情报工程 ISSN 2095-915X
第2卷 第1期 103-108 2016年2月

TECHNOLOGY INTELLIGENCE ENGINEERING
ISSN 2095-915X Vol.2 No.1 103-108 Feb 2016

doi:10.3772/j.issn.2095-915x.2016.01.013

多领域视角下的知识标注研究与实现

李鹏

（中国科学技术信息研究所 北京 100038）

摘要：领域主题词表提供了某个领域或者行业某个视角下的关注的重点。利用领域主题词表进行标注能够揭示某个视角下文档的语义，而多个领域主题词表能够从多个视角联合揭示和挖掘文档的语义。文章通过将多个领域主题词表联合起来进行语义标注，提出了多领域视角下的知识标注方案。文章以皮肤病领域下文档标注为例进行了设计和实现。多领域视角下的知识标注为挖掘文档的知识提供了参考，并为进一步知识库的构建等奠定了基础。

关键词：叙词表，知识组织系统，知识标注，自动标注

Design and Implementation of Knowledge Tagging System on Multiple Domain Perspective

LI Peng

(Li Peng ISTIC, Beijing 100038)

Abstract: Thesaurus with domain concepts focuses on a perspective in a key field or industry. Semantic annotation with thesaurus can reveal the semantics of a document in some perspective, and document semantics can be revealed by tagging with concepts of multiple thesauri. This paper provides the design scheme of tagging system model with multiple thesauri. Taking the documents' tagging in the skin disease field for example, the design and implementation are completed. The system provides reference for mining document's knowledge in the different perspective, and could help the construction of knowledge base.

Keywords: Thesaurus, knowledge organization system, knowledge tagging, automatic labeling

作者简介：李鹏（1979- ），硕士，高级工程师。研究方向：智能信息处理。E-mail: lipeng_cn@istic.ac.cn。

1 引言

标注是一个相对主观和灵活的行为，是将文本资源中相对有意义的内容和特征进行描述和标记，并存取下来作为进一步加工的处理过程。标注不仅是信息过滤的必要组成部分，也是对原信息的精炼与提升，可以使检索更有效率，更为精准。全文索引的广泛采用，大大提高了文献的检索效率，但对于语义等反应知识的标注，全文索引难有用武之地。尤其是随着自动摘要、文本分析、知识挖掘的进一步需要，标注的作为愈发重要[1]。

国内外对于标注进行了一系列的研究。李素建等[2]通过建立了最大熵模型的特征集合进行分类标引，章成志[3]将抽词标注转换为序列标注问题，利用标注对象的特征提出基于条件随机场的自动抽词标注模型。朱嘉贤等[4]按树根结点、分支结点、叶子结点及资源信息元为粒度单位对 Web 资源进行组织管理进行多粒度语义标注研究。

本文通过知识组织系统尤其是主题词表进行知识标注，探讨多领域视角下的知识标注，并介绍以皮肤病为领域的知识标注研究及具体实现。

2 知识组织系统

2.1 知识组织系统简介

知识组织是以知识为对象的诸如整理、加工、表示、控制等一系列组织化过程及其方法。知识组织系统是用于进行知识组织的各类规范和方法的统称，是获取、利用知识的重要手段。知识组织系统是将文献、标引人员或用户的自然语言转换成规范化语言的一种术语控制工具；它是概括各门或某一学科领域并由语义相关、族性相关的术语组成的可以不断扩充的规范化的词表。早期的知识组织系统是词单（同义词环、权威文档、地名表等），后面逐渐发展为分类与大致归类（图书分类法、知识分类法），关联组织（实用分类法，语义网络，概念地图以及叙词表等。

2.2 范畴

范畴是指某指定知识领域按照特定方式对知识的分类。词表的范畴是按照树状目录方式组织的。一个范畴可包含多个下级（下位）范畴。本系统中一个范畴只能有一个上位范畴（顶级范畴除外）。如遇到一个范畴包括多个上级的情况（如一个词条：失眠，它的上级范畴是神经系统疾病(nsd)，再上级范畴是疾病(di)。再有一个词条：脑积水，它的上级范畴也是神经系统疾病(nsd)，再上级范畴是解剖(an)。在此例中神经系统疾病(nsd）不仅属于疫病范畴(di)也属于解剖范畴(an)，按照在两个顶级范畴下建立同名称、编码不同的范畴进行处理。如上例中：在疫病范畴下建立神经系统疾病(dinsd)、在解剖范畴下建立同名不同编码范畴神经系统疾病(annsd）。

范畴代码是一个范畴的索引值。在一部词表中，范畴代码是唯一的。

3 知识标注系统整体设计

本系统最底层为用户管理部分，用户在注册并通过审核后，可管理属于自己的文档、同时可管理术语自己的词表。词表维护完毕后，可进行标注工作。标注完成后，客户端系统可利用 API 对标注的文档进行解析，或访问服务器端的资源。

系统中各功能操作需要遵循本系统的技术规范，否则数据不能导入或者不能正常操作。系统共定义了：词表规范、标注算法规范、词表导入规范、文档导入导出规范（导入 txt 规范、导出 xml 格式规范、导出 html 规范、导入 xml 规范等）。

系统还提供日志管理、信息发布等辅助功能。

领域词表构建支持词表的管理、类型管理、概念管理、概念的导入、概念导出、概念浏览、用户词表维护、用户词表审核以及用户词表共享等功能。详细系统架构如图 1 所示。

图 1 系统架构图

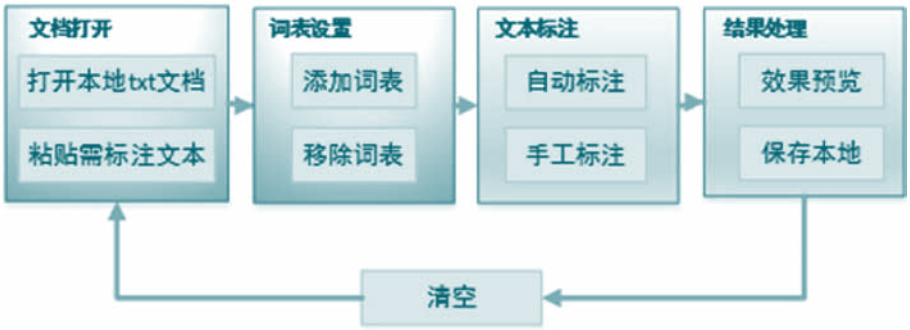

图 2 知识标注流程

知识标注流程包括文档管理、词表管理、文本标注以及结果处理四个过程。详细见图 2 所示。

3.1 文档管理

为了提供更快捷的文档管理，系统提供文档库。用户可建立多个文档库，存放文件。在建立文档库后，用户可在文档库中添加文档。针对网络平台，每个用户具有相对独立的文档管理系统，因此，系统设计相关的文档导入、导出规范。

在文档管理中：

1）文档库不能嵌套

2）每个用户的文档库不能重名。不同用户的文档库可以重名。

3）文档必须在文档库中

3.2 词表管理

词表管理包括词表的添加、修改、删除与相关分类管理、词条管理，以及待标注词表的设置等功能。

词条管理主要有以下几个方面的管理：添加新词条、编辑词条信息、删除词条、修改词条状态，导入词条改变词条类型。

分类描述是将词条划分到某个范畴内，便于浏览和检索。因此，词条分类描述的准确性显得十分必要。

系统允许用户选择用来标注的词表，如皮肤病，新能源汽车等专业性强的词表，也可以选择人员表、机构表、地名表等通用词的词表。

3.3 文本标注

文本标注，是指利用自动标注算法或者手工标注方式对原文进行标注。

自动标注，是指利用自动标注算法，对待标注文档利用包括的多部词表进行自动标注。如果文档原来标注过，则须先将原来的标注信息删除后，才能进行再次自动标注。手工标注，是指用户选择叙词表以及相应的类型属性，然后使用鼠标选择相关的待标注文本内容进行标签的创建、修改和删除。文本标注详细流程如图 3 所述。

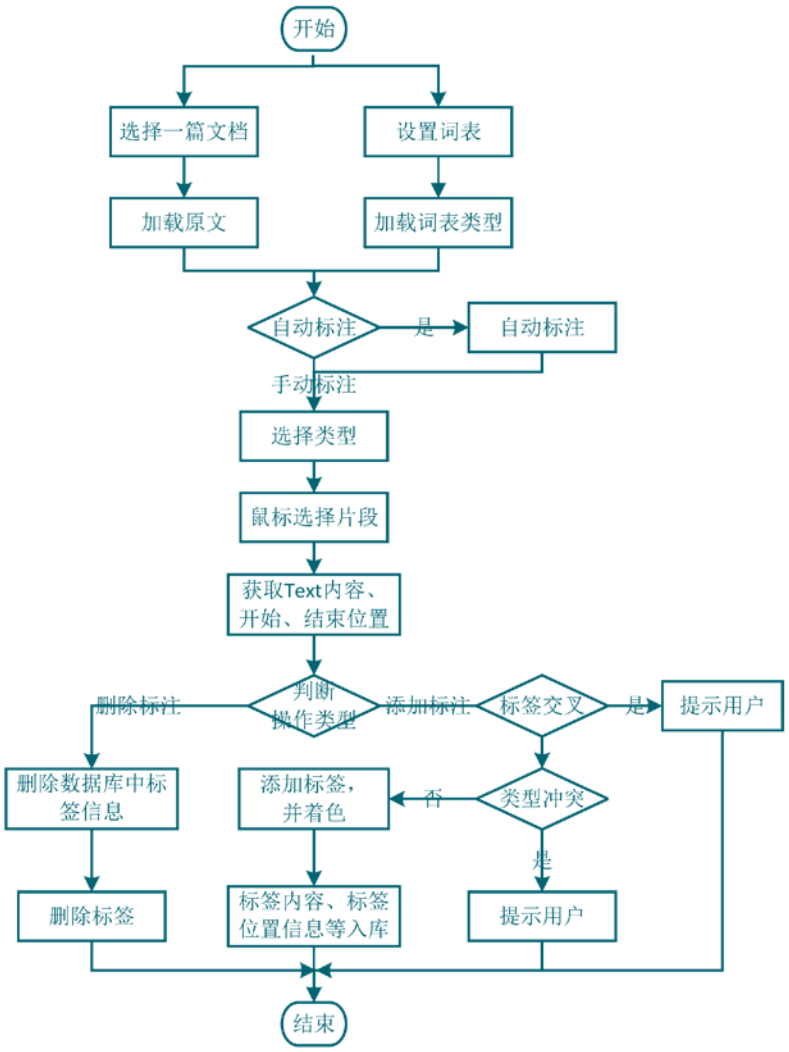

图 3 标注详细流程图

3.4 结果处理

结果处理，包括标注效果预览、保存至本地以及标注结果复制和清空等功能。系统支持多视角展示标签。多领域视角下的知识标注能够增加标注的维度，从而能够揭示原文的语义。结果处理中要求能够通过词表分组的方式显示标签结果，并进行按词表、按分类对标注结果进行统计和汇总。

4 知识标注系统实现

多领域视角下的知识标注系统采用基于JAVA的分层结构设计，使用SSH框架，JDK版本为1.6；数据库为MySQL 5.0。系统以皮肤病领域为例，介绍其实现。

多领域视角下的知识标注系统标注页面如图4所示。页面上部是功能菜单区，菜单内容有

图4 标注主页面

图5 范畴管理

文档管理、设置词表、标注和统计等功能，中间为文档及标注结果展示，右侧是根据标签的分类汇总。

词表范畴管理如图 5 所示。系统展示当前词表的范畴树型目录以及该目录对应的词条，并允许进行相应的修改、删除等操作。

多领域视角下的知识标注系统标注结果单独存储，通过记录标签所在的位置和标签内容存储到相应的数据库表中。标注后的预览效果如图 6 所示。

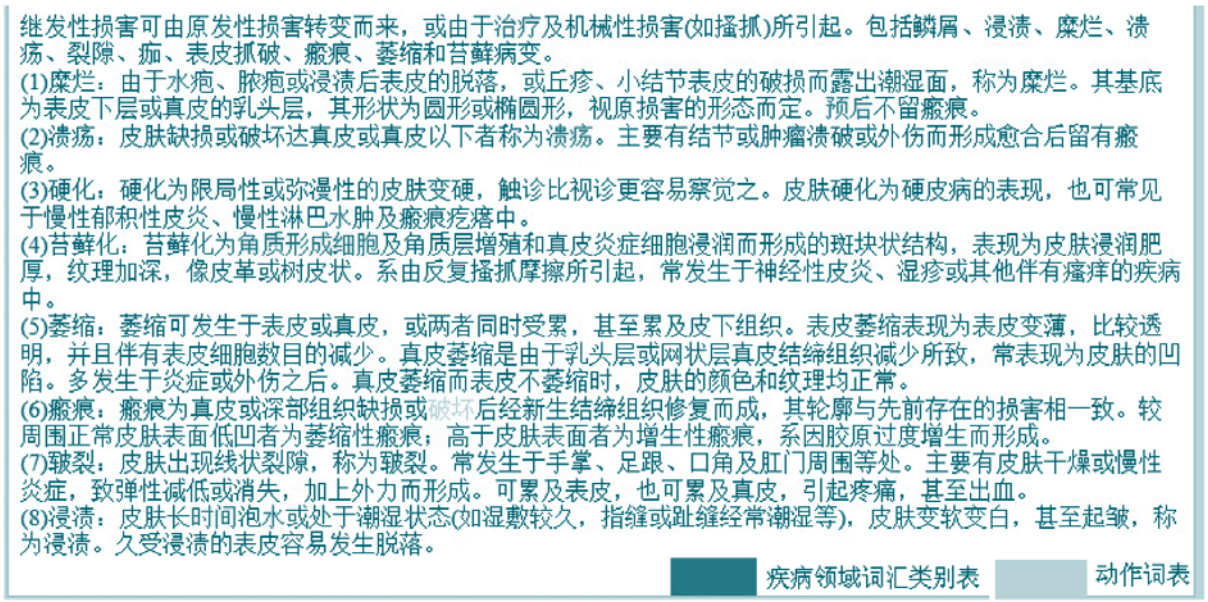

图 6 标注后的预览效果

5 结语

本文介绍了叙词表多领域视角下的知识标注系统设计和实现，包括多领域视角下的知识标注系统的整体设计，详细标注流程。多领域视角下的知识标注系统可以作为深度检索系统（知识标注系统检索能够支持深度检索系统，实现专业化检索，以提供精确的检索结果）、用户自定义标注引擎（原文上进行联合标注，形成组合的标注方案）等未来应用的基础。但是，多领域多视角的知识标注是一个复杂的问题，涉及各种标注算法的对比，标注指标等个性化化需求，这些是实际系统设计与实现中有待进一步深入的问题。

参考文献

[1] 周雪虹 . 制定文献编目著录细则若干问题的探讨 [J]. 高校图书馆工作 ,2003,23(6) :42-43.

[2] 李素建 , 王厚峰 . 关键词自动标注的最大熵模型应用研究 [J]. 计算机学报 ,2004,27(9):1192-1197.

[3] 章成志 . 基于条件随机场的自动标注模型研究 [J]. 中国图书馆学报 ,2008(5):89-94.

[4] 朱嘉贤 , 白伟华 , 李吉桂 .Web 资源的多粒度语义标注及其应用技术研究 [J]. 计算机科学 ,2011,38(8):83-87.