面向电动汽车领域的专利文献加工和术语抽取方法研究

曾 文
（中国科学技术信息研究所，北京 100038）

摘要：随着国家科技战略规划发展的进一步深化，知识产权战略已经提升到国家层面，未来国家重点产业持续发展和新兴产业创新开拓都与知识产权战略息息相关。本文以电动汽车领域专利文献为基础，从专利文献加工和解决专利文献术语抽取的研究问题入手，提出专利文献再处理的基本流程以及一种基于专利术语语言特点和统计计算相结合的专刊文献术语抽取识别方法，并在电动汽车专利文献数据集上进行了验证和测试。测试结果表明，本文提出的方法是有效的。

关键词：电动汽车；专利文献；专利分析；专利加工；术语抽取
中图分类号：G356.8 **文献标识码：**A DOI：10.3772/j.issn.1674-1544.2014.05.009

Research of Processing and Term Extraction Based on Electric Automobile Patent Documents

Zeng Wen
(Institute of Scientific and Technical Information of China, Beijing 100038)

Abstract: With the further development of China's strategic planning of technology, the importance of intellectual property has been growing at the national level. In the future, the sustainable development of China's key industries will be closely related to its strategy of intellectual property. Based on the electric automobile patents' document, the paper proposed basic process of reprocessing patent documents. The paper also proposed a automatic extraction method based on patent's term characteristics and statistical computing. The algorithm was verified on the Electric automobile's test data set. Experimental results showed that the proposed method was effective.

Keywords: electric automobile, patent literature, patent analysis, patent processing, term extraction

1 引言

专利文献是技术、产品、应用和法律状态信息的混合载体，是具有技术价值和商业价值的知识蓄水池。与其他科技文献（图书、期刊、研究报告、会议论文、技术标准、学位论文）相比，专利文献的特点和情报分析价值主要表现在6个方面：（1）内容相对新颖、广泛；（2）信息密度大，针对性和实用性强；（3）叙述详尽，但语言表述上具有较强的技术性；（4）文献结构格式统一；

作者简介：曾文（1973- ），女，博士，中国科学技术信息研究所副研究员，研究方向：智能信息处理。

基金项目：“十二五”国家科技支撑计划课题“基于多源信息的电动汽车数据挖掘关键技术研究”（2013BAG06B01）；国家社会科学基金项目“基于事实型科技大数据的情报分析方法及集成分析平台研究”（14BTQ038）；中国科学技术信息研究所科研项目预研资金项目“基于领域的科技文献重要度评价方法研究”（YY-201416）。
收稿日期：2014年5月12日。

（5）报道相对及时，时效性相对较强；（6）专利文献数量庞大，重复出版量大。专利文献信息的特点和价值，使得专利文献的分析与应用成为国家管理部门、科研机构和企业等进行技术分析、技术创新和发展的重要手段之一[1-4]。

电动汽车技术的研究符合资源节约型和环境友好型社会的建设要求，因此，我国“八五”期间启动了电动汽车的研究和开发工作，在“九五”期间启动了“空气净化工程，到了“十五”期间，科技部提出了我国发展新能源汽车的实施方案，电动汽车重大专项被国家科教工作领导小组批准为国家“十五”期间重点组织实施的12个重大科技专项之一。基于此，本文围绕电动汽车领域的专利文献，开展电动汽车领域专利文献的基础性研究工作。

2　电动汽车领域专利文献数据的加工研究

专利文献的数据内容相比其他类型的科技文献更具技术性和创新性，其数据资源蕴含的科技信息价值最高。例如，电动汽车领域专利文献中的标题、文摘、权利要求项、正文等文本信息含有重要技术细节和技术保护等内容，如何从这些非结构化文本内容中抽取潜在的技术信息，分析领域技术的发展趋势，对于科学技术创新，辅助技术发展决策等具有重要的意义。

对电动汽车领域专利文献数据进行技术分析的前提是拥有良好质量的电动汽车领域专利数据资源作为基础。因此，构建高质量的电动汽车领域专利文献数据资源成为关键问题之一。电动汽车领域专利文献的数据与其他科技文献的数据加工在一定程度上具有异曲同工之处，即对于数据源首先均需要进行再次加工处理和存储的基本过程，其原因在于，电动汽车领域专利文献数据资源包括中外文电动汽车领域专利文献。此外，电动汽车领域专利文献数据资源规模庞大，由于国内外电动汽车领域专利文献数据信息采集和存储方式不同，存在数据信息存储和组织方式不一致的问题。目前，我们进行电动汽车领域专利文献数据加工的基本流程如图1所示。其中，加工模块实现以下功能：（1）识别数据元素，即自动识别电动汽车领域专利文献数据信息；（2）清洗数据内容，由于电动汽车领域专利文献资源的数据质量良莠不齐，因此在数据资源存储之前，首先需要对数据资源进行必要的自动“清洗”处理，去除不规范的字符和符号等，否则导入数据库的过程中会出现不必要的数据导入错误，而且影响日后数据整合和分析质量。存储模块的任务主要是：（1）建立数据库，用于存储处理后的数据；（2）将自动识别的数据资源内容与存储的数据库中的字段实现自动匹配，并自动存储在相应的数据库字段内。按照图1的处理流程，我们开发了相应的软件工具，实现电动汽车领域专利文献的加工和存储，示例图分别见图2和图3。通过电动汽车领域专利文献的数据加工技术，可以提供诸如标题、权利要求说明和摘要等二次电动汽车领域专利文献数据信息。

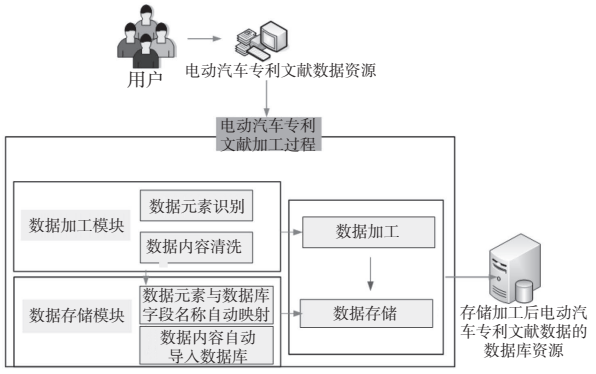

图1　电动汽车领域专利文献加工的基本流程

此外，除了针对电动汽车领域专利文献自身的数据内容进行加工处理外，还需要可用于专利分析的其他数据内容，如被引次数、同族专利数、法律状态、消歧后的作者姓名和单位名称、正文等详细数据信息，并不能实现单独提供，需要借助国外专利数据库或商业分析软件进行获取。对于这些数据的加工，本文认为在一定程度上需要借助于网络抓取引擎工具，从而减轻人工成本，即将信息的抓取过程抽象为统一的抓取工具。以电动汽车领域专利文献为例，需要通过抓取工具的配制参数，控制抓取数据的来源及与电动汽车之间的关联度，并可以使用不同的模板来

规定抓取引擎如何抽取不同的关键词和表属性以及如何清理数据和入库。针对电动汽车领域专利文献数据，网络抓取引擎定制不同的抓取代理。每个抓取代理包含一个抓取模板、一个抓取引擎和一个抓取探测器。抓取模板根据需要的数据格式以及少量网页样本来学习该资料的抓取模式。目前，这部分的工作还处于研究和测试的阶段。

3　电动汽车领域专利术语的识别与抽取方法研究

3.1　电动汽车领域专利术语的识别与抽取方法

国内已有的专利术语抽取研究工作基本是采用统计计算和构建专利术语信息抽取模板的方法，实现对专利文献主题词的抽取[5]。国内外现有的其他术语抽取技术方法则以利用统计计算方法居多[6−9]，但是统计计算的方法需要依赖于语料库的规模来保证抽取结果的准确度，需要解决的问题是构建语料库的成本和质量；由于模板规则的覆盖面小，基于模板规则的方法就需要构造相应的规则库，而构建术语信息抽取模板是十分耗费人力和物力的。因此，本文采用的基本策略是基于专利术语的语言规则和统计计算相结合的术语抽取策略，实现专利文献术语的抽取，具体方法如下。

通过对中文国际专利分类表及专利文献进行抽样分析并结合科技文献术语特点，可以发现：专利术语未出现语气词、状态词、叹词、拟声词和代词；专利术语的首词未出现助词、连词；专利术语的末尾词中未出现方位词、连词和助词；专利术语中包含名词、动词或形容词的数量占多数。

根据上述分析，本文制定专利文献术语抽取的基本语言规则是：专利术语中至少含有一个动词、名词或名词性成分；专利术语最后一个词为动词、名词或名词性成分；专利术语第一个词不为介词、量词；专利术语中无连词、代词和语气词。

为了提高专利文献术语自动抽取的准确性，

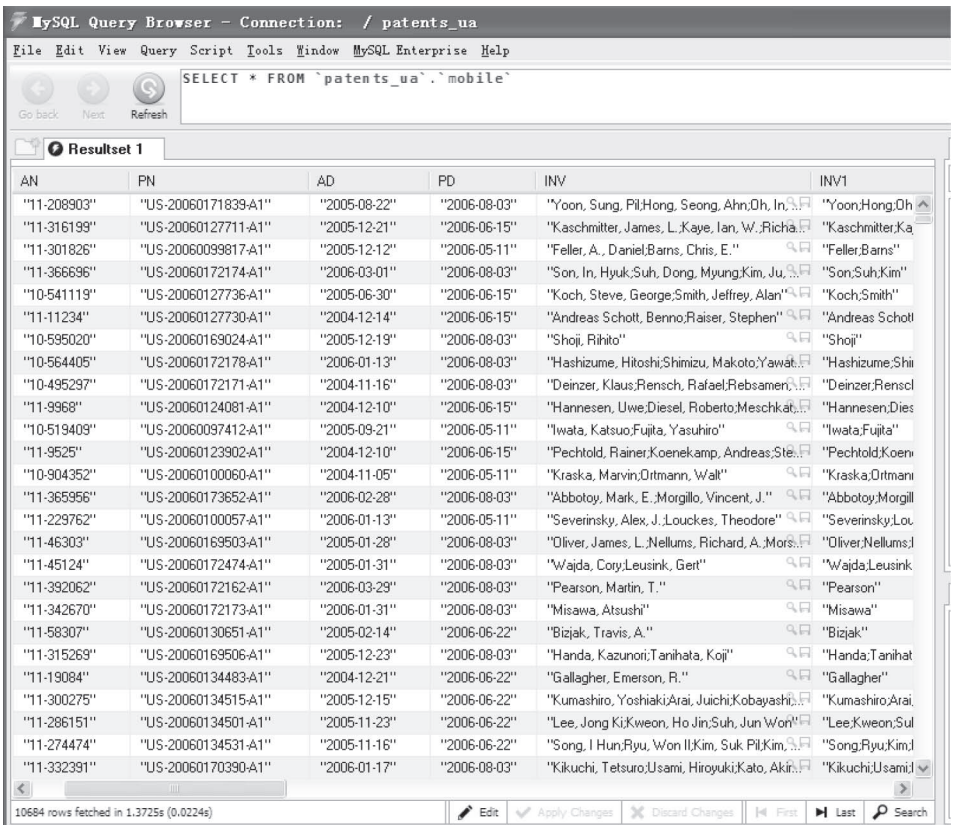

图2　电动汽车领域专利文献加工之后的数据库存储状态示例

本文将专利术语词语的自动抽取过程分为两部分：一是基于语言特点进行术语的自动抽取；二是基于统计算法对专利术语进行二次抽取识别和过滤，以完成整个专利术语的自动抽取过程。

具体专利术语抽取流程见图3。

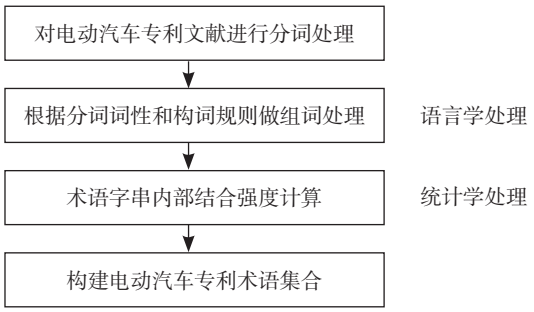

图3　电动汽车专利文献术语抽取基本流程

在图3中的语言学处理是指基于语言特点，将分词后的字进行字串组合，形成长度为2至10的字串，这些字串即是候选专利术语词。这些候选专利术语词经过统计学计算处理，得到最终的专利文献术语。统计计算的数学模型如下。

假设专利术语T是(t_i, t_j)的组合，词语t由t_1，t_2 ...t_n组成，字串记为$t_i = t_1t_2...t_{n-r}, t_j = t_rt_{r+1}...t_n$，对于词$t = t_1t_2...t_n$，则字串直接的结合强度关系可以表示为：

$$Strength(t_i, t_j) = \log_2 \frac{P(t_i, t_j)}{P(t_i) \times P(t_j)} \qquad (1)$$

$P(t_i)$：词t_i单独在所有专利文献中出现的概率；

$P(t_j)$：词t_j单独在所有专利文献中出现的概率；

$P(t_i, t_j)$：词t_i和t_j共同出现在同一专利文献中的概率。

$Strength(t_i, t_j)$值越大，则词t_i和t_j组合成专利文献术语的概率越大。

3.2　实验结果和分析

为了验证本文提出的专利文献术语的识别与抽取方法的效果，进行了相关的实验，实验数据是电动汽车领域的专利文献数据1226篇，相关实验结果见表1和表2。

表1　抽取的部分术语示例

术　语	强度值	术　语	强度值
电动汽车	0.82	纯电动	0.56
电动	1.0	混合动力	0.55
电池	1.0	能源	1.0
发动机	1.0	磁悬浮	0.46
燃料电池	0.75	负载	1.0

表2　实验数据结果

数据来源	采用语言学处理后专利术语抽取数目	采用统计学处理后专利术语抽取数目	术语的准确率
电动汽车专利文献（1226篇）	188432篇	76532篇	62.1%

对于实验结果的评估，本文采用的是人工识别判定的方法，在不同区域连续随机抽取800个词语样本，之后经过人工判定若干次800个样本中正确的术语词语个数，最终得到的平均准确率约为62.1%。

$$\text{precisionsubjectterms} = \frac{Number\_accurcy\_terms}{Number\_terms} \times 100\% \qquad (2)$$

从以上统计的结果可以发现：通过执行本文设计的专利文献术语识别抽取方法，获取的术语词语平均准确率可以达到62.1%左右，其主要原因首先是由于方法本身需要数据语料的规模和质量的保证，而非算法本身所能完全确定的客观事实，其次方法本身仍需要进一步的改进。

4　研究设想和展望

专利文献的加工和术语识别抽取方法是构建高质量专利文献数据，实现专利文献深层次数据挖掘的基础。因此，本文以电动汽车领域专利文献作为研究切入点，重点研究专利文献加工和术语识别抽取的技术和方法。实验分析和结果均表明本文的方法是有效的，但其在数据质量和术语抽取的准确度方面由于数据集选择规模的大小或数据集内容质量的不同而降低，达不到人工识别的精确和智能，在专利文献术语自动抽取的具体

（下转第64页）

agement Review,1998(41):464−476.
[8] Bell G G. Clusters, Networks, and Firm Innovativeness[J]. Strategic Management Journal, 2005 (26): 287−295.
[9] 解雪梅,左蕾蕾.企业协同创新网络特征与创新绩效:基于吸收能力的中介效应研究 [J].南开管理评论,2013,16(3):47−56.
[10] Cohen W M, Levinthal D A. Absorptive Capacity: A New Perspective on Learning and Innovation [J]. Administrative Science Quartelry,1990,35(1):128−152.
[11] Freeman C. Networks of Innovators: A Synthesis of Research Issues[J]. Research Policy,1991,20(5):499−514.
[12] Freeman L. Centrality in the Social Networks: Conceptual Clarification[J]. Social Networks, 1979(1):215−239.
[13] Borgatti S P, Everett MG, Freeman LC, et al. Software for Social Network Analysis[M]. Natick: Analytic Technologies,1999.
[14] 潘松挺,蔡宁.网络关系强度与组织学习:环境动态性的调节作用 [J].科学决策,2010,4(4):48−54.
[15] Dyer J H, Singh. The Relation Review: Cooperative Strategy and Sources of International Competitive Advantage [J].Academy of Management Review, 1998, 23(4):660−679.
[16] 潘雄锋.高技术产业集群创新机理研究[M].吉林:吉林大学出版社,2011.
[17] 付尧,刘红丽.社会网络结构与互动对知识转移的影响 [J].商场现代化,2009(1):394−395.
[18] 蔡莉,朱秀梅.科技型新创企业集群形成与发展机理研究[M].北京:科学出版社,2008.
[19] Alan D Meyer,James B Goes. Organizational Assimilation of Innovations: A Multilevel Contextual Analysis [J]. Academy of Management Review, 1988,31(4):897−923.
[20] Ari Jantunen. Knowledge−processing Capabilities and Innovative Performance: An Empirical Study [J]. European Journal of Innovation,2005,8(3):336−349.
[21] Spender J C. Marketing Knowledge the Basis of a Dynamic Theory of the Firm [J]. Strategic management Journal,1996(17):45−62.
[22] Hammady Ahmed Dine Rabeh,Daniel Jimene'z−Jimene'z,Micaela Martı'nez−Costa. Managing Knowledge for a Successful Competence Exploration [J]. Journal of Knowledge Management, 2013,17(2):195−207.
[23] 中国汽车技术研究中心,中国汽车工业协会.中国汽车工业年鉴[M].北京:中国汽车工业年鉴出版社,2012.
[24] Granovertter M. The Strength of Weak Ties[J]. American Journal of Sociology,1973,78(6):1360−1380.
[25] Uzzi B. Social Structure and Competition in Interfirm Networks: The Paradox of Embeddedness[J]. Administrative Science Quarterly, 1997,42(1):37−69.
[26] Capaldo A. Network Structure and Innovation: The Leveraging of a Dual Network as a Distinctive Relational capability [J]. Strategic Management Journal,2007,28(6):585−605.

（上接第56页）

算法设计上有待进一步的细化和设计，以提高实验结果的质量。

参考文献

[1] Marc Krier,Francesco Zacca.Automatic Categorization Applications at the European Patent Office[J].World Patent Information,2002, 24(3): 187−196.
[2] 李振亚,孟凡生.基于四要素的专利价值评估方法研究[J].情报杂志,2010(8):87−90.
[3] 郭婕婷,肖国华.专利分析方法研究[J].情报杂志,2008(1):12−14.
[4] 李建蓉.专利信息与利用[M].北京:知识产权出版社,2011:8−10.
[5] 王裴岩,张桂平,蔡东风,等.一种用于专利主题词抽取的模板自动生成方法[J]. 沈阳航空工业学院学报,2010(3):46−49.
[6] 常鹏,马辉.高效的短文本主题词抽取方法[J].计算机工程与应用,2011(20):126−128,154.
[7] 李鹏,王斌,石志伟等.Tag−TextRank:一种基于Tag的网页关键词抽取方法[J].计算机研究与发展,2012,49(11): 2344−2351.
[8] 张榕.术语定义抽取、聚类与术语识别研究[D].北京:北京语言大学,2006:35−41.
[9] Frantzi K T,Ananiadou S,Mima H.Automatic Recognition of Multi−word Terms:The C−value/NC−value method[J].International Journal on Digital Libraries,2000,3(2):115−130.