Digital Library Forum 数字图书馆论坛
http: // www.dlf.net.cn

广义后缀树及其在汉语科技词系统中的应用研究

□ 徐硕 乔晓东 朱礼军 张运良 / 中国科学技术信息研究所 北京 100038
薛春香 / 南京理工大学经济管理学院 南京 210094

摘要：科技词汇知识是科技信息智能处理的基石，如何加速汉语科技词系统的构建是目前研究的热点问题之一。考虑到中文术语构词的特点，文章引入了一种灵活的数据结构——广义后缀树，从字面的角度提出了关系辅助构建、任务分配以及输入提示等辅助工具，使得知识工程师的工作更加高效。

关键词：广义后缀树，汉语科技词系统，关系构建，任务分配，输入提示
DOI：10.3772/j.issn.1673—2286.2013.04.005

1 引言

科技词汇知识是科技信息智能处理的基石，长期以来，以科技类主题词表为代表的科技词汇知识体系存在着编制过程中知识丢失、维护机制落后、词更新周期长、开放程度不够、非面向机器使用等突出问题，难以满足信息加工和软件开发需求。为应对这一问题，07年在国家“十一五”科技支撑计划课题资助下，中国科学技术信息研究所组织开展了汉语科技词系统的研究和开发工作，并以新能源汽车领域为试点，进行了新能源汽车词系统建设实践，详见文献[1,2]或直接访问词系统管理加工平台：http://www.vocgrid.org/。

与英文不同，中文构词有其自身的特点，陆汝占教授认为汉语是义符文字，词语结构与意义以名词为中心，构造方式为毗连组合，直接对应概念耦合，实体类分类由实体本质属性标识[3]。比如对于图1[3]中各种面包、糕饼和甜食的中英文表达，从字面上看英文表达几乎没有任何的相似性，但对中文表达就不同了，大部分都含有“面包”这两个字，也就是说中文术语从字面上传递了很多有用信息。正因如此，本文主要从字面的角度讨论汉语科技词系统内容的辅助构建方法。

图1 各种面包、糕饼和甜食中英文表达

* 本研究受“十二五”国家科技支撑计划“面向外文科技知识组织体系的大规模语义计算关键技术研究”（编号：2011BAH10B04）、“基于STKOS的知识服务应用示范”（编号：2011BAH10B06）、中国科学技术信息研究所预研项目“基于词系统的领域深层主题规律揭示分析研究”（编号：YY201216）、江苏省社会科学基金项目“数字报纸的自动标引研究”（编号：09TQC011）以及教育部人文社会科学研究项目“电子报纸内容深加工研究”（编号：09YJC870014）资助。

2 广义后缀树简介

尽管后缀树概念的提出独立于TRIE的概念，但为了更容易理解后缀树，让我们首先来解释一下TRIE树[4]。“TRIE”这个单词来自于“retrieval”，TRIE树是一个简单但实用的数据结构，通常用于实现字典查询。TRIE树的每条分支代表一则子串，树的叶节点代表完整的字符串，和普通树不同之处在于相同的字符串前缀共享同一条分支。

为叙述方便，首先给出后缀的定义。给定长度为n的字符串S=$S_1S_2\cdots S_i\cdots S_n$#（#号表达字符串的结束），和整数i（1≤ i≤n），子串S[i: n] = $S_iS_{i+1}\cdots S_n$#称为字符串S的后缀，一般将空字符串（记为#）也算其后缀，通常称子串S[i: n]为起始位置为i的后缀。后缀树实际上是包含字符串所有后缀的压缩TRIE树[5]。

后缀树是针对一个字符串构建的，而广义后缀树[5,6]是针对一个字符串集合构建的。比如在表1所示的字符串集合上构建的广义后缀树如图2[6]所示。为了区分方便，表1中为每个字符串添加了一个不同的特殊字符$\#_i$，在广义后缀树的实际构建中只需一个特殊字符即可。

表1 8个字符串示例

ID	字符串	ID	字符串
1	ABC$\#_1$	5	CD$\#_5$
2	BDE$\#_2$	6	E$\#_6$
3	BC$\#_3$	7	AB$\#_7$
4	E$\#_4$	8	D$\#_8$

构建后缀树的第一个线性时间算法是由Weiner[7]于1973年提出的，另外一个不同但空间效率更高的算法是由McCreight[8]于1976年提出的。几乎在20年之后，Ukkonen[9]给出了一个概念上完全不同的线性时间构建算法，该算法允许后缀树的在线构建，而且更容易理解。尽管Ukkonen和McCreight的算法在高层组织上差别很大，但二者的联系被Giegerich和 Kurtz[10]详细解释。

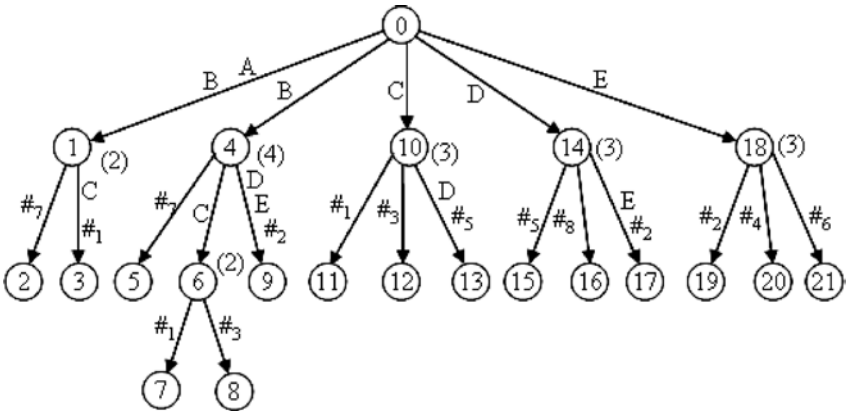

图2 表1所示字符串集合上构建的广义后缀树

3 广义后缀树在汉语科技词系统中的应用研究

本节主要考虑如何借助广义后缀树，辅助知识工程师完成词系统中知识单元的构建。

3.1 关系辅助发现

根据中文术语构词特点，从字面的角度发现语义关系通常需要对术语列表进行排序，比如前端一致的正/倒排序，或后端一致的正/倒排序。但该方法只能发现具有共同前缀或共同后缀的术语间语义关系，对于“能源”与“新能源汽车”间的语义关系并不能发现，本小节提出的方法可发现具有共同子串的术语间语义关系，只要子串定义得当。

表2 10条新能源汽车领域术语示例

ID	字符串
1	<燃料, $\#_1$>
2	<甲醇, 汽车, $\#_2$>
3	<燃料, 汽车, $\#_3$>
4	<专用, 燃料, 汽车, $\#_4$>
5	<两用, 燃料, 汽车, $\#_5$>
6	<双, 燃料, 汽车, $\#_6$>
7	<汽车, 动力, $\#_7$>
8	<汽车, 动力, 系统, $\#_8$>
9	<汽车, 动力, 学, $\#_9$>
10	<汽车, 动力, 装置, $\#_{10}$>

假设我们有一部收词粒度通常比较细的词典D，本节参考文献[11-13]，首先给出几个定义。语义知识库中存在的词汇为原子术语（Primitive Term，PT）；词典D中不存在，但可由两个或更多原子术语组合而成的词汇称为组合术语（Combined Term，CT）；原子术语与组合术语统称为术语（Term）。

严格来说，就是给定一部词典D = {PT_1, PT_2, …, PT_K}，则D中每个元素都是一个原子术语，而符合下式定义的词汇CT为组合术语：

$CT=PT_{i_1}PT_{i_2}\cdots PT_{i_n}, CT \in D, PT_{i_j} \in D, j=1,2,\cdots,n, n\geq 2$ (3-1)

对于任意一个组合术语CT，由于构成它的原子术语的位置是确定的，因此每个组合术语都可以表示为一个由原子术语构成的字符串，即

$CT=<PT_{i_1}, PT_{i_2}, \cdots, PT_{i_n}>$. (3-2)

为一致起见，原子术语PT也可类似地表示为<PT>。表2给出了10条新能源汽车领域术语，并且将每条术语均表示成了原子术语构成的字符串，在这些字符串的基础上，按照前文所述方法可构建如图3所示的广义后缀树。

从图3可以看出，含有原子术语“燃料”的术语ID集合{1, 3, 4, 5, 6}全部位于以节点1为根节点的子树下，而含有“燃料汽车”的术语ID集合{3, 4, 5, 6}全部位于以节点2为根节点的子树下。这样只要深度优先（DFS）遍历图3所示的广义后缀树，即可得到术语间的上下位关系、相关关系等。我们针对新能源汽车领域的术语已经构建了广义后缀树，并输出了所有的遍历结构给知识工程师，表3为部分结果展示。表3中每个区的开始为几个术语所共同包含的原子术语子串，以及包含这个原子术语子串的术语数量，比如包含原子术语子串“动力电动汽车”的术语共有5条，其ID及相应的术语逐个列于其后。

3.2 任务分配

为了加快词系统内容的协同构建，通常需要为每个用户分配不同的加工任务。对任务的划分通常有五种方式：按范围分、按字顺分、按关系类型分、按词族分以及按分面分等。但如果构建图3所示的广义后缀树，还可按原子术语进行任务划分。比如可以将含有原子术语“燃料”的术语ID集合{1, 3, 4, 5, 6}分配给一个用户来完成。

表3 新能源汽车领域术语的广义后缀树部分遍历结果

动力电动汽车(5)	
2284	复合动力电动汽车
2283	并联式混合动力电动汽车
2282	串联式混合动力电动汽车
2281	混合动力电动汽车
27	混联式混合动力电动汽车
动力汽车(9)	
5080	复合动力汽车
4830	可外接充电式混合动力汽车
4778	轻度混合动力汽车
4756	燃料电池混合动力汽车
4704	并联混合动力汽车
1748	并串联式混合动力汽车
1747	串联式混合动力汽车
1746	并联式混合动力汽车
动力系统(7)	
4851	混合动力汽车ISG系统
5182	完全混合动力系统
5181	中混合动力系统
5180	轻混合动力系统
5179	微混合动力系统
4799	汽车动力系统
4703	并联混合动力系统
1441	混合动力系统
动力机械(3)	
906	往复式动力机械
885	动力机械
879	旋转式动力机械
动力装置(2)	
5365	车载动力装置
4797	汽车动力装置

3.3 输入提示

在许多实际应用中，为了改善用户体验，提高用户完成各种任务的速度，经常需要根据用户的输入给出一定的提示。比如我们在搜索引擎中输入“新能”时，Google和百度通常会给出以“新能”为前缀的几个用户关键词提示，分别见图4中（a）和（b）。这一方面可以加快用户的输入速度，另一方面也可对那些不太清楚自己输入什么关键词的用户起到提示的目的。然而，Google和百度提供的输入提示，只能提示只输入字符串为前缀的术语集合。如果用户希望查找“完全混合动力系统”，但他/她记不清全称了，可能只记得“混合动力系统”或“动力系统”，那么Google和百度的输入提示就显得有点苍白无力了。

考虑表4所示的术语集合，此时原子术语为单个的汉字，它的广义后缀树如图5所示。每个内部节点包括三个数组，分别用于记录相应字符串出现于术语首部、中间及尾部的术语集合。比如节点1，字符串“燃料”出现在首部的术语ID集合为{1, 4}，出现在尾部的术语ID集合为{1}，出现在中间的术语ID集合为{}。有了这三个数组，我们就可以很轻松地完成各种输入提示了。

4 结论

考虑到中文术语构词的特点，本文主要从字面的角度考虑如何辅助汉语科技词系统内容的构建。具体来说，通过引入了一种灵活的数据结构——广义后缀树，从字面的角度提出了关系辅助构建、任务分配以及输入提示等辅助工具，使得知识工程师的工作更加高效。

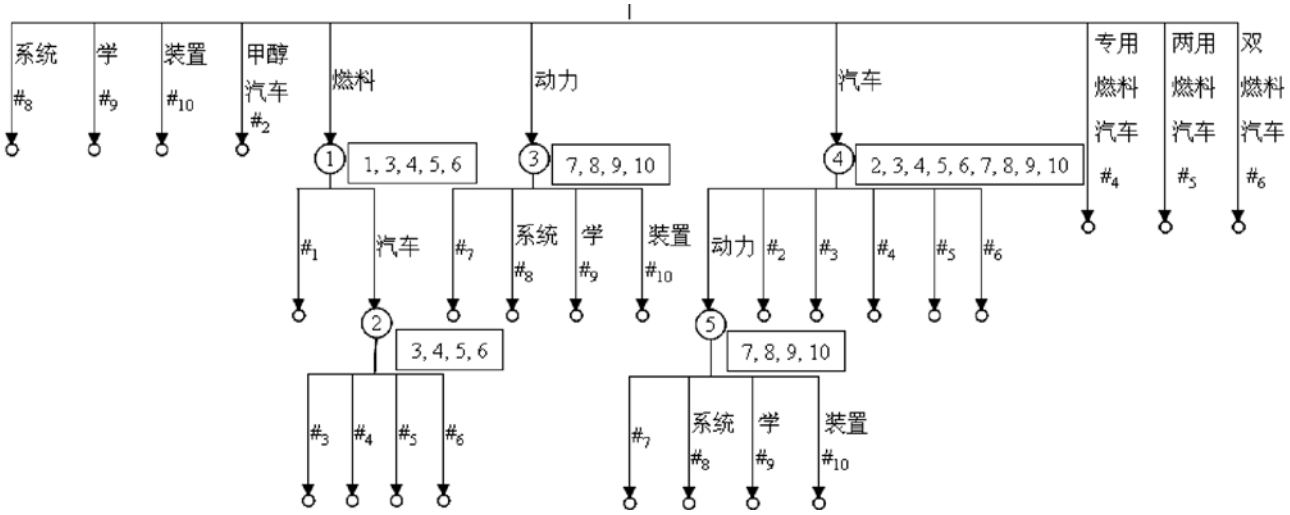

图3 表2所示术语集合的广义后缀树

图4 输入提示的成功案例

表4 5条新能源汽车领域术语示例

ID	字符串
1	<燃, 料, #1>
2	<甲, 醇, #2>
3	<甲, 醇, 汽, 车, #3>
4	<燃, 料, 汽, 车, #4>
5	<汽, 车, 动, 力, #5>

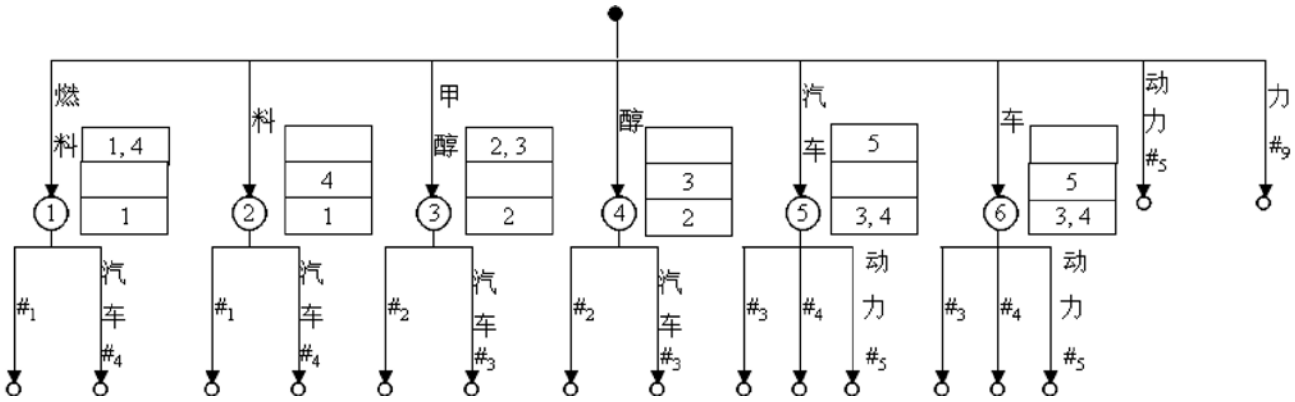

图5 表4所示术语集合的广义后缀树

参考文献

[1] 朱礼军,乔晓东,张远良.汉语科技词系统建设实践——以新能源汽车领域为例[J].情报学报,2010,29(4):723-731.
[2] 乔晓东,张远良,朱礼军.汉语科技词系统建设与应用进展[J].情报学报,2010,29(5):978-986.
[3] 陆汝占.关于汉语语义概念的一点思考[M]//朱小健,张全,陈小盟.HNC与语言学研究(第4辑),北京:北京师范大学出版社.
[4] KNUTH D E. The Art of Computer Programming. vol. 3, Sorting and Searching [M]. Addison-Wesley, 1972.
[5] GUSFIELD D. Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology [M]. USA: Cambridge University Press, 1997.
[6] XU S, QIAO X-D, ZHU L-J, et al. Deep Analysis on Mining Frequent & Maximal Reference Sequences with Generalized Suffix Tree [J]. Journal of Computational Information Systems, 2010, 6(7): 2187-2197.
[7] WEINER P. Linear Pattern Matching Algorithms [C]// Proceedings of the 14th Annual Symposium on Switching and Automata Theory, 1973: 1-11.
[8] MCCREIGHT E M. A Space-Economical Suffix Tree Construction Algorithm [J]. Journal of the ACM, 1976, 23(2): 262-272.
[9] UKKONEN E. On-line Construction of Suffix Trees [J]. Algorithmica, 1995, 14(3): 249-260.
[10] GIEGERICH R, KURTZ S. From Ukkonen to McCreight and Weiner: A Unifying View of Linear-Time Suffix Tree Construction [J]. Algorithmica, 1997, 19(3): 331-353.
[11] 夏天.汉语词语语义相似度计算研究[J].计算机工程,2007,33(6):191-194.
[12] 王文荣.词汇知识系统动态构建方法研究与工具实现[D].中国科学技术信息研究所,2008:58-75.
[13] 徐硕,朱礼军,乔晓东,等.基于双序列比对的中文术语语义相似度计算的新方法[J].情报学报,2010,29(4):701-708.

作者简介

徐硕（1979-），男，博士，中国科学技术信息研究所信息技术支持中心副研究员。研究方向：数据挖掘、信息抽取、生物信息等。E-mail: xush@istic.ac.cn
乔晓东（1965-），男，研究员，硕士生导师，中国科学技术信息研究所总工程师，中国互联网协会理事，中国情报学会计算机应用分会副主任委员，CILIP会员，研究方向：为信息资源管理工作和信息服务。E-mail: qiaox@istic.ac.cn
朱礼军（1973-），男，研究员，硕士生导师，中国科学技术信息研究所信息技术支持中心副主任。研究方向：Semantic Web、Web Service和知识技术在科技信息服务、电子政务/商务中的应用以及知识组织系统的集成与服务体系。E-mail: zhulj@istic.ac.cn
张远良（1979-），男，博士，中国科学技术信息研究所信息技术支持中心副研究员。研究方向：文本自动分类，概念层次网络，知识组织。E-mail: zhangyl@istic.ac.cn
薛春香（1979-），女，博士，南京理工大学经济管理学院副教授。研究方向：智能信息组织，知识组织系统构建。E-mail: xuechunxiang@gmail.com

Generalized Suffix Trees with Its Applications in Chinese Scientific Technical Vocabulary System

Xu Shuo, Qiao Xiaodong, Zhu Lijun, Zhang Yunliang / Institute of Scientific and Technical Information of China, Beijing, 100038
Xue Chunxiang / Nanjing University of Science and Technology, School of Economics and Management, Nanjing, 210094

Abstract: The scientific and technical words are basis of the scientific and technical information processing. Currently, there are extensive attentions on how to speed up constructing the Chinese scientific and technical vocabulary system. Due to the characteristics of word formation for Chinese terms, a flexible data structure, generalized suffix tree, is introduced. Based on the generalized suffix tree, the paper proposes some assistant tools for relationship construction, task allocation, input prompt, and so on from the viewpoint of literal meaning, which enables knowledge engineers to be much more efficient.

Keywords: Generalized suffix tree, Chinese Scientific and Technical Vocabulary System, Relationship construction, Task allocation, Input prompt

（收稿日期：2012-11-01）