基于语义的非相关文献知识发现研究*

[1]刘耀 [2]段慧明 [2]穗志方 [1]王惠临 [3]周扬 [3]王振国

1（中国科学技术信息研究所 北京 100038）

2（北京大学 计算语言学研究所 北京 100871）

3（山东中医药大学 文献研究所 济南 250014）

摘要：本文利用自然语言处理（NLP）技术与方法，针对中文非相关文献知识发现所需的数据基础进行研究，采用以词义为主轴的综合型语言知识库构建思路与方法，结合中医药文献特点，开发相应工具，实现了文本的自动切分与词性标注。在此基础上构建专业语义词典、针对性停用词表，研发基于语义的中文非相关文献的知识发现辅助系统，利用语义网络、语义词典、显性关系排除、语义限制、频率过滤等技术和方法，对非相关文献之间的关联性进行智能化初选，成功地模拟了非相关文献知识发现的过程，达到了帮助科研人员揭示中文文献中的隐性关联，引导并实现知识发现的预期目标。

关键词：非相关文献；自然语言处理；知识发现

Semantics-based Non-interactive Literature-Based Knowledge Discovery

[1]Liu Yao [2]Huiming Duan [2]Sui Zhifang [1]Wang Huilin [3]Zuo Yang [3]Wang Zhenguo

[1]Institute of Scientific and Technical Information of China, Beijing, 100038,China

[2]Institute of Computational Linguistics, Peking University, Beijing, 100871,China

[3]Institute of Chinese Medical History and Literature,Shandong University of Traditional Chinese Medicine,Jinan, 250001,China

Abstract: Using natural language processing (NLP) technique and method, this paper focuses on the research of data base required by Chinese non-interactive literature-based knowledge discovery. It adopts the construction method of comprehensive language knowledge-base with a principal axis of word sense, and develops a corresponding tool to realize automatic Chinese text segmentation and POS tagging, according to the features of Chinese medical literature. Based on the above research, it constructs a pertinence stop list and develops an assistant system of Chinese non-interactive literature-based knowledge discovery. Using some techniques and approaches such as semantic web, explicit relation exclusion, semantic constraint, frequency filtering, etc. the system intelligently chooses the relations among non-interactive literatures. It simulates the process of non-interactive literature-based knowledge discovery successfully, aiming at the anticipative target to help scientific researchers to disclose implicit relation in Chinese literatures

and introduce them to achieve knowledge discovery.

Keywords: non-interactive literature; knowledge discovery; natural language processing

1. 引言

随着学科专业的逐步细化，专业文献的研究范围开始逐渐缩小，专业间的沟通变得越来越困难。原本在专业文献间有价值的关联信息，由于专业文献的高度分化，日益被专业内部海量的信息掩盖。美国芝加哥大学的DonR. Swanson教授于 1986 年创立的一种纯情报学研究方法：基于非相关文献的知识发现[1]，Swanson将他称之为ABC模式。与传统的由A→C的知识发现方法相比，Swanson的知识发现方法明显增强了目的性和方向性，他使科研人员找寻这种隐藏关系的过程不再盲目。B的出现为科研人员提供有益的启发和关键性的引导，帮助专业研究人员认识和发现潜在有用的知识片断间的关联，进一步证实科学假设的可行性。

2. 总体设计

本文采用的解决方案总体原则是：运用自然语言处理（NLP）技术与方法，针对中文非相关文献知识发现的数据基础进行深入研究，在实现文本的自动切分与标注的基础上，研发针对性停用词表；利用语义网络、语义词典、显性关系排除、语义限制、频率过滤等技术手段和方法，对非相关文献之间的关联性进行智能化初选，找到能表达两类非相关文献间关联性的概念、词语列表；帮助科研人员揭示文献中的隐性关联，形成一个有价值的科学假设，从而引导并实现知识发现。 其整体框架如图1所示：

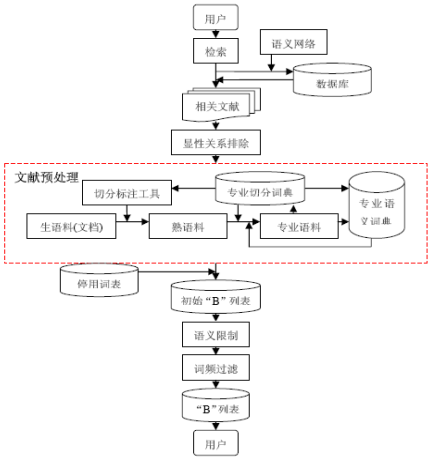

图1：整体框架图

3. 具体方案

3.1 预处理方案及相关技术研究

中文非相关文献知识发现需要的预处理过程，实际上是许多中文信息处理研究的第一步，具体方案是在作者所在单位北京大学计算语言学研究所，俞士汶教授提出的以词义为主轴的综合型语言知识库建设思路与方法的基础上构建而成的。[2]具体思路如图2所示。

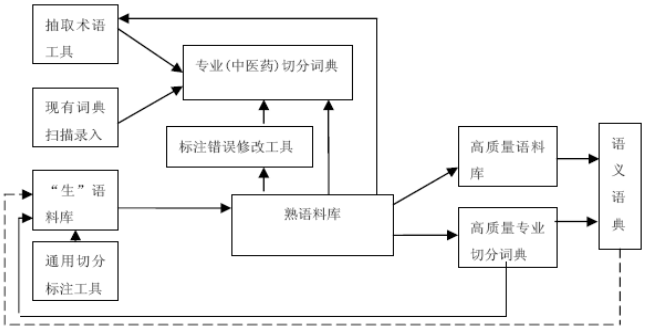

图 2： 专业语料库与专业切分词典关系及生成示意图

功能实现：

该系统以北京大学计算语言学研究所自动切分与标注软件为基础，对语料加工所需的多种软件进行了开发与集成，形成了集加工、辅助修改及词典生成为一体的专业语料加工系统，主要有文件、编辑、检索、切分程序、词表替换、整理词典、抽词程序、环境设置、帮助等

主要功能，系统界面如图 3 所示：

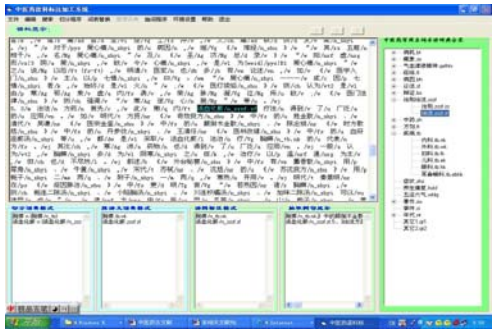

图 3：专业领域语料标注加工系统界面

工作流程及结果:在以上工具的有效支持下，我们进行了大规模中医药古文献语料库的标注工作。其工作流程及结果示例如下：

a 生语料（原文）：

活血通窍法

用于瘀血阻络之头痛。症见头痛如锥如刺,痛有定处,头胀而痛剧,时轻时重,日久不愈,舌质暗,或有瘀斑、瘀点,脉涩。

b 汉语文本切分与词性标注工具标注后的语料：

活/v 血/n 通窍法/n

用于/v 瘀/Ag 血/n 阻/Vg 络/Ng 之/u 头/n 痛/a 。/w 症/Ng 见/v 头痛/a 如/v 锥/Ng 如/v 刺/n ，/w 痛/a 有/v 定/v 处/n ，/w 头/n 胀/a 而/c 痛/a 剧/Ag ，/w 时/Dg 轻/a 时/Dg 重/a ，/w 日/q 久/a 不/d 愈/Vg ，/w 舌/Ng 质/Ng 暗/a ，/m 或/c 有/v 瘀斑/n 、/w 瘀点/n ，/w 脉/Ng 涩/a 。/w

c 运用专业切分词典再次加工后的语料：

活血通窍法/n_zf

用于/v 瘀血阻络/n_bj 之/u 头痛/n_bm 。/w 症/Ng 见/v 头痛/n_bm 如/v 锥/Ng 如/v 刺/n ，/w 痛/a 有/v 定/v 处/n ，/w 头胀/n_zz 而/c 痛/a 剧/Ag ，/w 时/Dg 轻/a 时/Dg 重/a ，/w 日/q 久/a 不/d 愈/Vg ，/w 舌质暗/n_zf ，/m 或/c 有/v 瘀斑/n_zz 、/w 瘀点/n_zz ，/w 脉涩/n_zf 。/w

原始语料经过基本加工之后，文本由汉字串变成了“切分单位”的序列。“切分单位”基本是词，即语言学家所指的“句法词”。斜杠后的字母是根据该词语所表现的句法特性而加的标记。 [3]

3.2 语义词典的构建与利用

利用专业药语料库与专业切分词典构建专业语义词典，并将其植于非相关文献知识发现辅助系统内，从而实现“B列表”语义限制。

3.3 语义网络限定

形成假设结论之前，用户可以对 B 的语义范围进行限定，因为通过对 AB 与 BC 的分析，用户头脑中会形成一个大致的思路，这时 B 大致的语义范围可以设定，由于该辅助工具是基于数据库系统的，所以可以依据数据库分类进语义网络限定，如基础理论、中药、方剂、内科、外科、妇科、儿科、病因病机等，这些范畴之下还可设立子范畴，以进一步增强专指度，用户也可根据不同需求进行选择性组配。

3.4 非相关文献的自动识别

文献间非相关关系的确定主要依赖引文分析法，排除文献间存在的互引、共引等关系。目前，在“中文科技期刊数据库(引文版)”、“中国期刊全文数据库”等数据库中，对文献间的引文关系都有所揭示，为确定非相关文献提供了良好的数据基础。

运用遍历的方法，将收集到的论文集，与中国期刊全文数据库等数据库提供的有关集合进行比对，将具有互引、共引、被引等显性关系的文献进行排除。

3.5 中文停用词表的研制开发

在非相关文献处理过程中，所需要抽取的词汇，并不是题名、主题词、文摘中所包含的全部词语。因此，应该建立针对性中文停用词表，收录一些没有检索意义的助动词、连词、副词等，也可收录专业内不具备检索意义的通用词，如日期、时间、研究、测试等，在抽词的过程加以删除，降低运算的复杂程度，提高结果的准确性。

3.6 频率过滤

B列表中会显示每个B词的频率，用户可按自己的要求对B进行频率过滤，频率过高的词表明A、C在B领域的研究已经比较成熟，因而进行新的知识发现的可能性不大；而如果B的频率过低，则可能包含偶然性的因素较大。所以，对频率如何取值，目前还处在探索阶段。

4．实验方案

4.1　资源准备

实验将充分利用北京大学计算语言所已有的资源与成果。语言资源包括：5000 万字的大规模词性标注语料库、现代汉语语法信息词典、现代汉语语义词典、现代汉语短语结构规则库、中文概念词典 CCD 等；软件工具包括：汉语切分标注工具、语料库辅助工具、古代诗词计算机辅助研究系统等。

除语言资源外，我们还有大量专业资源可以利用，包括以本人为主要成员(第 2 位)开发完成的“中医药文献保障系统”等。

4.2　实验范围

实验范围以“中医药文献保障系统”为主，以“中国期刊全文数据库(医药卫生)”为辅。

4.2.1　中医药文献保障系统[4]

该系统分为“原籍文献数据库”“专题文献数据库”“原籍图像数据库”“标引数据库”“现代期刊数据库”等五大部分，由六十多个数据库组成。涉及到了中医药学的方方面面，数据庞大，检索、管理功能齐备，特别值得一提的是系统根据古文献的特点，以自然段作为基本单位进行自动入库贮存，实现了某些特定功能：

①摘要式显示输出:可以随意控制摘要显示的字符数,尽可能的排除与检索词无关内容。

②自然段显示输出：根据古文献的特征，设定了自然段显示输出，可以实现检索词所在段落的显示与输出。

通过以上特定功能的实现，该系统不但可以全文显示输出，也基本实现了文本的任意输出，为非相关文献知识发现的数据处理提供了极大的方便。

4.2.2　中国期刊全文数据库（医药卫生）

由于“中医文献保障系统”中只收录了与中医药有关的期刊文献，对现代生物医学没有收录，可以利用“中国期刊全文数据库（医药卫生）”的数据进行弥补。

4.3　实验步骤

该设计方案是基于概念的自然语言处理系统的，可以形成假设并检验假设，分析的单元是“中医药文献保障系统”中规范的主题概念，通过对主题概念的语义过滤，可以大幅度缩小用户分析数据的空间。其情报分析主要包括三个过程：C－>B、B－>A 和 A－>B<－C。

前两步是形成假设的开放过程，第三步是检验假设的闭合过程。

在 C－>B 过程中，用户在“中医药文献保障系统”中输入检索的主题概念 C，系统自动

在各数据库中搜索并下载相关记录，选择包含概念 C 的句子，并将句子中出现的所有概念放入临时的数据库表中形成概念 C 的集合。对这些概念可以进行语义过滤，缩小查找空间，并从中选择所需的概念 B。

在 B—>A 过程中，利用选择的概念 B，重复数据库中检索过程，并对出现概念 B 的所有句子的概念进行分析；去除与概念 C 的集合中重复的概念，得到概念 A 的集合。可对集合进行语义过滤，根据用户的知识来选择合适概念 A。

在检验假设（即 A—>B<— C）过程中，用户在“中医药文献保障系统”中输入概念 A，到相关数据库中进行检索，对出现概念 A 的句子进行概念分析，选择与概念 C 的集合中共有的概念，并采用与 C—>B 过程中相同的语义过滤标准，得到 A 与 C 之间的联系，即概念 B。

4.4 实验类型划分

针对不同的文献类型，实验方案略有调整：

4.4.1 文献类型不同：

古文献与现代期刊文献：

中医药古文献是两千多年的积累沉淀，是中医的精华之所在，且数量相对有限。所以方案中所涉及的古文献一律采用全文处理方式，而现代期刊内容庞杂而繁多，加之各种期刊数据库对全文的贮存、处理差异较大，数据收集多有不便，所以，目前只对标题、关键词、摘要进行处理。

4.4.2 概念描述体系不同：

a 中医药古文献与中医药现代文献

由于两者在概念上一脉相承，在文献的处理之初不必进行概念的转换。

b 中医药文献与现代生物医药文献

两者在概念有着较大的差异，在处理之初首先对双方概念的进行转换处理，如：根据“中西医病证名对照大辞典”等工具进行处理，以确保文献查全率。

5. 系统实现

利用自然语言处理(NLP)理论和技术方法，针对中文非相关文献知识发现所需的数据基础进行研究，采用以词义为主轴的综合型语言知识库构建思路与方法，结合中医药文献特点，开发相应工具，实现了文本的自动切分与词性标注。在此基础上构建专业语义词典、针对性停用词表，成功开发出了研发基于语义的中文非相关文献的知识发现辅助系统，其主要功能如下：

①专业语料的切分与标注：集加工、辅助修改及词典生成为一体的专业语料加工系统，主要有文件、编辑、检索、切分程序、词表替换、整理词典、抽词程序、环境设置、帮助等主要功能。

②停用词表构建：用户可以根据需要自由添加，主要提供单个与批量两种添加方式。

③闭合式搜索：分为基于语义的与非语义的两种，界面如图 4 所示。

④开放式搜索：分为基于语义的与非语义的两种，主界面如图 5 所示。

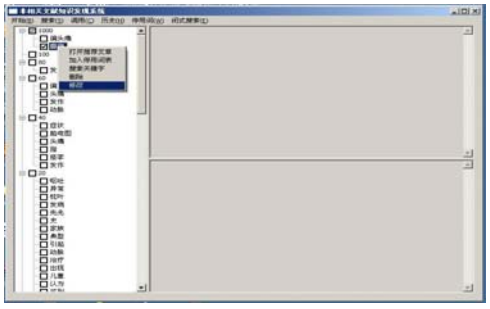

图 4：开放式搜索界面图

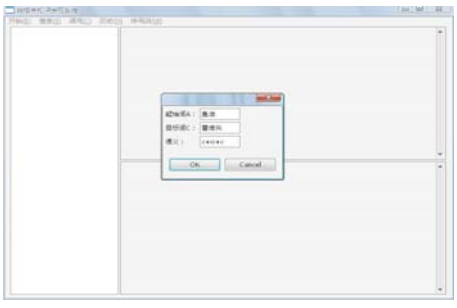

图 5：闭合式搜索界面图

6. 结语

本文利用自然语言处理（NLP）技术与方法，针对中文非相关文献知识发现的数据基础进行深入研究，在实现文本的自动切分与词性标注的基础上，研发专业语义词典、针对性停用词表；利用语义网络、显性关系排除、语义限制、频率过滤等技术手段和方法，对非相关文献之间的关联性进行智能化初选，成功地模拟了DonR.Swanson教授的知识发现过程。其意义：利用NLP技术，创建专业语义词典，并利用其提供的多种词法及句法结构，进行词性、语义类以及共词排除等多种控制，这在国际上尚未见到相关报道，不但是对非相关文献知识发现理论本身的一种创新与发展，而且也是对专业语义词典应用及研究进行了拓展与探索。

参考文献

[1] Swanson,D R.Fish oil,Raynaud ’ s syndrome,and undiscovered public knowledge.Perspect Biol Med[J],1986,30(1):7～18

[2]俞士汶等.以词义为主轴的综合型语言知识库.全国第八届计算语言学联合学术会议论文集,2005

[3]俞士汶，段慧明等．北京大学现代汉语语料库基本加工规范[J]．中文信息学报，2002(5):49-64

[4]王振国，刘耀．对古代科技文献信息构建的理论与方法－－中医药古文献的开发与利用．情报资料工作，2005：(2)

*本文得到国家科技支撑计划项目（2006BAH03B03）、国家 973 项目（2007CB512601）、教育部人文社科项目(06JC870001)、山东省中医药科技专项项目（2003-14）、中国科学技术信息研究所预研基金（YY-200721）的支持。