情报工程 ISSN: 2095-915X
第1卷 第2期 31-37 2015 年 4 月
TECHNOLOGY INTELLIGENCE ENGINEERING
ISSN: 2095-915X Vol.1 No.2 31-37 Apr 2015

doi:10.3772/j.issn.2095-915x.2015.02.004

面向技术机会发现 TOD 的专利信息抽取 *

——韩国科学技术信息研究院 KISTI 语义服务

周雷[1,2]，李颖[1]，石崇德[1*]

（1. 中国科学技术信息研究所 北京 100038；2. 北京万方数据股份有限公司 北京 100038）

摘要：技术机会发现 (TOD,Technology Opportunity Discovery) 是面向新技术进行监测，并提供机会的一种服务；所谓“基于专利的信息”是指采用自然语言技术对专利进行抽取的结果。本研究的目标资源覆盖过去 20 年间发表的所有专利，目标信息则是其中产品名称及其部分 – 整体关系 (Part−of relations)。应用基于词典和相似度的命名实体识别、基于模式的关系抽取、以及基于机器学习的信息过滤几项技术，本研究取得了令人鼓舞的效果。

关键词：技术趋势分析，自然语言处理，信息抽取，部分 – 整体关系，机器学习

中图分类号：TP391.2

Patent Information Extraction for Technology Opportunity Discovery

——A KISTI's Semantic Service

ZHOU Lei[1,2], LI Ying[2] , SHI Chongde[2]

(1.Institute of Scientific and Technical Information of China, Beijing 10038, China; 2.Wanfang Data, Beijing 100038, China)

Abstract: Technology Opportunity Discovery is a service to detect and provide opportunities for the new technologies. Patent-based information is extracted by natural language processing techniques. All patents published during the past 20 years are target resources and product names and their Part-of relations are target information. A dictionary and similarity-based named entity recognition, a pattern-based relation extraction, and a machine learning-based filtering have been used and showed an encouraging performance.

Keywords: Technology trend analysis, natural language processing, information extraction, Part-of relation, machine learning.

* 本文章素材 (Detection of Technology Opportunities from Patents. 2014;KISTI TechnologyOpportunity Discovery. ICSTI2014) 由中国科学技术信息研究所国际合作机构 KISTI 的全弘宇 (Hong - Woo Chun) 博士提供并授权使用。

基金项目：本研究得到国家自然科学基金项目“面向科技监测的实体识别与关系抽取研究”（编号：71403257）资助。

作者简介：周雷（1987），助理翻译，研究方向：自然语言处理，email：zhoulei@wanfangdata.com.cn，联系电话：010 - 58882726；李颖（1964），博士，副研究员，研究方向：自然语言处理，email：liying@istic.ac.cn，联系电话：010 - 58882470；石崇德 (通讯作者)（1979），博士，助理研究员，研究方向：自然语言处理，email：shicd@istic.ac.cn，联系电话：010 - 58882447。

0 引言

KISTI(Korea Institute of Science andTechnology Information，韩国科学技术研究院) 成立于 1962 年，现役员工 590 名。其使命为充当韩国先端科技信息服务的门户，促进技术的市场化。围绕这一使命，针对科技新项目与新业务，KISTI 启动了一系列与技术机会发现（Technology Opportunity Discovery, TOD）有关的研发活动。TOD 是针对某一项技术，进行监测并提供机会的服务。TOD 的架构如图 1 所示，它可以理解为机会搜索，针对某一产品，搜索目标主要包含共现产品、产品有关专利、上游产品、下游产品、同类产品（它们具有同一上游或下游产品）、与产品布局相近的专利受让人分布、受让人名下的项目等等。

本研究作为 TOD 一个重要环节，探讨基于专利的技术机会监测。

1 概述

每年技术相关专利的发表量都稳定在 150000 项以上，80% 的技术信息可以从专利中获取 [1]，因此大量专利是技术机会发现（TOD）的充足的素材资源。本研究包含两个假设：(1) 产品趋势分析在技术机会发现中扮演着重要角色；(2) 所有产品的目标都是面向交易的，因此它们需要分配商标。基于自然语言处理技术（NLP）的信息抽取是分析产品趋势的第一步，而本研究进行信息抽取的对象即是专利中的产品名称及名称间关系。

本研究的目标是：利用多种自然语言处理技术对美国专利与商标局 (USPTO) 的专利进行信息抽取，以实现高效的技术机会监测。

2 相关研究

此前，已有较多研究将自然语言处理技术用于产品趋势的分析。S.Wu 等人致力于从对话文本中识别产品名称，这种方法在 CPROD2012（Consumer PRODucts contest，消费品竞赛）中脱颖而出，他们融合了字符串匹配、启发式规则识别和机器学习三种方法对产品名称进行识别，取得了 F 值为 0.22041 的实验效果 [2]。Z.Shaik 等人提出了一种聚类方法 [3]，他们对 1976 至 2011 年间发表在 SUGI/SAS 全球论坛上的研究论文中的术语进行聚类分析，而后利用同一类别内有关术语的频率信息来分析技术生命周期。S.C.Choi 等人则尝试利用专利句子的“主谓宾”（SVO）结构来提升产品和技术趋势分析的准确率 [4]，但由于“主谓宾”结构的抽取需要花费大量时间，这种方法在处理大规模数据时有一定局限性。

虽然，将自然语言处理技术应用于产品趋势分析已有较多尝试，但其中所涉及的自然语言处理技术往往相对简单，并且一些研究甚至未提及实验的效果如何。因此，本研究意在将信息抽取技术引入产品趋势分析，利用专利自身特征（characteristics）对专利中的产品信息进行抽取，以期提出一种有效的产品趋势分析方法。

图 1 TOD 架构

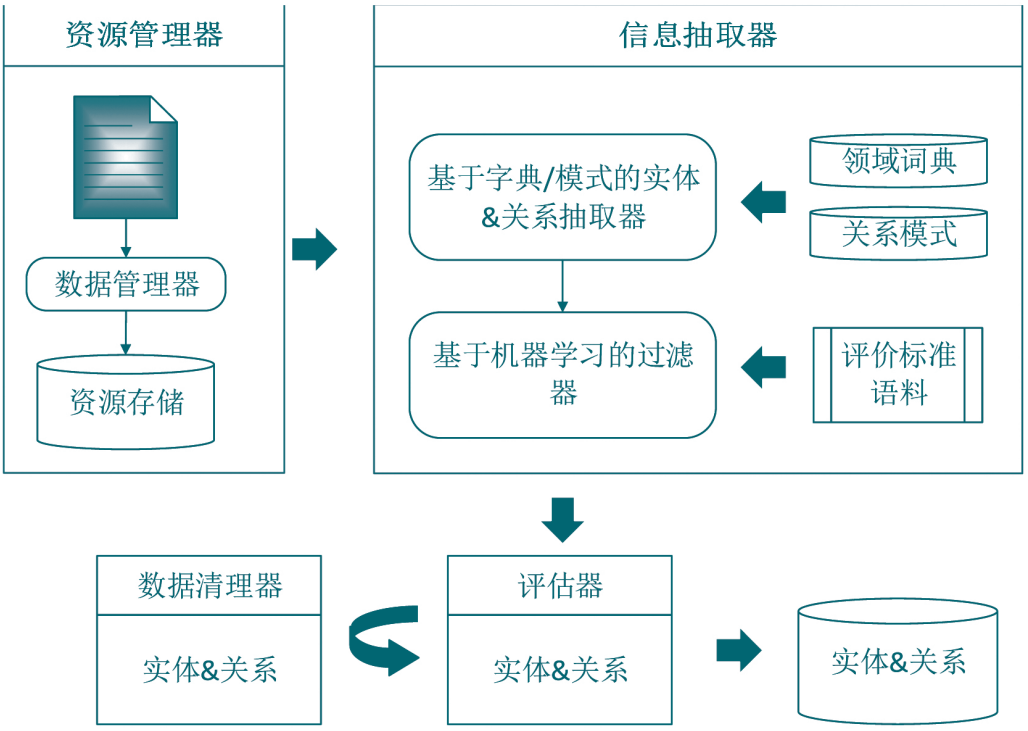


图 2 基于专利的技术机会监测系统架构

3 提出方法

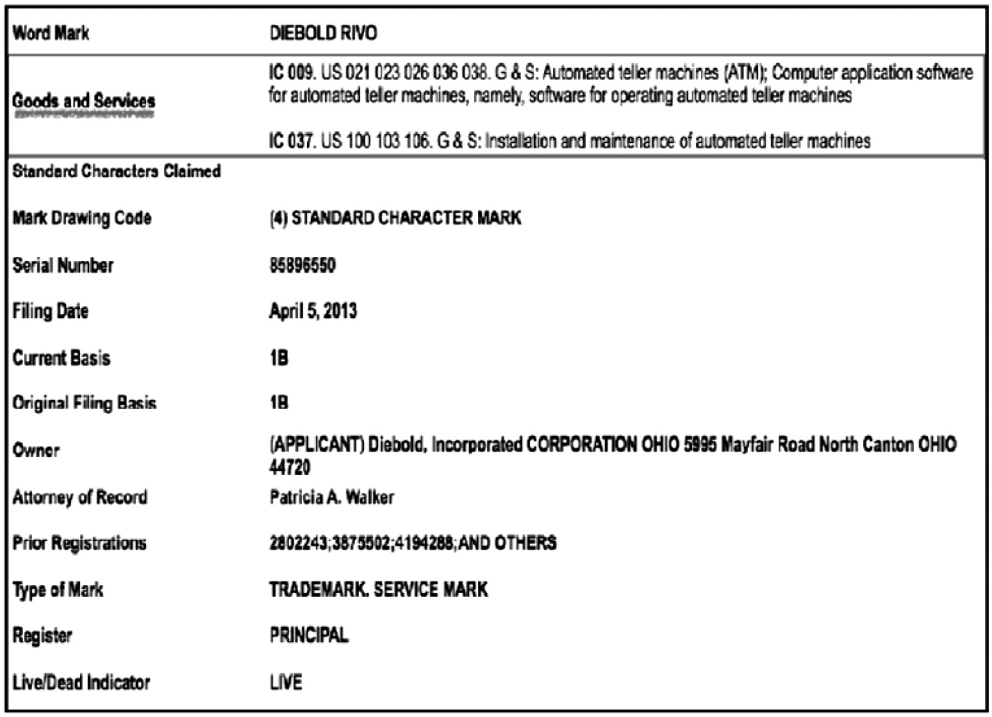
图 2 是对系统架构的描述。USPTO 专利信息首先被存储在资源管理器中，而后系统通过两个环节来实现信息抽取：第一个环节包含一个基于词典的实体识别器和一个基于模式的关系抽取器；由于基于词典和模式的方法通常会产生大量误报，因此需要在第二个环节采用基于机器学习的过滤器来过滤掉错误实体和错误关系。为了评测抽取结果，本研究构建了评价标准数据体系，评价指标包括准确率、召回率和 F 值。

3.1 目标数据

专利是非常适合做产品趋势分析和技术机

表 1 专利数量

IPC 分类号	说明	专利
A	人类生活必需	526402
B	作业；运输	555772
C	化学；冶金	289189
D	纺织；造纸	24591
E	固定建筑物	78498
F	机械工程；照明；采暖；武器；爆破	214917
G	物理	934492
H	电学	687,032



Word Mark	DIEBOLD RIVO
Goods and Services	IC 009. US 021 023 026 036 038. G & S: Automated teller machines (ATM); Computer application software for automated teller machines, namely, software for operating automated teller machines IC 037. US 100 103 106. G & S: Installation and maintenance of automated teller machines
Standard Characters Claimed	
Mark Drawing Code	(4) STANDARD CHARACTER MARK
Serial Number	85896550
Filing Date	April 5, 2013
Current Basis	1B
Original Filing Basis	1B
Owner	(APPLICANT) Diebold, Incorporated CORPORATION OHIO 5995 Mayfair Road North Canton OHIO 44720
Attorney of Record	Patricia A. Walker
Prior Registrations	2802243;3875502;4194288;AND OTHERS
Type of Mark	TRADEMARK. SERVICE MARK
Register	PRINCIPAL
Live/Dead Indicator	LIVE

图 3 美国商标数据库案例·

[1] 图 3 中文参考翻译

文字商标	DIEBOLD RIVO
商品和服务	IC 009.US 021 023 026 038.G&S; 自动提款机（ATM）; 自动提款机的电脑应用软件，即用于操作自动提款机的软件 IC 037.US 100 103 106.G&S: 自动提款机的安装与维护
规范字体声明	
商标绘图代码	（4）标准文字商标
序列号	85896550
提交日期	2013 年 4 月 5 日
类别	1B
原始归档类别	1B
所有者	（申请人）俄亥俄迪堡股份有限公司，俄亥俄州（邮编: 44720），北坎顿，梅尔菲尔路 5995 号。
记录在案的律师	帕特丽夏 .A. 沃克
优先注册	2802243；3875502；4194288；AND OTHERS
商标种类	商标，服务商标
注册记录	主簿
是否有效	有效

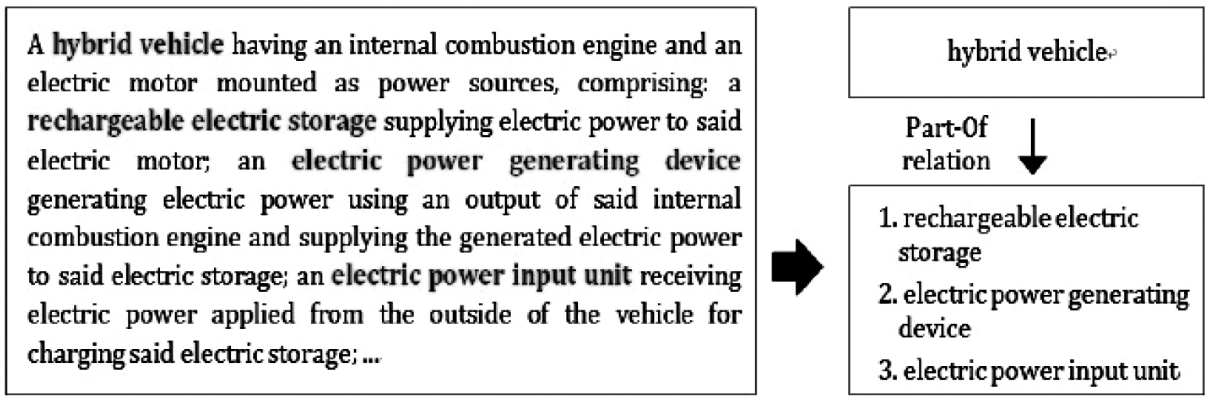
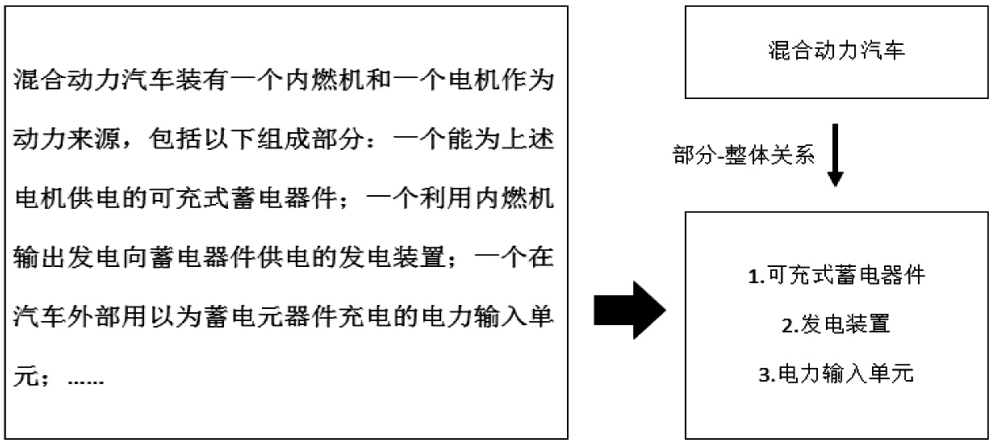


图 4 专利声明中的关系抽取实例 [2]

[2] 图 4 中文参考翻译



会分析的一类资源，这一点在此前的许多论文中都有所论述。因此，我们从谷歌专利网站搜集了自 1992 年 1 月 22 日至 2013 年 7 月 16 日间，USPTO 发布的全部专利信息，其中共包含 118908059 条句子和 3624693988 个记号 (tokens)。本研究仅使用其中的标题、摘要和声明项。表 1 是依据国际专利分类（IPC）对收集到的专利按部统计的数据结果。

3.2 词典构建

我们利用美国商标数据库来构建产品名称词典。美国商标数据库包含“商品和服务”这一字段（见图 3），我们对这一字段所包含的信息进行噪声过滤，提取其中复合名词，通过在这种方法共收集了 328288 个产品名称。如图 3 和图 4 中的“自动提款机”和“ATM”(Automated teller machines) 便是需要提取的产品名称。

3.3 命名实体识别

本研究采用基于词典的最长匹配算法和基于机器学习的过滤方法对专利中的产品名称进行识别，并通过简单字符串的相似度计算对新词（即新产品名称）进行识别。计算字符串的相似度，需要对词典实体和专利名词短语（见公式 1 命名实体识别的相似度计算方法）中相同的词进行统计，并对词典实体和专利名词短语中的中心词进行比对。由于中心词是最能描述实体特征的词，因此当两个实体的中心词相同时，其间相似度应

乘以 2。公式 1 中，实体 1 来自词典，实体 2 来自专利。这一算法需要对专利和词典中的所有词条进行比对，因此需要强大的运算能力和高效的算法。

$$\text{相似度} = \frac{\text{相同词数量}}{\text{最长词数量(实体 1，实体 2)}}\text{相似度} = \text{相似度} \times 2$$

如果最后一个词(实体 1)=最后一个词(实体 2)（公式 1）

基于机器学习的过滤器可以过滤掉基于词典的命名实体识别过程中产生的误报名称，而候选词条的词性特征和候选词条前后相邻的两个单词的特征均可作为机器学习过滤器的判断依据。

3.4 关系抽取

本研究意在使用基于模式的关系抽取和基于机器学习的过滤技术，对产品名称间“部分－整体”模式的关系进行抽取，并据此构建出产品的供应链条。为此，我们人工构建了 3039 个语义模式，根据专利中英文句式特点对这些语义模式进行优化，并将专利的上下文特征用于关系抽取过程。图 4 和图 5 是专利声明中的一个语句，从中我们可以抽取三个“部分－整体”的关系：“混合动力汽车”－“可充式蓄电器件”(Hybrid vehicle−rechargeable electric storage)，“混合动力汽车”－“发电装置”(Hybrid vehicle−electric powergenerating device)，“混合动力汽车”－“电力输入单元”(Hybrid vehicle−electric power input unit)。这种抽取方法需要对专利语言的特征加以利用，比如在专利声明部分，包含“comprising”的语句就是比较常见的关于“部分－整体”关系的描述。

基于机器学习的过滤器在关系抽取的环节中同样扮演着重要角色，它可以从基于模式的关系抽取的结果中过滤掉误报信息。而诸如两个备选词条中所有词的词性、词袋、两个备选词条的距离、构词模式等则都可以作为特征信息应用于机器学习当中。

4 实验结果

为了验证这一方法的有效性，我们人工建立了一个评价标准数据体系，共包含 2400 条专利中的 509322 个句子，应用了准确率、召回率、F 值三个评价指标。表 2 展示了实验的结果。基于规则的方法是指基于词典匹配的命名实体识别和基于模式的关系抽取两种方法。机器学习方法能合理的降低召回率并提高准确率，从美国商标数据库提取的规则也同样是有效的。由于基于规则方法和机器学习方法是先后进行的，所以在评价整体效果时取的是两者的乘积。

5 结束语

本研究将专利信息抽取应用于技术机会发现（TOD），并进行了一系列探索，包括利用自然语言处理技术对专利语句进行分析，并在 2400 项专利的基础上构建用于监督学习和有效性评价的语料库。我们在利用机器学习过滤技术过滤误报

表 2　实验结果

	规则方法		机器学习方法		整体	
	实体	关系	实体	关系	实体	关系
准确率	0.651	0.506	0.975	0.809	0.635	0.409
召回率	0.683	0.552	0.874	0.895	0.597	0.494
F 值	0.667	0.528	0.922	0.850	0.616	0.447

信息、利用美国商标数据库中产品名称进行专利产品名称识别两个方面取得了重要进展。目前而言，将基于词典的命名实体识别和基于机器学习的信息过滤两项技术相结合取得了可观的收效，在未来的研究中，我们还将就产品名称的词义消歧和基于大数据的自然语言处理这两个问题进行更为深入的研究。

参考文献

[1] LEE C Y, JEON J H, PARK Y T. Monitoring Trends of Technological Changes Based on the Dynamic Patent Lattice: a Modified Formal Concept Analysis Approach[J]. Technological Forecasting and Social Change, 2011:690-702.

[2] WU S, FANG Z P, TANG J. Accurate Product Name RecognitionfromUser Generated Content[C]. IEEE 12th International Conference on DataMining Workshops, 2012: 874-877.

[3] CHOI S C, KIM H B, YOON J Y, et al.An SAO-based Text-mining Approach for Technology Roadmapping Using Patent Information [J]. R&D Management, 2013:52-74.

[4] SHAIK Z, GARLA G, CHAKRABOTY G. SAS® Since 1976: an Application of Text Mining to Reveal Trends [C] // SAS Global Forum, 2012: 1-10.