

汉语科技词系统研究进展

□ 朱礼军 张运良 徐硕 乔晓东 / 中国科学技术信息研究所 北京 100038

摘要: 汉语科技词系统是科技词汇知识深层次组织与应用服务工具的集成体系。文章首先概述了词系统的内涵, 然后总结了三年来在词汇自动获取、词空间构建和词聚类以及词系统的复杂网络分析评价上所取得的技术突破, 并展望了科技词系统下一步的重点研发方向。

关键词: 科技词系统, 知识组织系统, 新词发现, 词聚类, 评价

DOI: 10.3772/j.issn.1673-2286.2011.03.008

1 汉语科技词系统概述

汉语科技词系统综合吸收叙词表、词典、本体等优势, 立足现实条件, 在过去三年中, 进行了理论方法、工程实践方面的探索。与传统的汉语主题词表相比, 汉语科技词系统从设计目标、应用场景、词汇知识内容、技术体系和应用服务方式上均有重大的变化。汉语科技词系统的设计目标是构建一个汉语科技词汇知识内容的创建、管理、维护和应用的技术框架体系, 实现科技词汇知识的开放加工和共享使用, 为中文科技信息资源的内容处理提供词汇层面的语义支撑^[1]。

从应用场景方面分析, 汉语科技词系统主要面向计算机应用程序, 提供有关汉语科技词汇的各类信息。从词汇知识内容层面来看, 汉语科技词系统除了提供主题词表的用、代、属、分、参五种主要关系外, 还要根据具体领域, 提供更为复杂的关系。例如: 在用、代关系上, 将细化出时变关系、不同域之间的同义关系等; 在属分关系上, 将细化结构上的部分与整体、学科上的包容关系等; 在参关系上, 将细化出逻辑上的因果关系、实证关系等。从技术体系上来看, 汉语科技词系统一方面通过自动构建和管理维护工具保持词汇内容的动态、及时更新, 另外一方面, 通过规范的程序访问接口和独立于计算平台的词汇输出格式来实现词汇构建过程和应用过程的封装, 使得汉语科技词系统更容易被其他研究机构和软件开发方重用。

目前, 通过汉语科技词系统的研发, 形成了一套

可行的知识结构设计原则和方法, 形成了汉语科技词系统的建设流程, 并建设了汉语科技词系统协同构建平台 (<http://168.160.18.252/vocabulary/>), 是对词系统数据模型的实践。通过汉语科技词系统的研发, 在已有平台和资源基础上, 构建了新能源汽车汉语科技词系统, 其中包含54831条词条, 其中5712条为核心词, 其余49120条为基础词, 设计包含推荐关系类型在内的76种关系类型, 以及57821个关系实例, 有52种属性类型, 并建设了18362个实例。面向新能源汽车的NEV分类法有4层154个类目, 并且构建了5431个类目实例, IPC和CLC概率分类信息还在处理当中。每一个核心词都包含对应的英译, 系统中包含有5431条定义。所有的5712条核心词和另外的4548条重要基础词拥有HNC概念描述^[2]。

2 汉语科技词系统研发过程中突破的主要技术难点

历时三年的研发过程中, 主要突破的技术难点如下:

其一, 确立了汉语科技词系统的数据模型。吸收ISO25964标准, 复杂程度高于叙词表, 并结合多种知识描述方式的知识组织系统, 系统面向计算机应用, 不严格遵循本体公理, 具有一定的容错性。

其二, 词系统辅助构建的关键算法。词系统自动构建主要是利用基于统计的机器学习技术和基于规则的自然语言处理技术来实现新词和新关系发现、语义相似度计算以及主题词自动推荐等功能, 进一步推动

中文词汇知识组织系统的自动构建及应用。我们在基于构词能力和构词模式的新词发现技术、基于双序列比对的中文术语语义相似度计算方法、基于最大频繁项集挖掘的主题词推荐算法,以及根据Proximity数据构建词汇向量空间模型方法以及词系统的可视化分析方面做了大量的探索工作,开发了多种词系统辅助构建工具,比如术语抽取工具、术语间关系辅助构建工作、词条索引工具等等,并且许多工具嵌入了词系统服务平台^[3]。

其三,基于过程的知识组织系统的评价技术。我们结合网络环境下知识组织系统构建、管理和应用的特点采用一些新的评价指标,对所构建的新型知识组织系统进行了全面的评价分析。首先,在传统分类法、叙词表、分类主题一体化词表评价方法和评价指标的基础上,结合词系统的特点,从功能、内容和结构三个方面构建了新的评价指标体系,做到从宏观和微观两个层面对汉语科技词系统进行定性评价和定量评价。其次,由于传统知识组织系统评价方式存在缺乏宏观测度、时滞过长、不能动态揭示等缺陷,我们结合复杂网络理论,从基本特性、动态特性和鲁棒性三个角度着手,对汉语科技词系统网络特性进行分析。实验证明:汉语科技词系统同时具有小世界特性和无尺度特性,在建设过程中应采用有导向的建设,在系统管理和维护的过程中要注重词汇间相互关系,保证系统的连通性。

3 科技领域的新词发现

新词发现常用的方法分三种:基于词频统计的方法、基于规则的方法、规则和统计结合的方法。我们以构词能力和构词模式作为规则与基于统计的方法相结合,在科技新词发现方面作了一些探索。

在科技领域,意义专指的复合词占很大的比例。在实际的新词发现中,我们通过字和词的构词能力、构词模式构建一系列的规则,提高新词发现的效果和速度。构词能力实际反映了一个字或词与其他的字或词联合构成新词的能力大小^[4]。设WFP(c)为字串c的构词(复合词)的能力,则:

$$WFP(c) = \frac{\text{count(含C的多字词)}}{\text{count}(c)} \quad (1)$$

WFP(c)=0表示在该领域c全是单字词; WFP(c)=1表

示在该领域c全是多字词。式(1)实际上反映了单个汉字在该领域的构词能力大小,则某一字串构成新词w的概率为:

$$P_{FWP}(w) = \begin{cases} 1-WFP(c), & |c|=1 (w \text{ 是单字词}) \\ \prod_{c \in w} WFP(C_i), & |c|>1 (w \text{ 为多字词}) \end{cases} \quad (2)$$

显然,从概率论的角度可以认为构词能力为零或接近于零的字或词,与其他字或词联合构成新词的概率为零。

Yao Yuan博士曾提出基于head-middle-tail结构的构词模式。任何一个字,在构成新词时有词首(head)、词尾(tail)、词中(middle)三种模式^[4]。设pttn(c)表示字的构词模式,则

$$Pr(pttn(c)|c) = \frac{\text{count}(pttn(c))}{\text{count}(c \text{ 位于多字词})} \quad (3)$$

其中,表示汉字C在某一多字词中取模式pttn(c)时的条件概率。显然对于同一汉字,存在

$$\sum_{pttn(c)=H,M,T} Pr(pttn(c)|c) = 1$$

跟构词能力一样,构词模式的结果跟所用语料有直接关系。针对某一领域语料计算构词模式,主要是词首和词尾的概率。用此结果作为规则,对新词进行判断、筛选,以此提高新词的质量。

针对《科技日报》这种形式的语料,本文用二元语法和三元语法作为统计模型,用已有的构词能力和构词模式作为规则,决定是否把其作为候选新词进行词频统计。从系统测试来看,在目前的条件下系统的准确率达到90%左右,覆盖率能达到40%以上^[5]。

4 词空间与词聚类的理论研究

词类是一个相对封闭的系统,而且词类的数量远小于词汇的数量。在汉语科技词系统的建设过程中,如果首先对分散的词汇进行聚类处理,然后在各个聚类中,进行词间关系的分析、校正,将大大地降低词系统内容加工的工作量。词聚类过程通常分为两个步骤:(1)构造一个词空间。在这个空间中,可以判别词汇之间的距离关系(将相似性或者相关性转化为可度量的距离关系),这是词聚类的前提;(2)选择合适的聚类方法完成聚类分析,其中常用的方法包括划

分聚类法及层次聚类法。

在相似关系的词空间构建方面,我们对词集 $S=\{T_1, T_2, \dots, T_n\}$, 建立词空间模型, 即通过计算为每个词语 T_i 确定一个 p 维实数空间向量与之对应, $p \leq n$ 。

具体步骤:

(1) 根据双序列比对算法^[6]和结合了同义词林同义关系信息的计算公式^[7], 计算得到 S 的词相似度矩阵:

$$S=(s_{ij})_{n \times n}$$

其中, $s_{ij} = \text{Sim}(T_i, T_j)$

(2) 把相似度矩阵 S 转换为度量型距离矩阵:

$$D=(d_{ij})_{n \times n}$$

(3) 对 D 进行经典多维标度分析, 得到 n 个正交特征向量和对应的特征值。

我们在实践中, 针对新能源汽车领域, 运用双序列比对算法计算词语相似度, 通过多维标度法构建词空间, 并在词空间中进行 k -means 算法聚类实验; 从词语相似度矩阵出发, 应用 AP 算法进行词语聚类实验; 设计实验对 k -means 聚类 and AP 聚类进行对比, 证明基于词空间的词聚类的有效性^[8]。从实验结果可以看出, k -means 聚类结果总体要优于 AP 聚类结果, 并得出以下结论:

(1) 利用多维标度法构建词空间会造成少量的语义信息丢失, 但是并没有对词聚类结果产生严重影响。

(2) 用多维标度法构建的词空间模型很好地反映了词语间的语义相似度关系。从词语语义相似度矩阵出发, 利用多维标度法构建词空间的方法是可行的。

(3) 词空间构建完成之后, 许多常用的聚类分析方法 (如 k -means 算法) 可以直接应用于词聚类分析。

(4) 通过词聚类可以大幅度减少语义关系提取的工作量。词聚类的质量与词簇个数的确定密切相关。

5 词系统网络的评价

词系统的内容质量控制对特定领域的汉语科技词系统建设成功与否至关重要。传统主题词表采用一些定量指标, 例如词量、语词先组度、语词清晰度、语词分布的均衡度、入口率、语词关联比和参照度等来进行评价。汉语科技词系统是一种新型的知识组织系统, 在结构、功能、构建、管理和应用等方面有别于传统的主题词表, 具有以下的特殊分析评价需求:

(1) 要及时考察领域词系统宏观状态, 即各项网络参数。将领域词系统中的“词汇”抽象成结点, 将

“词汇间的关系”抽象成边, 其结构就像一张网络。网络特征参数的大小, 能够定量地、形象地反映出领域词系统建设的合理程度。

(2) 要能揭示领域词系统构建过程中的动态变化。新词汇产生、传播和普及迅速, 领域系统也必须不断完善和更新。从领域词系统建设的初期发展到较为完善的稳定期, 整个系统的拓扑结构、功能都在动态发展。通过揭示其动态变化的规律, 能够为相似的领域词系统的建设提供参考。

(3) 要能体现网络化环境下的领域词系统的安全性以及维护和管理成本问题。有意或恶意地删除一定量的词汇会对领域词系统的可用性造成多大影响?

(4) 要能反映建设过程中领域词系统与科研人员 (或者科技语料中) 关注的重点、热点的相符程度, 这是领域词系统作为领域分析工具实用性的关键。

我们将新能源汽车领域词系统中的词汇抽象成“结点”, 将语义关系抽象成“边”, 就构成了一张张“网络”, 并利用复杂网络理论对其特性进行研究^[9]。将截止到2010年3月12日数据, 通过 pajek 软件构建了有向网络 $G=(V, E)$ 。其中, $V(G)=\{1, 2, 3, \dots, N\}$, 是由具体的词语抽象成的点, N 是结点总数, $E(G)=\{(u, v) | u, v \in V(G)\}$, 并且 $(u, v) \neq (v, u)$, 是将词语之间的语义关系抽象成的边, 在有向网络中常称为弧。如图1, 是其中1月8日备份的数据绘制的拓扑结构图。从图中, 可以了解词系统网络的宏观布局、重点词汇结点等信息。

分别选取了10月30日、1月8日和3月12日的新能源汽车词系统网络作为代表, 它们的参数值见表1。事实表明新能源汽车词系统网络具有小世界特性。

1月8日词系统网络结点度分布符合幂律分布 $\gamma=1.75$, $R^2=0.888$, 说明词系统网络具有无尺度特性, 并且该词系统网络的 γ 范围为1.6-1.8。1月8日词系统结点度分布确实具有不均匀的特点, 结点度高的结点所占数量极少, 多数分布在结点度1~30的范围。在词系统网络中, 结点度越高, 说明该词语在网络中越核心, 也预示着相关领域被关注的热度高。例如, 1月8日该期网络结点度最高的是结点“继电器”, 该结点出、入度之和为215。网络中心结点会随着网络的进一步建设和研究热点的变迁而不断变化。与常见的复杂网络相比, 词系统网络的结点度分布更为不均匀, 这意味着该系统的词汇间关系建设可能不够充分。

同时, 我们设计了词系统的损毁实验。研究表明, 词系统网络的抗损毁性较强。本次实验中, 词系

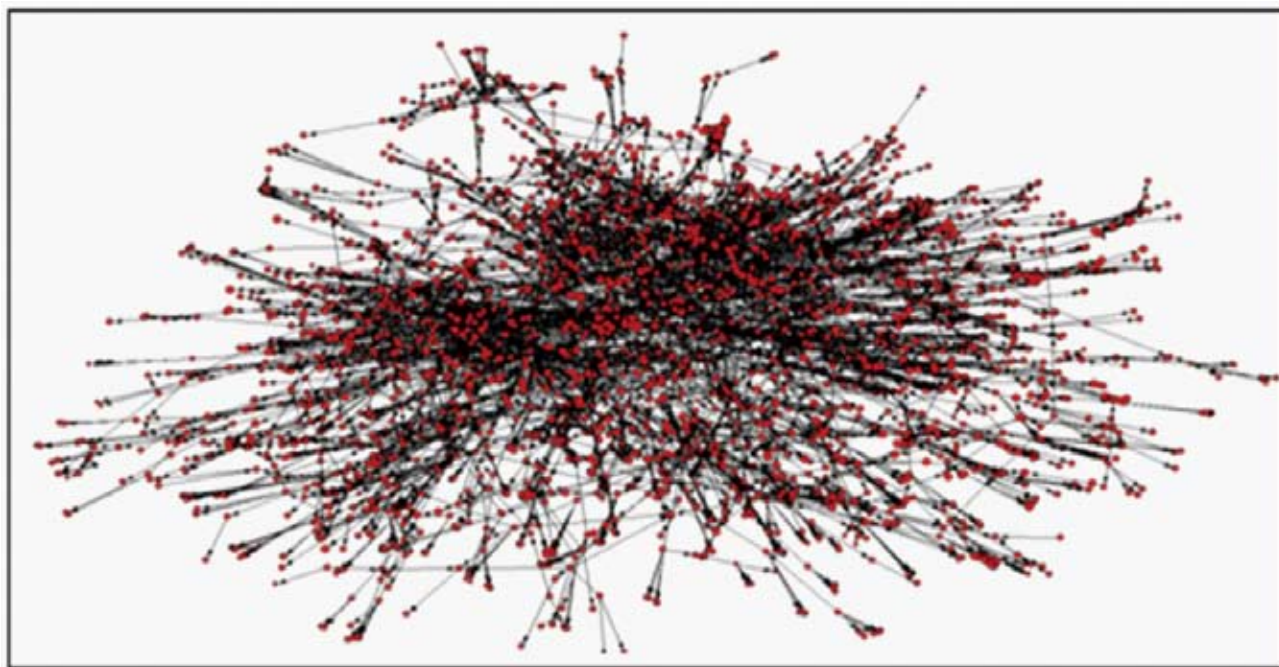


图1 词系统网络拓扑图

表1 词系统网络与ER随机网络对比情况

网络	结点数	平均结点数	聚集系数	直径	平均路径长度	密度
10月30日	4439	4.421	0.0448	22	8.69934	0.0003
随机网络	4439	4.421	0.0006	19	7.75662	0.0006
1月8日	7376	4.421	0.1198	22	8.18492	0.0005
随机网络	7376	4.421	0.0005	13	6.16704	0.0006
3月12日	9681	6.185	0.2042	24	6.57075	0.0003
随机网络	9681	6.185	0.0007	10	5.23564	0.0006

统网络一共接受了18次删除操作。当蓄意删除的结点比例达到0.0275，即203个最大度数结点被删除后，最大连通网络规模降为初始网络的47.1%，此时词系统网络崩溃。但是对词系统网络的随机删除效果不明显，当删除结点比例达到0.0275时，最大连通网络的结点数目比例为94.5%，攻击比例为0.03时，结点数目仍为93.9%。从图14-19中可以看，词系统网络在面对恶意删除时，最大连通网络的规模、网络分支的数量、弧的

数量、散点数量、平均路径长度以及聚集系数的变化都比随机删除情况下变化迅速，这几项参数很好地衡量了词系统网络被损毁的程度。网络直径和密度的变化也在一定程度上揭示了网络损毁的情况。

损毁实验表明：在对词系统进行维护，删除已被淘汰和替代的旧词汇时，不仅仅要考虑词汇的文献保障率以及用户保障率，还必须考虑该词汇结点在整个词系统网络中的地位和重要性。

6 下一步的研究计划

一个动态灵活、具有生命力的领域词系统，应该是在领域语料（素材库）基础上，通过以自动工具为主的计算机辅助构建模式，在相对科学的过程监控和内容质量评价模式的指导下，循环迭代进行建设。前期由于时间、经费和人力等多方面制约，我们在领域语料库建设、自动工具研发、过程监控和内容质量评估上，尚未形成一个循环迭代的闭环。下一阶段，我们将在对多种语料的处理、以词汇为核心围绕相似度、相关度、维度和粒度展开语义辅助工具的开发、集成，形成针对领域词系统建设的闭环迭代。通过高效的辅助构建工具集，可以将成功的词系统构建经验

快速复制，推广到其它领域，加快其它领域词汇知识的构建速度，并在术语服务、辅助教学、科技监测等应用方面进行深入实施。

致谢

除署名作者外，特别感谢参与汉语科技词系统建设与研究工作的历届硕士研究生与博士后工作站研究人员。他们包括：闫莹莹在收集和整理语料方面做了大量工作，薛春香博士在传统知识组织系统的评价方面做出了非常好的总结与分析；王文荣、郭怀恩和胡长爱同学分别在术语发现、词空间和聚类以及基于复杂网络的词系统评价上做出了卓有成效的研究工作。

参考文献

- [1] 朱礼军,乔晓东,张运良. 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,张运良,朱礼军. 汉语科技词系统建设与应用进展[J]. (情报学报录用).
- [3] 张运良,梁健,朱礼军,乔晓东. 基于术语定义的科技知识组织系统自动丰富关键技术研究[J]. 现代图书情报技术,2010(7/8):66-71.
- [4] 王晓龙. 计算机自然语言处理[M]. 北京:清华大学出版社,2005:39-41.
- [5] 王文荣,乔晓东,朱礼军. 词汇知识系统动态构建方法研究[J]. 现代图书情报技术,2007,年刊:100-102.
- [6] XU SHUO, ZHU LI-JUN, QIAO XIAO-DONG. Comparative Study on the Double-Array Structure for Large English & Chinese Lexicons [C]// Proceedings of the second International Conference on Intelligent Computation Technology and Automation,2009:268-272.
- [7] XU SHUO, ZHU LI-JUN, QIAO XIAO-DONG, et al. A Novel Approach for Measuring Chinese Terms Semantic Similarity based on Pairwise Sequence Alignmen [C]// 2009 Fifth International Conference on Semantics, Knowledge and Grid, 2009,6(S):92-98.
- [8] 郭怀恩. 词空间模型构建及其在词间语义关系发现中的应用. 中国科学技术信息研究所,2010.
- [9] 胡长爱. 基于复杂网络理论的知识组织系统分析与评价. 中国科学技术信息研究所,2010.

作者简介

朱礼军 (1974-), 博士, 副研究员。目前主要研究语义网、智能信息处理技术在科技信息服务、电子政务、电子商务中的应用。E-mail: zhulj@istic.ac.cn
 张运良 (1979-), 博士, 副研究员, 主要研究方向: 文本自动分类, 概念层次网络, 知识组织。
 徐硕 (1979-), 博士后。研究方向为: 数据挖掘、智能信息处理、科技热点监测。
 乔晓东 (1965-), 研究员, 主要研究方向: 知识服务技术、数字图书馆和信息资源管理方向研究。

Research Progress in the Chinese Science and Technology Vocabulary System

Zhu Lijun, Zhang Yunliang, Xu Shuo, Qiao Xiaodong / Institute of Scientific and Technical Information of China, Beijing, 100038

Abstract: Chinese Scientific & Technical Vocabulary System (CSTVS) is an integration system of term knowledge database and sets of construction tools. This report is organized as follows: Section 1 introduces the background of CSTVS. Section 2 describes the breakthroughs of key technologies in past three years including new term discovery, vocabulary space construction, term clustering visualization and complex network based evaluation. Section 3 indicates the further development of this project.

Keyword: Scientific & Technical vocabulary system, Knowledge Organization System, New Term Discovery, Term Clustering, Evaluation

(收稿日期: 2011-02-16)