基于形式概念分析的情报学领域本体构建

赵洋飞

(大连海事大学 信息科学与技术学院 辽宁 大连 116026)

摘 要 随着语义网的快速发展，本体构建已经成为了语义网应用的重要研究领域。但目前该领域研究还处于探索阶段，没有形成成熟、统一的方法作为指导。本文引入形式概念分析的理论方法，探讨如何将形式概念分析应用于本体构建中。首先分析了本体与形式概念分析的联系，接着对基于形式概念分析构建领域本体的代表方法进行了综合对比，最后结合情报学领域本体构建实例对形式概念分析在本体构建中的具体方法进行了说明。

关键词 本体 形式概念分析 概念格

中图法分类号 G350

Construction of Domain Ontology in Information Science Based on Formal Concept Analysis

Zhao Yangfei

（College of Information Science and Technology, Dalian Maritime University, Liaoning Dalian, 116026, China）

Abstract With the rapid development of semantic web and the ever increasing need for ontology,ontology development has been regarded as one of the most important fields in the semantic web related research work. However, there is no mature methodology to guide ontology construction in this paper we present an ontology learning method that is based on formal concept analysis, which is a theory of data analysis which identifies conceptual structures among data sets . Firstly, the relations between ontology and FCA are described, and then several methods of ontology building based on FCA have been analyzed. Finally we demonstrate the role of FCA in ontology construction on information science.

Key words Ontology; Formal concept analysis; Concept Lattice

1 引言

本体（ontology）是共享概念模型的明确的形式化的规范说明[1]，领域本体（domain ontology）是专业性的本体，提供了某个专业学科领域中概念的词表以及概念间的关系[2]。本体作为共享概念的形式化说明，用于表达数据源的语义、识别和建立概念间的语义关联达成语义一致，提供了语义异构问题的解决途径[3]。近年来本体的功能已经被人们逐渐认识，并得到越来越广泛的应用。但是本体的构建方法，尤其是自动或半自动的构建方法，目前尚处于探索阶段，仍未彻底消除构建过程中人的主观因素的影响。

种建立在数学基础之上，从形式背景进行数据分析和规则提取的工具[4]。FCA 强调用数学手段来表达客观知识，可以削弱开发者对领域本体构建过程的主观影响， 并能挖掘出领域中隐含的概念以及概念之间的层次关系。FCA 用数学符号从内涵和外延两方面表示所有概念，达到了形式化概念模型的效果[5]，因此基于 FCA 的领域本体构建方法越来越受到国内外相关学者的关注和青睐。运用 FCA,通过构造概念格（Concept Lattice），半自动地构建本体，可以综合领域的各种类型的资源，是一种有效的领域本体构建方法[6]。

由于领域本体构建的传统方法中，本体层次结构的形成受主观因素影响较大，而运用 FCA 可以较好地解决这个问题，有助于实现领域本体的半自动构建。因此本文就领域本体构建中 FCA的运用进行了实证研究，并获得良好的效果。

2 形式概念分析及其与本体的关联

形式概念分析是应用数学的一个分支，它建立在概念和概念层次的数学化基础之上，根据用二元关系表达的形式背景，从中提取概念层次结构，即概念格（ Concept Lattice ）[7]。概念格的每个节点就是一个概念，由两部分组成：外延（Extension），即概念所覆盖的实例；内涵（Intension），即概念的描述该概念覆盖实例的共同特征[8]。概念格是格结构的一种，格结构是一种典型的代数结构。也就是说，概念格就是一种典型的代数结构。本体对领域主题进行规范的、公认的描述，提供了用来表达和交流主题知识的概念以及概念间的关系。概念间的分类关系（ Kind-of 关系）呈现一种偏序结构，所有的分类关系集合呈现出格结构。由此可知，本体概念间的分类系也是一种典型的代数结构。综合形式概念分析与本体来看，在代数结构的视角下，本体概念的分类关系及领域本体的概念格都表现为一种代数结构，也就是格结构。二者之间的这种共性从本质上揭示了二者能有紧密联系的根本原因，即具有相同的代数结构。这种相同的代数结构使得二者之间很容易产生一种映射关系，即可以通过映射规则将概念格映射成本体概念分类关系集合的格结构， 而且将概念格节点、节点属性、节点对象和节点间联系分别映射成领域本体概念、概念属性集、概念实例和概念分类关系的映射规则将是最合理的映射规则。映射规则如图 1 所示。

综上所述，FCA 能与本体相联系的原因包括：

①相联系的根本原因：都是对概念及概念之间关系的描述；②能有紧密联系的本质原因： 都具有相同的代数结构，即格结构；③都是形式化的工具和方法；④在哲学、 信息科学、 概念知识处理及知识表示等各个层次上的应用中都有共同的应用领域等。在实际应用中，FCA 和本体作为两种形式化方法，具有一定的相似性。它们都强调概念主体间一致的重要性，也都强调模型形式说明的必要，因此可作为一种学习技术用于支持本体的构建。FCA 与本体结合的

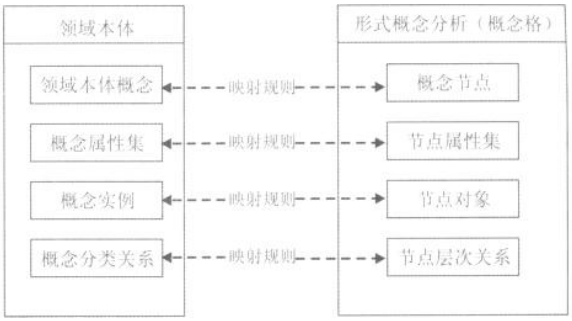

图 1 领域本体与概念格的映射关系

方式包括：作为描述工具来发掘和描述领域核心概念；作为分析工具来分析领域本体概念分类关系、进行领域本体概念间类属关系的推理、完成领域本体重构、实现领域本体复用；作为建模工具来构建领域本体概念层次的可视化模型以及本体原型。在运用 FCA 构建本体的过程中，虽然开发过程依然离不开人的因素，但概念格作为本体的构建方式，清楚表达了概念以及概念之间的关系，而且容易为人们所理解，在一定程度上消除了主观影响。

3 基于形式概念分析的本体构建方法

3.1 四种方法简介

本体应用的基础在构建本体，而构建本体的关键在于确定了领域之后，从领域中找到概念以及概念之间的关系。然而令研究人员感到困难的是，这种概念之间的关系大都是隐含在人的头脑之中的。如何通过机器自动或半自动地获取这些概念及其之间的相互关系，从而构建领域本体，许多国外学者已将目光投向形式概念分析，包括 Cimiano, GuTao, Haav, Obit-ko 等 。Cimiano 方法[9-10]是采用 FCA 分析词语在文本中的使用方式，来获得形式背景，从而生成领域本体。其基本思想为：①用自然语言解析器对领域文本的每一个句子进行解析，获得语法树；②由此语法树得到动词对象之间的依赖关系；③通过词典查询，对所提取的动词与对象用词进行规范化表示；④将 FCA 中的概念与领域本体的概念直接等同，获得概念格，概念格得到本体。GuTao 方法[11]是通过采用 Protege（可视化本体构建工具）、ConExp（概念格工具）及其本人开发的 FcaTab（ Protege 工具插件） 这三大工具，通过迭代的过程半自动地获得领域本体。其关键步骤为：①手动或者以自然语言处理技术，从领域文本获得概念和属性；②用 Protege 对上述概念和属性建模，以类（class）、槽（slots）、分面（facets）来表示领域本体；③ 用 FcaTab 产生形式背景，并转化成 ConExp 所要求的输入格式；④ 用 ConExp 建立概念格；⑤ 重复步骤 ③④，直到获得满意的本体。Haav 方法[6, 10]主要适用于领域文本内容较短的情况，且假设领域文本描述了某一实体，其中包含了

描述领域的术语。其基本思想是将 FCA 与基于规则的语言相结合，半自动地构建领域本体。关键步骤为：① 从领域数据或文本中提取形式背景，构建形式背景三元组；② 由形式背景生成概念格，并进行概念格约减，作为初始本体；③建立 FCA 与规则语言的映射关系，将初始本体移植成用一阶谓词逻辑表示的集合；④ 添加规则和事实扩充初始本体；⑤本体推理。Obitko 方法由 Marekobitko 等人在 GACR 项目中提出，其基本原理是，概念由属性描述；属性决定概念的层次结构；当两个概念的属性相同时，这两个概念也相同；本体的生成基于概念格对概念及属性的可视化，本体原型可由概念格映射得到。基本步骤为：①以概念和属性的空集开始；② 按需向概念表中添加概念和属性；③用 FCA 对概念及其属性的表格进行可视化——使得本体构建者以可视化的方式观察本体或其部分；④ 基于可视化结果，对其进行规范化，方法包括“ 直接” 编辑（添加或移除概念或属性，为概念分配属性）和根据本体构建工具的建议编辑（合并相同位置的概念或命名由属性产生的新概念）两种；⑤ 重复上述过程直到获得满意的本体。

3.2 Obitko方法的优势

Obitko 方法的优点包括：① 提供了分布式的本体编辑环境，实现了领域本体的分布式开发；② 按属性进行分类，倡导将领域本体概念分类关系作为领域本体概念层次关系的重点，克服了当前分类方法存在的问题； ③ 提出了一整套对形式背景和概念格的编辑修改机制，值得借鉴；④ 实现了可视化基础上的概念格编辑。将 Cimiano 方法、 GuTao 方法、 Haav 方法、Obitko 方法的优劣势进行对比，可以发现 Obitko 方法更有一些突出的优势：① 本体语义丰富度最高；② 可以支持分布式协作开发；③ 形式背景最为合理；④ 领域数据来源全面性最好，可来源于领域专家的隐形知识，而非仅为纯文本；⑤相关技术确定性高，这一点上优于 GuTao 方法、Haav 方法；⑥背景和概念格的编辑修改机制相对健全，这一点上优于 Cimiano 方法、Haav 方法；⑦ 具有循环反馈机制， 这一点上优于 Cimiano 方法、Haav 方法。综合来看，Obitko 的方法在这四种方法中相对较为科学。

4 Obitko 方法用于情报学领域本体构建

4.1 原理

领域本体的构建方法与领域的特征密切相关，与其他学科相比，我们所关注的情报学领域有其独特之处：① 情报学是一门学科交叉性显著的学科，其领域知识在现有的分类体系中分布较为分散。②不像其它自然学科，情报领域概念主要为具体的事物对象或自然过程，情报学领域概念以理论、方法、技术等名词为主。③情报学作为一种仍在继续发展的学科， 其领域的大多数概念很难有权威的定义。针对以上所列出的情报学领域独特之处， 我们找到了 Obitko 方法与情报学领域本体构建的结合点。

（ 1） 经调查发现，情报学领域的概念、术语等分布在与该领域相关的结构化、半结构化文档中，如情报学叙词表、分类主题词表、领域学术论文等。只有综合处理这些异构的领域知识库， 我们才能全面的查找出该领域的相关概念。Obitko 方法从空的属性和对象开始，这便于我们从各种异构的数据源中抽取领域概念。

（ 2） 情报学领域概念以理论、方法、技术等名词为主，通过阅读学术文档发现，其中表示“ 概念－属性”的语法现象并不明显。而通过统计的方法添加对象和属性， 构造形式背景则是一种可行的方法。

（ 3） 虽然情报学领域中的大多数概念，目前尚无明确的权威的定义，但是 Obitko 方法是通过属性来描述概念，由属性决定概念的层次结构， 这种思路对概念等级关系的确定提供了较大的帮助。

4.2 步骤

以上文分析和 Obitko 方法基本步骤为基础，我们采用 FCA 的方法构建情报学领域本体的步骤如下：

（ 1） 利用自然语言理解技术（ NLP） 对收集来的领域文本进行预处理，取得该领域的候选概念集。

（ 2） 利用概率统计或者规则的方法，获得能代表领域特征的关键概念词汇。

（ 3） 针对所找出的概念词汇，结合相应的文本集合形成词汇、 文件的二元关系表。对存在多值的二元关系表，转换成单值的二元关系表。再由单值形式背景按照造格方法来构造概念格。

（ 4） 从概念格转换成相应的本体。这里采用简化的方法，即用属性来代表所形成的形式概念，并且在标注时只让属性在概念格中出现一次，由于这里的属性都是词汇，而本体所描述的重点元素也都是词汇概念，因此，可以用概念格中的属性来表示本体概念。

5 实验

根据上述方法的指导，我们对情报学领域的相关资源进行搜集， 运用自然语言处理技术和统计方法抽取领域概念，并结合领域文本构建概念格，最终实现由概念格向本体的转化。以下分别对领域知识库的搜集和处理、领域核心概念的选择和概念格的生成与转换进行详细论述。

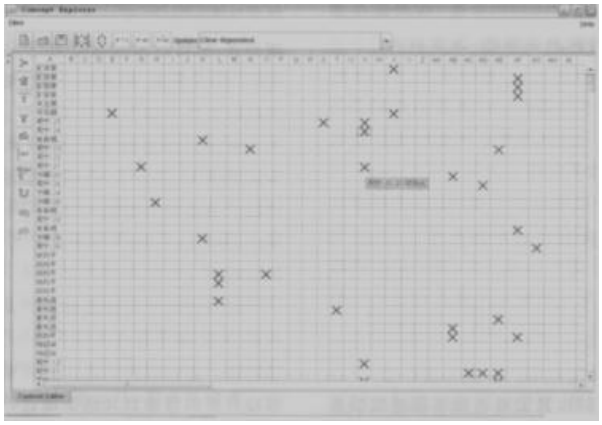

图 2　情报学领域文档 — 核心概念的概念格

5.1　领域知识库的搜集与处理

经文献调研发现，情报学领域知识库主要包括领域相关的学术论文、情报学叙词表、中国分类主题词表中的相关条目等。在科研论文中，关键词是每一篇论文的核心，作者往往倾向于使用简明扼要的科技用词或专业术语对文章内容进行有效的概括，期刊论文的关键词较全面地涵盖了该领域的主要研究内容，所以作为词库的主要来源。《 情报学叙词表》 提供了该领域基础且重要的词汇，是领域术语库的基础组成部分。中国分类主题词表中 G35 是对情报学的学科体系的框架性描述，G2527，G254，G256 等类目与文献检索、分类、标引相关，对词库的完善能起到一定的作用。通过对上述资料中的词语进行处理，共得到 6386 个概念，经过去重处理共计 3570 个术语。

5.2　领域核心概念的选择

经上一步处理的领域术语数量较多，不利于形式背景的构建，而且大多数术语，如信息、方法、管理等词并非领域专属概念，这对情报学领域形式背景的构建而言是一种噪音，因此，领域核心概念的选择对形式背景的构建具有重要意义，我们需要采用特征选择的方法筛选出质量较高的领域概念。对于情报学领域而言，恰当的领域核心概念，必须具备以下四个特征：①区分性，领域核心概念要有很强的领域性，能将目标文档与其他文档相区分；②领域通用性，领域核心概念必须是领域内通用、共同认可、不存在表达争议的概念；③全面性，确立领域核心概念时，要考虑领域内各研

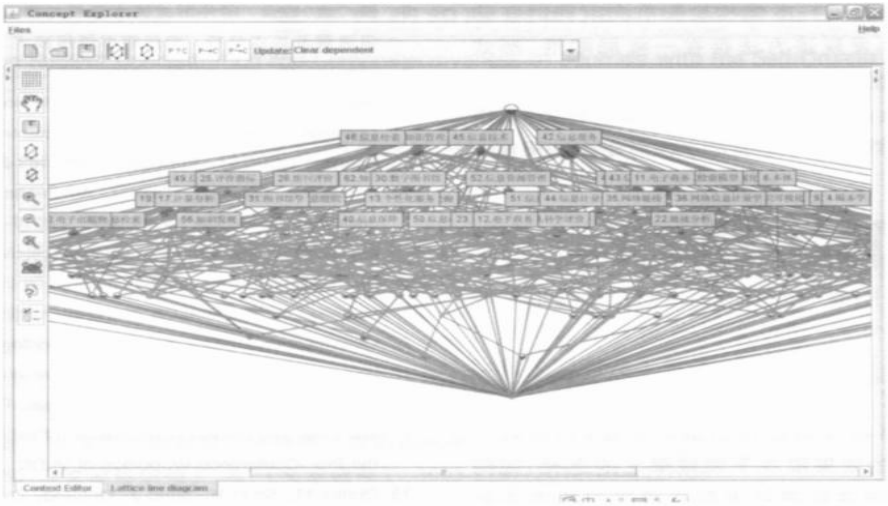

图 3　情报学领域概念格

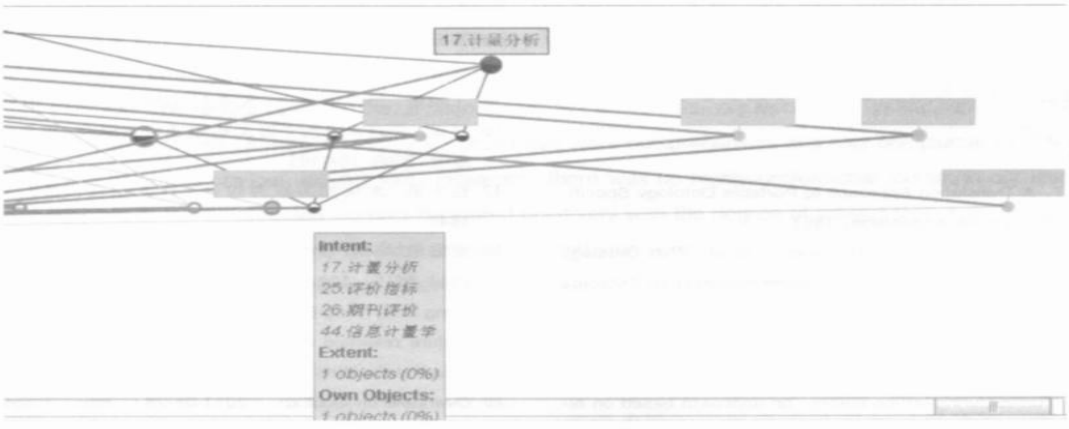

图 4　FCA 中于计量分析相关的概念

究方面的均衡分布；④精练性，领域核心概念的数量应当尽可能的小，这有助于领域概念关系的呈现。现有的关于特征选择的方法主要有三种：基于统计的方法、基于语言学的方法以及将两者相结合的混合方法。统计的方法由于对先验知识要求低，而且能够较方便的移植到其他领域进行使用，因此目前使用更加普遍。我们对统计方法中各种参数的选择效果和计算复杂度进行综合比较，最终采用 TF-IDF 作为特征选择的方法。TF 通过词语在文本集中出现的次数衡量词语的特征权重，根据齐夫定律，中频词是一个专业领域内最稳定的词汇，如果词语频次过高说明词汇在文本集中频繁出现，属于“ 常用词汇”， 例如信息、技术、管理、模式等词。如果频次过低，说明词汇属于文本集中的“稀有” 词汇，价值不高，例如质量受控、缩微复制品等词。IDF 通过词语在文本集中的分布范围来衡量词语的领域特性，文本集中分布较广或极稀疏的词语，均不适合作为领域概念。结合这两个指标， 我们给出词语特征值的计算公式如下：

$TF/IDF(ti)=tfimax(tf)\log\frac{N}{i}$（式 1）其中,tfi 表示词语 ti 在文档 dj 出现的次数，max(tf)表示在文档 dj 中出现频次最高的那个词的词频，ni 表示文档集中出现词语 ti 的文档数 N 表示文档总数。每个词语会计算得出 N（设共有 N 篇文章）个特征值，选择其中最大值作为该词的特征值。通过设定不同的特征值阈值对词语进行筛选， 我们设定阈值段分别为 0～1.5，1.5～2.5，2.5～3.5，3.5～4.5，4.5～5.5，5.5～6.5，6.5～7.5，查看各阈值段所包含的词语，结合上述的特征词选择标准， 最终确定阈值范围为 3.5－6.2，对于阈值以外的且符合筛选特征的词语， 通过人工进行添加。最终确定的领域特征词共计 194 个。进一步筛选确定领域核心概念为 81 个。

5.3 概念格的生成与转化

概念格的构建是生成本体的基础。在情报学领域概念格的构建中，我们通过统计核心概念在领域文档中的分布情况，即文档－特征词矩阵（在该矩阵中，将所有特征词在文档中出现的次数（无论几次），均设定为 1），由此构建形式背景（见图 2）在此基础上，采用概念格生成工具 Concept-Exploer 生成概念格，如图 3 所示。在由概念格转换为本体的过程中，我们采用属性来代表所形成的形式概念，由于这里的属性都是词汇，而本体所描述的重点元素也都是词汇概念，因此，可以用概念格中的属性来表示本体概念。概念之间的等级关系可以通过概念格中概念结点所处的层级获得，概念间的相关关系则通过概念节点之间的连线关系获得。如图 3、图 4 所示，点击概念结点“计量分析”，找到与其相连的其他特征词，通过结点的大小特征、节点间的位置关系与连线关系判断相连的其他特征词与该核心概念的关系，将其划分为该核心概念的子类或者相关类。与计量分析相连的概念有评价指标、期刊评价、信息计量学，它们之间存在相关关系，根据节点所处的层级差异，可以进一步确定信息计量学为计量分析的子概念。

5.4 实验结果与评价

本文选取维普资讯全文数据库（1989～2008 年），武汉大学信息资源研究中心 65 名研究人员的文章为主要数据源， 辅以情报学叙词表，以及中国分类主题词表中的部分相关类目，采用形式概念分析的方法进行情报学领域本体的构建。实验的核心在于采用 TF－IDF 方法抽取该领域的核心概念，并利用形式概念分析的方法确定概念间的层级关系。试验中，我们共筛选出领域核心概念 81 个，这些概念分布在 3 个层次，其中处于第一级的概念有知识管理、 信息服务、信息系统、信息检索、知识组织、信息技术、计量分析、科学评价、可视化、信息资源管理、知识产权。这些一级概念较大程度地符合我们对情报学分类的预期，与我们从其他渠道搜集整理的情报学概念分类基本吻合。此外，通过概念格中的节点位置和连线关系还能够挖掘出部分概念的子概念和相关概念，如知识可视化是知识管理的子概念，知识管理与信息系统互为相关概念。利用 Obitko 方法构建情报学领域本体：① 能在统一的形式背景下处理不同数据源中的概念，实现对异构数据源中概念的统一处理，有效地丰富了领域的形式背景；② 在确定概念间的等级关系时，主要依据概念在文档中的分布情况，减少了对领域专家的依赖性，在一定程度上避免了主观因素的影响，准确度较高；③ 具有较好的可扩充性，利用形式概念分析的方法构建本体可以根据语料进行自动更新———删减文档或者特征词，快速进行本体构建和扩充。同时在这个过程中，我们也发现该方法存在一些局限性：形式背景的规模对最终形成的本体有一定的影响，规模太大，关系过于复杂，不利于概念格向本体的转化；规模太小，则内容太稀疏，也无法满足本体对语义的需求。

6 结语

本文通过对形式概念分析的四种方法进行对比，结合情报学领域特点，选择 Obitko 方法构建情报学领域本体。在实验中，重点分析了领域特征选择的思路和领域概念格的构建，通过文档－关键词形式背景建立起概念和概念间的层次关系，实现了本体的半自动生成，在一定程度上解决了传统构建方法易受主观因素影响的问题，有较强的应用价值。但是，利用这种方法构建的本体的效果取决于领域概念的选择，如何有效地利用自然语言处理技术和机器学习的方法处理数据，使得既能够保证数据间丰富的语义关系，又降低生成概念格的时间复杂度，是需要考虑的重要问题；如何从概

念格中挖掘概念间的非等级关系也是亟待解决的问题。在理论和应用上建立和加强 FCA 与本体相结合的相关领域的研究是今后的研究方向。

参 考 文 献

[1] 杨强, 赵明清. 概念格研究进展[J]. 计算机工程与设计, 2008, 29(20): 5293-5296.

[2] 李云, 程伟, 蔡俊杰, 等. 基于消息传递的概念格并行构造[J]. 计算机应用与软件, 2006, 23(8): 3-5.

[3] 王俊红, 梁吉业. 概念格与粗糙集[J]. 山西大学学报(自然科学版), 2003, 26(4): 307-310.

[4] 李云, 李拓, 蔡俊杰, 等. 基于概念格提取简洁关联规则[J]. 南京邮电大学学报(自然科学版), 2007, 27(3): 44-48.

[5] 马俊. 概念格及其可视化研究[D]. : 河南大学, 2005.

[6] 李拓. 基于概念格的本体模型及其相关运算研究[D]. 扬州: 扬州大学, 2008.

[7] Ganter B, Wille R.. Formal concept analysis: mathematical foundations[M]. Berlin: Springer, 1999.

[8] Nourine L. A fast algorithm for building lattices[J]. Information Processing Letters, 1999, 10(71): 199-204.

[9] Godin R, Mineau G. Applying concept formation methods to sotfware reuse[J]. Int J Sotfware Engand Knowledge Eng, 5(1): 119-142.

[10] Gerd Stumme, Efficient Data Mining Based on Formal Concept Analysis. Institut f ¨ur Angewandte Informatik und Formale Beschreibungsverfahren AIFB,Universit ¨at Karlsruhe, D–76128 Karlsruhe, Germany