树算法；文本聚类

中图分类号：TP391.1　　**文献标志码**：A

Application of STC algorithm to internet public opinions clustering

PENG Jing[1], ZHAI Ying[2], FENG Shuang[3]

(1. College of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang Hebei 050018, China; 2. College of Information Technology, Hebei University of Economics and Bussiness, Shijiazhuang Hebei 050061, China; 3. Department of Teaching Affairs, Hebei University of Science and Technology, Shijiazhuang Hebei 050018, China)

Abstract: In answer to the requirement of internet opinions analysis, this paper discusses the STC algorithm for text clustering, in order to discover common phrases that can assign documents and form document clusters. Because web document titles and abstracts can express the main ideas, web document clusters are created by STC algorithm, and clusters of internet public opinions information are created by using this method.

Key words: internet public opinions; STC algorithm; text clustering

在网络舆情监控系统中，包括舆情采集、舆情存储、舆情分析和结果显示 4 个主要模块。舆情采集和存储模块对指定的网页进行抓取、解析、中文分词，按关键词索引并存储在舆情数据库；舆情分析和显示模块实现根据关键词进行舆情搜索并对搜索结果自动聚类显示[1]。

舆情聚类的过程是实时的，为了提高速度，并不是对整个网页文档的全部内容聚类，而是提取关键内容：文档标题、文档摘要、文档 URL。文档摘要能够全面准确地反映文档的简单连贯的短文，比较准确地体现文档的主题[2]，摘要提取本文不做讨论。舆情分析模块可以按标题、摘要和 URL 自动聚类。文档标题和文档摘要都看做短文本文档，研究应用后缀树算法实现短文本文档的自动聚类[3]。

1　后缀及后缀树定义

1.1　后缀定义

给定长度为 n 的字符串 $S=S_1S_2\cdots S_i\cdots S_n$，子串 $S_iS_{i+1}\cdots S_n$ 都是字符串 S 从 i 开始的后缀，记为 $S_{[i,\cdots,n]}$。

收稿日期：2011-06-27；修回日期：2011-11-17；责任编辑：陈书欣
基金项目：河北省科技支撑计划项目(10213557)
作者简介：彭　静(1970-)，女，河北定州人，副教授，硕士，主要从事文本挖掘方面的研究。

以字符串 S=abab 为例，它的长度为 4，所以 $S_{[1,\cdots,4]}$，$S_{[2,\cdots,4]}$，…，$S_{[4,\cdots,4]}$ 都是 S 的后缀，空字串 $ 也是后缀，字符串 S 后缀分别是：abab，bab，ab，b，$（空字串）。包含这个字符串的所有后缀的压缩 Trie[4]，就是字符串 S 的后缀树。

1.2 后缀树定义

一个具有 n 个字符的字符串 S 的后缀树 T，就是一个包含 1 个根节点的有向树，该树恰好带有 n 个叶子，这些叶子被赋予从 1 到 n 的标号。除了根节点以外的每一个内部节点，都至少有 2 个子节点，而且每条边都用 S 的一个非空子串来标志。出自同一节点的任意两条边的标志不会以相同的词开始。后缀树的关键特征是：对于任何叶子 i，从根节点到该叶子所经历的边的所有标志串联起来后恰好拼出 S 的从 i 位置开始的后缀，即 $S_{[i,\cdots,n]}$。把树中节点标志为从根到该节点的所有边的标志进行串联[5]，如图 1 表示字符串 S=abab的后缀树。

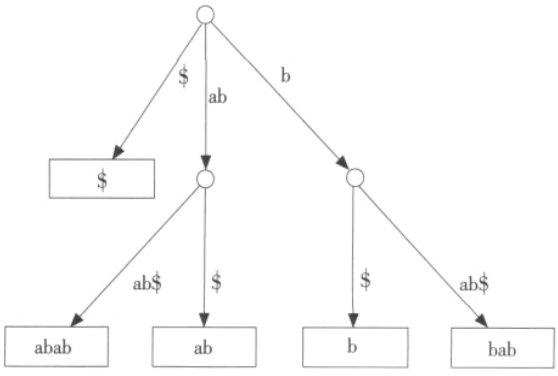

图 1 字符串 S=abab 的后缀树

Fig. 1 Suffix tree of string "abab"

2 后缀树算法实现文档聚类

2.1 文档表示为后缀树[6]

有 3 篇英文文档，其内容分别为 cat ate cheese，mouse ate cheese too，cat ate mouse too，在英文文档中，以空格分隔单词，文档中连续的多个单词为有一定含义的单词串，以下称为短语，短语能反映文档的含义，文档由多个短语构成。以这 3 个文档构建 1 棵后缀树如图 2 所示。

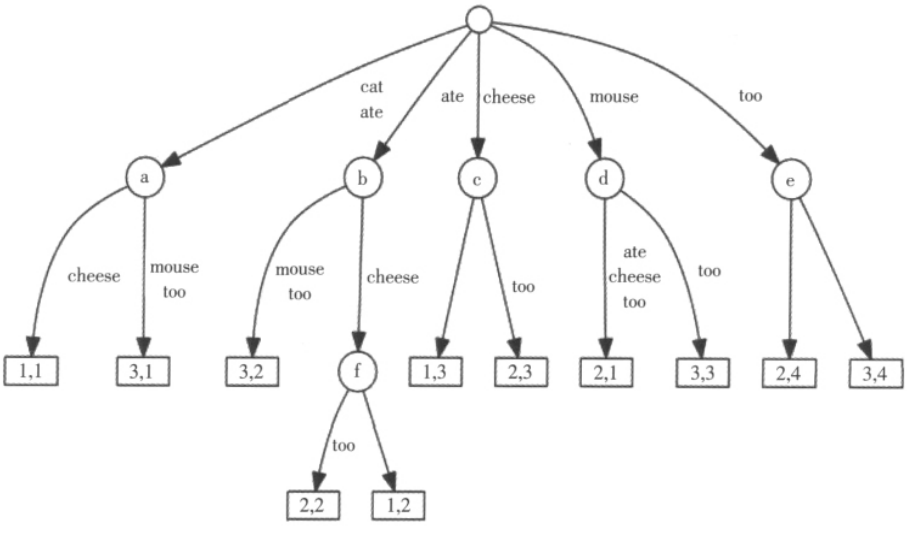

图 2 文档的后缀树表示

Fig. 2 Suffix tree of documents

树中有 10 个叶子节点，每个叶子节点用一个二元组来表示，如叶子节点 3,1 表示第 3 个文档，从位置 1 开始的后缀"cat ate mouse too"。a，b，c，d，e，f 为内部节点，从根节点到内部节点的短语在多个句子中被包含，这个单词串称为短语，比如内部节点 a，从根节点到 a 节点的边为短语"cat ate"，称节点 a 的短语为"cat ate"。从 a 节点出发有 2 个叶子节点 1,1 和 3,1，那么包含"cat ate"短语的文档有文档 1 和文档 3，后缀树节点、短语和包含短语的文档之间的关系如表 1 所示。

如果 2 个文档含有多个相同短语，说明这 2 个文档相似。通过计算得到文档之间的相似度，超过一定阈值，文档聚为一类。

2.2　应用后缀树实现文档聚类的步骤

1)文档解析:这一步骤主要实现对文档分词,比如文档 1 的内容是"cat ate cheese",以空格为分隔符,分解为 3 个单词 cat,ate,cheese。

2)文档表示为后缀树,按短语进行文档聚类:表 1 中所示的节点 a 对应的短语"cat ate",文档 1 和文档 3 含有该短语,文档 1 和文档 3 聚类为 1 个基类。每个聚类后的基类通过计算得到 1 个值,值较高(超过一定阈值)的再进行步骤 3)。

3)基类的合并,第 2 步得到的基类合并为更大的文档聚类。

表 1　后缀树节点、短语及文档之间的关系

Tab. 1　Suffix tree nodes, prases and documents

节点	短语	包含短语串的文档
a	cat ate	1,3
b	ate	1,2,3
c	cheese	1,2
d	mouse	2,3
e	too	2,3
f	ate cheese	1,2

在步骤 2)中按短语进行文档聚类,聚类的效果用一个分值来评价,这个分值相当于通常的文本分类的相似度,这里用 $S(m)$ 来表示:

$$S(m)=|m|\cdot f(|m_p|)\cdot\sum \text{tf-idf}(w_i), \tag{1}$$

式中:m 是一个短语对应的文档聚类;m_p 是短语;$|m|$ 是短语聚类 m 中包含文档的个数;w_i 是短语 m_p 中包含的单词;$\text{tf-idf}(w_i)$ 是在 m_p 中每个词的词频权重。$|m_p|$ 是 m_p 中包含词的个数,$f(|m_p|)$ 与 $|m_p|$ 成线性关系。

词频权重的计算公式为

$$\text{tf-idf}(w_i,d)=(1+\log(\text{tf}(w_i,d)))\cdot\log(1+N/\text{df}(w_i)), \tag{2}$$

式中:$\text{tf}(w_i,d)$ 是 w_i 在文档 d 中出现的次数即词频;N 是文档集合中文档的总数;$\text{df}(w_i)$ 是 w_i 所在文档的个数。

第 3 步是基类的合并,在第 2 步中,确定了含有共同短语的为 1 个基类,然而 1 个文档中包含多个短语,因此不同基类中包含的文档可能是重叠的,比如 c 节点的短语聚类有文档 1 和文档 2,f 节点的短语聚类也包含文档 1 和文档 2,STC 算法的第 3 步是把不同的基类(有重叠或交叉的文档经过计算相似度超过一定阈值)合并成更大的聚类。

定义相互交叉或重叠的 2 个短语聚类之间的相似度用 $\text{sim}(m_i,m_j)$ 表示,其值为

$$\begin{cases}\text{sim}(m_i,m_j)=1, & |m_i\cap m_j|/|m_i|>\alpha \text{ and } |m_i\cap m_j|\cdot|m_i|>\alpha,\\ \text{sim}(m_i,m_j)=0, & |m_i\cap m_j|/|m_i|<\alpha \text{ or } |m_i\cap m_j|\cdot|m_i|<\alpha,\end{cases} \tag{3}$$

α 取值在 0～1 之间,通常取 0.7。

在表 1 的基础上计算所有短语聚类的相似度,α 取值分别为 0.7 的时候,合并基类的结果见图 3,聚类结果的表格形式见表 2。

表 2　短语聚类的合并结果

Tab. 2　Clusters identified by the phrase cluster

聚类编号	节点	文档
1	a	1,3
2	b	1,2,3
3	d,e	2,3
4	c,f	1,2

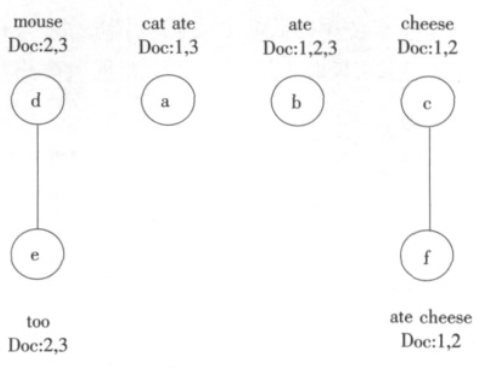

图 3　短语聚类结果

Fig. 3　Phrase cluster graph

3　实验结果及分析

聚类的数据源采用通过雅虎搜索的 100 篇关于"核爆炸"的网页文档。提取标题、摘要、URL,结果用 XML 文档格式表示[7]。

〈document id="4"〉

〈snippet〉

众所周知，火星因富含氧化铁的沙土呈现迷人的红色。但最新科学研究发现，这颗行星并非生来如此。据英国《每日邮报》4月2日的报道，一名俄罗斯科学家日前指出，巨大的核爆炸是火星如今呈现红色的真正原因，他还声称我们居住的地球在未来也可能会同样地“变脸”。

〈/snippet〉

〈url〉

http://www.chinadaily.com.cn/hqzx/2011-04/04/content_12272414.htm

〈/url〉

〈title〉

俄科学家称核爆炸导致火星红脸 地球也可能变色

〈/title〉

〈/document〉

〈document id="5"〉

…

〈/document〉

应用后缀树算法对标题和摘要自动聚类的结果如表3所示，聚类处理时间1 500 ms。

表3 网页文档聚类结果

Tab.3 Clusters of web document

聚类编号	短 语	文档编号
1	Powered by Discuz	13,21,38,56,62,68,83,92,97,99
2	1)据英国《每日邮报》4月2日的报道 2)他还声称我们居住的地球在未来也可能会同样地“变脸”	4,46,55,73,95
3	Nuclear Explosion	0,3,5,12,32,91
4	核爆炸	35,39,43,50,58,63,76,88,90,92,97
5	该次大爆炸造成欧洲30多万人受放射性伤害死去 在引起热能爆炸的第四核反应堆里	6,9,12
6	朝鲜试验场地上空未发现核爆炸微粒 不过在后期的报告中	7,48,96
7	Chinese	5,9,12,43,91
8	Other Topics	其他

4 结 语

结合网络舆情分析的应用背景，研究了后缀树算法。通过对文本文档建立后缀树，发现文档之间的公共短语串，实现文本文档的聚类，不需要中文分词，公共短语串比公共关键词更能体现文档之间的相似度。在舆情监控系统中，对Web文档的摘要、标题和URL应用后缀树模型实现聚类，相当于对短文档聚类，可以大大提高速度，有助于快速准确地发现舆情热点，为进一步实现话题跟踪打下基础。

参考文献：

[1] 汤寒青，王汉军. 改进的K-means算法在网络舆情分析中的应用[J]. 计算机系统应用(Computer Systems and Applications)，2011，20(3)：165-168.

[2] 廉 捷，刘 云. 网络舆情中的信息预处理与自动摘要算法[J]. 北京交通大学学报(Journal of Beijing Jiaotong University)，2010，34(5)：94-99.

[3] 郭 莉，张 吉，谭建龙. 基于后缀树模型的文本实时分类系统的研究和实现[J]. 中文信息学报(Journal of Chinese Information Processing)，2005，19(5)：16-23.

[4] GUO Xi，YANG Xiao-chun，YU Ge，et al. Choosing meaningful structure data for improving web search[J]. Journal of Southeast University (English Edition)，2008，24(3)：243-246.

[5] UKKONEN E. On-line construction of suffix trees[J]. Algorithmica，1995，14(3)：249-260.

[6] ZAMIR O E. Clustering Web docaments：A phrase-based method for grouping search engine results[D]. Washington：University of Washington，1999.

[7] SONG Ming-qiu，WU Xin-tao. Content extraction from Web pagesbased on Chinese punctuation number[A]. International Conference on Wireless Communications，Networking and Mobile Computing[C]. Shanghai：[s.n.]，2007.5 573-5 575.