



杭州电子科技大学硕士学位论文

面向 STKOS 的概念映射与
关联算法研究及其实现

研 究 生： 王 波

指导教师： 姚英彪 副教授

2012 年 12 月

**Dissertation Submitted to Hangzhou Dianzi University
for the Degree of Master**

Oriented STKOS Concept Mapping and Correlation Algorithm and Its Implementation

Candidate: Wang Bo

Supervisor: Prof. Yao Yingbiao

December, 2012

杭州电子科技大学

学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。申请学位论文与资料若有不实之处，本人承担一切相关责任。

论文作者签名：王波 日期：2013 年 3 月 26 日

学位论文使用授权说明

本人完全了解杭州电子科技大学关于保留和使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属杭州电子科技大学。本人保证毕业离校后，发表论文或使用论文工作成果时署名单位仍然为杭州电子科技大学。学校有权保留送交论文的复印件，允许查阅和借阅论文；学校可以公布论文的全部或部分内容，可以允许采用影印、缩印或其它复制手段保存论文。（保密论文在解密后遵守此规定）

论文作者签名：王波 日期：2013 年 3 月 26 日

指导教师签名：郑伟 日期：2013 年 3 月 27 日

摘 要

近年来,科技文献服务的需求日益知识化,对科技文献标引的深度和准确度提出了更高的要求。其中,知识抽取已经成为制约知识服务系统开发建设的关键环节,需要集中大规模语义计算的优势力量,展开攻关。该研究的总体目标是以科技文献概念空间关联为主线,面向科技文献词表和分类体系,研究多维度的科技文献概念空间构建方法,提出用于概念空间关联的多渠道概念匹配算法,设计并开发面向科技文献概念空间建立于关联的原型工具。

本文的研究内容,主要包括理论及其系统实现两个方面。

理论方面,所用到的科技文献词表素材是各领域的叙词表,涉猎的分类体系包括“国际专利分类表”、“国际标准分类法”、“中国标准分类法”、“杜威十进制分类法”、“国际十进制分类法”和“中国图书馆分类法”。根据对主题词表和几种分类法进行结构上的异同分析,构建一个有能力从语义层面统一描述各种异质知识体系的公共数据模型。公共数据模型主要实现两个作用:一方面,可以将各种分类法和主题词表投影到相同的知识概念空间;另一方面,实现从不同分类体系中取词来进行比较。基于统一的概念空间(用公共数据模型描述),可以进一步开展面向科技文献知识体系的语义映射算法的研究。本文中所研究的算法是基于字符的概念匹配、基于结构的概念匹配和基于语义推理的概念匹配方法的集成,这样可以更好的提高机辅映射发现的查全率和查准率。

在开发方面,公共数据模型用 IBM 的 Ecore 模型来描述,通过建立的 Ecore 模型驱动生成一个面向科技文献知识体系概念空间构建的可视化图形化编辑插件,进而编码实现了分类法和主题词表的自动导入和可视化显示,最后开发了面向科技知识体系映射机辅发现的语义比较插件,其中编码实现了模型中各结点之间比较的关联算法,并将上述工具集成为一个具有数据导入、数据可视化及数据导出三种功能的 RCP 产品。

本文研究了基于科技知识组织体系(Science and Technology Knowledge Organizing System,简称 STKOS)的概念映射及其关联算法,所实现的意义有以下几点:首先,本文的研究,以统一的概念空间为研究基准,实现科技文献概念与分类的共享,从而从不同的维度为科技概念建立关联,有着明显的现实意义。其次,从语义层建立概念和分类体系之间的关联,能够提高整个概念空间的知识化程度,使关联映射的复用性大幅度提高,得到更接近自然语言思考和表达的语义关联网络。最后,综合应用基于字符的概念匹配方法(字符串比较,编辑距离比较和数据字典查询等),基于结构的概念匹配方法,和基于语义的推理的概念匹配方法,以便达到在不同的概念空间中最大程度的自动地、准确地发现概念关联。

关键字: 主题词表, 分类法, 语义, 概念空间, 映射, 关联

ABSTRACT

In recent years, the demands of scientific literature services are increasingly knowledge-based, and higher requirements on the depth and accuracy of scientific literature indexing are put forward. Among these technologies, knowledge extraction has become a key link which is restricting the development and construction of the knowledge services system, thus, focusing on the superior forces of large-scale semantic computing has been the scientific research methods. The overall objective focusing on scientific literature space association and oriented towards vocabularies and classification system for scientific literature is to study construction method of multi-dimensional concept space of scientific literature, put forward multi-channel matching algorithm for the concept of spatial association, and design and develop a prototype tool for scientific literature concept space associated.

The research content is mainly implemented by two aspects of the theory and development.

On the theoretical aspects, the vocabulary which we use in the scientific literature is the thesaurus in various fields, and the classification system which we use includes “International Patent Classification”, “International Classification for Standards”, “Chinese Classification for Standards”, “Dewey Decimal Classification”, “Universal Decimal Classification”, “Chinese Library Classification”. According to the analysis of the similarities and differences in the structure of thesauri and several classification, this paper builds a common data model which has the ability of unified description of heterogeneous knowledge systems from the semantic level, the common data model mainly plays two roles, on the one hand, it can make various taxonomies and thesauri project onto the same concept space of knowledge, on the other hand, fetch words from different classification systems to compare. Basing on unified concept space (described using the common data model), we can further research the semantic mapping algorithm for the scientific and technical literature of knowledge. The algorithm in this article was the integration of concept matching which is based on character, structure and semantic reasoning, so we can better improve the recall rate and the precision rate of the computer-aided mapping discovery.

On the development aspects, the Ecore model is used to describe the common data model based on Eclipse, and we can drive the built Ecore model to generate the visual graphics plug-ins orienting towards knowledge system concept space of scientific literature, thus achieve taxonomies and thesauri automatic import and visualization after encoding, and finally we develop the semantic comparison plug-ins facing the mapping computer-aided discovery of scientific and technological

knowledge system, which includes coding and achieving the comparison associated algorithm between the various nodes in the model. The integration of the above tools has been the RCP product with three functions of the data import, data visualization and data export.

In this paper, the concept mapping based on Science and Technology Knowledge Organizing System and its associated algorithms are studied. The significance of it is as follows: First, this article's research, in the unified concept space as the research base, can make sharing concept and classification of scientific and technological literature come true, and establish the related concepts of science and technology from a different dimension which is of obvious practical significance. Second, establishing the link between the concept and classification system from the semantic layer can improve the knowledge degree of the whole concept space and the reusability of associated map and get semantic link network closer to natural language of thinking and expression. Finally, the comprehensive application of concept matching method based on character which is including string comparisons, editing distance to compare and data dictionary query and so on, structural and semantic reasoning is to furthest find concept association in different concept space automatically and accurately.

Keywords: Thesaurus, Taxonomy, Semantics, Conceptual Space, Mapping, Associate

目 录

摘 要	I
ABSTRACT.....	II
第 1 章 绪 论	1
1.1 研究背景	1
1.2 国内外研究现状	2
1.2.1 概念空间及其关联	2
1.2.2 语义本体库构建	3
1.2.3 语义本体软件工具	4
1.2.4 分类法及其之间的映射和叙词表	5
1.3 研究意义与研究内容	6
1.4 论文的主要工作及结构安排	7
第 2 章 面向科技知识组织体系的公共数据模型	9
2.1 常用分类法与主题词表概述	9
2.1.1 国际专利分类表	9
2.1.2 国际标准分类法	11
2.1.3 中国标准分类法	12
2.1.4 杜威十进制分类法	13
2.1.5 国际十进制分类法	14
2.1.6 中国图书馆分类法	16
2.1.7 主题词表	16
2.2 常用文献分类法和主题词表的特点分析	18
2.3 面向科技知识组织体系的公共数据模型	20
2.4 本章小结	24
第 3 章 基于公共数据模型的自动映射算法	25
3.1 统一概念空间下的语义映射发现算法	25
3.2 基于语法的映射发现算法	26
3.3 基于结构的映射发现算法	26
3.4 基于语义推理的映射发现算法	33
3.5 本章小结	35
第 4 章 系统实现与实证结果	36

4.1 系统实现 36

 4.1.1 系统框架 36

 4.1.2 系统的技术体系 37

 4.1.3 系统的功能体系 38

 4.1.4 GMF 框架概述..... 38

4.2 实证结果 39

 4.2.1 实验数据 39

 4.2.2 数据处理 41

 4.2.3 实验结果 47

4.3 本章小结 53

第 5 章 总结与展望 54

致 谢 55

参考文献 57

附 录 61

第 1 章 绪论

1.1 研究背景

从古到今，人类记录与传播知识的手段在不断的发生着重大的变革。从远古时代的结绳记事和烽火传信，一直发展到现今的高速率的、超大容量的存储设备和简便快捷的无线电，以及光纤通信设备，人类历史发展至今，可以说，大量的知识是愈来愈容易被记录下来和传播，进而得以保存下来。知识的记录和传播保证了人类文明的延续和发展。特别是当今，Web 的出现，人类组织和利用知识的形式得到了非常大的丰富和改变。这种现象主要体现在两个方面：一方面，Web 作为信息的载体，将信息用多种表现形式体现了出来；另一方面，人类可以利用 Web 所提供的各类检索和查询服务来方便快捷的获得各类信息^[1]。

人类对 Web 极其密切的关注以及极度频繁的使用，在很大程度上促进了这项新技术的发展和完善。例如 Web 的页面从最初始的静态发展到如今的动态页面；对信息内容的描述开始从最初始的简单描述发展到至今对内容的结构及表示进行分离处理。Web 的出现也为人工智能的研究和应用提供了一个前所未有的平台。由于 Web 能够帮助用户访问一些存放在任何已经介入到互联网的数字内容，它给全球的公共信息服务、科研和商务带来了巨大的变革。现如今，互联网已经成为我们获取信息最重要的途径，不仅积累了大量的信息，而信息的数量也在以指数级的速度增长。在互联网发展到 20 世纪 90 年代中期，由于信息的大量增加，人类也很快的意识到在处理和利用这些海量的互联网信息这一方面欠缺有效的技术和手段。

为了解决这些问题，Web 的创始人 Tim Berners-Lee 等人在 2000 年提出了下一代互联网的概念——语义网（The Semantic Web）。语义网的基本思想就是利用元数据（Metadata）语言对 Web 信息资源方面的内容进行语义方面的描述，通过向数据添加一些机器可以理解的语言和启发式的信息，进而使计算机可以利用这些语义信息对信息资源的内容进行理解 and 处理，在此基础上，实现更高层次的且基于知识的智能应用，最后在一定程度上实现机器对信息的自动处理。

本体（Ontology）源于哲学概念，从哲学范畴来说，本体是对客观存在的一个系统的解释和说明，关心的是客观现实的抽象本质^[2]。在人工智能领域和知识工程领域，最早给出的本体定义是在文献^[3]中，Neches 等人将本体定义为“构成相关领域词汇的基本术语和关系，以及利用这些术语和关系构成的规定这些词汇外延的规则的定义”。Gruber^[4]则给出了本体的一个最流行的定义，即“本体是概念模型的明确的规范说明”。应用于指定领域知识的本体是领域本体，该类本体中包含了相关领域的概念及概念间的关系，强调的是领域的本质概念及概念之间的关联，领域本体的这些特点为其广泛应用提供了条件。

本体自从被提出以后，引起国内外众多科研人员的关注，并在诸多领域得到了广泛应用，如计算机、生物、化学、医学、农业、历史、军事等领域。在计算机领域，本体被应用于知

识工程、数字图书馆、信息检索、信息过滤、自然语言处理、数据集成及语义 Web 等许多方面，并取得了显著效果。

近年来，随着科技文献知识服务需求的日益增长，需要后台的科技文献标引的深度和准确程度不断增加。知识抽取工作已经成为制约知识服务系统开发建设的关键环节，需要集中大规模语义计算的优势力量，展开攻关，获得突破。目前，国内外对不同分类法的映射研究，在很大程度上，是关注于间接映射这一方法的实现，而忽略了对各词表中复杂的类目的研究，这种研究方法导致了词表的映射效果不是特别理想。那么，在信息数量以指数级速度增长，网络规模更是以惊人的速度增长的现在，如何能在大规模语义计算的前提下，更好的提高知识抽取方面效率，并且更好的提高概念空间在机辅映射发现的查全率和查准率，成为了一个关键问题。本文研究了面向 STKOS 的概念映射，以及对统一概念空间实现关联算法，可以很好的解决该问题。

1.2 国内外研究现状

1.2.1 概念空间及其关联

概念，是反映客观事物的一般的、本质的特征。对概念词的统计以及对概念之间的关系的研究，是图书馆情报学理论研究的核心问题之一。为了表达用概念描述的各种事物之间的亲疏程度，在美国康奈尔大学教授 Gerard Salton 的信息检索系统代数模型的基础上，在《图书情报数学》这一书中提出了“概念空间”（Concept Space）这个概念^[5]。概念及其之间的联系是空间的，多维的，任何一个复杂概念可以由简单概念组配而成。如果我们将每个简单概念用一条坐标轴来表示，那么任意一个复杂概念，像一篇文献，一个提问，都可以用这些坐标轴所代表的空间中的一个点来表示。这个概念已在美国情报基础理论和情报检索教材中得到广泛的承认和应用。

概念空间的定义对于文献学、情报学和图书馆学这三方面的理论研究有着非常重要的意义。在还没有提出概念空间这个概念之前，没有任何一种方法能够表达、区分和计算文献与文献之间以及提问与文献之间的相关程度，包括传统检索方法中的主题加权在内也不能解决这个问题。而概念空间的定义为解决这个问题提供了一种十分有效的途径，尽管这个解决方案不能做到完美无缺。也正是因为这样，在美国的情报基础理论和情报检索教材当中，都是把概念空间当作一个基本概念和一个基本理论提出的。而概念空间在实践中的作用，体现最明显的是在它的基础上提出的信息检索系统的代数模型。

随着概念空间维度的不断增大，计算量迅速上升，即对高维度概念空间的操作付出的代价（大量的内存、大批的人力、缓慢的检索速度）与所得到的效益（检索效率的少量提高）基本不成比例。出于对实用性的考虑，往往将高纬度的概念空间分解为多个较低维度概念空间组成的集合，再建立概念空间的关联来达到代替高维度概念空间计算的目的。因此，概念空间的关联是概念空间代数模型能够得以应用关键问题。

概念空间的关联包括三个步骤：（1）概念空间的构建；（2）概念空间的关联；（3）概念

空间关联的检验。目前,用语义本体表达概念空间,再通过语义本体的映射、合并,以及对齐等技术实现概念空间的关联是主流的实现方案。

1.2.2 语义本体库构建

近年来,随着计算机科学与自动化技术的发展,各领域对基于语义的人机交互和系统间交互的应用与需求日益广泛和深入。语义交互的基础和核心问题是本体库的构建,因此如何高效地构建完备、规范的领域本体库一直是近 10 年的热点问题。被普遍引用的本体的定义是由 Gruber 提出的“本体是概念模型的明确的规范说明”^[6]。通俗地讲,本体就是用来描述某一个领域或者更大范围之内的概念以及概念间的关系,使得这些概念和关系在共享的范围内具有用户群体与系统共同认可的、明确的、唯一的定义,且它是实现人与机器之间或者系统之间进行语义交流的数据基础。而本体的本质其实就是概念模型,表达的是概念与概念之间的关系。目前,本体已经被广泛应用于语义 Web、智能信息检索、系统间集成、企业业务交互,以及共建资源库等领域^[7],而这些都需要完备的,规范的综合了众多用户语义概念的本体库的支持。因此,怎样能够既省时又省力的构建大规模的语义本体,是一个非常有意义的研究方向。

现今为止,国外在该方向的研究非常的活跃,主要有两条分支:一个是本体学习(ontology learning)技术,它的目标就是利用知识获取技术、机器学习和统计等技术自动或者半自动的从已有的数据资源中获取本体,该方向研究的核心问题,即研究如何处理大量已有数据,从而得到期望的语义本体^[2];另一个是本体的自组织构建技术,它的目标是以网络环境和无线技术为支撑建立开放的多用户交互平台,通过多个用户的共同参与建立本体库,最后结合统计优化等技术得到面向多用户交互的语义本体,该方向研究的核心问题,即研究如何由大量用户共同构造新的适合用户交互的本体库^[8]。

语义本体库的构建方法按照它的模式可以划分为三种^[9]:

1) 自顶向下(Top-down)法^[10]。该类方法的特点是:首先,要建立宏观通用的资源概念体系框架,然后根据应用的需求,逐步地细化和扩展应用本体。这种方法可以控制本体开发的细节,但本体的性能对上层框架的依赖性较强。不合理的上层概念直接影响本体模型的性能。

2) 自底向上(Bottom-up)法^[11]。即以具体的应用为出发点,按照通用的本体语言来构建概念模型以及形式化表示,并通过实际应用来形成较为综合的上层概念。这种方法的缺点是容易产生过多的细节,导致本体概念的重复定义和不一致。

3) 核心扩展(Middele-out)法^[12]。这种方法主要是从概念集合中选取一组重要的对象概念,作为本体层次结构的中间层,然后对其进行细化或泛化。这样一方面保证本体的细节规模,一方面保证本体模型的整体稳定性。该方法是自顶向下法和自底向上法的折中。

本体学习技术大多数采用的本体构造法是自顶向下法,而本体的自组织构建技术通常采用的是自底向上的本体构造方法^[13]。采用本体学习技术构建的语义本体一般具有较好的一致

性和较少的数据冗余^[14]，但由于在这种模式下，本体模型的复杂度有限，且本体模型的可扩展性也很差，故这种模式比较适用于从大量结构化数据源（非结构化和半结构化数据的知识获取技术尚不成熟）生成具有层次结构的语义本体。采用本体的自组织构建技术生成的语义本体，理论上讲本体模型的复杂程度可以无限增加，而且本体模型的可扩展性很好^[15]，因此，这种模式较适用于建立面向大量用户交互的本体库，分布式环境是对该模式的有力支持，但这种模式必须解决好两个问题：①能够在满足综合本体库语义信息完整的条件下，尽量减少客户端本体编辑器中本体映射的工作量。②解决由于多用户参与造成的数据冗余和语义不一致。

在具体的应用中，本体的表示方式是多种多样的，其中主要的有 4 个大类^[16]：非形式化、半非形式化、半形式化及形式化语言。而在具体描述本体上，也是有多种方法的，可以用自然语言对本体进行描述，也可以用框架、语义网络或者逻辑语言等来描述本体，其中使用最广泛的方法有 3 种，分别是 Ontolingua^[17]、CycL^[18]和 Loom^[19]。

1.2.3 语义本体软件工具

在过去的十多年里，相继出现了许多本体构建工具，从早期的 Ontolingua^[20]，OntoSaurus^[21]，WebOnto^[22]到 Protégé^[23]，WebODE^[24]，OilEd^[25]，OntoEdit^[26]，以及 KAON^[27]等。本体编辑工具的功能日趋完善，大体可以分为四个阶段：①提供友好的命令或图形编辑界面，屏蔽本体描述语言编写的细节。用户可以把精力集中到本体内容的组织方面，而不必去了解本体描述语言的细节方面，从而避免了大量错误的发生，同时方便了本体的构建。②异质数据的导入接口的开发。早期的本体编辑工具提供的仅仅是本体编辑功能，且支持的仍旧是人工构建本体的方式。即使用这些本体编辑工具，用户依然需要逐个地输入和编辑每个概念的名字、约束、属性等一系列内容。如 Cyc^[28]和 Mikrokosmos^[29]等。开发异质数据导入接口，试图实现从企业信息系统中提供的数据格式直接或经过少量修改生成用于语义交互的本体，主要的数据输入格式有：简单文本、XML Scheme、RDB Script、EDI、MS Excel 等。通过对数据转换接口的开发，简化了本体学习的过程，增加了本体编辑工具的实用性，使得用户构建各自本体的效率比单纯的手工方式提高很多。③研究本体样本的合并技术。企业构建各自本体库的最终目的是实现与其他用户进行语义交互，这需要将众多用户建立的本体合并为一个综合语义本体。但由于用户自定义的领域本体在概念和结构等方面存在很大歧义，当前本体编辑器中的本体合并工具并不成功。④基于网络的本体映射工具。由于本体合并技术的不成功，本体映射工具出现在本体编辑器中，用户以本体编辑器作为客户端，通过网络与其他用户定义的本体建立映射，从而形成用于多边语义交互的本体库，如：MAFRA^[30]、OISs^[31]、Madhavan and colleagues^[32]、OntoMapO^[33]、IFF^[34]等项目都是这种框架。

当前，国内基本还停留在对本体编辑器本身的研究和开发，对于分布式环境下多用户共建本体库的控制和优化问题研究较少；国外一些相关项目包括了这方面的研究内容，如本项目申请人近期参与并负责的欧盟第六框架项目——STASIS 就是针对分布式环境下共建语义

本体库而设立的。但目前为止在较大范围内成功构建领域本体库的案例尚未出现,这主要也是受到了两个因素的影响:①用户在客户端本体编辑器中完成本体映射时操作反复,工作量大,这大大降低了多用户参与共建本体库的积极性。②本体库的数据结构不够合理,难免出现大量的数据冗余和语义不一致,在系统间语义交互时大大降低了语义信息的检索效率。

很多具有语义交互功能的开放交互平台都采用了本体的自组织构建技术^[35],但如何保证用户自组织共建的语义本体的一致性,以及如何减少共建的本体库中存在的数据冗余,一直是一个研究很少又不可回避的问题^[21]。针对这一问题,目前的做法是,在开放交互平台中开发一个或几个用于清理和验证语义本体的独立组件,按一定周期扫描整个本体库的内容,从而清除或报告本体库中的数据冗余或逻辑不一致。但是,面对实时变化的、大规模的语义本体时,这种检验方法的缺点显而易见,甚至是不可实现的^[2]。

1.2.4 分类法及其之间的映射和叙词表

随着社会的快速进步,信息呈现爆炸式的急剧增长。而伴随着信息的剧增,出现了大量的不同标准的分类体系及各个学科领域的主题词表。分类是标准化工作的基础,是统一和交流的前提。世界各国对标准文献的分类都十分的重视,可以说是几乎所有的先进的工业国家都有自己的分类体系,而我国也在1983年编制了一部符合我国国情的,且操作性较强的《中国标准文献分类法》,简称CCS。目前,大范围内使用的分类法有《国际专利分类表》(International Patent Classification,简称IPC)、《国际标准分类法》(International Classification for Standards,简称ICS)、《中国标准分类法》(Chinese Classification for Standards,简称CCS)、《国际十进制分类法》(Universal Decimal Classification,简称UDC)、《杜威十进制分类法》(Dewey Decimal Classification,简称DDC)和《中国图书馆分类法》(Chinese Library Classification,简称CLC)等。与各分类体系不同的是,主题词表主要是国内在使用。其中,对国内影响较大的叙词表(主题词表)有:我国的第一部主题词表《航空科技资料主题词表》,简称《航空词表》;我国第一部大型综合性分类主题一体化主题词表《中国分类主题词表》;作为国家重点工程“汉字信息处理系统”(748工程)的配套项目《汉语主题词表》;作为专门用于存储和检索社会科学情报的主题词表《社会科学主题词表》;规范社会科学术语概念的主题词表《社会科学检索词表》;用于存储检索档案文献的专用主题词表《中国档案主题词表》;用于存储检索教育专业文献的专业主题词表《教育主题词表》;用于标引、存储和检索石油工业科技情报的专用主题词表《石油汉语主题词表》,简称《石油词表》;用于存储以及检索环境科学文献资料的专用主题词表《环境科学主题词表》,以及各个学科领域的专用主题词表等。

相对于国内来讲,国外对于分类法之间的映射的研究开始的比较早,例如:在1987年,瑞典皇家图书馆就通过直接映射实现了《瑞典图书馆分类法》(Klassifikations system for svenska bibliotek)与DDC19之间的映射;OCLC研究机构通过采用同现映射和直接映射组合的方法,研究了《杜威十进制分类法》DDC和《美国国会图书馆标题表》^[36](Library of Congress Subject Headings,简称LCSH)的兼容操作^[37];Northwestern大学的研究项目LCSH和《医

学主题词表》(Medical Subject Headings, 简称 MeSH) 的映射^[38]; 而在国内, 分类法的互操作研究就起步比较晚了, 最初在国内分类法的直接映射的研究中取得了一定的成绩, 如: 在 1997 年, 山西图书馆研究出了“计算机文献标引对照系统”, 这一成果实现了《中图法》、《科图法》和《人大法》之间的相互转换; 在 2001 年, 黄筱玲对《科图法》和《中图法》进行了逐级比较的研究, 发现其中的 2/3 的类目可以建立一种对应关系^[39]。最近几年, 间接映射的研究一直在增多, 侯汉清等一些学者从 DDC 和《中图法》类目体系的比进行研究, 目的在于根据类目的相似值、类目之间对应概念因素的差额、相等的概念因素的类别制定了一系列详细的类目映射规则, 从而实现自动映射^[40-41]。

1.3 研究意义与研究内容

近年来, 随着科技文献知识服务需求的日益增长, 需要后台的科技文献标引的深度和准确程度不断增加。知识抽取工作已经成为制约知识服务系统开发建设的关键环节, 需要集中大规模语义计算的优势力量, 展开攻关, 获得突破。当前, 国内外对于不同分类法的映射研究, 在很大程度上, 只是关注于间接映射这种方法的实现, 且忽视了对各个词表中复杂的类目进行研究, 这种研究方法导致的结果就是词表的映射效果不是非常理想。

本课题的总体目标是以科技文献概念空间关联为核心, 基于科技文献词表和分类体系, 研究多维度的科技文献概念空间构建方法, 提出用于概念空间关联的多渠道概念匹配算法, 设计并开发面向科技文献概念空间建立与关联的原型工具。对于本课题的研究, 可以通过概念空间的视角研究, 从而实现科技文献概念与分类的共享, 这样, 便于从不同纬度为科技概念建立联系, 有着明显的意义。而从语义层建立概念和分类体系之间的关联, 能够提高整个概念空间的知识化程度, 使关联映射的复用性大幅度提高, 得到更接近自然语言思考和表达的语义关联网络。最后, 综合应用基于字符的概念匹配方法(字符串比较, 编辑距离比较和数据字典查询等), 基于结构的概念匹配方法, 和基于语义的推理的概念匹配方法, 以便达到在不同的概念空间中最大程度的自动、准确地发现概念关联。本文所研究的内容构架如图 1.1 所示:

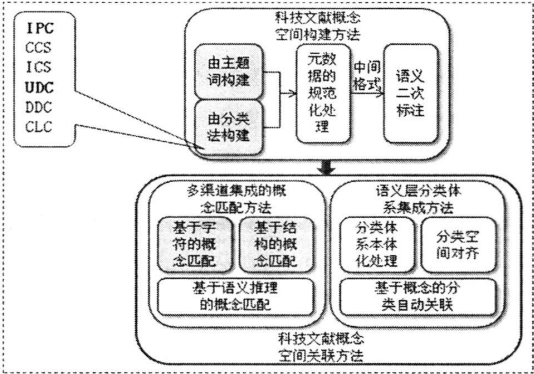


图 1.1 研究内容构架图

1.4 论文的主要工作及结构安排

本文所做的主要工作可分为理论和开发两个方面：

理论方面，对多种常用文献分类法和叙词表的结构异同进行比较分析，建立一个有能力从语义层面统一描述各种异质知识体系的公共数据模型，即将各种分类法和叙词表等知识体系投影到相同的知识概念空间；进而在统一的概念空间（用公共数据模型描述）中研究面向科技文献知识体系的语义映射算法。该算法是综合 3 个方面实现的，即基于字符的概念匹配、基于结构的概念匹配和基于语义推理的概念匹配方法的集成，这样有助于提高机辅映射发现的查全率和查准率。

开发方面，基于 Eclipse 的公共数据模型用 Ecore 模型描述，由建立的 Ecore 模型驱动生成面向科技文献知识体系概念空间构建的可视化图形插件，编码实现分类法和叙词表的自动导入和可视化显示，进而开发面向科技知识体系映射机辅发现的语义比较插件，其中编码实现模型中各结点之间比较的关联算法（其中涉及了 WordNet 和 jena 的二次开发），最后集成出一个具有数据导入、数据比较以及数据导出功能的 RCP 产品。

论文结构安排如图 1.2 所示：

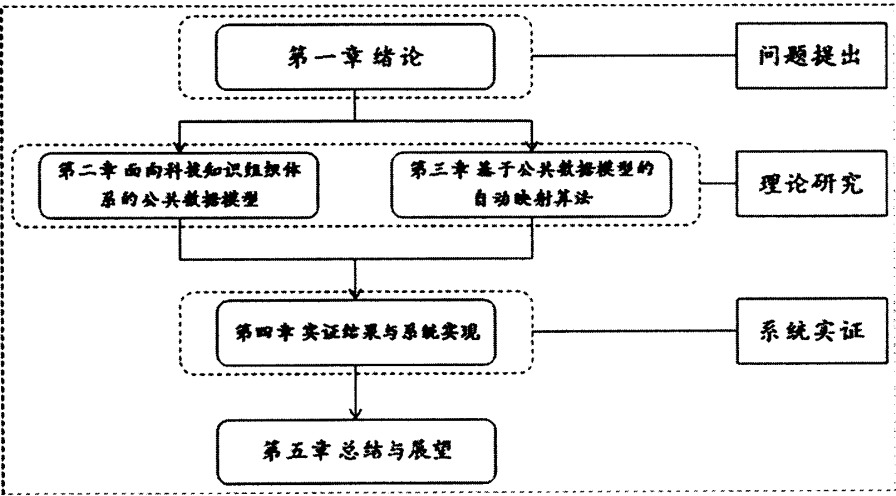


图 1.2 论文结构图

所有章节内容如下：

第一章绪论主要介绍了课题的研究背景、概念空间及其关联，语义本体库构建、语义本体软件工具和分类法及其之间的映射和叙词表的研究现状和完成本课题的思想及意义，并提出本文所要研究的问题。

第二章和第三章是对第一章所提出的问题的理论方面的研究。

其中，第二章首先详细介绍了常用文献分类法和叙词表的概念及结构体系，然后是对常用文献分类法和叙词表的特点分析，最后，在此基础上构建面向科技知识组织体系的公共数据模型。

第三章的内容主要是设计基于 CDM 的自动映射算法，由以下三个方面组成：1. 基于语法的映射发现算法；2. 基于结构的映射发现算法；3. 基于语义推理的映射发现算法。

第四章是对第一章所提出的问题的开发方面的研究，根据第二章和第三章的分析做出的实证结果以及系统实现，其中系统实现主要实现在开发方面，通过软件的开发和算法研究进行实现。

第五章，即最后一章，首先是对全文内容进行总结，并指出现有研究工作中的不足之处，最后是对接下来的工作进行展望。

第 2 章 面向科技知识组织体系的公共数据模型

2.1 常用分类法与主题词表概述

在分类法中，类目的内涵和外延是通过多个语义元素进行表述和限制的，包括分类号、类名、上下位类、注释、主题词等。不同的分类法编制类表的原则、划分类目的依据都是有所不同的，从而导致了分类法之间的映射存在着很大的困难^[41]。在本文的研究中，主要用到的几种分类法是：《国际专利分类表》（International Patent Classification，简称 IPC）、《国际标准分类法》（International Classification for Standards，简称 ICS）、《中国标准分类法》（Chinese Classification for Standards，简称 CCS）、《杜威十进制分类法》（Dewey Decimal Classification，简称 DDC）、《国际十进制分类法》（Universal Decimal Classification，简称 UDC）和《中国图书馆分类法》（Chinese Library Classification，简称 CLC）。文章中所要研究的主题词表在其结构上，表述及限制方面都是相同的，并没有因为其领域不同而在其他方面有所不同。下面是对不同的分类法以及主题词表的概述。

2.1.1 国际专利分类表

《国际专利分类表》（International Patent Classification，简称 IPC）^[42]是使各国专利文献获得统一分类的一种工具。IPC 的应用范围相当广泛，目前包括中国在内，已经有 50 多个国家在正式使用了。此表是根据 1971 年签署的“关于国际专利分类的斯特拉斯堡协定”编制的。且《国际专利分类表》是目前唯一的国际通用的专利文献分类和检索工具。经过多年的发展，国际专利分类表已经是非常成熟了。《国际专利分类表》的第一版是根据 1954 年所签订的“关于国际发明专利分类欧洲协定”的规定而进行编制，并在 1968 年 9 月 1 日正式生效。在之后的发展中，为了能够更好的适应全球科技信息的发展，不断地改善分类体系和技术，从而进行了多次的修订。为了能够更好的适应计算机和通讯网络等新技术的高速发展，将 IPC 称作专利文献检索的有效工具，IPC 联盟大会的成员国以及世界知识产权组织（WIPO）在 1999 到 2005 年对《国际专利分类表》进行了改良，将第 8 版 IPC 分成基本版本和高级版本两级结构。第 8 版 IPC 的基本版本约有 20000 条左右，其中包括部、大类、小类、主组和某些技术领域的个别的多点组的分组，而高级版本约有 70000 条左右，其中包括了基本版本及对基本版本进一步细化的类目。为了更加适应社会的发展，第 8 版的 IPC 也将比以往修订的更加频繁，其中基本版会每 3 年修订一次，高级版将每 3 个月进行修订一次。目前，我们所使用的《国际专利分类表》为第 8 版的。为了便于公众化专利文献著录项目，同时也为了便于计算机管理，巴黎联盟专利局间情报检索国际合作委员会（ICIREPAT）为专利文献著录项目制定了统一的代码，简称 INID，国际专利分类号在专利文献上著录的 INID 代码为“(51)”，

标注成 Int.cl.m××××。其中 Int.cl. 表示国际专利分类，m 表示修订的版次。如 Int.cl.4H05KL/00 表示第四版 IPC 的一个类号；而 Int.cl. A01B1/02 表示第一版 IPC 的一个类号。

在专利的分类中，分类原则有功能分类原则和应用分类原则两种。功能分类原则是从发明的功能或内在性质这两个方面出发，从而将发明进行分类的，同时在分类表中设定了对应的分类位置，将这一分类位置定义为功能分类位置。这种分类原则在将发明进行分类时，这项被分类的发明在技术上是不会受到其使用范围影响的。应用分类原则是从发明物的特殊目的或者应用的范围出发，对其进行分类的，同时在分类表中设置了与其相互对应的分类位置，将这一分类位置称为应用分类位置。《国际专利分类表》的分类原则正是结合功能分类和应用分类这两种原则，同时也是世界各国的专利局在一个多世纪的分类工作经验的基础之上进行总结，最后在通过专家进行详细且周密的规划之后，所制定出来的。

IPC 分类的条理性非常强，同时在专业类目的安排方面基本上也能够反映出当代技术的全部内容。IPC 所采取的分类标准为等级分类，是把技术内容按照部（Section），分部（Subsection），类（Class），主组（Main group）和分组（Subgroup）进行逐级分类，最终生成一个相对比较完整的分类体系。IPC 把全球已有的专利技术领域通过 8 个部进行了总体上的分类，其中各部都涵盖了相当广泛的技术内容，这 8 个部分别用 A-H 八个大写字母来表示。八大部分的技术范围如表 2.1 所示：

表 2.1 IPC 的八大部的技术范围

部	专利技术领域
A 部	人类生活需要 (Human Necessities)
B 部	作业、运输 (Operations; Transporting)
C 部	化学、冶金 (Chemistry and Metallurgy)
D 部	纺织、造纸 (Textiles and Paper)
E 部	固定建筑物 (Fixed Construction)
F 部	机械工程、照明、加热、武器、爆破 (Mechanical Engineering; Lighting; Weapons; Blasting)
G 部	物理 (Physics)
H 部	电学 (Electricity)

在八大部的下面设置了分部，分部只有标题，是不用分类符号表示的。例如：A 部下面设置有农业；食品、烟草；个人或家用物品；保健、救生、娱乐四个分部。故“部”之后的细分类目即为大类了，大类的类号是由其所在的部的编号加上两个数字组成，如 A22，B08 等。之后在大类的下面还设有很多小类，而小类的类号是由其所在的大类的类号在后面加上一个大写字母来表示的，但是其中有六个字母是不能用的，分别是 A，E，I，O，U 和 X，例如 B05B、D01F 和 G06K 等。从小类再向下细分，还设有主组和分组，其中主组号是由其

所在的小类的类号在后面加上 1-3 个阿拉伯数字, 其后再加上/00 来表示, 如 A01B1/00、G05G7/00 等。如果斜线的后面用的是 2 到 5 个数字表示, 而且也不是 00, 如 A01B1/02, G05G5/06 等, 则这个类号就是分组的类号。分组是主组的展开项目。并且在分组类号后面还可以加一些圆点, 而圆点的个数则表示各个分组之间的从属关系。例如: A01B3/00: 装有固定式犁铧的犁, 3/36 · : 拖拉机悬挂犁, 3/40 · · : 双向犁, 3/42 · · · : 翻转犁壁双向犁, 3/426 · · · · : 带有由两个或两个以上构件组成的悬挂架的。斜线后面的数字在分类表中不是表示任意进一步的细分类的等级分类, 即 3/426 并不是 3/42 的细分类, 而 3/40 后的两个圆点决定了它是 3/36 的细分类, 所表达的意思就其是离它的位置最近且少了一个圆点的类号的细分类。下面说明了一个完整的 IPC 分类号, 如表 2.2 所示:

表 2.2 完整的 IPC 分类号

分类号	A	47	L	5/00			至主组级
类目	部	类	小类	主组			
分类号	A	47	L	5	/	06	至分组级
类目	部	类	小类	主组		分组	

IPC 在科技查新的过程中必须要确认 IPC 分类号, 其方法有很多, 常用的主要有三种: (1) 直接查找法; (2) 《关键词索引》法; (3) 检索结果统计分析法^[42]。直接查找法是确定将要检测课题的 IPC 分类号的最直接、最基础的方法。在查找分类号时既可以按照“部、大类、小类、主组、分组”的顺序逐级查找, 也可以利用网站所提供的检索功能, 根据主题词来直接查询分类号。《关键词索引》法是 IPC 和《关键词索引》结合在一起使用的, 这个索引是独立出版的, 且版次与《专利分类表》的版次是相互对应的。而检索结果统计分析法则是依据一些网站对已经存在的检索结果进行统计分析, 该方法是快速补全相关的分类号非常有效的一种方法。

2.1.2 国际标准分类法

《国际标准分类法》, 英文名称是 International Classification for Standards, 简称 ICS, 是由国际标准化组织制定的标准文献分类法。其主要的范围是国际标准、区域标准和国家标准以及相关标准化文献的分类、编目、订购和建库, 进而使国际标准、区域标准、国家标准以及其他标准化文献在全球范围内得到了更好的传播。《国际标准分类法》的形成是非常有必要的。由于分类的重要性以及各种优点, 所以标准文献的分类在世界很多国家都非常受重视, 最终形成了很多发达的工业国家都有了各自的分类法。但是随着标准化事业的发展, 各行其是的分类法就不能再满足世界标准化信息交流的需求, 特别是在世界贸易不断涉及到的情况下, 统一标准的分类法对于国际交流与合作就愈加显得迫切和重要了。世界贸易组织 (WTO) 委托国际标准化组织 (ISO) 负责 TBT 协议 (贸易技术壁垒协议) 中有关标准通告事宜的具体实施, 在其实施过程中就必须使用《国际标准分类法》。故在 1991 年, ISO 组织就完成了 ICS 分类法的指定, 并在 1994 年在其颁布的标准汇总采用 ICS 分类号。由

于《国际标准分类法》具有国际性，所以其应用是相当广泛的，它的应用范围有：（1）ICS 可以作为国际、区域性和国家标准以及其他标准文献的目录结构，并且作为国际、区域性和国家标准的长期订单系统的基础，也可以应用到数据库和图书馆中的标准以及标准文献的分类。（2）ICS 也可以当作信息数据的排序工具，例如目录、选择清单、参考文献、磁介质和光介质上的数据库等。

《国际标准分类法》所采用的是数字分类法，其他使用数字分类法的如 DIN。数字分类法与字母分类法（例如 ASTM）和字母与数字混合分类法（例如 JIS、GB）这两种分类法相比，其具有扩充方便、方便计算机管理等一系列的优点，而且也没有文种障碍，这些优点都便于进行交流与推广。ICS 类目的设置有以下 4 个特点：（1）有些类目设置比较新颖，如：35，办公室机械设备（*包括开放系统互连（OSI）和计算机图形技术）。（2）受欧洲传统分类思想影响较大，由于 ICS 中某些类目的设置主要是围绕德国分类法的思想来进行的。例如：39，精密机械、珠宝，而在此类目下却只有手表和珠宝，其他的仪表、仪器无法写入类中，所有不具有合理性。（3）有些类目分得过于细致；有些类目定界又不太明显；有些类目设置比较陈旧，难以入类等各种问题。（4）第一级中的 41 个大类的编码全是单数，没有双数。ICS 的分类原则有以下几个特点：（1）根据标准文献主题内容所在的学科和专业进行分类，较为概括的内容分为第一级，较为具体的内容分为第二级，更加具体的则分为第三级；

（2）术语标准和图形符号标准可以优先分到术语和图形符号类中，接着按照标准化对象所在的专业分到专业类；（3）一个标准可以根据其主题或者主题侧面分到多个不同的二级类或三级类当中；（4）一个标准所涉及到的主题范围涵盖了一个二级类以下的所有的三级类目，但标准应分到此二级类；（5）为了便于计算机进行检索，涉及到某个二级类的全部主题的标准，可以用“×× . ××× . 00”来表示，这样在检索的时候用“×× . ××× . 00”提问，将得到“×× . ×××”下的综合标准；而用“×× . ×××”提问将得到“×× . ×××”下以及“×× . ××× . 01”到“×× . ××× . 99”所有类下的全部标准；（6）用户可自行扩充类，多个分类号中间以分号隔开。

从 ICS 的体系结构方面来看，ICS 所采取的分类法是层累制分类法，它是由三级类目组成的。第一级有 41 个大类，例如：材料储运设备、货物的包装和调运、服装工业等。每一个大类都是由两位阿拉伯数字表示的，例如：73 就是“采矿和挖掘”这一个大类的类号。全部的一级类目又分成 387 个二级类目。二级类目的类号是由一级类目的类号、一个圆点和一个三位数的阿拉伯数字三个部分组成，例如：73.120，矿产加工设备。二级类目的下面又细分为三级类目，总共有 789 个，三级类目的类号由二级类目的类号、一个圆点和一个两位数的阿拉伯数字组成，例如：73.100.30，表示“钻井和挖掘设备”。

2.1.3 中国标准分类法

《中国标准分类法》（Chinese Classification for Standards，简称 CCS）又叫《中国标准文

献分类法》，简称《中标分类》。《中国标准文献分类法》的类目设置标准是以专业领域划分为主，适当的结合科学分类。序列采取的是从总到分，从一般到具体的逻辑系统。《中国标准分类法》所采取的是更为简单的二级分类法，一级主类的设置标准主要是从专业领域划分的，二级类目的设置标准是采取非严格的等级制度的列类方法进行划分的。其中一级分类是有 24 个大类组成，每一个大类都有 100 个二级类目；一级分类的分类号是由单独一个拉丁字母组成的，二级分类的分类号是由两个两位数的数字组成的。比如：一级分类“G 化工”；二级分类“70/79 化学助剂、表面活性剂、催化剂、水处理剂”，其中 70 是“化学助剂基础标准与通用方法”、71 是“化学助剂”、72 是“表面活性基础标准和通用方法”等。

《中国标准分类法》中有通用标准和专用标准两种划分。所谓的“通用标准”，就是指两个或者两个以上的专业共同使用的标准，而“专用标准”是指某一专业特殊用途的标准。在中国标准文献分类法中对这两类标准的使用是采取“通用标准”相对来说比较集中，“专用标准”适当分散为原则的，例如：“医药”这一标准是入“C：医药、卫生、劳动保护”的，而“兽用药”这一标准是入“B：农业、林业”的。但是对于各类有关基本建设、环境保护、金属与非金属材料等方面的标准文献采取的是相对集中列类的方法的，例如：“园林绿化与市容卫生、水利，水电工程综合、建筑工程施工机械”等是入“P：工程建设类”等。

2.1.4 杜威十进制分类法

《杜威十进制分类法》(Dewey Decimal Classification, 简称 DDC 或 DC) 受美国圣路易斯市图书馆哈利斯分类法的影响，根据 17 世纪英国哲学家培根关于知识分类的思想，将人类知识分为记忆（历史）、想象（文艺）和理性（哲学，即科学）三大部分，并且将其进行倒置排放，展开分为 10 个大类别。美国 M. 杜威编制的综合性等级列举式分类法，其分为详细和简单两种版本，简版的内容基本上是详细版本的十分之一。详细的版本出现于 1876 年，为其取名叫做《图书馆图书小册子排架及编目适用的分类法和主题索引》，在 1951 年，把第 15 版的名称改作《杜威十进制分类法》，接着在 1989 年的时候，出版了第 20 版。而简版主要应用于中小型图书馆，到目前为止，已经出版到底 12 版了。《杜威十进制分类法》是世界现代文献分类法史上的一个十分重要的里程碑。它是世界上在现行文献分类法中最流行且最有影响力的一部分分类法，现在已经出版了有 30 几种语言了，同时世界上有 135 个国家与地区的图书馆采用该分类法。在 1910 年，我国就已经对《杜威十进制分类法》开始进行翻译和介绍了，并陆续的出现了一批仿照《杜威十进制分类法》体系，同时符合中国特点的分类法。

第 20 版的《杜威十进制分类法》一共分 4 卷，除了主表之外，还包括附表，即辅助表、索引以及使用手册等。附表包括：(1) 标准复分表；(2) 地区复分表；(3) 文学复分表；(4) 语言复分表；(5) 人种、种族、民族复分表；(6) 语种复分表；(7) 人物复分表。

《杜威十进制分类法》的分类标准也是按照等级制来进行分类的，它采用阿拉伯数字作为标记符号，并采用小数制（即十进制）的层累标记制。以三位数（000-999）形成前三级的等级结构。在这三位数当中，凡是带有“0”的号码都表示总论性类目，也就是说，后两位数

字为“0”的三位数号码表示一级类（大类），最后一位数字为“0”的号码则表示二级类。凡是最后一位数字不是“0”的三位数字，其号码均属于三级类。而凡是在三位数之后再展开的号码，都必须在三位数字后面加上一个小数点来隔开。例如：“600”表示的是“应用科学”；“630”表示“农业”；“631”表示“农业经营”；接着是“631”之后的细分类目“631.5”表示的是“作物栽培”。《杜威十进制分类法》是分为 10 个大类的，也就是一级类目。一级类目的目录体系如表 2.3 所示：

表 2.3 一级类目的目录体系

分类号	类目内容
000	总论
100	哲学
200	宗教
300	社会科学
400	语言
500	自然科学和数学
600	技术（应用科学）
700	艺术、美术和装饰艺术
800	文学
900	地理、历史及辅助学科

其中，《杜威十进制分类法》的这种分类方法有着不少的优点：（1）体系结构完整、严谨，类目详尽，层次清楚，易于理解，便于按学科进行检索；（2）采取非常严格的层累标记制，类号等级非常明确，具有比较强的助记性，但有的时候号码又太过于冗长；（3）首创仿分、复分等具有组配性质的编号法；（4）最先为分类表编配详细的相关索引，提供一条简洁方便的字顺检索的方法；（5）拥有实力雄厚的管理机构，定期修订，使分类法不断得到更新。

2.1.5 国际十进制分类法

《国际十进制分类法》（Universal Decimal Classification，简称 UDC），又称为《通用十进制分类法》。它是属于国际通用的多文种的综合性文献分类法，这部分类法至少拥有 39 中不同语言的版本，应用范围非常之广泛，也延伸到了很多以英语作为其官方语言或者是工作语言的国家当中，甚至对其他的不是以英语为主的国家也有比较大的影响。《国际十进制分类法》是由比利时人 P.-M.-G.奥特莱和 H.-M.拉封丹在《杜威十进制分类法》第 6 版的基础上编制的。在 1936 年，英国的国家标准局决定采用《国际十进制分类法》作为国家标准，并且将其翻译成英文本，然后以分册的形式进行出版，最终在 70 年代中期将其全部出齐，称之为第 4 版。

UDC 可以说是包含了人类所知道的所有知识体系，其中自然科学与技术类在 UDC 中所占的比例要大于艺术类以及社会科学类。《国际十进制分类法》是有主表和辅助表两部分组成，其中主表涵盖的大类有 10 个，辅助表中则包含了很多的辅助符号，其作用是连接多个概念，

最终目的是为了体现出概念与概念之间的一种或多种关系。《国际十进制分类法》的类号一般情况下都是用的阿拉伯数字，特殊点的就有一些数字和字母符号。因为《国际十进制分类法》采用的是十进制进行分类，所以在类目的扩充上是可以按照原来的标识符体系的。按照《国际十进制分类法》的详简程度来看，它的版本就是有口袋版（Pocket Edition）、以“主参考文档”（Master Reference File，简称 MRF）为其内容参考的节版（Medium Edition）和足版（Full Edition）三种。其中口袋版大约有 4100 个类目，它是为了教学而出版的。而《国际十进制分类法》的应用范围也是相当之广泛的，它可用在将各种不同形式的文件以及一些信息资源给组织和记录下来，也可以将不同内容的图片进行分类放置，例如：画册、地图、照片等。到目前为止，《国际十进制分类法》的主要应用领域是在各个国家的图书馆以及博物馆、学科书目以及书目类别的数据库，还有以信息服务为主的机构和网络上。

现在，《国际十进制分类法》一直在为了适应时代的发展和科技技术的进步而不断地发展，而发展的最为突出的就是在计算机化与网路化这方面了。在最近的几年中，《国际十进制分类法》在组织和检索电子数据这两个方向上的价值都已经被权威给证实了。主参考文档（MRF）已被授权使用到来自国际十进制分类法联盟（Universal Decimal Classification Consortium，简称 UDCC）的会员国地区或网络上。在 2005 年，UDCC 编辑委员会出版了最全新的全版《国际十进制分类法》，而这个新版本的修订是建立在 MRF 的基础之上的。它包括了 65000 个类目，是以两个卷的形式来出版的。而《国际十进制分类法》的最为重要的部分，即 MRF 的数据库，也经历了很多次的修订，在 2005 年进行了最新的修订，这其中包括了有大约 67000 个资料文档，和《国际十进制分类法》的类目数是一致的，大小有 13000kb，它的检索文档则占了有 5900kb 的大小。

由于 UDC 在等级列举制的基础之上采取了多种符号进行组配，因而发展成为了一部结合等级列举与组配相的混合型分类法。UDC 采用的是单纯的阿拉伯数字作为标记符号的。它用一位数（0-9）标记一级类，作为一级类号；两位数（00-99）标记二级类，作为二级类号；三位数（000-999）标记三级类，作为三级类号。以下每扩展（细分）一级，就添加一位数。每三位数字后面加上一个小数点。下面一个例子如表 2.4 所示：

表 2.4 UDC 的分类形式

类号	分类内容
6	应用科学、医学、技术
62	工程、技术（总论）
621	机械工程总论、核技术、电气工程、机械制造
621.3	电工程、电技术、电气工程
621.39	电信技术
621.396	无线电通讯设备和方法
621.396.9	雷达

2.1.6 中国图书馆分类法

《中国图书馆分类法》(Chinese Library Classification, 简称 CLC), 原名《中国图书馆图书分类法》, 简称《中图法》是我国建国以后编制出版的一部具有代表性的大型综合性的分类法。它在我国主要应用于图书馆和情报单位。《中国图书馆分类法》是由中国图书馆图书分类法编辑委员会编写的。《中国图书馆分类法》的第一版是在 1975 年出版的, 且在 1999 年出版了第四版。修订后的《中国图书馆分类法》第四版增加了类分资料的类目, 并与类分图书的类目以“+”的标识来进行区分, 因此正式改名为《中国图书馆分类法》, 简称不变。《中国图书馆分类法》的发展过程有将近 30 年的时间。在 1971 年, 由北京图书馆提出了倡议, 文化部图书出版局组织全国 36 个单位赞助而组成《中国图书馆分类法》的编辑组, 来集中编辑《中国图书馆分类法》, 经过长达 2 年的奋斗, 终于在 1973 年完成了初稿, 并以《试用本》的形式来出版。在 1975 年, 出版了第一版。1979 年, 《中国图书馆分类法》编辑组在长沙召开了《中国图书馆分类法》的修订工作会议, 并成立了《中国图书馆分类法》编辑委员会, 后又在太原召开了编委会议, 接替《中国图书馆分类法》编辑组的工作。1980 年修订出版了第二版, 1989 年出版了第三版, 最终在 1999 年修订出版了第四版。《中国图书馆分类法》第四版全面补充新主题、扩充类目体系, 使得分类法跟上了科学技术发展的脚步。同时规范类目, 完善参考系统、注释系统, 调整类目体系, 增修复分表, 明显的加强了类目的扩容性和分类的准确性。

在《中国图书馆分类法》的体系结构中, 是按照“五分”法进行分类的, 将所有的文献知识所包含的内容划分为五个基本部类, 这五个部类分别是: 马列主义、毛泽东思想, 哲学, 社会科学, 自然科学, 综合性图书。之后还有 22 个基本大类, 《中国图书馆分类法》使用字母和数字相结合的混合号码, 基本采用层累制。《中国图书馆分类法》除了分出来的 22 个基本大类之外, 还分了附表。《中国图书馆分类法》总共有 8 个通用复分表, 它们分别是: (1) 总论复分表; (2) 世界地区表; (3) 中国地区表; (4) 国际时代表; (5) 中国时代表; (6) 世界各族与民族表; (7) 中国民族表; (8) 通用时间、地点表。

《中国图书馆分类法》有着比较多的特点: (1) 充分反映了新学科、新事物、新技术、新主题; (2) 注意解决图书的集中与分散问题; (3) 妥善分排类目组织中一般和个别的问题; (4) 考虑到分类工作的特点, 采用了一些特殊方法; (5) 采用了字母和数字相结合的混合号码; (6) 采用了可选择的可伸缩的配号方法; (7) 充分考虑和应用计算机技术。基于以上《中国图书馆分类法》所具有的特点, 故《中国图书馆分类法》的使用范围是相当广泛的, 包括国家图书馆, 省、市、自治区、县图书馆, 甚至于其他系统的图书馆, 如: 高校、科研等的图书馆。

2.1.7 主题词表

主题词表, 又被称之为叙词表或者检索词典。主题词表的一种概括了某一学科领域, 以规范化的、受控的、动态的主题词(叙词)作为基本成分和以参照系统来说明词间关系, 作

用于标引、存储以及检索文献的词典。

从主题词表的宏观结构来看,主题词表是由一个主表和若干个附表和辅表所组成的。主表(也称为字顺表)是将全部的主题词按字顺排列,并添加一些必要的标注和显示词间的等同、等级或者相关关系的参照项,是主题词表的主体结构。主表的功能是供标引员和用户按照字顺直接对主题词进行查找,并通过参照系统判定词汇的含义,进行相关词的查检。主表是标引和检索的主要依据。附表是将众多专有的主题词款目按照字顺排列的一览表,为专有主题词索引,如地区索引、机构索引、人名索引、产品索引等,是从主表中分离出来的专用词汇表,是主表的重要组成部分。其中专有名词是指表达某一单一事物对象的名称。常用的专有名词包括人名、地名、机构名、产品型号名等,其数量大,范围广。附表的作用有三点:(1)缩小主表的篇幅;(2)分类显示作用;(3)有利于管理或者增补。附表的款目结构和排列方式和字顺表是相近的,一般不收入到范畴索引,且无范畴号。而且只有在有必要的时候才会编制附表。

辅表一般有词族索引(族系表)、范畴索引(分类表)、轮排索引(轮排表)、双语种对照索引等,它们都是主题词表的辅助部分。词族索引也称等级索引,是利用概念成族原理(根据概念的等级关系)将有关主题词汇集在一起成为一族,构成一个从泛指主题词到专指主题词的等级系统,从而可从一族中外延最广的主题词(族首词)出发,找到一系列同族的主题词,并且可以明确它们之间的层层隶属关系。词族索引是机器检索时实现自动扩检、满足族性检索要求的重要手段。范畴索引又被称之为分类索引,是将主题词按其概念所在学科或范畴分成多个大类,之后在大类之下再分成多个小类,接着在小类的下面则将主题词按照字顺进行排列,最终形成一个像体系分类表一样的概念分类系统,以便于从学科或专业这一视角来选用主题词。按词汇所属的学科范畴编制的概略分类系统,又称分类索引、范畴表。便以学科分类角度来查找与某一范畴内容有关的主题词。轮排索引是利用字面成族原理,将含有相同单词的词组主题词汇集在一起,排列在该单词之下,可从它出发查到任何含有该单词的词组主题词。一个词组主题词由几个单词构成便可轮排几次,能在几处查到它。轮排索引便于查找,并在某种程度上起族性检索作用。双语种对照索引可提供从另一种语言字顺入手查词的途径,有助于在标引外文文献时选准主题词或利用本国主题词表查阅外国的检索工具。双语种对照索引还是一种将全部叙词、非叙词与相应外语译名相互对应,按外文译名字顺排列的词汇表,是通过外文译名使用词表的辅助工具。

从主题词表的微观结构来看,主题词表就是具有不同的主题词款目。其中包括款目词项、标记项、注释项和参照项。款目词项是有款目词、汉语拼音以及外文译名组成。标记项就是款目词的序号以及所属范畴号。注释项包括有含义注释、用法注释和历史注释。参照项有代、属、分、参和用,且还有特殊的一项“族”。在几乎所有的主题词表中,词汇之间都是有三种词间关系的。其中汉语主题词表的三种词间关系如表 2.4 所示:

表 2.4 汉语汉语主题词表的三种词间关系

词间关系	符号	含义	作用	备注
等同关系	Y	用	指引相应的正式主题词	正式主题词
	D	代	指引相应的非正式主题词	非正式主题词
等级关系	F	分	指引所含的下位主题词	下位词
	S	属	指引所从属的上位主题词	上位词
	Z	族	指引缩从属的族首词	族首词
相关关系	C	参	指引相关关系的主题词	相关词

从表 2.4 可以看出，Y 是主题词指引符，只在主题词下使用，其后所列的词都是与款目词等同的主题词；D 是非主题词指引符，只在非主题词下使用，其后所列的词都是款目词所代替的非主题词；F 是下位主题词指引符，其后所列的主题词是款目词的下位主题词；S 是上位主题词，其后所列的主题词是款目词的上位主题词；Z 是族首词指引符，其后所列的主题词是款目词所属词族的族首词；C 是相关主题词指引符，其后所列的主题词是款目词的相关主题词。

对于上述三种词间关系，它们都是有着各自的作用的。等同关系的作用是为了帮助用户来区分规范和非规范检索词的；等级关系是帮助用户扩检或者缩检的；相关关系的作用是为了帮助用户获得目前检索主题的相关资料的。等同关系是指正式主题词和非正式主题词之间的同义、近义或者准同义关系。例如：工业废水处理，D：工业污水处理。等级关系是指不同专指度的上位概念主题词和下位概念主题词之间的一种关系。例如：放射性污染：“S：环境污染，F：海洋放射性污染”。

2.2 常用文献分类法和主题词表的特点分析

通过对几种分类法和主题词表的概述，不难看出，它们的分类标准和体系结构存在差异性，然而，它们之间也存在一定的共性。它们之间的共性包括两点，首先 IPC、ICS、CCS、UDC、DDC 和 CLC 都是按照等级来进行分类的，虽然主题词表中没有明确的按照等级分类，但是主题词表的三种词间关系存在等级关系；其次，《国际标准分类法》、《中国标准分类法》和主题词表都是以专业和学科的领域为主进行分类的。

然而，对于几种分类法和主题词表的异同，本文在下面进行了具体的分析，主要是通过三个方面着手：（1）每一级分类中包含的类目名称；（2）每一级分类的类目所包含的属性；（3）每一级的类目与其下级之间的关系。

在此，所有概念间的关系不外乎是等价、继承、相关、“是一种特征”和“是一部分”五种。其中继承关系可以细分出“是一种”、“应用于”及“具备特征的”三类关系；“是一种特征”可细分出“是一种属性”和“是一种表现”两类关系。

对各分类体系和主题词表的具体分析如下：

（1）IPC：从其分类结构分析，在专利技术领域将信息逐级分为部、类、小类、主组和

分组,且上下级间没有明确说明从属关系。在分组类号后,根据圆点数来表明其从属关系。在各级分类中,共同包含了中英文名称和编号三种属性。部包含的信息是具有分类性质的名称,其属性编号。“部”与“大类”的关系定义为“是一种”;“大类”包含的信息是一系列的具有列表性质的名称,其属性包括说明。它与“小类”的关系定义为“应用于”;“应用于”为“是一种”的特殊情况;“小类”包含的信息是具有描述和列表两种性质的名称,其属性有说明和编号,它与“主组”的关系有两种,分别定义为“应用于”和“是一部分”;“主组”包含的信息是具有列表性质的名称,其属性有说明,它与“分组”的关系也是两种,分别定义为“是一种”和“是一种属性”;“分组”中包含的信息是分类名称及各种属性特征,其属性有特征,与其后面带有圆点的信息之间的关系定义为“具备的特征”。同时,“分组”作为最后一级,属于叶结点。

(2) ICS:按等级划分为三级,一级分类是按学科、专业进行分类的。其中二级分类和三级分类存在有注释和无注释两种情况。ICS的一级分类所含的信息是具有列表性质的名称,其属性有说明和编号,它与二级分类的关系有两种,分别定义为“是一种”和“应用于”;二级分类包含的信息也是具有列表性质的名称,其属性有说明、描述和编号,它与三级分类的关系也是两种,分别定义为“是一种”和“应用于”;三级分类,属于叶结点,它包含的信息也是具有列表性质的名称,其属性有说明、描述和编号。

(3) CCS:分为两级,如同ICS一样,CCS的一级分类也是按学科和专业进行分类的,且它们之间没有加注释或是指引注释。一级分类包含的信息是具有列表性质的名称,其属性有说明和编号,对下级的关系有两种,分别定义为“是一种”和“应用于”;二级分类包含的信息同一级分类,其属性有说明和编号,因为是最后一级,属于叶结点。

(4) UDC:属于综合性分类,按等级划分为三级。DDC的三个分类级包含的信息也都是具有列表性质的名称,它们的属性也相同,是说明和编号。一级分类对二级分类的关系定义为“是一种”,二级分类对三级分类的关系有两种,分别定义为“是一种”和“是一部分”。且三级分类是叶结点。

(5) DDC:同《杜威十进制分类法》一样,也属于综合性分类,按等级分为三级。一级、二级分类和三级分类包含的信息都是具有列表性质的名称,它们的属性有说明和编号。一级分类对二级分类的关系定义为“是一种”;二级分类对三级分类的关系定义为“是一种”。且三级分类为叶结点,与其后面细分的信息之间的关系定义为“应用于”。

(6) CLC:按照等级来划分,首先是分为5个大部,在部的基础上分为三级。其一级分类、二级分类和三级分类包含的信息都是具有列表性质的名称,它们的属性有说明和编号。一级与二级之间的关系定义为“是一种”;二级和三级分类之间的关系定义为“是一种”。三级分类为叶结点。

(7) 主题词表:没有明确的分类划分,按照字顺排列,是通过添加必要的注释项和词间关系进行检索的,还包含了各叙词的英文名称。因此,可以把所以的叙词定义为一种结点,

即叙词结点。而主题词表的词汇之间存在三种关系，分别同级关系、等级关系和相关关系，因此，可以看作叙词结点之间同样是这三种关系，分别定义为等价关系、继承关系和相关关系。其中，继承关系包括“是一种”和“是一种表现”两类。

2.3 面向科技知识组织体系的公共数据模型

根据对各分类法及叙词表的各方面的分析，在忽略其分类视角的情况下，总结得出以下内容：

结点：IPC 的部是一类结点，命名为部结点，记为 A；IPC 的类和组与 ICS 的一级分类、CCS 的一级分类、UDC 的一级分类和二级分类、DDC 的一级分类和二级分类以及 CLC 的一级分类，可以归为一类结点，定义为目录结点，记为 B；IPC 的小类和 ICS 的二级分类可以视为一类结点，定义为分类结点，记为 C；IPC 的分组为一种结点，定义为专利叶结点，记为 D；CCS 的二级分类、UDC 的三级分类、DDC 的三级分类和 CLC 的二级结点可以视为一类结点，定义为学科叶结点，记为 E；ICS 的三级分类是一种结点，定义为 ICS 叶结点，记为 F；主题词单独作为一种结点，定义为叙词结点，记为 G。共计 7 种结点。

关系：很明显的，有 8 种关系：分别为“是一种”，记为 L₁；“应用于”，属于 L₁ 的一种，记为 L₂；“具备的特征”，也属于 L₁ 的一种，记为 L₃；“是一部分”，记为 L₄；“是一种属性”，记为 L₅；“是一种表现”，记为 L₆；等价关系，记为 L₇；相关关系，记为 L₈。

通过以上的分析与总结，可以得出一个面向各分类体系和主题词表的公共数据模型，这样可以让我们更直观的了解整个体系结构。其中的 7 种结点和 8 中关系如图 2.1 所示：

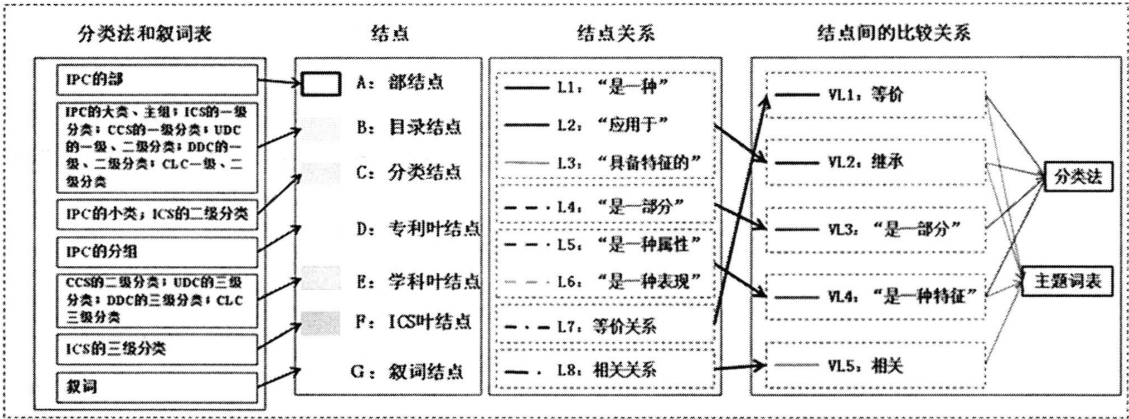


图 2.1 体系结构中结点及各种关系

对于图 2.1 中的各类结点和关系，其中 7 个结点的应用范围及表示图标如下：

- A 结点：IPC 的八个部，用白色矩形表示；
- B 结点：IPC 的类、主组，ICS 的一级分类，CCS 的一级分类，DDC 的一级、二级分类，UDC 的一级、二级分类，CLC 的一级、二级分类，用淡黄色矩形表示；
- C 结点：IPC 的小类，ICS 的二级分类，用淡青色矩形表示；

- D 结点: IPC 的分组及其后面带圆点的信息, 用白烟色矩形表示;
- E 结点: CCS 的二级分类, DDC 的三级分类, UDC 的三级分类, CLC 的三级分类, 用淡灰色矩形表示;
- F 结点: ICS 的三级分类, 用淡红色矩形表示;
- G 结点: 主题词, 即叙词, 用天蓝色矩形表示。

八种关系的表示图标:

- L₁: 黑色实线;
- L₂: 红色实线;
- L₃: 青色实线;
- L₄: 黑色虚线;
- L₅: 红色虚线;
- L₆: 青色虚线;
- L₇: 黑色点划线;
- L₈: 黑色双点划线。

关于图 2.1 中的结点之间的 5 种比较关系, 是从上述总结的内容中推断出来的, 下面是详细分析。在此, 对各种分类体系和叙词表之间应该分开进行分析。

对于图 2.1 中的“等价”来说, 其实是一种等价关系的比较, 因为在所用的分类体系和主题词表中的很多结点之间都可能会存在或多或少的等价相似性, 当然, 也有一部分结点之间不存在等价性, 故在分类法和叙词表中都存在着等价关系的相似度比较, 此关系记为 VL₁; 在上述分析中分类法各结点间存在有“是一种”、“应用于”、“具备特征的”、“是一部分”以及“是一种属性”这 5 种关系, 其中“是一种”、“应用于”和“具备特征的”这 3 种关系都是属于继承这一类的, 而“是一种属性”是归类于某结点是另一结点的一种特征, 因此, 分类法中各个结点间还存在着继承关系的相似程度、“是一部分”关系的相似程度以及“是一种特征的”关系的相似程度 3 种比较关系, 即继承关系的相似度比较、“是一部分”关系的相似度比较和“是一种特征的”关系的相似度比较, 分别记为 VL₂、VL₃ 和 VL₄。而在叙词表中, 存在着 4 种关系: “是一种”、“是一种表现”、同级关系和相关关系。同分类法一样, 其中“是一种”是属于继承的一类, 而“是一种表现”可以归类于某结点是另一结点的一种特征, 故这两种关系可分别归类到 VL₁ 和 VL₄ 中。另外, 相关关系在比较中可以单独视为一种关系, 称之为相关关系的相似度比较, 记为 VL₅。这样, 可以得出, 当只有各种分类法进行比较的时候, 只存在着 VL₁、VL₂、VL₃ 和 VL₄ 这四种比较关系; 当只有主题词表进行比较时, 只存在有 VL₁、VL₂、VL₄ 和 VL₅ 四种关系; 当分类法和主题词表进行比较时, 则五种关系全部存在。五种比较关系的表示图标如下:

- VL₁: 黑色实线;
- VL₂: 红色实线;

- VL₃: 蓝色实线;
- VL₄: 绿色实线;
- VL₅: 青色实线。

根据各结点的特点及它们之间的关系,可以得到一种结点与结点间的映射关系。对于 7 种结点来说,有的结点之间是没有任何关系的,但它们本身之间可能都存在着同级关系,相当于 A 和 A、B 和 B、C 和 C、D 和 D、E 和 E、F 和 F 以及 G 和 G 各自之间都存在关系 L₇; IPC 的部和类的关系为“是一种”,即 A 和 B 之间有关系 L₁; IPC 的类和小类的关系为“应用于”,即 B 和 C 有关系 L₂; IPC 的小类和主组的关系为“应用于”和“是一部分”,即 C 和 B 有关系 L₂ 和 L₄; IPC 的主组与分组的关系为“是一种”和“是一种属性”,即 B 和 D 有关系 L₁ 和 L₅; IPC 的分组间的关系为“具备特征的”,即 D 和 D 有关系 L₃; ICS 的一级与二级分类的关系为“是一种”和“应用于”,即 B 和 C 有关系 L₁ 和 L₂; ICS 的二级与三级的关系为“是一种”和“应用于”,即 C 和 F 有关系 L₁ 和 L₂; CCS 的一级与二级的关系为“是一种”和“应用于”,即 B 和 E 有关系 L₁ 和 L₂; DDC 的一级与二级的关系为“是一种”,即 B 和 B 有关系 L₁; DDC 的二级与三级的关系为“是一种”和“是一部分”,即 B 和 E 有关系 L₁ 和 L₄; UDC 的一级与二级、二级与三级以及三级与三级的关系都为“应用于”,即 B 和 B 有关系 L₁, B 和 E 有关系 L₁, E 和 E 有关系 L₂; CLC 的一级与二级、二级与三级的关系为“是一种”,即 B 和 B 有关系 L₁, B 和 E 有关系 L₁; 主题词表中除了同级关系以外,还有“是一种”、“是一种表现”及相关三种关系,即 G 和 G 有关系 L₁、L₆ 和 L₈。

经总结后得出的映射关系以图的形式表现出来,可以更直观,更清晰的知道结点间都存在哪种关系。结点间的关系映射模型如图 2.2 所示:

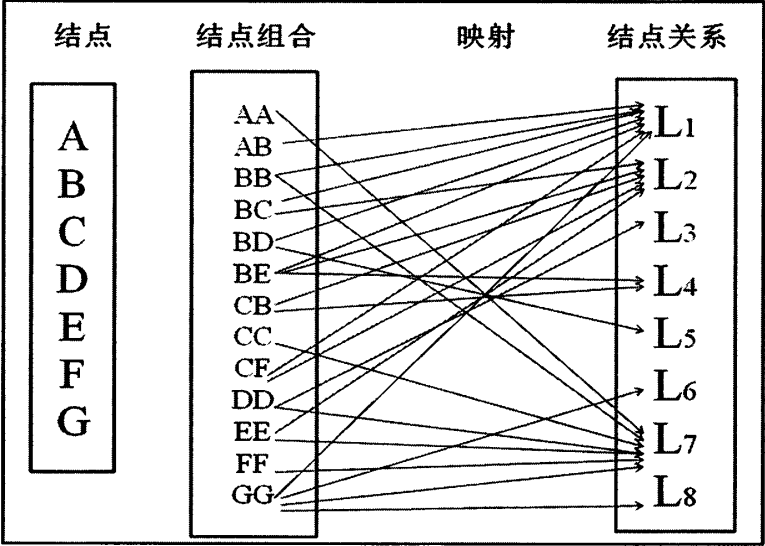


图 2.2 结点间的关系映射模型

对于 7 种结点,它们都是有各自的属性的,属性一:名称,它们都有名字或列表性名称(说明),其名称都有中、英文之分;属性二:来源,每个结点都有自己从属的分类法或主题

词表，来源就说明了它们分别来自哪种分类法或主题词表；属性三：编号，是 IPC、ICS、CCS、DDC、UDC 和 CLC 中存在的属性；属性四：描述，该属性是针对那些有注释的等级分类或者是包含有描述性名称的结点。各结点的结构图如图 2.3 所示：

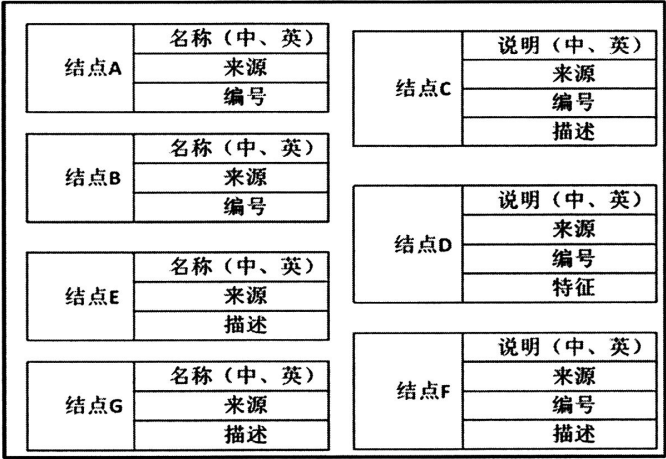
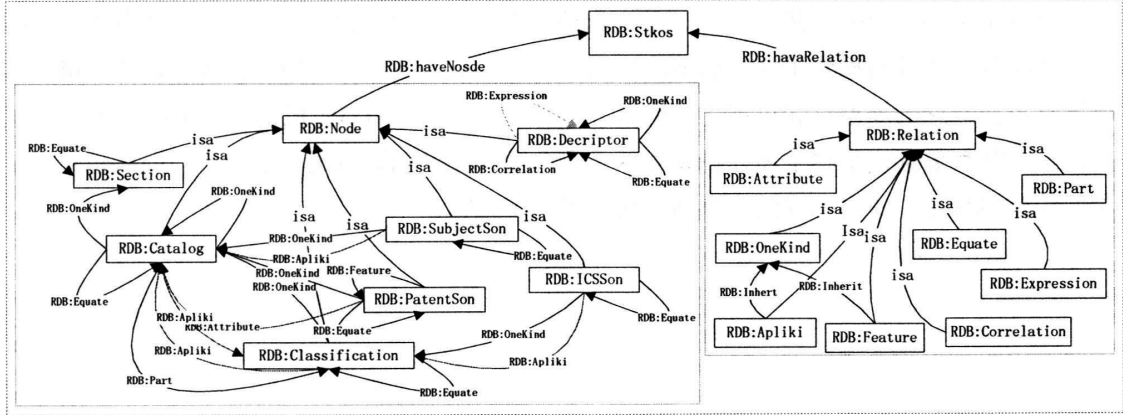


图 2.3 各结点的结构图

将图 2.1、图 2.2 与图 2.3 相结合，利用平台 Eclipse 中的 GMF（Graphical Modeling Framework）工具将语义公共数据模型用 Ecore 模型描述。在此，将 Ecore 模型所描述的语义公共数据模型按照结点间存在的关系和比较关系两种功能分别呈现出来。如图 2.4 所示，其中图（a）充分体现了 7 种结点和 8 种关系之间的联系，图（b）则体现了 7 种关系和 5 种比较关系之间的联系。



图（a） 结点之间的关系

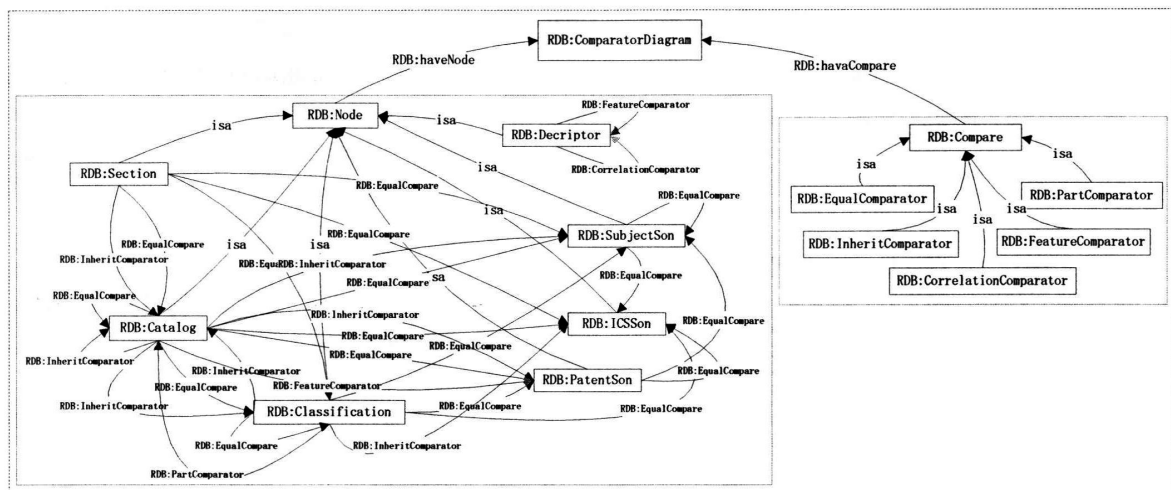


图 (b) 结点之间所具有的比较的关系

图2.4 GMF中的Ecore模型

在GMF工具中，建立如图2.4所示的Ecore模型，是为了实现在开发方面的内容：驱动生成面向科技文献知识体系概念空间构建的可视化图形插件，编码实现分类法和主题词表的自动导入和可视化的显示，进而实现面向科技知识体系映射的语义比较，其中编码来实现模型中各结点之间比较的关联算法。

2.4 本章小结

本章的主要内容是概括描述了IPC、ICS、CCS、DDC、UDC和CLC几种分类体系及主题词表的体系结构和分类机制。首先,对这几种分类体系以及主题词表做出了具体分析,根据它们之间存在的异同构建了统一的概念空间,即公共数据模型。该模型实现了两个功能:实现了对数据概念的模型化和可视化。其次,利用Eclipse的插件GMF把该模型用Ecore模型进行描述。该公共数据模型将上述几种分类体系和主题词表的相同点进行了融合,同时也体现了它们各自的特点,使得不同的分类体系形成了一个整体。

第 3 章 基于公共数据模型的自动映射算法

3.1 统一概念空间下的语义映射发现算法

本文在所构建的公共数据模型的基础之上，设计出了一套映射算法，所要达到的是，在有大量的信息这个前提下，可以更为直接、方便的对信息进行检索。这套映射算法可以从以下三个方面来实现：（1）基于语法的映射发现算法；（2）基于结构的映射发现算法；（3）基于语义推理的映射发现算法。如图 3.1 所示，这是一个关于整套算法的大概框架图。

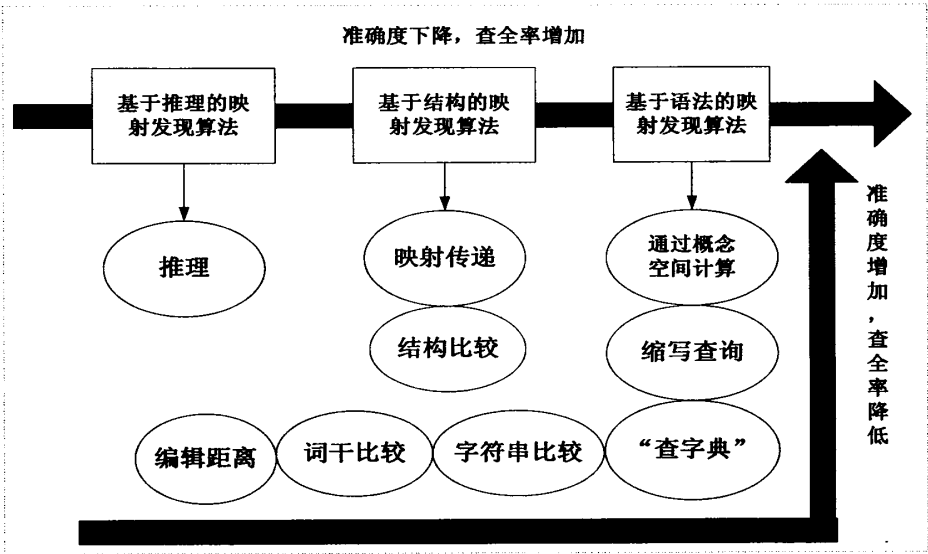


图 3.1 算法框架图

从图 3.1 中可以看出，在本文中，基于语法的映射发现算法的基本算法有通过概念空间计算、缩写查询、“查字典”、字符串比较、词干比较和编辑距离；基于结构的映射发现算法所用到的基本算法有映射传递和结构比较；基于语义推理的映射发现算法所用到的方法就是语义上的推理。这三种发现算法的准确度在逐级下降，查全率在逐级增加。在基于语法的映射发现算法中，按照图 3.1 中的顺序，从编辑距离经过其中的几种基本算法，到通过概念空间计算，在准确度上是逐渐增加的，在查全率上是逐渐降低的。在基于结构的映射发现算法中，从结构比较到映射传递在准确度上是逐渐增加的，在查全率上述逐渐降低的。

对于整套映射算法，三种映射发现算法中间是通过一定的关系及层次结构来实现的，其算法实现图如图 3.2 所示：

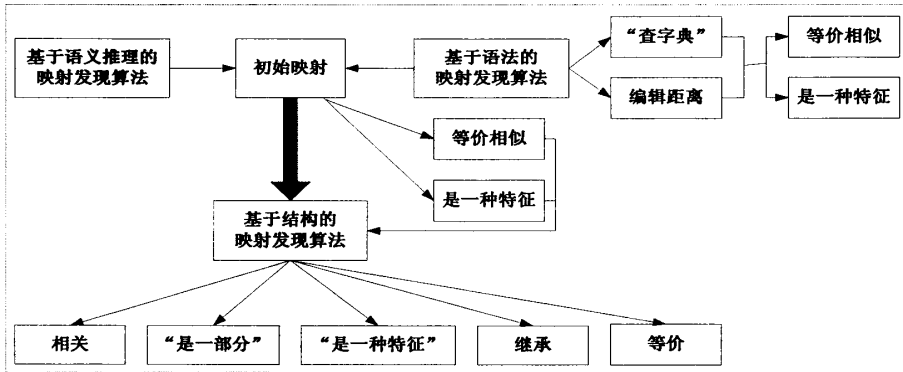


图 3.2 算法实现图

根据图 3.2 所显示，首先通过基于语法的映射发现算法和基于语义推理的映射发现算法来得到一个初始映射，其中基于语法的映射发现算法是根据“查字典”和编辑距离来得到等价相似和“是一种特征”这两种关系。其次，根据得到的初始映射通过基于结构的映射发现算法以及由初始映射得出的等价相似和“是一种特征”，来推断出等价、继承、“是一种特征”、“是一部分”和相关五种结点间的比较关系。

3.2 基于语法的映射发现算法

通过语法方面来映射得到的算法常用的基本算法有 6 种，它们是通过概念空间计算、缩写查询、“查字典”、字符串比较、词干比较和编辑距离。其中常用的有“查字典”的比较方法和编辑距离（edit distance，又叫 Levenshtein Distance，简称 LD）。“查字典”中主要用到的是基于 WordNet 的相似度计算的算法，WordNet 不光是一个把单词以字母顺序排列的简单的英语词典，而是按照单词的概念意义组成了一个“单词的网络”。WordNet 是按照单词的拼写的正字法原则来组织的，其中涵盖了各种的描述对象，如复合词，成语等。其比较方法的原理是：计算在 WordNet 中所有词节点之间的上下位关系所构成的最短路径来计算词与词之间的相似度。编辑距离的原理是通过两个字符串的距离来计算它们的相似性的，两个字符串的距离就是一个字符串转换成另一个字符串这一过程中的添加、删除和修改的数值大小。

通过“查字典”和编辑距离所计算出来的词语之间的相似度，就是所谓的等价相似度，故本文中提到的 5 种关系比较中的等价关系的相似度是可以用“查字典”和编辑距离这两种方法之间计算得出。由于“查字典”相对于编辑距离来说，准确度要高一些，所以最终得出的相似度在这两种方法得到的值中取最大的。

3.3 基于结构的映射发现算法

通过结构的映射得到的算法是由两种算法组成的，它们是映射传递和结构比较，在这里用到的是映射传递这种方法。其中，映射传递，顾名思义，从字面上就可以看出它是具有传递性的，就像数学中的一个原理一样，如果 $A > B$ ，且 $B > C$ ，那么 $A > C$ 。通过一个例子可以很清晰的理解映射传递的意思。例如：Sachbearbeiter 和 Contact person 看起来是没有任何关

系的，但是 Sachbearbeiter 和 Staffer 是等价的，且 Staffer 和 Contact person 是等价的，那么，Sachbearbeiter 和 Contact person 就是等价的。更直观的可以通过图 3.3 来观察。

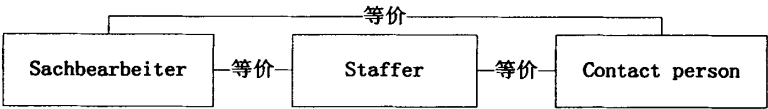


图 3.3 映射传递

映射传递这种算法是比较简单的，除了等级，继承、“是一种特征”、“是一部分”和相关这四种结点间关系的比较也是可以通过映射传递的方法来进行传递的。以下是对各种关系的比较的详细内容。

因为映射传递是要逐层进行传递的，故要考虑其传递过程中是否要衰减。对于各种分类法和主题词表来说，它们呈现的结构都是树形结构，所以在是否衰减这个问题上要分两种情况进行分析：（1）向上传递；（2）向下传递。而分向上传递和向下传递这两种情况的目的是为了为了更好的实现在算法上的查全率，即召回率。在这个树形结构中，其从上往下大体上呈现的是属性增多、代表范围收敛的趋势，因此，在向上传递时是不需要衰减的，即传递系数为 1；在向下传递时就需要进行传递上的衰减了，对于衰减程度来说，还需要具体情况具体分析。很明显的，向上传递和向下传递所用的策略是不同的，其目的是在于为了更好的实现在算法上的查准率。

对于等价，继承、“是一种特征”、“是一部分”和相关这五种结点间关系的比较来说，其他的四种比较跟等价关系的比较是不同的，它们之间有一个共同的特点，都是在任意两个概念通过字面来看没有任何相似性的情况下，可以通过映射传递来发现这两个概念之间是否存在四种关系中的某一种的相似度。举一个例子来说明吧，假如 M 和 N 之间存在着一定的相似度，但 M 和 P 之间从字面来看没有任何的相似性，而 P 是 N 的父类概念，且 P 和 N 之间的关系为“是一种特征”，那么 P 和 M 之间也是“是一种特征”的关系，且 P 和 M 之间也存在着一定的相似度。所以，另外的四种比较都是在等价关系的比较的基础上触发的。

向上传递过程中，由于传递系数为 1，对于等价，继承、“是一种特征”、“是一部分”和相关这五种结点间关系的比较，其结果就是传递前的原来两个概念的等价相似度的值。通过图 3.4 这个例子可以清晰的理解这种情况。

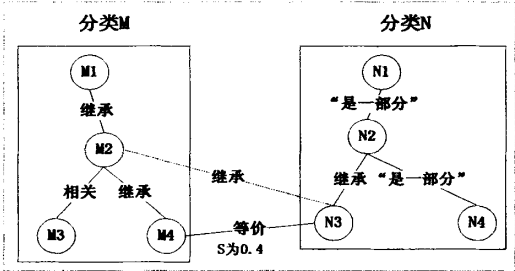


图 3.4 向上传递图

如图 3.4 所示, M_4 与 N_3 是等价的, 其等价相似度为 0.4, 且 M_4 是继承于 M_2 的, 所以 M_2 与 N_3 之间具有继承关系, 由于其传递系数为, 故其相似度为 0.4, 也就是说 M_2 与 N_3 之间的继承相似度为 0.4。

向下传递过程中, 由于这五种结点间关系的传递系数是不同的, 所以要每种关系进行单独的分析。下面是对各种关系进行的详细分析。

每一个信息元 (结点) 都是有很多的属性的, 当然, 在本文中所有的概念的属性的数量都是可以明确知道的, 包括中、英文名称 (说明项), 来源, 编号和描述五种。在这些属性中, 中、英文名称相对于其他属性来说, 是比较重要的, 在比较过程中所占的比重相对来说要大一些。图 3.5 是进行比较的两个概念的内容。

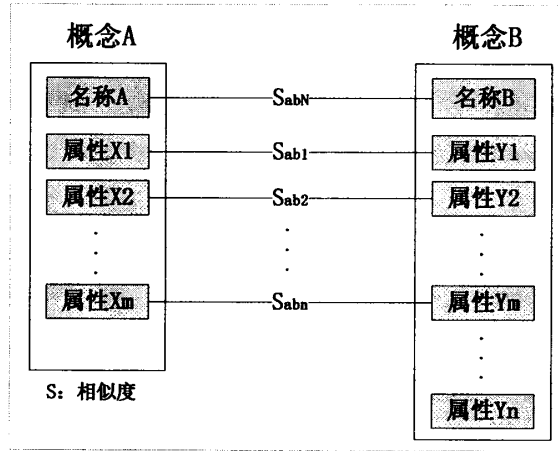
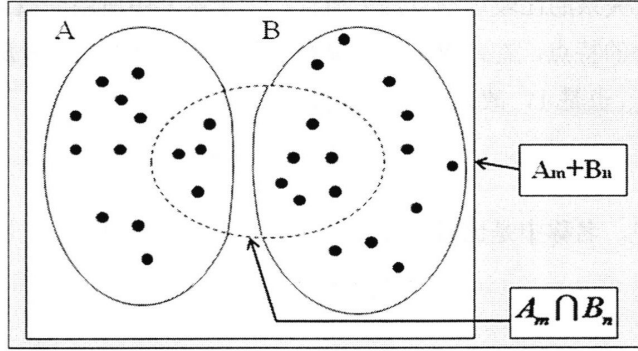


图 3.5 进行比较的两个概念的内容

关于等价关系的比较, 任意两个概念之间有关系就是有关系, 没有关系就是没有关系, 非常的明确, 因此, 等价关系的比较是单独考虑的, 同其他的四种关系是有区别的。对于每一个概念来说, 都可以对其各种属性进行比较的, 最终来得出每两个概念之间的相似度。其中, 概念间的相似度与其名称及属性的之间的约束条件有: (1) 相似度在 0 到 1 之间, 可以等于 0 且可以等于 1; (2) 当概念名称及属性完全相同时, 即概念的名称的相似度及所有属性的相似度为 1 时, 概念之间的相似度为 1。因此, 计算任意两个概念之间的等价相似度可以通过式 (3.1) 来得出结果。

$$S_{abs} = \alpha_s S_{abN} + (1 - \alpha_s) \sum_{r=1}^R \frac{1}{A_m + B_n - A_m \cap B_n} S_{abr} \quad R=1, 2, \dots, A_m \cap B_n \quad (3.1)$$

式 (3.1) 中 S_{abs} 表示概念 A 和 B 的相似度, α_s 表示权重, S_{abN} 和 S_{abr} 分别表示概念的名称和除名称之外其他属性 r 各自比较后得到的等价相似度, A_m 和 B_n 分别表示概念 A 和 B 的除名称以外的属性个数, $A_m \cap B_n$ 表示概念中除名称之外相同属性的个数。式 (3.1) 中的第二项多项式只有在 A_m 、 B_n 和 $A_m \cap B_n$ 都不等于 0 的情况下才会出现。其中, $A_m + B_n$ 和 $A_m \cap B_n$ 如图 3.6 所示:

图 3.6 $A_m + B_n$ 和 $A_m \cap B_n$ 所表示的内容

在本文中，关于所有概念的几种属性有以下几个特点：（1）编号项和来源项没有比较的作用，可以去掉不考虑；（2）名称项（说明项）虽然分成中、英两项，但只比较了一项；（3）描述信息在分类法和主题词表中不全，暂时不考虑。根据这些特点，两个概念的比较就只有名称的比较了。为了符合本文的特点，故式（3.1）可以简化为式（3.2）：

$$S_{abs} = \alpha_s S_{abN} \quad (3.2)$$

在这种情况下，很明显的，名称的相似度就是概念的相似度，即 S_{abs} 等于 S_{abN} ，故可以视 α_s 为 1。即得到最终公式为：

$$S_{abs} = S_{abN} \quad (3.3)。$$

五种关系的比较中，其他四种关系的比较不同于等价关系的比较。由于传递不是单单的一层传递，而是多层传递的，且在传递的过程中，可能不是由单一的关系进行传递的，故在此过程中，是根据传递路径的最后的一层关系来确定首尾两个概念的关系的。

第一，关于“是一种特征”这种关系的比较，这个不像等价，从名称就可以看出很多的内容，而是将概念的包括名称在内的所有的属性进行比较，从而进行传递的。然而，“是一种特征”的比较和其他的三种稍有不同，它不仅可以通过两个概念之间的等价相似度进行传递，也可以通过两个概念的“是一种特征”的相似度来进行传递。通过分析，“是一种特征”的比较的相似度可以通过式（3.3）计算得出。

$$S_{abF} = \sum_{r=1}^R \frac{1}{A_{m+1} + B_{n+1} - A_{m+1} \cap B_{n+1}} \alpha_f^{L_1} \alpha_l^{L_2} \alpha_p^{L_3} \alpha_c^{L_4} f_r, \quad R=1,2,\dots, A_{m+1} \cap B_{n+1} \quad (3.4)$$

在式（3.4）中， S_{abF} 表示“是一种特征”的传递相似度， α_f 、 α_l 、 α_p 和 α_c 分别表示传递过程中的“是一种特征”、继承、“是一部分”和相关四种关系的传递权重， L_1 、 L_2 、 L_3 和 L_4 分别表示在一条传递路径中“是一种特征”、继承、“是一部分”和相关四种关系分别出现的次数， f_r 表示包括概念名称在内的所有属性的相似度， A_{m+1} 和 B_{n+1} 分别表示概念 A 和 B 的包括名称在内的所有属性个数， $A_{m+1} \cap B_{n+1}$ 表示概念中包括名称在内的相同属性的个数。

然而，如同等价关系的比较这一算法中所说，由于本文中所用的概念都是比较特殊的，为了符合本文中概念的特点，在此过程中， R 简化到了 1，由于 A_{m+1} 和 B_{n+1} 都为 1， $A_{m+1} + B_{n+1} - A_{m+1} \cap B_{n+1}$ 也是 1，故式 (3.4) 可简化为式 (3.5)：

$$S_{abF} = \alpha_f^{L_1} \alpha_l^{L_2} \alpha_p^{L_3} \alpha_c^{L_4} f_r \quad (3.5)$$

而在这个算法中，名称中是比较的英文名称，故 f_r 可以写成 f_{EN} ，可得到式 (3.6)：

$$S_{abF} = \alpha_f^{L_1} \alpha_l^{L_2} \alpha_p^{L_3} \alpha_c^{L_4} f_{EN} \quad (3.6)$$

在式 (3.6) 中，权重 α_f 是根据“是一种特征”这种关系分别在两种分类法（也可以是一种分类法和主题词表或者两个主题词表）中占有所有关系的比重 P_{Af} 和 P_{Bf} 之和的平均值 $\overline{P_f}$ 以及概念 A 和概念 B 的等价相似度 S 来确定的。权重 α_l 是根据继承这种关系分别在两种分类法（也可以是一种分类法和主题词表或者两个主题词表）中占有所有关系的比重 P_{Al} 和 P_{Bl} 之和的平均值 $\overline{P_l}$ 以及概念 A 和概念 B 的等价相似度 S 来确定的。权重 α_p 是根据继承这种关系分别在两种分类法中占有所有关系的比重 P_{Ap} 和 P_{Bp} 之和的平均值 $\overline{P_p}$ 以及概念 A 和概念 B 的等价相似度 S 来确定的。权重 α_c 是根据相关这种关系分别在任意两个主题词表中占有所有关系的比重 P_{Ac} 和 P_{Bc} 之和的平均值 $\overline{P_c}$ 以及概念 A 和概念 B 的等价相似度 S 来确定的。其中

$$\overline{P_f} = \frac{P_{Af} + P_{Bf}}{2} \quad (3.7)$$

$$\overline{P_l} = \frac{P_{Al} + P_{Bl}}{2} \quad (3.8)$$

$$\overline{P_p} = \frac{P_{Ap} + P_{Bp}}{2} \quad (3.9)$$

$$\overline{P_c} = \frac{P_{Ac} + P_{Bc}}{2} \quad (3.10)$$

且权重 α_f 、 α_l 、 α_p 和 α_c 都必须符合一定的准则：首先， α_f 、 α_l 、 α_p 和 α_c 都必须单调的；其次，当 S 等于 0 时， α_f 、 α_l 、 α_p 和 α_c 分别为 0，当 S 等于 1 时， α_f 、 α_l 、 α_p 和 α_c 分别为 1；最后， S 的范围是从 0 到 1。其中， S 是通过公式 (3.3) 计算得出的等价相似度。从而分别得出权重 α_f 、 α_l 、 α_p 和 α_c 的计算公式：

$$\alpha_f = \frac{(\overline{P_f} + 1)S}{\overline{P_f} + S} \quad (3.11)$$

$$\alpha_l = \frac{(\overline{P_l} + 1)S}{\overline{P_l} + S} \quad (3.12)$$

$$\alpha_p = \frac{(\overline{P_p} + 1)S}{\overline{P_p} + S} \quad (3.13)$$

$$\alpha_c = \frac{(\bar{P}_c + 1)S}{\bar{P}_c + S} \quad (3.14)$$

将式 (3.11)、(3.12)、(3.13) 和 (3.14) 分别代入到式 (3.6) 中可得到最终的计算公式:

$$S_{abF} = \left[\frac{(\bar{P} + 1)S}{\bar{P} + S} \right]^{L_1} \left[\frac{(\bar{P}_I + 1)S}{\bar{P}_I + S} \right]^{L_2} \left[\frac{(\bar{P}_p + 1)S}{\bar{P}_p + S} \right]^{L_3} \left[\frac{(\bar{P}_c + 1)S}{\bar{P}_c + S} \right]^{L_4} f_{EN} \quad (3.15)$$

最后, 在用式 (3.15) 计算“是一种特征”的相似度的时候, 由于只是对概念的英文名称进行比较的, 所以 f_{EN} 可由等价相似度 S 来代替。

为了更直观的说明此算法, 现举例如下:

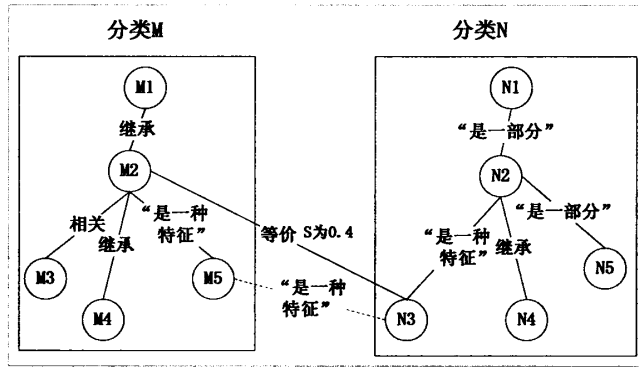


图 3.7 一个实例

如图 3.7 所示: 这个实例是已知了主题词表 M 中概念 M_5 与分类法 N 中的概念 N_3 的等价相似度, 计算 M_5 与 N_3 的“是一种特征”的相似度。在图中可以看出, M 和 N 中的关系总数分别为 4 和 4, 其中“是一种特征”这种关系的数量都只有 1 个, 故可以得出 P_M 和 P_N 都是 0.25, 根据式 (3.7) 计算出 $\bar{P}$ 为 0.25。式 (3.15) 中的 f_{EN} 就是 S , 即 0.4; 很明显的, L_1 为 1。将以上所得到的值代入到式 (3.15) 中, 即可计算出 M_5 与 N_3 的“是一种特征”的相似度 S_{abF} 等于 0.308。

第二, 关于继承的比较, 也是通过是将概念的包括名称在内的所有的属性进行比较来传递得出结果的。其计算公式如下:

$$S_{abl} = \sum_{r=1}^R \frac{1}{A_{m+1} + B_{n+1} - A_{m+1} \cap B_{n+1}} \alpha_f^{L_1} \alpha_I^{L_2} \alpha_p^{L_3} \alpha_c^{L_4} i_r \quad R=1, 2, \dots, A_{m+1} \cap B_{n+1} \quad (3.16)$$

式 (3.16) 中, S_{abl} 表示继承关系比较的传递相似度, α_f 、 α_I 、 α_p 和 α_c 分别表示传递过程中的“是一种特征”、继承、“是一部分”和相关四种关系的传递权重, L_1 、 L_2 、 L_3 和 L_4 分别表示在一条传递路径中“是一种特征”、继承、“是一部分”和相关四种关系分别出现的次数, i_r 表示包括概念名称在内的所有属性的相似度, A_{m+1} 和 B_{n+1} 分别表示概念 A 和 B 的包括名称之内的所有属性个数, $A_{m+1} \cap B_{n+1}$ 表示概念中包括名称在内的相同属性的个数。

根据本文中概念的特点, 式 (3.16) 可以进行简化, 其简化过程同“是一种特征”的比

较一样，简化公式为：

$$S_{abl} = \alpha_f^{L_1} \alpha_l^{L_2} \alpha_p^{L_3} \alpha_c^{L_4} i_r \quad (3.17)$$

将式 (3.11)、(3.12)、(3.13) 和 (3.14) 分别代入公式 (3.17) 中可得最终的计算公式：

$$S_{abl} = \left[\frac{(\bar{P}+1)S}{\bar{P}+S} \right]^{L_1} \left[\frac{(\bar{P}_l+1)S}{\bar{P}_l+S} \right]^{L_2} \left[\frac{(\bar{P}_p+1)S}{\bar{P}_p+S} \right]^{L_3} \left[\frac{(\bar{P}_c+1)S}{\bar{P}_c+S} \right]^{L_4} i_r \quad (3.18)$$

在这里，由于只是对概念的英文名称进行比较的，所以 i_r 可由等价相似度 S 来代替。

第三，关于“是一部分”这种关系的比较，同样是通过是将概念的包括名称在内的所有的属性进行比较来传递得出结果的。其计算公式如下：

$$S_{abp} = \sum_{r=1}^R \frac{1}{A_{m+1} + B_{n+1} - A_{m+1} \cap B_{n+1}} \alpha_f^{L_1} \alpha_l^{L_2} \alpha_p^{L_3} \alpha_c^{L_4} p_r \quad R=1,2,\dots,A_{m+1} \cap B_{n+1} \quad (3.19)$$

式 (3.19) 中， S_{abp} 表示“是一部分”这种关系的比较的传递相似度， α_f 、 α_l 、 α_p 和 α_c 分别表示传递过程中的“是一种特征”、继承、“是一部分”和相关四种关系的传递权重， L_1 、 L_2 、 L_3 和 L_4 分别表示在一条传递路径中“是一种特征”、继承、“是一部分”和相关四种关系分别出现的次数， p_r 表示包括概念名称在内的所有属性的相似度， A_{m+1} 和 B_{n+1} 分别表示概念 A 和 B 的包括名称之内的所有属性个数， $A_{m+1} \cap B_{n+1}$ 表示概念中包括名称在内的相同属性的个数。

根据本文中概念的特点，式 (3.19) 可以进行简化，其简化过程同“是一种特征”的比较一样，简化公式为：

$$S_{abp} = \alpha_f^{L_1} \alpha_l^{L_2} \alpha_p^{L_3} \alpha_c^{L_4} p_r \quad (3.20)$$

将式 (3.11)、(3.12)、(3.13) 和 (3.14) 分别代入公式 (3.20) 中可得最终的计算公式：

$$S_{abp} = \left[\frac{(\bar{P}+1)S}{\bar{P}+S} \right]^{L_1} \left[\frac{(\bar{P}_l+1)S}{\bar{P}_l+S} \right]^{L_2} \left[\frac{(\bar{P}_p+1)S}{\bar{P}_p+S} \right]^{L_3} \left[\frac{(\bar{P}_c+1)S}{\bar{P}_c+S} \right]^{L_4} p_r \quad (3.21)$$

在这里，由于只是对概念的英文名称进行比较的，所以 p_r 可由等价相似度 S 来代替，最终计算得出 S_{abp} 。

第四，关于相关的比较，同样是通过是将概念的包括名称在内的所有的属性进行比较来传递得出结果的。其计算公式如下：

$$S_{abc} = \sum_{r=1}^R \frac{1}{A_{m+1} + B_{n+1} - A_{m+1} \cap B_{n+1}} \alpha_f^{L_1} \alpha_l^{L_2} \alpha_p^{L_3} \alpha_c^{L_4} c_r \quad R=1,2,\dots,A_{m+1} \cap B_{n+1} \quad (3.22)$$

式 (3.22) 中， S_{abc} 表示相关比较的传递相似度， α_f 、 α_l 、 α_p 和 α_c 分别表示传递过程中的“是一种特征”、继承、“是一部分”和相关四种关系的传递权重， L_1 、 L_2 、 L_3 和 L_4 分别表示在一条传递路径中“是一种特征”、继承、“是一部分”和相关四种关系分别出现的次数， c_r 表示包括概念名称在内的所有属性的相似度， A_{m+1} 和 B_{n+1} 分别表示概念 A 和 B 的包括

名称之内的所有属性个数, $A_{m+1} \cap B_{n+1}$ 表示概念中包括名称在内的相同属性的个数。

根据本文中概念的特点, 式 (3.22) 可以进行简化, 其简化过程同“是一种特征”的比较一样, 简化公式为:

$$S_{abC} = \alpha_f^{l_1} \alpha_l^{l_2} \alpha_p^{l_3} \alpha_c^{l_4} c_r \quad (3.23)$$

将式 (3.11)、(3.12)、(3.13) 和 (3.14) 分别代入式 (3.23) 中可得最终的计算公式:

$$S_{abC} = \left[\frac{(\bar{P}+1)S}{\bar{P}+S} \right]^{l_1} \left[\frac{(\bar{P}_l+1)S}{\bar{P}_l+S} \right]^{l_2} \left[\frac{(\bar{P}_p+1)S}{\bar{P}_p+S} \right]^{l_3} \left[\frac{(\bar{P}_c+1)S}{\bar{P}_c+S} \right]^{l_4} c_r \quad (3.24)$$

同样的, 在这里, 由于只是对概念的英文名称进行比较的, 所以 c_r 可由等价相似度 S 来代替, 从而计算得出 S_{abC} 。

基于结构的映射发现的算法也会出现一些特殊的情况, 会同时有多个概念与某个概念存在等价相似性, 且这多个概念向上都是同一个父结点, 故在向上传递的过程中, 会出现多个相似度的值, 在这里, 我们可以简单的取它们中的最大值。如下面一个例子很好的说明了这种特殊的情况。如图 3.8 所示:

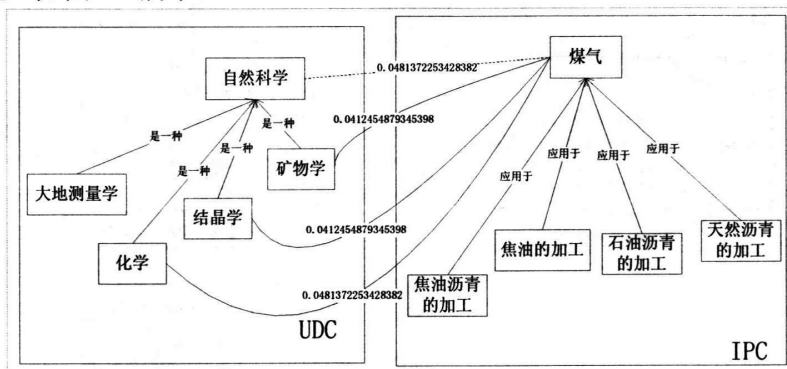


图 3.8 特殊情况的一个例子

3.4 基于语义推理的映射发现算法

描述逻辑 (Description Logic) 是基于知识表示的形式化, 也称为概念表示语言或术语逻辑, 是一阶谓词逻辑可判定的子集, 能够提供可判定的推理。描述逻辑的基本组件有概念 (Concepts), 关系 (Roles) 或者称属性, 个体 (Individuals) 三部分^[43]。概念是用来描述资源对象集合的, 在本文中, 概念可视为一个类; 关系 (属性) 是对资源对象的性质和对象之间的关系的描述和判定; 个体是资源对象的具体化, 也就是类中的一个实例。概念、关系和个体组合在一起构成了对信息的描述。语义推理就是建立于描述逻辑的基础之上的, 其在本文中主要研究的是简单推理和复杂推理。下面通过两个例子来简单描述一下这两种推理, 如图 3.9 和图 3.10 所示:

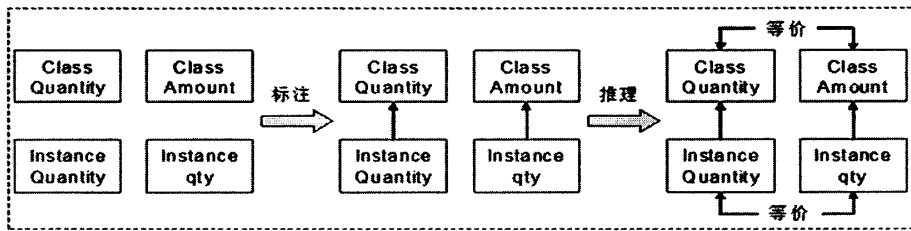


图 3.9 简单推理

图 3.9 所描述的是简单推理。其过程为：首先，定义两个类和两个实例，两个类分别命名为 Class Quantity 和 Class Amount，两个实例分别命名为 Instance Quantity 和 Instance qty；其次，通过标注发现，其中实例 Instance Quantity 是属于类 Class Quantity 的，实例 Instance qty 是属于类 Class Amount 的；接着，根据推理得出，由于类 Class Quantity 和类 Class Amount 等价，所以，实例 Instance Quantity 和实例 Instance qty 也是等价的。

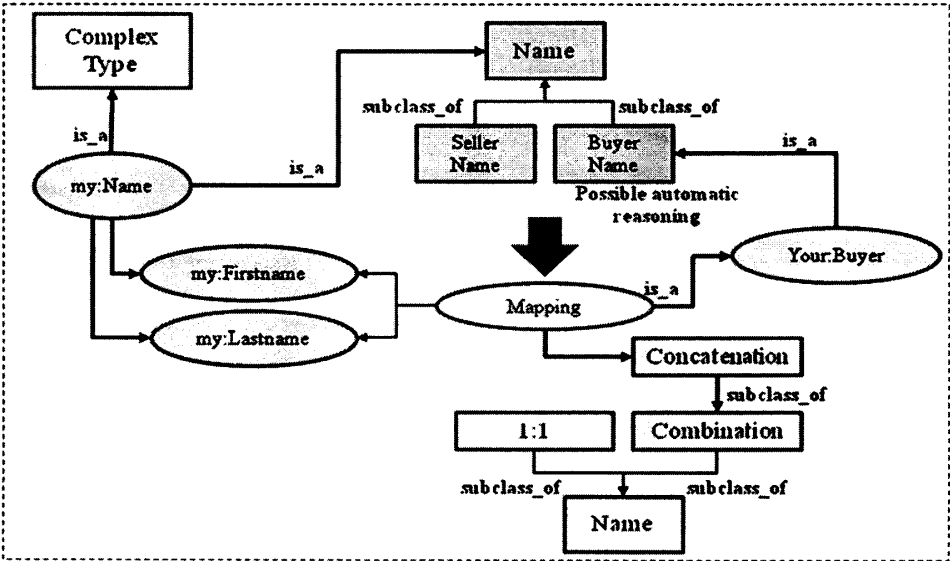


图 3.10 复杂推理

在图 3.10 中，描述的是复杂推理。其中 is_a 表示“是一个实例”，如 A is_a B，其意思为 A 是 B 的一个实例；subclass_of 表示“是子类”，如 C subclass_of D，即 C 是 D 的子类。首先，定义一个实例 my.Name，它是复杂类型（Complex Type）的一个实例，也是类 Name 的一个实例，而 my.Name 又分为姓氏和名字，即 my.Firstname 和 my.Lastname。其次，定义了两个类 Seller Name 和 Buyer Name，它们是类 Name 的两个子类；再定义一个属于 Buyer Name 的实例 Your:Buyer；第三，根据一定的可能性自动推理，Your:Buyer 与 my.Firstname 和 my.Lastname 进行映射；最后经过聚合（Concatenation），合并（Combination）得出 Name，在这里，需要的约束条件就是必须是一对一的。

3.5 本章小结

本章的主要内容是在上一章节的基础上提出了基于公共数据模型的自动映射算法，所实现的目标是：在信息大量化的前提下，更方便、高效的对信息进行检索。该算法是从三个方面来完成的，分别是：基于语法的映射发现算法、基于结构的映射发现算法和基于语义推理的映射发现算法。综合这三种算法目的在于提高查询的查全率。

其中，基于语法的映射发现算法有多种实现方法，其目的是提高查询的准确度。基于语法的映射发现算法包括常用的编辑距离和“查字典”法。其得到的计算结果是各信息之间的等价相似度。而基于结构的映射发现算法是在得到的等价相似度的基础上计算其他几种关系的相似度，包括等价、继承、“是一种特征”、“是一部分”和相关五种比较关系。

第 4 章 系统实现与实证结果

本节所描述的是本文研究内容的开发方面的内容，在这方面的实现主要是应用 Eclipse 平台及其一些插件，其中包括 GMF 和 RCP 等。在开发过程中，包括系统方面的实现和实例认证两部分。在系统方面，需要了解一些关于 Eclipse、GMF 和 RCP 方面的知识和技巧。其中利用 GMF 来实现上文所提到的公共数据模型的统一描述，利用 RCP 将所要实现的功能集成为一个产品。在实例认证方面，所需要的数据主要是主题词表、UDC 和 IPC 三种体系。其整体构架主要有以下几块，系统的系统结构图如图 4.1 所示：

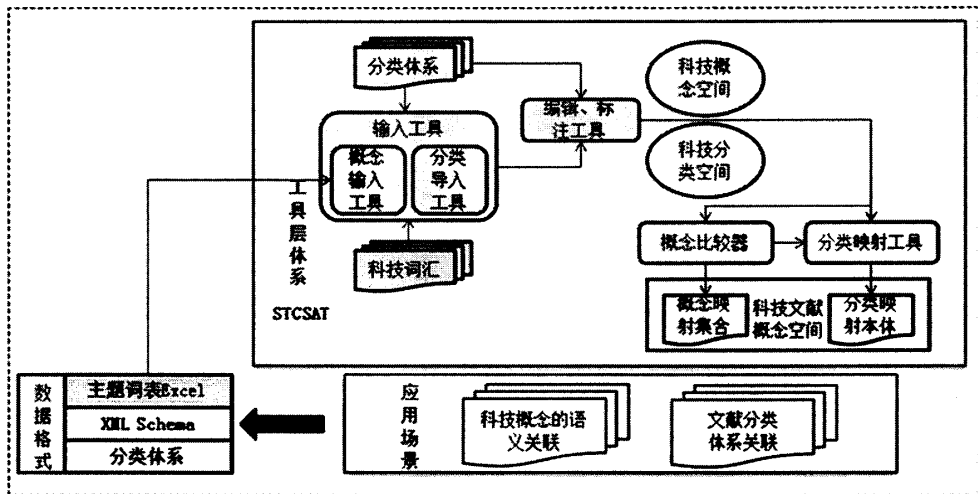


图 4.1 系统结构图

本文所研究的内容，主要的应用场景是在科技概念的语义关联和文献分类系统关联两个方面。集成的 RCP 产品主要有概念输入工具和概念比较器，即产品的导入功能和比较功能。首先，在数据格式方面，需要将主题词表及的内容用 Excel 格式来保存，这样方便导入数据。其次，是导入数据，主要是将以 excel 格式保存的数据内容导入到所开发的系统软件内，实现可视化图形，在导入过程中，分类体系需要进行编辑、标注等。再次，是对概念的比较。

4.1 系统实现

在系统方面的实现，通过 Eclipse 平台及其插件，编码实现其各方面的功能，数据导入功能、数据比较功能和数据导出功能。

4.1.1 系统框架

在这里，系统的实现分为三大模块：功能层、模型层和数据层。其中功能层有三个方面：数据导入、数据比较和数据导出。模型层的实现是先由上文分析得到的公共数据模型在 Eclipse 中用 Ecore 模型描述，驱动生成 CDM 构架，接着将所需要的数据导入，实现数据的模型化，

最后通过编码进行算法的设计，以实现其数据比较的功能。在算法设计方面，是基于语法的映射发现、基于结构的映射发现和基于语义推理的映射发现三方面的集成。数据层主要是一些在系统开发所需要的数据，其中包括分类体系中的 IPC、ICS、CCS、DDC、UDC 和 CLC 以及主题词表。其系统框架图如图 4.2 所示：

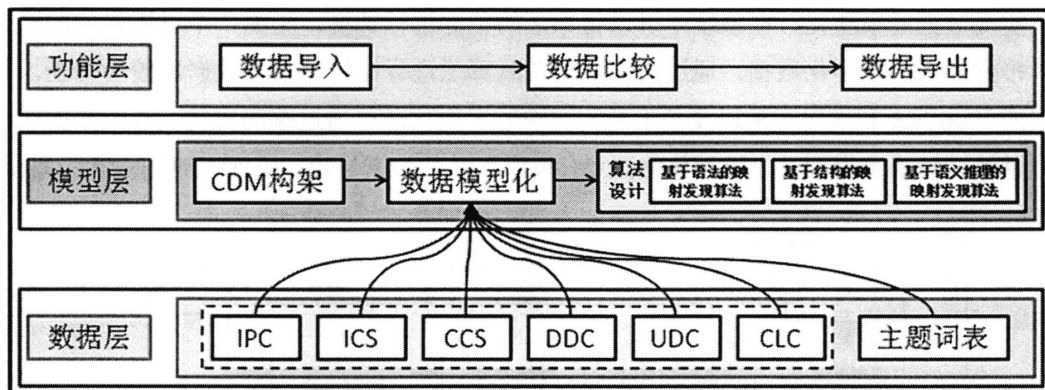


图 4.2 系统框架图

4.1.2 系统的技术体系

对于开发本身来说,需要集多种技术于一身,才能开发出一个具有多功能的工具。在本文中,所用到的技术有 Eclipse 的插件 GMF 和 RCP 以及其他插件。设计算法的技术上,综合多种关联算法的应用。通过将原始数据经过编码分析,使数据标准化,然后经由 GMF 来实现数据的可视化,同时,也将数据模型化,之后通过设计的关联算法使数据进行比较,再经由 GMF 来实现数据的可视化。其系统的技术体系图如图 4.3 所示:

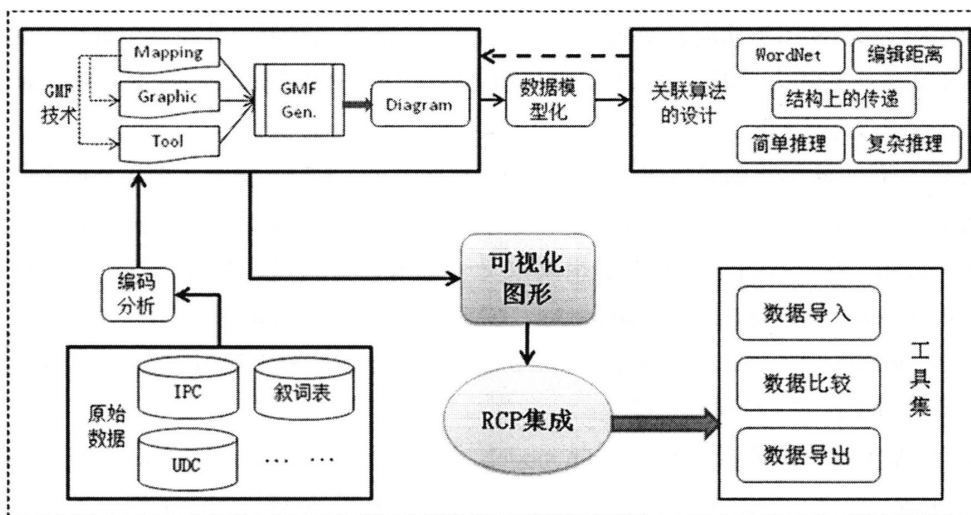


图 4.3 系统的技术体系图

其中 GMF 的技术方面, 通过 Graphic (图形化定义) 模型与 Tool (工具定义) 模型合并生成 Mapping (映射定义) 模型, 通过创建 GMF Gen 模型来进一步生成一个图形化编辑工具。

4.1.3 系统的功能体系

在系统开发方面，系统的功能实现是很重要的一个模块。这里，主要是实现三个功能：数据导入、数据比较和数据导出。首先，根据上文提到的公共数据模型，在 Eclipse 的 GMF 插件中用 Ecore 模型描述并驱动出 CDM 模型，将主题词表和分类体系通过 CDM 模型导入系统，实现数据的模型化，在系统中可以将导入的数据进行转换，生成一个可视化的图形。其次，根据导入的模型化数据，通过所设计的关联算法进行比较，实现数据比较的功能，其算法中的 WordNet 和编辑距离以及基于结构的映射发现算法可通过编码实现，其中数据比较后也可以通过转换来得到一个可视化的图形。最后可进行数据导出。系统的功能体系图如图 4.4 所示：

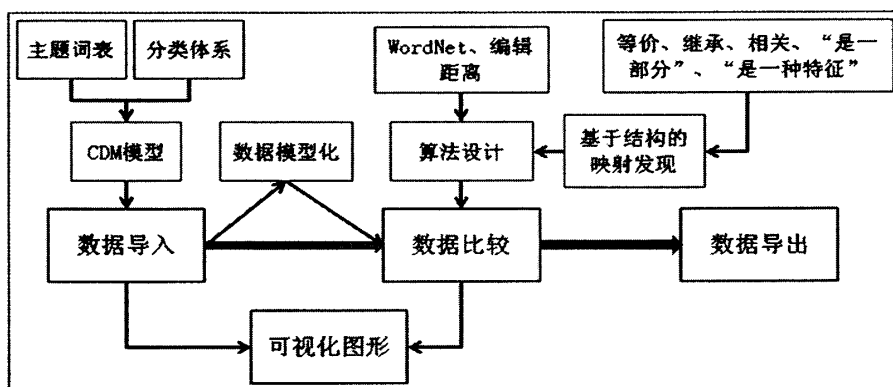


图 4.4 系统的功能体系图

4.1.4 GMF 框架概述

在没有 GMF (Graphical Modeling Framework) 之前，是使用 GEF (Graphical Editing Framework) 来创建图形化编辑器的，由于使用 GEF 既慢又痛苦，且在这个过程中包括了很多理解复杂的框架和大量的冗余代码，故 IBM 在 GEF 和 EMF (Eclipse Modelling Framework) 的基础之上开发了 GMF。

GMF 提供了图形化编辑器的开发环境和运行时框架。GMF 框架主要包括两个部分：运行时框架，图形化开发环境。图形开发环境依赖底层的运行时框架。在 GMF 的图形化开发环境中，GMF 编辑器的主要功能是用来定义一组用来生成代码框架的模型，主要包括：graphical model, tool model 和 map model。GMF 的创建向导图如图 4.5 所示：

其中，graphical model, tool model 和 map model 分别是图 4.5 中的图形化定义模型、工具定义模型和映射定义模型三个模块。

GMF 编辑器首先需要 domain model，通常是一个 *.ecore 文件，GMF 根据这个 domain model 做进一步的操作，这个 model 是实施 GMF 后续工作的基础。其中：

graphical model: 定义了 ecore model 元素的图形信息，如形状，颜色，边框等等。严格的说这个图形信息是和 ecore model 无关的(以便于领域模型和用户界面模型的映射)，但是在具体的操作过程中我们都是根据 ecore model 来生成其所对应的 graphical model。

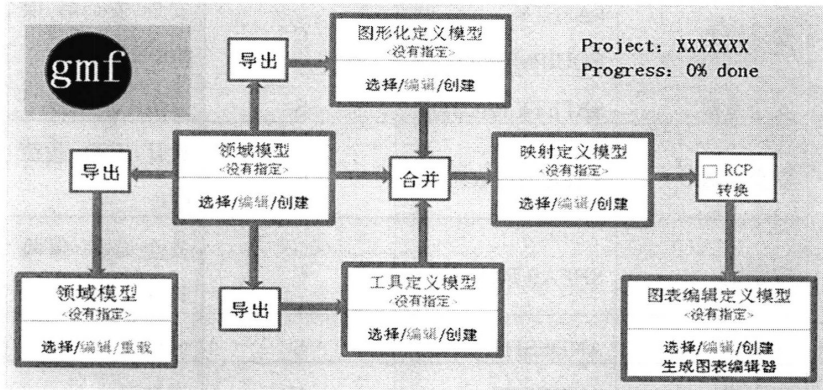


图 4.5 创建 GMF 的向导图

tool model: 定义了在未来生成的图形编辑器中 palette view 里面的 palette entry 信息。如名字, 图标。

map model: 定义了 ecore model, graphical model 和 tool model 三者间的映射关系, 通过 map model 来调整其他 3 个 model 的映射关系可以增强 model 的可重用性和灵活性。

4.2 实证结果

根据上述的理论分析和总结及系统开发结果, 需要一些实例数据来认证。因此, 在这里, 我们需要一些实验数据来进行实例认证。

4.2.1 实验数据

本文中, 用于实例认证的数据主要是主题词表、IPC 和 UDC。主要选取了石油方面的一些概念。由于概念数据的信息量过大, 故只截取其中的一部分来引用。

其中石油的主题词表见表 4.1。

表 4.1 石油主题词表数据

中文	英文	属性	值(中文)
.....			
安全接头	SAFETY JOINT	C	钻柱
安全卡子	SAFETY CLAMP	S	安全设备
安全卡子	SAFETY CLAMP	C	钻井设备
安全设备	SAFETY EQUIPMENT	F	安全卡子 保险设备 报警装置 防碰天车装置 水击防护设备 消防设备
安全设备	SAFETY EQUIPMENT	S	设备

安全设备	SAFETY EQUIPMENT	C	安全 安全销 保护 预防措施
安全系数	SAFETY FACTOR	S	安全
安全系数	SAFETY FACTOR	C	设计 系数 载荷 (力)
安全销	SHEAR PIN	C	安全设备 保险 设备 闭锁
安山岩	ANDESITE	S	火山岩
安山岩	ANDESITE	C	中性岩
安装	MOUNTING	D	吊装
.....			

石油主题词表的数据来源是万方数据库。

石油 IPC 的原始数据¹如图 4.6 所示：

C01 无机化学
C01B 非金属元素；其化合物
C01C 氮；氨；其化合物
C01D 碱金属，即锂、钠、钾、铷、铯或铯的化合物
C01F 金属铍、镁、铝、钙、锶、钡、镭、钍的化合物，或稀土金属的化合物
C01G 含有未列入C01D或C01F小类之金属的化合物
C02 水、废水、污水或污泥的处理（沉淀池，过滤，如砂滤池或筛选装置入B01D）
C02F 水、废水、污水或污泥的处理
C03 玻璃；矿棉或渣棉
C03B 制造、成型及辅助工艺
C03C 玻璃、釉或搪瓷釉的化学成分；玻璃的表面处理；由玻璃、矿物或矿渣制成的纤维或细丝的表
C04 水泥；混凝土；人造石；陶瓷；耐火材料
C04B 石灰；氧化镁；矿渣；水泥；其组合物，例如砂浆、混凝土或类似的建筑材料；人造石；陶瓷
C05 肥料；肥料制造
C05B 磷肥
C05C 氮肥
C05D 未列入C05B、C小类的无机肥料；产生二氧化碳的肥料
C05F 未列入C05B、C小类的有机肥料、如用废物或垃圾制成的肥料
C05G 分属于C05大类下各小类中肥料的混合物；由一种或多种肥料与无特殊肥效的物质，例如农药、
C06 炸药；火柴

图 4.6 石油 IPC 原始数据

石油 UDC 的原始数据²如图 4.7 所示：

¹ <http://www.hczl.com/ipc/ipc.htm>

² <http://www.udcc.org/udccsummary/php/index.php?lang=chi>

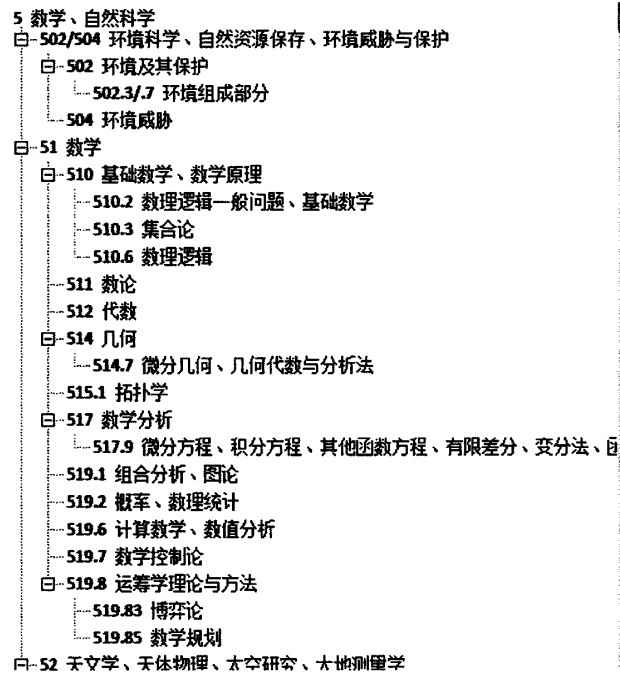


图 4.7 石油 UDC 原始数据

4.2.2 数据处理

将(1)中所提供的原始数据在开发系统中进行一系列的处理之后，得到标准化处理的数据。由于数据信息量过大，故在这里截取其中的一部分进行实验。处理之后得到的数据分别见表 4.1、表 4.2 和表 4.3。

表 4.1 石油主题词表数据

中文	英文	属性	值(中文)	值(英文)
爱克斯射线	X RAY	Y	X 射线	X RAY
果油井	OIL PRODUCING WELL	Y	油井	OIL WELL
安山岩	ANDESITE	S	火山岩	VOLCANIC ROCK
地热能	GEOTHERMAL ENERGY	C	热能	HEAT ENERGY
地理学	GEOGRAPHY	C	地貌学	PALEOGEOMORPHOLOGY
地理学	GEOGRAPHY	C	地质学	GEOLOGY
地沥青	ASPHALTUS	S	沥青	ASPHALT
石油开采	OIL PRODUCTION	Y	采油	OIL PRODUCING

油气开采	OIL AND GAS PRODUCTION	F	采气	GAS RECOVERY
油气开采	OIL AND GAS PRODUCTION	S	开采	EXPLOITATION
油气开采	OIL AND GAS PRODUCTION	C	高含水期	HIGH WATER CUT STAGE
油气开采	OIL AND GAS PRODUCTION	F	采油	OIL PRODUCING

表 4.2 石油 IPC 数据

中文	英文	pid (A 列的 id)	属性	值（中文）	值（英文）	id
煤气	GAS	C10	应用于	焦油的加工	WORKING-UP TAR	C10C
煤气	GAS	C10	应用于	焦油沥青的加工	WORKING-UP PITCH	C10C
煤气	GAS	C10	应用于	石油沥青的加工	WORKING-UP ASPHALT	C10C
煤气	GAS	C10	应用于	天然沥青的加工	WORKING-UP BITUMEN	C10C
煤气	GAS	C10	应用于	焦木酸	PYROLIGNEOUS ACID	C10C
焦油的加工	Working-up tar	1/00	是一种	水的去除	Removal of water	1/02
焦油的加工	Working-up tar	1/00	是一种	用蒸馏方法	by distillation	1/04
焦油沥青的加工	Working-up pitch	3/00	是一种	熔化	Melting	3/10
石油沥青的加工	Working-up asphalt	3/00	是一种	熔化	Melting	3/10
天然沥青的加工	Working-up bitumen	3/00	是一种	熔化	Melting	3/10
熔化	Melting	3/10	具备特征的	所用设备	Devices therefor	3/12

天然气	NATURAL GAS	C10L	是一部分	气体燃料	Gaseous fuels	3/00
气体燃料	Gaseous fuels	3/00	是一种 属性	含乙炔的组 合物	Compositions containing acetylene	3/02

表 4.3 石油 UDC 数据

中文	英文	pid (A 列的 id)	属性	值 (中文)	值 (英文)	id
自然科学	NATURAL SCIENCES	5	是一种	化学	Chemistry	54
自然科学	NATURAL SCIENCES	5	是一种	结晶学	Crystallography	54
自然科学	NATURAL SCIENCES	5	是一种	矿物学	Mineralogy	54
自然科学	NATURAL SCIENCES	5	是一种	大地测 量学	Geodesy	52
大地测量学	Geodesy	52	是一种	测量学	Surveying	528
测量学	Surveying	528	应用于	工程测量	Engineering survey	528.4
测量学	Surveying	528	应用于	各领域 测量	Special fields of surveying	528.4
化学	Chemistry	54	是一种	物质状态	State of substance	54.1
化学	Chemistry	54	是一种	化合物 种类	Particular kinds of compound	54.3
化学	Chemistry	54	是一种	化学品	Chemicals	54.4
化学	Chemistry	54	是一种	试剂	Reagents	54.4
化学	Chemistry	54	是一种	化学物质与 系统	Chemical substances and systems	54.01
矿物学	Mineralogy	549	应用于	硫化物	Sulphides	549.3
矿物学	Mineralogy	549	应用于	硫化盐	Sulpho-salts	549.3
矿物学	Mineralogy	549	应用于	卤化物	Halides	549.4
矿物学	Mineralogy	549	应用于	氧化合物	Oxygen compounds	549.5

将表 4.1、表 4.2 和表 4.3 的数据分类导入开发系统中，可分别得到一个可视化的图形和一个后缀名为“.stkos”的文件，其中“.stkos”是在系统开发中自定义的一个文件，即一个后缀名为“.diagram”的文件。其中开发系统所实现的可视化图形的工具如图 4.8 所示：

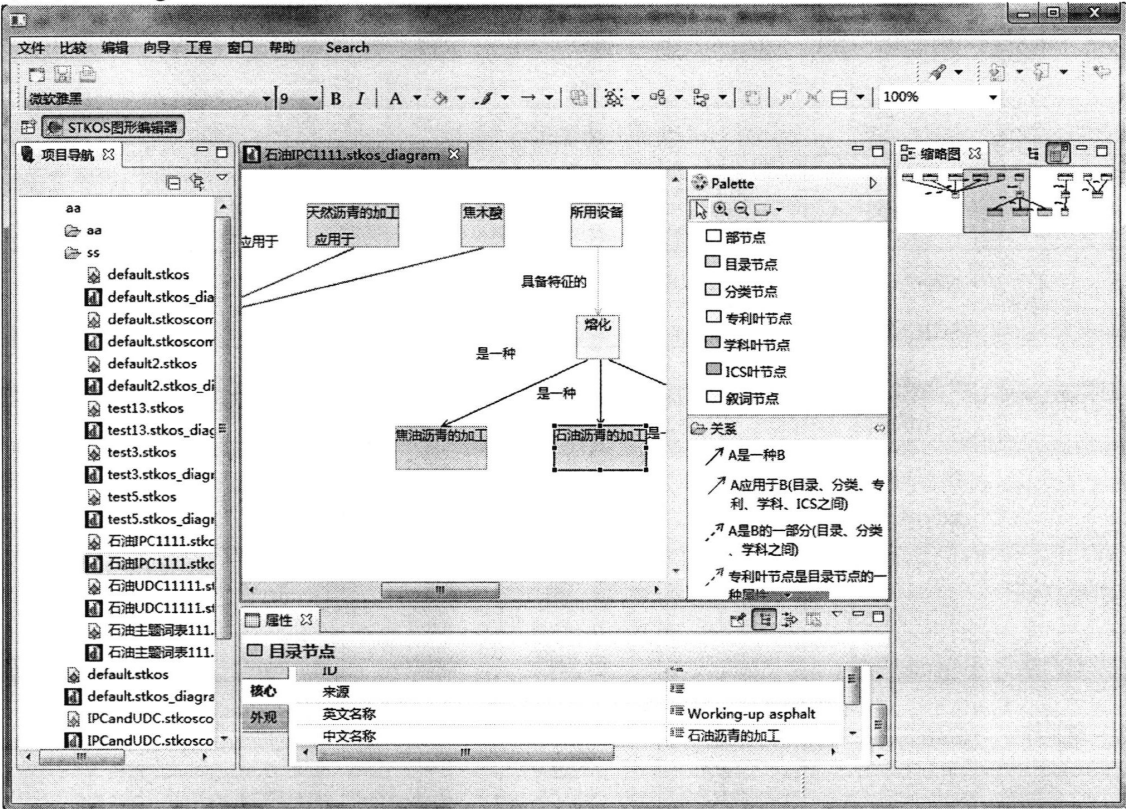


图 4.8 开发系统的可视化图形工具

图 4.8 中，在可视化编辑界面中，每个结点和关系的属性都会在编辑界面下方的“Properties”中显示其所有的属性，如中文名称、英文名称等。编辑界面的右方是显示的是每种结点和关系所表示的图形和箭头。例如：在编辑界面的右下角有一个被选中的结点，这样，我们可以看到在该结点的属性都在“属性”中。

其中的可视化图形分别如图 4.9、图 4.10 和图 4.11 所示：

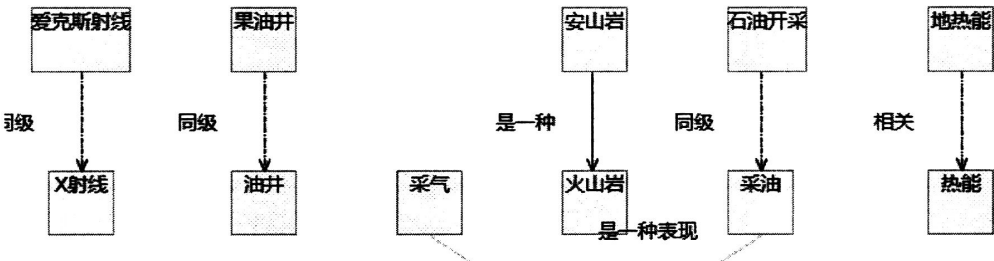


图 4.9 石油主题词表的可视化图形

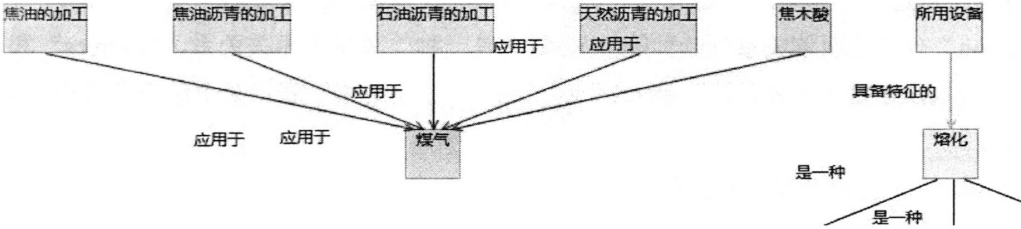


图 4.10 石油 IPC 的可视化图形

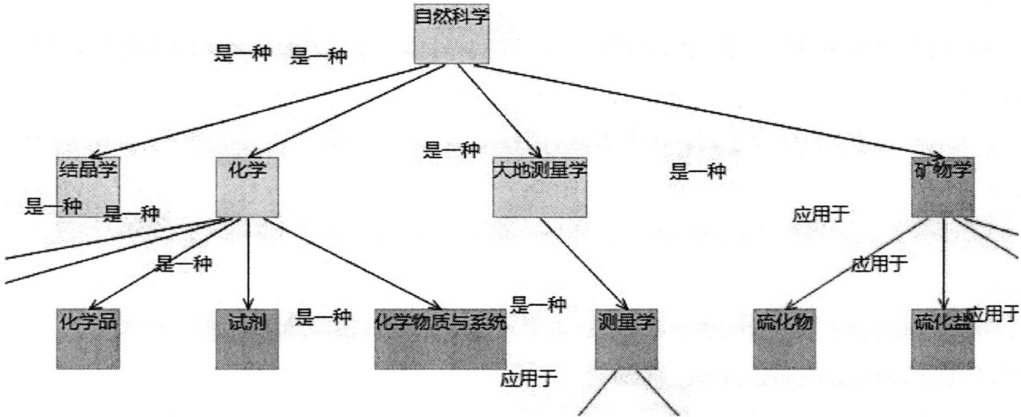


图 4.11 石油 UDC 的可视化图形

石油主题词表的“.stkos”文件片段如下：

.....

```
<haveNodes xsi:type="org.stkosproject.stkos:Descriptors" nameCN="采油"/>
```

```
<haveNodes xsi:type="org.stkosproject.stkos:Descriptors" nameCN="油气开采"/>
```

```
<haveNodes xsi:type="org.stkosproject.stkos:Descriptors" nameCN="采气"/>
```

```
<haveNodes xsi:type="org.stkosproject.stkos:Descriptors" nameCN="开采"/>
```

```
<haveNodes xsi:type="org.stkosproject.stkos:Descriptors" nameCN="高含水期"/>
```

```
<haveRelations xsi:type="org.stkosproject.stkos:Equivalence" name="同级" sources="//@haveNodes.0"
```

```
targets="//@haveNodes.1"/>
```

```
<haveRelations xsi:type="org.stkosproject.stkos:Equivalence" name="同级" sources="//@haveNodes.2"
```

```
targets="//@haveNodes.3"/>
```

```
<haveRelations xsi:type="org.stkosproject.stkos:KindOf" name="是一种" sources="//@haveNodes.4"
```

```
targets="//@haveNodes.5"/>
```

```
<haveRelations xsi:type="org.stkosproject.stkos:Correlation" name="相关" sources="//@haveNodes.6"
```

```
targets="//@haveNodes.7"/>
```

.....

其中，“nameCN”表示概念信息的中文名称，“haveNodes”说明该概念信息是一个结点，“xsi:type=‘org.stkosproject.stkos:Descriptors’”说明该结点是属于主题词表的，“havaRelations”

说明该概念信息是一个关系,“xsi:type= ‘org.stkosproject.stkos:Equivalence’”、“KindOf”和“Correlation”分别说明该关系是一种等价关系、“是一种”关系和相关关系,“sources”和“targets”分别表示这个关系的出发点和终点,“haveNode.+数字”表示某个结点在可视化图形中的 ID。

石油 IPC 的“.stkos”文件片段如下:

```
.....
<haveNodes xsi:type="org.stkosproject.stkos:PatentSons" nameCN="所用设备" nameEN="Devices
therefor"/>
<haveNodes xsi:type="org.stkosproject.stkos:Classifications" nameCN="天然气" nameEN="NATURAL
GAS"/>
<haveNodes xsi:type="org.stkosproject.stkos:Catalogs" nameCN="气体燃料" nameEN="Gaseous
fuels"/>
<haveNodes xsi:type="org.stkosproject.stkos:PatentSons" nameCN="含乙炔的组合物"
nameEN="Compositions containing acetylene"/>
<haveRelations xsi:type="org.stkosproject.stkos:Apliki" name="应用于" sources="//@haveNodes.1"
targets="//@haveNodes.0"/>
<haveRelations xsi:type="org.stkosproject.stkos:KindOf" name="是一种" sources="//@haveNodes.10"
targets="//@haveNodes.12"/>
<haveRelations xsi:type="org.stkosproject.stkos:Feature" name="具备特征的"
sources="//@haveNodes.13" targets="//@haveNodes.10"/>
<haveRelations xsi:type="org.stkosproject.stkos:PartOf" name="是一部分" sources="//@haveNodes.15"
targets="//@haveNodes.14"/>
<haveRelations xsi:type="org.stkosproject.stkos:Attribute" name="是一种属性"
sources="//@haveNodes.16" targets="//@haveNodes.15"/>
.....
```

其中,“xsi:type= ‘org.stkosproject.stkos:PatentSons’”、“Classifications”和“Catalogs”分别表示 IPC 中的专利叶结点、分类结点和目录结点,“Apliki”、“Feature”、“PartOf”和“Attribute”分别表示该关系是应用于、“具备特征的”、“是一部分”和“是一种属性”几种关系。

石油 UDC 的“.stkos”文件片段如下:

```
.....
<haveNodes xsi:type="org.stkosproject.stkos:SubjectSons" nameCN="卤化物" nameEN="Halides"/>
<haveNodes xsi:type="org.stkosproject.stkos:SubjectSons" nameCN="氧化合物" nameEN="Oxygen
compounds"/>
```

```
<haveRelations xsi:type="org.stkosproject.stkos:KindOf" name="是一种" sources="//@haveNodes.0"
targets="//@haveNodes.1"/>
```

```
<haveRelations xsi:type="org.stkosproject.stkos:KindOf" name="是一种" sources="//@haveNodes.0"
targets="//@haveNodes.2"/>
```

.....

其中，“xsi:type= ‘org.stkosproject.stkos:SubjectSons’ ”表示学科叶结点。

4. 2. 3 实验结果

在这里所说的结果即数据进过映射发现比较之后的结果。首先是三种体系的内容进行两两比较，石油主题词表和石油 IPC、石油主题词表和石油 UDC 及石油 IPC 和石油 UDC 的比较结果分别见表 4.4、表 4.5 和表 4.6。

表 4.4 石油主题词表和石油 IPC 的比较结果

中文	英文	ID	相似度	属性	中文	英文	ID
煤气	GAS	C10	0.511347	等价	爱克斯射线	X RAY	Descriptor
煤气	GAS	C10	0.511347	等价	X 射线	X RAY	Descriptor
煤气	GAS	C10	0.219645	等价	油井	OIL WELL	Descriptor
煤气	GAS	C10	0.374266	等价	火山岩	VOLCANIC ROCK	Descriptor
煤气	GAS	C10	0.583948	等价	热能	HEAT ENERGY	Descriptor
煤气	GAS	C10	0.142857	等价	地质学	GEOLOGY	Descriptor
煤气	GAS	C10	0.206922	等价	沥青	ASPHALT	Descriptor
煤气	GAS	C10	0.083333	等价	开采	EXPLOITATION	Descriptor
熔化	Melting	3/10	0.357062	等价	爱克斯射线	X RAY	Descriptor
熔化	Melting	3/10	0.357062	等价	X 射线	X RAY	Descriptor
熔化	Melting	3/10	0.081687	等价	油井	OIL WELL	Descriptor
熔化	Melting	3/10	0.076898	等价	火山岩	VOLCANIC ROCK	Descriptor
熔化	Melting	3/10	0.399455	等价	热能	HEAT ENERGY	Descriptor
熔化	Melting	3/10	0.076898	等价	沥青	ASPHALT	Descriptor
天然气	NATURAL GAS	C10L	0.181818	等价	爱克斯射线	X RAY	Descriptor
天然气	NATURAL GAS	C10L	0.181818	等价	X 射线	X RAY	Descriptor

天然气	NATURAL GAS	C10L	0.090909	等价	油井	OIL WELL	Descriptor
天然气	NATURAL GAS	C10L	0.269702	等价	火山岩	VOLCANIC ROCK	Descriptor
天然气	NATURAL GAS	C10L	0.181818	等价	热能	HEAT ENERGY	Descriptor
天然气	NATURAL GAS	C10L	0.272727	等价	沥青	ASPHALT	Descriptor
煤气	GAS	C10	0.219645	等价	果油井	OIL PRODUCING WELL	Descriptor
煤气	GAS	C10	0.374266	继承	安山岩	ANDESITE	Descriptor
煤气	GAS	C10	0.583948	相关	地热能	GEOHERMAL ENERGY	Descriptor
煤气	GAS	C10	0.142857	相关	地理学	GEOGRAPHY	Descriptor
煤气	GAS	C10	0.206922	继承	地沥青	ASPHALTUS	Descriptor
煤气	GAS	C10	0.083333	继承	油气 开采	OIL AND GAS PRODUCTION	Descriptor
熔化	Melting	3/10	0.081687	等价	果油井	OIL PRODUCING WELL	Descriptor
熔化	Melting	3/10	0.076898	继承	安山岩	ANDESITE	Descriptor
熔化	Melting	3/10	0.399455	相关	地热能	GEOHERMAL ENERGY	Descriptor
熔化	Melting	3/10	0.076898	继承	地沥青	ASPHALTUS	Descriptor
天然气	NATURAL GAS	C10L	0.090909	等价	果油井	OIL PRODUCING WELL	Descriptor
天然气	NATURAL GAS	C10L	0.269702	继承	安山岩	ANDESITE	Descriptor
天然气	NATURAL GAS	C10L	0.181818	相关	地热能	GEOHERMAL ENERGY	Descriptor

表 4.5 石油主题词表和石油 UDC 的比较结果

中文	英文	ID	相似度	属性	中文	英文	ID
化学	Chemistry	54	0.755751	等价	地质学	GEOLOGY	Descriptor
化学	Chemistry	54	0.165534	等价	开采	EXPLOITATION	Descriptor
结晶学	Crystallography	54	0.64447	等价	地质学	GEOLOGY	Descriptor
结晶学	Crystallography	54	0.13664	等价	开采	EXPLOITATION	Descriptor
矿物学	Mineralogy	54	0.971519	等价	地质学	GEOLOGY	Descriptor
矿物学	Mineralogy	54	0.13664	等价	开采	EXPLOITATION	Descriptor
测量学	Surveying	528	0.115268	等价	地质学	GEOLOGY	Descriptor
测量学	Surveying	528	0.281023	等价	开采	EXPLOITATION	Descriptor
化学	Chemistry	54	0.755751	相关	地理学	GEOGRAPHY	Descriptor
化学	Chemistry	54	0.165534	继承	油气开 采	OIL AND GAS PRODUCTION	Descriptor
结晶学	Crystallography	54	0.64447	相关	地理学	GEOGRAPHY	Descriptor
结晶学	Crystallography	54	0.13664	继承	油气开 采	OIL AND GAS PRODUCTION	Descriptor
矿物学	Mineralogy	54	0.13664	继承	油气开 采	OIL AND GAS PRODUCTION	Descriptor
测量学	Surveying	528	0.115268	相关	地理学	GEOGRAPHY	Descriptor
测量学	Surveying	528	0.281023	继承	油气开 采	OIL AND GAS PRODUCTION	Descriptor
自然科学	NATURAL SCIENCES	5	0.755751	继承	地质学	GEOLOGY	Descriptor
自然科学	NATURAL SCIENCES	5	0.165534	继承	开采	EXPLOITATION	Descriptor
物质状 态	State of substance	54.1	0.663908	继承	地质学	GEOLOGY	Descriptor
化合物 种类	Particular kinds of compound	54.3	0.663908	继承	地质学	GEOLOGY	Descriptor
化学品	Chemicals	54.4	0.663908	继承	地质学	GEOLOGY	Descriptor
试剂	Reagents	54.4	0.663908	继承	地质学	GEOLOGY	Descriptor
化学物 质与系 统	Chemical substances and systems	54.01	0.663908	继承	地质学	GEOLOGY	Descriptor

物质状态	State of substance	54.1	0.052412	继承	开采	EXPLOITATION	Descriptor
化合物种类	Particular kinds of compound	54.3	0.052412	继承	开采	EXPLOITATION	Descriptor
化学品	Chemicals	54.4	0.052412	继承	开采	EXPLOITATION	Descriptor
试剂	Reagents	54.4	0.052412	继承	开采	EXPLOITATION	Descriptor
化学物质与系统	Chemical substances and systems	54.01	0.052412	继承	开采	EXPLOITATION	Descriptor

表 4.6 石油 IPC 和石油 UDC 的比较结果

中文	英文	ID	相似度	属性	中文	英文	ID
煤气	GAS	C10	0.048137225	等价	化学	Chemistry	54
煤气	GAS	C10	0.041245488	等价	结晶学	Crystallography	54
煤气	GAS	C10	0.041245488	等价	矿物学	Mineralogy	54
煤气	GAS	C10	0.041245488	等价	测量学	Surveying	528
焦油的加工	WORKING-UP TAR	C10C	0.00472482	继承	化学	Chemistry	54
焦油沥青的加工	WORKING-UP PITCH	C10C	0.00472482	继承	化学	Chemistry	54
石油沥青的加工	WORKING-UP ASPHALT	C10C	0.00472482	继承	化学	Chemistry	54
天然沥青的加工	WORKING-UP BITUMEN	C10C	0.00472482	继承	化学	Chemistry	54
焦木酸	PYROLIGNEOUS ACID	C10C	0.00472482	继承	化学	Chemistry	54
水的去除	Removal of water	1/02	0.0004641	继承	化学	Chemistry	54
用蒸馏方法	by distillation	1/04	0.0004641	继承	化学	Chemistry	54
熔化	Melting	3/10	0.0004641	继承	化学	Chemistry	54

所用设备	Devices therefor	3/12	0.00002421	是一种特征	化学	Chemistry	54
煤气	GAS	C10	0.048137225	继承	自然科学	NATURAL SCIENCES	5
煤气	GAS	C10	0.00472482	继承	物质状态	State of substance	54.1
煤气	GAS	C10	0.00472482	继承	化合物种类	Particular kinds of compound	54.3
煤气	GAS	C10	0.00472482	继承	化学品	Chemicals	54.4
煤气	GAS	C10	0.00472482	继承	试剂	Reagents	54.4
煤气	GAS	C10	0.00472482	继承	化学物质与系统	Chemical substances and systems	54.01
煤气	GAS	C10	0.003493279	继承	硫化物	Sulphides	549.3
煤气	GAS	C10	0.003493279	继承	硫化盐	Sulpho-salts	549.3
煤气	GAS	C10	0.003493279	继承	卤化物	Halides	549.4
煤气	GAS	C10	0.003493279	继承	氧化合物	Oxygen compounds	549.5

将表 4.1、表 4.2 和表 4.3 的数据在开发系统中两两进行映射比较，得出了表 4.4、表 4.5 和表 4.6 的内容，且在开发系统中可分别得出一个可视化的图形和一个后缀名为“.stkoscomparator”的文件。

其中的可视化图形分别如图 4.12、图 4.13 和图 4.14 所示：

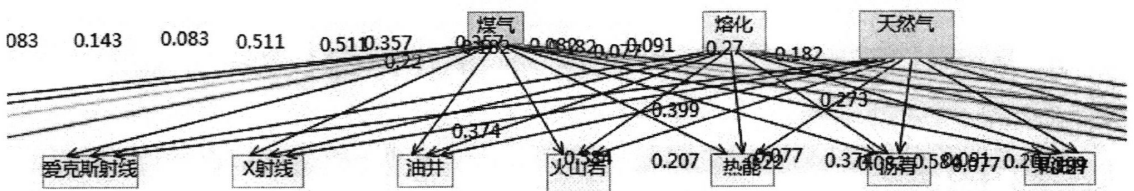


图 4.12 石油主题词表和石油 IPC 比较后的可视化图形

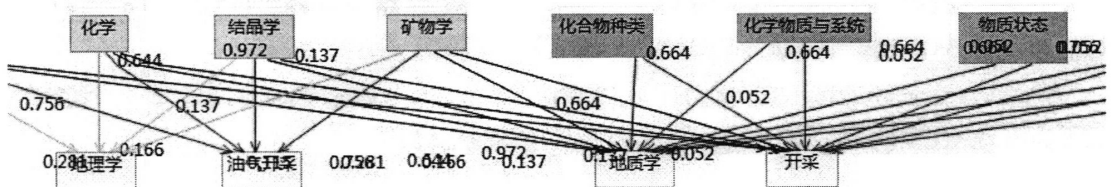


图 4.13 石油主题词表和石油 UDC 比较后的可视化图形

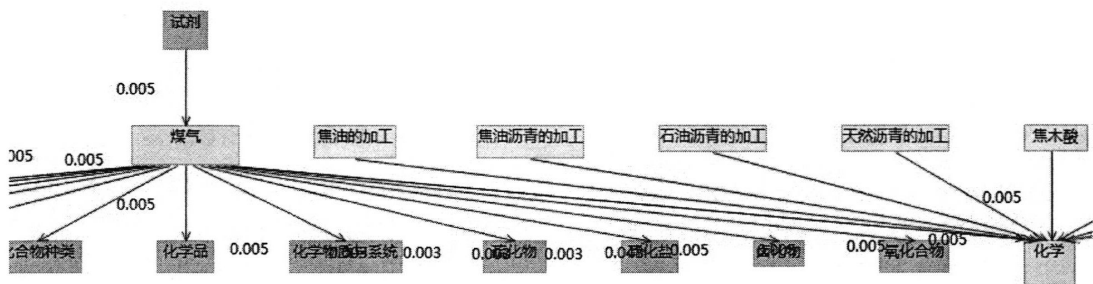


图 4.14 石油 IPC 和石油 UDC 比较后的可视化图形

石油主题词表和石油 IPC 比较后的“.stkoscomparator”的文件片段如下：

```
.....
<haveCompareNodes xsi:type="org.stkosproject.stkoscomparator:Descriptors" nameCN="地理学"
nameEN="GEOGRAPHY" description=""/>
<haveCompareNodes xsi:type="org.stkosproject.stkoscomparator:Descriptors" nameCN="地沥青"
nameEN="ASPHALTUS" description=""/>
<haveCompareNodes xsi:type="org.stkosproject.stkoscomparator:Descriptors" nameCN="油气开采"
nameEN="OIL AND GAS PRODUCTION" description=""/>
<haveCompares xsi:type="org.stkosproject.stkoscomparator:EqualCompare" similarity="0.511347094"
sources="///@haveCompareNodes.0" targets="///@haveCompareNodes.1"/>
<haveCompares xsi:type="org.stkosproject.stkoscomparator:EqualCompare" similarity="0.511347094"
sources="///@haveCompareNodes.0" targets="///@haveCompareNodes.2"/>
.....
```

其中，“haveCompareNodes”表示该概念是一个结点，“haveCompares”表示该信息是一个比较的关系，“EqualCompare”表示该关系是等价关系。

石油主题词表和石油 UDC 比较后的“.stkoscomparator”的文件片段如下：

```
.....
<haveCompareNodes xsi:type="org.stkosproject.stkoscomparator:SubjectSons" nameCN="试剂"
nameEN="Reagents" source=""/>
<haveCompareNodes xsi:type="org.stkosproject.stkoscomparator:SubjectSons" nameCN="化学物质与系统"
nameEN="Chemical substances and systems" source=""/>
<haveCompares xsi:type="org.stkosproject.stkoscomparator:EqualCompare" similarity="0.755751299"
sources="///@haveCompareNodes.0" targets="///@haveCompareNodes.1"/>
<haveCompares xsi:type="org.stkosproject.stkoscomparator:EqualCompare" similarity="0.165534031"
sources="///@haveCompareNodes.0" targets="///@haveCompareNodes.2"/>
.....
```

石油 IPC 和石油 UDC 比较后的“.stkoscomparator”的文件片段如下：

```

.....
<haveCompareNodes xsi:type="org.stkosproject.stkoscomparator:PatentSons" nameCN="用蒸馏方法"
nameEN="by distillation" feature=""/>
<haveCompareNodes xsi:type="org.stkosproject.stkoscomparator:PatentSons" nameCN="熔化"
nameEN="Melting" feature=""/>
<haveCompareNodes xsi:type="org.stkosproject.stkoscomparator:PatentSons" nameCN="所用设备"
nameEN="Devices therefor" feature=""/>
<haveCompares xsi:type="org.stkosproject.stkoscomparator:EqualCompare"
similarity="0.0481372253428382" sources="//@haveCompareNodes.0" targets="//@haveCompareNodes.1"/>
<haveCompares xsi:type="org.stkosproject.stkoscomparator:InheritCompare"
similarity="0.0481372253428382" sources="//@haveCompareNodes.0" targets="//@haveCompareNodes.5"/>
<haveCompares xsi:type="org.stkosproject.stkoscomparator:FeatureCompare" similarity="2.421E-5"
sources="//@haveCompareNodes.23" targets="//@haveCompareNodes.1"/>
.....

```

4.3 本章小结

本章节主要是通过实验来认证本文的研究结果的。由于分类体系繁多以及数据的信息量过大，本章只是选取了一个领域并截取了部分数据来进行实例认证，所选取的领域是石油领域，截取的数据是石油主题词表以及 IPC 和 UDC 在石油领域的的数据。在这里，不仅得出了数据的模型化格式，各数据的比较结果，也实现了它们的可视化图形，可以说是很好的达到了研究内容的预期效果。

第 5 章 总结与展望

由于科技文献知识服务需求的日益增长,需要后台的科技文献标引深度和准确程度不断增加,如何进行高效率的知识抽取成为了目前知识服务系统开发建设的难点。近年来,构建统一的概念空间成了克服这一问题的关键环节。在本文中,所研究的内容如下:

根据主题词表及分类体系来构建统一的概念空间,并在此基础之上设计映射算法。主要从理论研究和开发研究两个方面来解决问题的。

在理论方面,由于现在信息量呈指数级增加,出现了繁多的分类体系。而在本文中,挑选了其中用途范围比较广的分类体系和主题词表来分析研究的。根据分析各分类体系和主题词表的异同,研究多维度的科技文献概念空间,并进行总结其特点,得出不同的结点和关系,构建统一的概念空间,以便更好的设计用于概念空间关联的多渠道概念匹配算法,从而设计并开发面向科技文献概念空间建立和关联的原型工具。设计关联的匹配算法,并综合应用基于字符的概念匹配方法(数据字典查询、编辑距离比较等)、基于结构的概念匹配方法和基于语义推理的概念匹配方法,可以在构建的统统一的概念空间中最大程度的自动的、准确的发现概念之间的关联。此算法在开发系统上得到了验证,其效果良好。

在开发方面,根据 Eclipse 及其插件所提供的功能,将构建的公共数据模型在 Eclipse 中用 Ecore 来描述,并驱动生成 CDM 模型。编码将概念模型化,并实现数据导入功能和关联概念的可视化图形。根据设计的关联算法来对导入的数据进行比较,得出关联概念的等价相似度、继承相似度等,以实现数据比较功能,关联 Eclipse 平台的数据导出功能,利用 Eclipse 的插件 RCP 集成出一个可实现数据导入、数据比较和数据导出三个功能的简化的 RCP 产品。在集成的 RCP 产品中,实现了本文研究的几个功能,其中可视化的图形可以很直观的在产品的图形编辑界面中看到。

本论文在理论方面和开发方面都取得了一定的成果,但由于本人的精力和能力有限,研究内容中还有很多不足的地方,有待进一步提高。首先,由于时间有限,而本文所研究分析的分类体系又比较多,故在构建统一的概念空间上,仅根据各分类体系的基本属性进行的分析和总结,从而构建的统一数据模型基本可是代表所提到的主题词表和分类体系。而在设计关联算法上,由于各概念都比较自然化,且信息量大,可以实现基本精确,但在时间复杂度上不够,故该关联算法有待进一步优化。

致 谢

岁月如梭，如歌。转眼之间，两年半的研究生求学生活即将结束，站在毕业的门槛上，回首往昔，奋斗和辛劳成为丝丝的记忆，甜美与欢笑也都尘埃落定。回顾近三年研究生时光，本人在学习、生活和思想上都有了很大的提高，这里就要感谢我的老师、同学和家人，是他们教导和帮助下，自己才有了如此大的进步。在毕业论文即将完成之际，我谨向所有关心、爱护、帮助我的人们表示最诚挚的感谢与最美好的祝愿。

首先，我要感谢平日里指导我的导师姚英彪教授。他拥有丰厚的教学经验，浓厚的专业学术造诣，能及时把握最新的学术动态，并指导我对最新课题展开深入研究。同时，他经常与我们在一起进行科研工作，有时工作到很晚，这种崇高的敬业精神和忘我的工作热情深深感染了我，给我留下了深刻的印象。在科研中遇到困难的时候，他时常鼓励我们要有信心，不要放弃，要想办法解决问题。只有坚持到最后，才能笑到最后。在生活上，他很和蔼可亲，给予我无微不至的关怀。尽管自己在异地学习，但也让我感受到如家人般的温暖，在精神生活上感受到极大的鼓舞。姚老师的悉心指导、热情鼓励及科研上严格要求，促使我更加奋发学习，钻研学术课题，并在理论研究和实践等各方面都取得了极大提高，很大程度上调动了我的科研积极性。这里，我要向姚老师表达最诚挚的谢意。

同时，我也要感谢李光球、章坚武、孙文胜、张品等曾经教导过我的老师们。你们在我读研期间传授给我很多专业领域的知识，教会我如何独立思考问题，让我受益匪浅。

其次，还要感谢我的室友李伟，及好友彭其胜、陈宏炳、吴超、郑立等。尽管在一起只有短短两年半时间，但你们在学习和生活上对我的关心和帮助却是如何也无法忘记的。因为有你们在身边支持我、鼓励我，才使研究生生活更加丰富。我们的友谊弥足珍贵，相信这份友谊一定会永远持续下去。

另外，感谢实验室里与我一起共事的同学。赵广佩、曾宪宾、徐强、王博、李成霞、池景秀，我们一起学习、一起进步，就像一个大家庭一样。平日里积极组织课外活动，相互沟通和交流，共同成长！

在感谢老师和同学的同时，我还要感谢我所在实习单位的老师和同事。在研二期间，通过姚老师，我很庆幸的能在中国科学院科技政策与管理科学研究所实习。在这里，我接触到了很多在学校没有的事和人，同时参加了面向外科技知识组织体系的大规模语义计算关键技术研究“十二五”科技支撑计划（项目编号：2011BAH10B04-04）这一项目。很好的应用了在学校所学的知识，在郭剑锋副研究员的悉心指导下，我在理论研究和实践等各方面都取得了极大提高。在生活方面，得到了郭老师、室友及其他同事的很多帮助，使我很快的、很好的融入了课题组这个大家庭中。故在此，我要向郭老师、同事和室友表达最诚心的谢意！

最后，感谢我的家人。你们一直支持我勇攀知识高峰。在失意时鼓励我，在得意时教导我，教我如何做人，如何面对人生，如何为自己目标而不懈努力。感谢你们对我二十多年来的养育之恩，你们把希望寄托在我身上，我不会辜负你们。我会在今后工作生活中自强不息、勇于开拓，回报你们无私的爱和奉献！

再一次感谢给予我指导和帮助的老师、同学、亲友们！

另外，还必须要向在百忙之中抽出宝贵时间来评阅和参加答辩的所有专家、教授表示由衷的感谢！

参考文献

- [1] 陆建江, 张亚非, 苗壮等. 语义网原理与技术. 北京: 科学出版社, 2007, 151-160.
- [2] 邓志鸿, 唐世渭, 张铭, 杨冬青, 陈捷. Ontology 研究综述. 北京大学学报 (自然科学版), 第 38 卷, 第 5 期, 730-738, 2002.
- [3] Neches R, Fikes R E, Gruber T R, et al. Enabling Technology for Knowledge Sharing. AIMagazine, 1991, 12(3):36-56
- [4] Gruber T R. A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition, 1993, 5:199-220
- [5] 邓珞华 等. 图书情报数学. 东北师范大学出版社, 长春, 1983: 26
- [6] Gruber TR. A translation approach to portable ontology specifications. Technical Report, KSL 92-71, Knowledge System Laboratory, 1993.
- [7] Deng ZH, Tang SW, Zhang M, Yang DQ, Chen J. Over View of ontology. Acta Scientiarum Naturalium Universitatis Pekinensis, 2002, 38(5):730-738(in Chinese with English abstract).
- [8] 杜小勇, 李曼, 王珊. 本体学习研究综述. 软件学报. 2006, 17(09): 1837-1847.
- [9] Stojanovic L, Maedche A, Motik B, Stojanovic N. User-Driven Ontology Evolution Management. LECTURE NOTES IN COMPUTER SCIENCE, 2002, 2473: 133-140
- [10] Niles I, Pease A. Towards a standard upper ontology. Proceedings of the international conference on Formal Ontology in Information Systems-Volume 2001. 2001: 2-9.
- [11] Rodríguez H, Climent S, Vossen P, Bloksma L. The Top-Down Strategy for Building EuroWordNet: Vocabulary Coverage, Base Concepts and Top Ontology. Computers and the Humanities, 1998, 32(03): 117-152
- [12] Stumme G, Maedche A. FCA-MERGE: Bottom-Up Merging of Ontologies[J]. INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE. 2001, 17(1): 225-234.
- [13] Niles I, Pease A. Origins of the IEEE Standard Upper Ontology[J]. Working Notes of the IJCAI-2001 Workshop on the IEEE Standard Upper Ontology. 2001: 37-42.
- [14] Agirre E, Ansa O, Hovy E, Martinez D. Enriching very large ontologies using the WWW. In: Staab S, Maedche A, eds. Proc. of the ECAI 2004 Workshop on Ontology Learning. 2000. <http://ol2000.aifb.uni-karlsruhe.de/>
- [15] Missikoff M, Navigli R, Velardi P. Integrated approach for web ontology learning and engineering. IEEE Computer, 2002, 35(11): 60-63.

- [16] Uschold M. Building Ontologies. Towards A Unified Methodology. In expert systems 96, 1996
- [17] Ontolingua. <http://www.ksl.stanford.edu/software/ontolingua/>
- [18] CycL. <http://www.cyc.com/tech.html#cycl>
- [19] Loom. <http://www.isi.edu/isd/LOOM/LOOM-HOME.html>
- [20] Farquhar A, Fikes R, Rice J. The Ontolingua server: A tool for collaborative ontology construction. *Int'l Journal of Human-Computer Studies*, 1997, 46(6):707-727.
- [21] Swartout B, Ramesh P, Knight K, Russ T. Toward distributed use of large-scale ontologies. In: *Proc. of the AAAI Symp. on Ontological Engineering*. 1996.
http://ksi.cpsc.ucalgary.ca/KAW/KAW96/swartout/Banff_96_final_2.html
- [22] Duineveld AJ, Stoter R, Weiden MR, Kenepa B, Benjamins VR. Wonder tools? A comparative study of ontological engineering tools. *Int'l Journal of Human-Computer Studies*, 2000, 52(6):1111-1133.
- [23] Noy NF, Fergerson RW, Musen MA. The knowledge model of protégé-2000: Combining interoperability and flexibility. In: Dieng R, Corby O, eds. *Proc. of the EKAW 2000*. Heidelberg: Springer-Verlag, 2000. 17-32.
- [24] Arpirez JC, Corcho O, Fernandez-Lopez M, Gomez-Perez A. WebODE: A scalable ontological engineering workbench. In: Gil Y, Musen M, Shavlik J, eds. *Proc. of the K-CAP 2001*. New York: ACM Press, 2001. 6-13.
- [25] Bechhofer S, Horrocks I, Goble C, Stevens R. OilEd: A reason-able ontology editor for the semantic Web. In: Baader F, Brewka G, Eiter T, eds. *Proc. of the KI 2001, Joint German/Austrian Conf. on AI*. Heidelberg: Springer-Verlag, 2001. 396-408.
- [26] Sure Y, Angele J, Erdmann M, Staab S, Studer R, Wenke D. OntoEdit: Collaborative ontology engineering for the semantic Web. In: Horrocks I, Hendler JA, eds. *Proc. of the ISWC 2002*. Heidelberg: Springer-Verlag, 2002. 221-235.
- [27] Bozsak E, Ehrig M, Handschuh S, Hotho A, Maedche A, Motik B, Oberle D, Schmitz C, Staab S, Stojanovic L, Stojanovic N, Studer R, Stumme G, Sure Y, Tane J, Volz R, Zacharias V. KAON-Towards a large scale semantic web. In: Bauknecht K, Mintjoa A, Quirchmayr G, eds. *Proc. of the 3rd Int'l Conf. on E-Commerce and Web Technologies*. Heidelberg: Springer-Verlag, 2002. 304-313.
- [28] Lenat DB. CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 1995, 38(11):33-38.
- [29] Nierenburg S, Beale S, Mahesh K, Onyshkevych B, Raskin V, Viegas E, Wilks Y, Zajac R. Lexicons in the mikrokosmos project. In: Busa F, Johnston M, eds. *Proc. of the AISB Workshop on Multilinguality in the Lexicon*. Brighton: Ulrich Heid, 1996. 26-33.

- [30] Maedche and S. Staab. Semi-automatic engineering of ontologies from texts. In Proceedings of the 12th International Conference on Software Engineering and Knowledge Engineering (SEKE 2000), Chicago, IL, USA, pages 231–239, July 2000.
- [31] Calvanese, G. De Giacomo, and M. Lenzerini. Ontology of integration and integration of ontologies. In Proceedings of the 9th International Conference on Conceptual Structures (ICCS'01), Stanford, CA, USA, August 2001
- [32] Madhavan, P.A. Bernstein, P. Domingos, and A. Halevy. Representing and reasoning about mappings between domain models. In Proceedings of the 18th National Conference on Artificial Intelligence (AAAI'02), Edmonton, Alberta, Canada, August 2002
- [33] Kiryakov, K. Simov, and M. Dimitrov. OntoMap: Portal for Upper-Level Ontologies. In Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS'01), Ogunquit, Maine, USA, October 2001.
- [34] Kent. The Information Flow Foundation for Conceptual Knowledge Organization. In Proceedings of the 6th International Conference of the International Society for Knowledge Organization (ISKO), Toronto, Canada, August 2000.
- [35] Han J, Kamber M, 著. 范明, 孟小峰, 等, 译. 数据挖掘概念与技术. 北京: 机械工业出版社, 2001.
- [36] 曹树金. 《杜威十进制分类法》的电子化及未来研究重点. 图书馆, 1995(5)
- [37] 曹树金, 严丽君, 汪东波. DDC、LCC、UDC 网络版评价[J]. 中国图书馆学报, 2002(6): 61-65. (Cao Shujin, Yan Lijun, Wang Dongbo. DDC, LCC and UDC: a Comparative Study of Their Online Versions[J]. Journal of library science in China, 2002(6): 61-65.)
- [38] Tony Olson. The Integration of Information languages and Interoperability[EB/OL]. [2011-07-20]. <http://www.ala.org/ala/lita/litamembership/litaigs/2002authcontrol.pdf>.
- [39] 黄筱玲. 高校图书馆文献分类工作改革的思考——兼谈《中图法》和《科图法》的兼容[J]. 图书馆学研究, 2001(4): 29-31. (Huang Xiaoling. Some thoughts of document classific reforms of the university library[J]. Researches in Library Science, 2001(4): 29-31.)
- [40] 戴剑波, 侯汉清. 文献分类法自动映射系统的构建——以《中国图书馆分类法》和《杜威十进制分类法》为例[J]. 情报学报, 2006, 25(5): 594-599. (Dai Jianbo, Hou Hanqing. Construction and Use of Automatic Mapping System Between CLC and DDC [J]. Journal of The China Society For Scientific and Technical Information, 2006, 25(5): 594-599.)
- [41] 戴剑波, 侯汉清. 图书分类法映射系统设计原理——以《中国图书馆分类法》和《杜威十进制分类法》为例[J]. 情报学报, 2005, 24(3): 299-303. (Dai Jianbo, Hou Hanqing. Principle of the Automatic Mapping System of Library Classifications—Take CLC and DDC as the Example[J]. Journal of The China Society For Scientific and Technical Information, 2005, 24(3): 299-303.)

- [42] 马磊, 宋建玮. IPC 分类法在科技查新工作中的应用. 图书馆学刊
- [43] F. Baader, D. Calvanese, D. Nardi, and P. F. Patel-Schneider, editors. The Description Logic Handbook: Theory, Implementation, and Applications. Cambridge University Press, 2003
- [44] Tim Berners-Lee. Mark Fischetti. 张宇宏, 萧风译. 编织万维网——万维网之父谈万维网的原初设计与最终命运[M]. 上海: 上海译文出版, 1999:154-171
- [45] SA McIlraith, TC Son, H Zeng. Semantic web services[J].New York:IEEE Intelligent Systems, 2001(2):46-53
- [46] 王继成, 萧嵘. Web 信息检索研究发展[J]. 计算机研究与发展, 2001:38 (2) 187-193
- [47] J. Farrugia. Model-theoretic semantics for the Web. In: Proc of the twelfth international conference on World Wide Web. New York:ACM Press, 2003,82-85
- [48] 唐杰, 梁邦勇, 李涓子等. 语义 Web 中的本体自动映射. 计算机学报, 2006, 29(11):1956-1976
- [49] A. Doan. Learning to Map between Structured Representations of Data: [dissertation]. Washington: University of Washington, 2002,18-21
- [50] 邹瑾, 张燕飞. UDC、DDC 网络化发展对中图法电子版应用及推广的启示[J]. 图书情报知识, 2004, (4):69-70.

附 录

参加的科研项目

- [1] 面向外科技知识组织体系的大规模语义计算关键技术研究“十二五”科技支撑计划（项目编号：2011BAH10B04-04）2011.10-2014.06

已录用论文

- [1] 王波. 面向 STKOS 的概念映射及其系统实现. 南阳理工学院学报

面向STKOS的概念映射与关联算法研究及其实现

作者:

王波

学位授予单位:

杭州电子科技大学

本文链接: http://d.g.wanfangdata.com.cn/Thesis_Y2289846.aspx