情报工程 ISSN: 2095-915X
第1卷 第4期 88-94 2015年8月

TECHNOLOGY INTELLIGENCE ENGINEERING
ISSN: 2095-915X Vol.1 No.4 88-94 Aug 2015

doi:10.3772/j.issn.2095-915x.2015.04.010

基于论文与专利整合数据的研究方向发现模型研究

任智军[1,2]，詹淑琳[2]，范婷婷[2]

（1. 中国科学技术信息研究所 北京 100038；2. 国家知识产权局中国专利信息中心 北京 100088）

摘要：海量信息分析及基于分析结果得出相应的结论是一项非常重要的工作，该工作需要投入大量的时间和精力。本文涉及了基于论文与专利整合数据的研究方向发现模型，从论文和专利共同研究方向和新兴技术两个方面来揭示论文和专利之间关系，最后，我们对电动汽车论文和专利数据进行了实例验证。

关键词：论文，专利，知识挖掘，文本挖掘，知识发现

Research of Analysis and Mining Method Based on Integration of Papers and Patents

REN Zhijun[1,2], ZHAN Shulin[2], FAN Tingting[2]

(1.Institute of Scientific and Technical Information of China, Beijing 100038;
2.The China Patent Information Center, the State Intellectual Property Office, Beijing 100088)

Abstract: Analyzing mass information and supporting insight based on analysis results are very important work but it needs much effort and time. Therefore, in this paper, a model based on the integration of the data in paper and patent is proposed to reaveal the relationship between paper and patent by analyzing their mutual researching area and mutual emerging technologies with knowledge mining and text mining.And then we examine the empirical research by the data of electric vehicle.

Keyword: Papers,patents,knowledge mining,text mining,knowledge discovery

基金项目：中国博士后科学基金资助项目（基于论文与专利整合的分析和挖掘方法研究 2013M540125）
作者简介：任智军，男 (1977-)，高级工程师，研究方向：数据挖掘、技术预测；詹淑琳，女（1988-），研究实习员，研究方向：专利分析，竞争情报；范婷婷，女（1984-），助理研究员，研究方向：数据分析、数据挖掘。

1 概述

随着知识时代的来临，科学技术对经济和综合国力的影响越来越大，世界各国都在努力发展科学技术。对科技资源进行检测，分析评价科学技术发展的状态和趋势，选择合理优先发展的科技领域、合理配置科技资源、有效开展科技合作的方法和途径已经成为国内外学者普遍关注的热点[1]。论文是基础研究成果的表现形式，专利文献是技术创新成果的表

现形式[2]，随着基础研究与应用研究的界限渐趋模糊，期刊论文与专利文献之间相互参考与共享的价值日益凸显，两者之间的有效链接除了有助于用户便捷、高效的获取科技文献信息外，还能加强基础研究和应用研究之间的联系，有助于实现知识增值目的[3]。然而，当前的文献服务体系中，论文与专利文献资源表现出明显的局部有序但整体无序的孤岛特征[4]，国内学者大都局限于将论文或者专利数据进行单独研究，将结论进行综合分析，这样以单一视角和不全数据进行的分析得到的结果相对比较片面。

因此，论文和专利资源的融合分析对于理解科学和技术之间的联系、提高技术创新水平有着重要的意义，而且大量的实证研究也表明这两种资源的集成揭示分析有助于理解技术发展趋势、产学政关系、度量创新水平等。因此，研究如何进行论文和专利数据整合，同时改变现有数据孤岛式的深加工的加工方式，形成论文和专利数据整合理论和方法是具有重要实际意义的。本文提出一种新的研究方法，将论文和专利两种异构化的数据进行结构化整合，然后分析与挖掘可以得到更加全面多元化的分析结果，文章最后使用了电动汽车产业的期刊论文数据和专利数据对本方法进行了实证研究，结果也相对比较理想。

2 论文与专利整合策略

目前科技资源的整合实践主要是基于OPAC整合以及基于引文链接的整合，学者赖院根认为上述两种整合方式都不甚理想，于是提出了通过分类体系之间的映射关系实现期刊论文与专利文献在领域层面的对接，利用主题词表，建立期刊论文与专利文献在主题层面的链接[5]。赖院根提出的整合方法在一定程度上更加高效的实现了论文文献数据与专利文献数据的整合，在专利技术方面的分析较为有效，但是还不够细化，因此本文提出更为细化的整合策略，从信息资源整合的角度入手，通过两类数据中都共同存在的人名数据字段将两类资源进行整合，使得分散无序的资源变得集中有序。

首先，基于大规模科技论文和专利文献数据收集和整理。在数据收集整理阶段要对各种类型（中国专利著录项目及全文数据、美国专利著录项目及全文数据、中国科学技术信息研究所NSTL外国期刊和论文数据、中国期刊和论文数据）数据进行归并整理；在数据预处理阶段，先将科技人员进行同名处理等清理，得到唯一姓名标识，然后采用前期研究积累中已有的针对专利文献和论文定制的关键词提取工具对关键词进行提取，并将专利数据中的发明人进行英文转化，以便进行将中英文专利和论文同时纳入分析研究范围，提高数据全面性，以期得到更为全面的分析结果。

然后，本文选取电动汽车论文数据和专利数据进行实证研究，进行论文和专利的整合，从信息资源整合的角度入手，通过挖掘其理论深度，把无序分散的资源集中起来，把无序的资源变为有序，使之有效实现论文资源和专利资源的知识化组织、方便后继的分析与挖掘。

3 共同研究方向论文与专利识别分析

在论文和专利整合基础上，基于相似度算法进行论文和专利技术相关度度量，对论文和专利技术融合程度进行分析，研究论文和专利两种数据之间存在的相互影响和排斥关系，从而探索科学和技术之间的发展规律，确定技术研究方向。挖掘论文和专利之间的共同研究方向的首要问题是发现属于共同研究方向的文献，本文采用的方法则是利用整合数据做聚类，在聚类结果中论文和专利相似文献的基础上，再次聚类，研究论文和专利共同研究方向，具体如图 1 所示，模型说明如下：

(1) 从论文和专利整合数据库中抽取所有文献的关键词，并将低频关键词剔除。

(2) 利用关键词在论文专利高频矩阵中的位置分为论文关键词、高频共词、专利关键词。

(3) 根据关键词构件论文和专利特征。

(4) 首先将以上关键词建设为一篇文献作为 KMeans 聚类算法的初次核心，分别对论文和专利聚类，其中论文分为共同研究方向类和非相似论文类，专利分为共同研究类和非相似专利类。每次迭代使用共同研究方向文献计算新的核心，然后再次将所有论文或者专利进行聚类。

(5) 根据聚类结果将每篇文献比较可标识为共同研究方向论文和专利、非相似论文非相似专利。

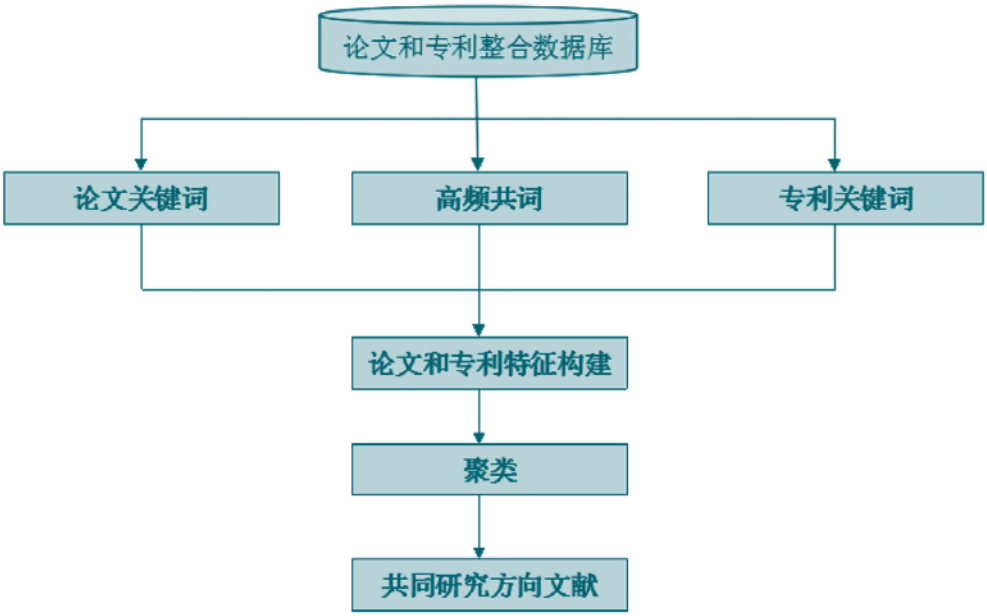

图 1 论文和专利共同研究方式识别模型图

4 技术研究方向挖掘模型

在找到相似专利和论文的基础上，如何确定共同研究方向及新的研究点，是本文要重点解决的问题。通过聚类分析可以使相似性更高的专利和论文聚为一类，在聚类中，具有相同关键词的文献聚为一类，可以通过基于关键词来标注该类别。由于新的研究方向通常由新兴技术词汇的文献来体现，因此对于论文和专利中可能的新方向可以利用新兴技术识别获得。为此，从共同的研究方向和新的研究方向两个方面来构建技术研究方向挖掘模型。

4.1 技术研究方向挖掘模型

在技术研究方向研究中，通过聚类分析的目标是挖掘论文和专利的共同研究方向，研究方向

的发现需要对论文和专利进行深度挖掘，以深入地了解论文和专利共同研究方向。图 2 给出了技术研究方向挖掘模型，具体如下：

(1) 分别在非相似论文、共同研究方向文献和非相似专利分别抽取关键词，并将低频关键词剔除。

(2) 关键词作为聚类操作的特征词，采用 KMeans 聚类算法进行聚类。由于聚类的类别数目确定了共同研究方向数量，因此对类别数目事先采用最佳类别评估方法。

(3) 研究方向的标注方法可以用基于互信息对每一个类别进行标注，该标注即为论文和专利的共同研究方向。

(4) 在共同研究方向中进行新兴技术识别。

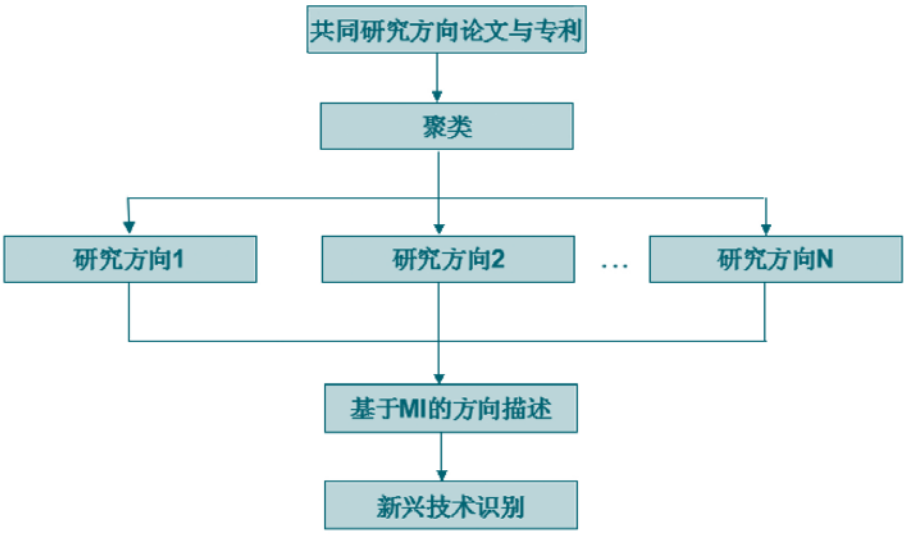

图 2 技术研究方向挖掘模型图

4.2 研究方向标注

在得到研究方向后，需要对研究方向进行标注，研究方向的标注主要通过科技术语。互信息是表示两个变量之间关联程度有用的度量，因此可以通过关键词与研究方向的互信息来进行研究方向的标注。在对共同研究方向论文和专利聚类后，具有相同研究方向的论文和专利聚为一类，这些论文和专利在关键词上具有共性，提取这个共性词汇就是对研究方向的标注。由于与同一个研究方向上互信息最大的关键词不只一个，因此共同研究方向标注有可能由一组词汇组成。基于互信息的研究方向描述的算法如下：

(1) 统计聚类结果中总的文献数 n 与各类别文献数 $ni,i=1,2,...,K$（k 为聚类数目）。

(2) 根据聚类结果，利用互信息公式计算每个关键词与所有类别的互信息，MI 公式如下所示

$$MI(t_k,c_i)=P(t_k,c_i)\log\frac{P(t_k,c_i)}{P(t_k)P(c_i)} \quad ①$$

(3) 分类别统计最大互信息值及对应的主关键词。

4.3 新兴技术发现模型

新兴技术是建立在科学基础上的革新，它们可能创立一个新行业或改变某个老行业。新兴技术代表了一个新的研究方向，被用来分析科技的前沿领域和发展趋势。新兴技术对于研究方向研究意义重大：一方面，论文与专利包括了科学和技术，在论文和专利的共同研究热点上，最有可

能涌现新兴技术术语。另外，新兴技术也会推动科学和技术之间的发展，吸引科研机构和商业公司投入研究，从而提高论文与专利的产出和提高技术创新水平有着重要的意义。由于新兴技术是技术创新的源泉，因此在研究方向中发现新兴技术是非常必要的，本文研究的新兴技术是指在研究年份最近两年刚刚出现，且技术增长率高于平均水平的新兴技术，公式如下所示：

新兴术语的识别可以通过与人工标识对比进行验证，识别准确率 = 人工标注 / 自动识别总数 *100%。

5 实验及分析

为验证基于论文与专利整合数据的研究方向发现模型的可行性，本文选择了电动汽车的论文数据和专利数据进行实例研究，在万方论文和 CPRS 获取了数据，包括中国电动汽车论文 65468 篇，中国电动汽车专利 145645 篇，利用论文与专利整合之后构建了电动汽车论文和专利整合数据库，然后对电动汽车领域进行论文与专利术语，研究方向发现和新兴技术发现。

$$N_{Pp}^{k}(t_1)+N_{Pp}^{k-1}(t_1)=AN_{Pp}^{k-2}(t_1),N_{Pp}^{k-2}(t_1)=0$$

$$N_{Pt}^{k}(t_1)+N_{Pt}^{k-1}(t_1)=AN_{Pt}^{k-2}(t_1),N_{Pt}^{k-2}(t_1)=0$$

$$FS_{RelativeGrowthRate}^{k}(Pp)>\sum_{r=2}^{Years}FS_{RelativeGrowthRate}^{k}(Pp)/(Years-1)$$

$$FS_{RelativeGrowthRate}^{k-1}(Pp)>\sum_{r=2}^{Years}FS_{RelativeGrowthRate}^{k-1}(Pp)/(Years-1)$$

$$FS_{RelativeGrowthRate}^{k}(Pt)>\sum_{r=2}^{Years}FS_{RelativeGrowthRate}^{k}(Pt)/(Years-1)$$

$$FS_{RelativeGrowthRate}^{k-1}(Pt)>\sum_{r=2}^{Years}FS_{RelativeGrowthRate}^{k-1}(Pt)/(Years-1)$$

表 1 论文和专利关键词频次（TOP10）

序号	关键词	论文频次	专利频次	频次	序号	关键词	论文频次	专利频次	频次
1	充电	4121	34382	38503	6	电动机	6899	6362	13261
2	发动机	4692	16224	20916	7	传感器	2542	8281	10823
3	蓄电池	2076	15379	17455	8	锂离子	1148	9077	10225
4	燃料电池	6303	10681	16984	9	控制系统	5655	4491	10146
5	控制器	4151	11966	16117	10	发电机	3575	6118	9693

5.1 论文与专利术语选择

论文与专利整合数据中共包含 61017 个关键词。分析论文和专利中频次较高的共同的关键词，构成相似关键词，频次排名前 10 的关键词见表 1。

5.2 研究方向标注分析

利用共同研究方向论文与专利识别方法对电动汽车论文与专利整合数据进行识别之后，得到共同研究方向论文与专利 37314 篇，并通过第 4.2 节中公式①对类别进行标注，结果见表 2。从表 2 可以看出，电动汽车论文与专利的研究中，共同的研究方向为“充电”、“燃料电池”、“外观设计”、“汽车电子”、“控制电路”、“锂离子电池”、“发动机”、“铅酸蓄电池”、“电动机”与“发动机仿真”等 10 个方向。

5.3 新兴技术发现结果分析

根据第 4.3 节中新兴技术术语的识别方法，在电动汽车的论文和专利的各个研究方向上对 2004 年到 2013 年共十年间出现的新兴技术术语进行检测，其中频次出现较高的共有 57 个，具体

表 2 电动汽车研究方向标注

类别	互信息	类别	互信息
充电	充电 ,2.11；蓄电池 ,1.85；太阳能电池 ,1.75；	锂离子电池	锂离子 ,2.04；离子电池 ,1.86；锂离子电池 ,1.86；
燃料电池	燃料电池 ,4.84；氧化物 ,0.83；燃料电池系统 ,0.78；	发动机	发动机 ,2.04；变速器，1.036；离合器，0.89；
外观设计	外观设计 ,5.67；设计产品 ,5.38；	铅酸蓄电池	铅酸蓄电池 ,0.18；酸蓄电池 ,0.18；
汽车电子	汽车电子 ,0.94；电动汽车 ,0.73；汽车工业 ,0.72；	电动机	电动机 ,1.04；定子 ,1.03；永磁同步 ,0.70；同步电机 ,0.60；
控制电路	输出端 ,0.60；控制电路 ,0.59；变压器 ,0.57；输入端 ,0.49；	发动机仿真	有限元 ,0.59；发动机 ,0.31；仿真结果 ,0.30；

表 3 电动汽车新兴技术术语表

序号	时间	新兴技术术语	序号	时间	新兴技术术语
1	2004	Saber 仿真；直接醇类燃料电池家用燃料电池；电动汽车动力电池；多能源动力系统	6	2009	路径识别；纯电动乘用车；发动机悬置支架；触媒金属
2	2005	电化学充放电；混合动力城市客车；燃料电池氢源系统；高容量锂电池；氢动力燃料电池	7	2010	电动汽车充电桩；交流充电桩；全钒液流电池；共烧结技术；圆筒永磁直线同步电机；光伏电池发电
3	2006	废弃电池；虚拟仪器开发平台；气液两相流动；无刷直流电机驱动器；固体氧化物燃料电池；储氢金属；微型甲醇燃料电池	8	2011	充换电服务；充换电服务网络；充换电设施；智能充换电；智能充换电服务网络；钛酸锂电池；路径诱导系统；电池箱更换设备；增程式电动车
4	2007	甲酸燃料电池；电控机械式自动变速器；全固态锂离子电池；双机械端口电机；磷酸铁锂电池；永磁直线同步电机伺服系统；直接二甲醚燃料电池；	9	2012	永磁辅助同步磁阻电机；电站监控系统；热网络模型；智能车载终端
5	2008	LEM 电子；新能源车；无传感器磁场定向控制；甲醇串流	10	2013	亚硝酸盐氮；机油压力传感器；FOC(磁场定向控制)；增程发电系统；自动报警系统

如表 3 所示。

通过与人工标注对比，57 个词中有 46 个词是真正的新兴术语，具有 80.7% 的准确性。

6 总结与展望

本文提出了一种基于论文与专利整合的分析与挖掘的研究方法，构建了基于论文与专利整合的分析与挖掘知识挖掘模型，从论文与专利共同研究方向和新兴技术发现两个方面来揭示论文与专利关系，并通过电动汽车的论文数据和专利数据进行了实例验证，为学技术发展趋势、产学政关系、度量创新水平等提供了一个新的研究思路。下一步的工作将在本文提出的模型基础上，进行基于论文和专利整合的技术生命周期研究。

参考文献

[1] 张忠华 , 刘云 . 基于 WOS 和 DII 数据的科技资源监测分析——以二氧化碳排放统计监测领域为例 [J]. 现代情报，2014, 34(8):105-111

[2] RICHARD S. Patent Trends as a technological Forecasting Tool [J]. World Patent Information,1983,5(3):137-143

[3] 赖院根 . 期刊论文与专利文献的链接研究 [C]// 数据图书馆高层论坛 2010 年年会论文集 .

[4] WRIGHT G, GOODWIN P. Decision making and planning under low levels of predictability: Enhancing the scenario method [J]. International Journal of Forecasting,2009,25(4):813-825

[5] 赖院根 , 曾建勋 . 期刊论文与专利文献的整合框架研究 [J]. 图书情报工作 ,2010, 54(4):109-112.

情报工程 ISSN: 2095-915X
第1卷 第4期 95-103 2015年8月

TECHNOLOGY INTELLIGENCE ENGINEERING
ISSN: 2095-915X Vol.1 No.4 95-103 Aug 2015

doi:10.3772/j.issn.2095-915x.2015.04.011

面向专利的化合物和生物实体识别系统

赖鸿昌，朱礼军，徐硕

（中国科学技术信息研究所信息技术支持中心 北京 100038）

摘要：探索专利文献中的化合物和生物知识变得至关重要。为了识别化合物实体和生物实体，开发了面向专利的化合物和生物实体识别系统。系统基于开源的机器学习和自然语言工具进行开发。系统按照流水线模式进行，本文将详细阐述其三个主要过程：预处理（句子分割、词条化），识别（基于条件随机场的方法），后处理（基于规则的方法）。最后，利用系统在已标注的化合物专利语料库进行大量实验，进行十折交叉验证，得到了69.20%的F值。但是，从结果可以看到，在专利文献上的实验表现，要低于论文和新闻语料库中的表现。

关键词：条件随机场，化合物和生物实体，专利挖掘，交叉验证

中图分类号：G350，TP311

Chemical and Biological Entity Recognition System from Patent Documents

LAI Hongchang, ZHU Lijun, XU Shuo

(Center of Information Technical Support, Institute of Scientific and Technical Information of China, Beijing 100038, China)

Abstract: It is crucial to explore the chemical and biological space covered by patent documents. In order to recognize chemical and biological entities, a recognition system is developed on the basis of open-source machine learning and natural language processing (NLP) toolkits. The system processing pipeline consists of three major components: pre-processing (sentence detection, tokenization), recognition (conditional random field (CRF) based approach), and post-processing (rule-based approach). The paper introduces each part in detail.

基金项目：本研究得到国家自然科学基金项目“基于论文和专利资源的技术机会发现研究”（项目编号：71403255）、中国科学技术信息研究重点工作项目“大数据环境下融合多源信息的科技文献智能分析服务平台建设及应用示范”（编号：ZD2014-7-1）的资助。

作者简介：赖鸿昌（1992-），硕士研究生，研究方向：知识工程；朱礼军（1973-），博士，副研究员，研究方向：知识组织、语义检索等；徐硕（1979-），博士，副研究员，研究方向：数据挖掘、信息抽取等。通讯作者：徐硕，E-mail: xush@istic.ac.cn。

Finally, extensive experiments on annotated chemical patent corpus are conducted, and the balanced-F measure is 69.20% with 10-fold cross validation. The results indicates that the performance on patent documents is slightly lower than that of counterpart on paper and news corpus.

Keywords: Conditional Random Field (CRF), chemical and biological entity recognition, patent mining, cross validation

1 引言

探索专利文献中的化合物和生物知识变得至关重要。在早期的药用化学活动的研究中，能起到很好的加速作用 [1, 2]。专利文献中包含大量有价值的化合物和生物实体，例如一些化合物、基因、药物、药物靶点等。但是，面向专利文献的自动识别系统仍然非常有限。

然而，针对论文语料和新闻语料，已经提出了很多识别方法，也开发了较多识别系统。对于专利语料库和论文、新闻语料库的反差，本文认为可能存在两个原因：(1) 已经标注的专利文献对公众而言不容易获取；(2) 专利文献具有法律效应，内容复杂繁琐，难于理解。但是对专利挖掘兴趣的不断提升，使得专利文献命名实体识别的情况正在不断好转，面向专利挖掘的会议也越来越多，例如 BIOINFORMATICS [3], BioCreative[4], JNLPBA [5] 和 iPaMin[6]。Akhondi 等人发布了一个已标注的化合物专利语料 [8]，有助于进行化合物和生物实体识别系统的开发。多篇专利语料先进性自动预标注，再人工将化学名词按照不同的子分类、疾病、靶标、作用方式等进行分类标注。语料全集（Full Set）包含 198 篇专利文档共 400,125 个标注，统一集（Harmonized Set）包含 47 篇专利文档共 37,776 个标注。而且专利文档具有相当的复杂性 [7]，有些内容可能多达上百页。当前环境下，对于无结构文档，尤其是专利文档，自动识别其中的化合物和生物实体仍然是非常具有挑战性的任务。

本文采取了和 Xu S 等人类似的方法 [11]，开发了面向专利文献的化合物和生物实体识别系统。文章的第二部分对数据集进行概述，第三部分介绍了系统的构成和所使用的方法，第四部分是介绍了实验的相关信息，并介绍了实验所使用的数据集。

2 数据概述

Akhondi 发布了一个黄金标准的化合物专利语料，包含两个数据集：统一语料集（harmonized corpus）和全语料集（full corpus）。统一语料集由 47 篇专利文档组成，包含了 9,813 个唯一术语的 36,537 个标注，另外，还含有 1,239 个 OCR（Optical Character Recognition，光学字符识别）错误，包括 1189 个拼写错误和 50 个换行错误；全语料集由 198 篇专利文档组成，包含了 80,977 个唯一术语的 400,125 个标注，其中含有 5,096 个 OCR（Optical Character Recognition，光学字符识别）错误，包括 4,403 个拼写错误和 693 个换行错误。专利的全文和实体标注文档公开发布在 www.biosemantics.org. 另外，还包括了只含两篇文档的训练集。在实验中，主要对统一语料集进行处理和实验。

在对训练集和统一语料集进行分析之后，

在统一语料集中发现了一些化合物和生物实体具有两个或者多个标注，具有嵌套关系。由于在系统中，将 CRF++ 作为处理序列标注问题的具体实现，而 CRF++ 无法处理和识别嵌套的实体，因此我们对此进行了一些相对合理的处理，使每个实体都只具有一个标签。在统一语料集中，标注为 OCR 错误的实体（OCRERRORSPELL，OCRERRORLINE）中，有些还在同一篇文档中标注为其他类型的实体，为保证准确性，我们将其统一为 OCR 错误。最后，我们移除了 488 个嵌套实体，将标注数量从 37,776 减少到 37,288。最后，由 47 篇文档所构成的统一语料集，一共包含了 14 个类别，含有 9857 个唯一实体的 37288 个标注（见表格 1）。

如表格所示，在统一语料集中，标注为 M（IUPAC，International Union of Pure and Applied Chemistry，国际理论和应用化学联合会）和 G（Generic）的实体明显多余其他类别的实体。而另一方面，类别为 Y (InChI, International Chemical Identifier，国际化合物标识), C (CAS(Chemical Abstracts Service，化学文摘社）注册码）and I (SMILES, Simplified molecular input line entry specification, 简化分子线性输入规范）在化合物专利文档中出现极少。由于在数据预处理中，移除了一些嵌套实体的标签（含有多个 tag 的实体），因此术语数量和 Akhondi 在文献中提到的 [8] 会有不同。

3 系统描述和相关方法

在对 MUC−4(Mucin 4) 系统中使用的基本方法进行总结后，Hobbs 提出了一个包含十个基础模块的信息抽取的一般方法模型 [10]。这是基于大量实践所提出的理论模型，也是我们系统设计的理论基础之一。另一方面，也参考了 Xu S 等人发表的化合物实体识别系统 [11]。如图 1 所示，系统由三个主要模块组成，犹如构成一个序列化的流水线。首先，将会对专利文档原文进行句子分

表格 1 黄金标准语料集中的各类术语数量（移除嵌套实体后）

	描述	标注术语量	唯一术语量
M	IUPAC	13943	4592
I	SMILES	20	20
Y	InChi	0	0
D	Trademark	2355	897
B	Abbreviation	2087	146
C	CAS number	6	5
F	Formula	1115	160
R	Registry Number	140	95
G	Generic	8381	811
T	Target	3221	654
Disease	Disease	3765	1205
MOA	MOA	1016	197
OCRERRORSPELL	Spelling error	1189	1029
OCRERRORLINE	Spurious line break	50	46
	Total	37288	9857