《智能信息处理》课程作业

基于形式概念分析的协同信息检索框架

曹宇航

作业	分数[20]
得分	

2021 年 11 月 28 日

基于形式概念分析的协同信息检索框架

曹宇航

（大连海事大学 信息科学技术学院, 大连 116026）

摘 要 形式概念分析(Formal Concept Analysis，FCA)是由 R. Wille 于 1982 年提出一种从形式背景进行数据分析和规则提取的强有力工具，在信息检索、数据挖掘、软件工程等领域上的应用十分广泛。在协同信息检索中，通过增量更新算法来构造形式概念格来生成大量的子文本数据库。然后，通过测量形式概念之间的相似度，将来自不同格的临时形式概念进行合并，得到新的形式概念，然后对新概念进行匹配，最终返回满足用户需求的结果。整个框架易于部署在分布式环境中，并以不精确的方式匹配查询词，它更好地反映了人类的需求。

关键词 协同信息检索；形式概念分析；概念格

Collaborative Information Retrieval Framework Based on FCA

Cao Yuhang

(School of Information Science and Technology, Dalian Maritime University, Dalian 116026)

Abstract Formal Concept Analysis (FCA) was developed by R. In 1982, Wille proposed a powerful tool for data analysis and rule extraction from formal context, which is widely used in information retrieval, data mining, software engineering and other fields. In collaborative information retrieval, we construct formal concept lattices by incremental updating algorithm to generate a large number of sub-text databases. Then, by measuring the similarity between the formal concepts, the temporary formal concepts from different lattices are merged to get the new formal concepts, and then the new concepts are matched, finally returns the result that meets the user's requirements. The entire framework is easy to deploy in a distributed environment and matches query terms in an imprecise manner. It better reflects human needs.

Key words Collaborative information retrieval; formal concept analysis; concept lattice

1 引言

网络上的信息日益增长，分类和结构也日趋多样化。文本信息尤其庞大，因此信息检索领域已成为众多专家学者最关注的领域。国内外对信息检索提出了许多方法和理论。其中，利用形式概念分析（Formal Concept Analysis，FCA）构建信息显示模型，以更准确地表示原始检索模型，并利用所构建的概念格进行信息检索。基于不同抽象级别的不同检索将在一个独特的级别上获得特定的搜索结果。此外，在分布式环境下，利用形式概念分析理论和

技术协同进行信息检索，可以大大提高检索时间和效率。

为了更快、更有效地获得更精确的结果，采用了一组基于语义的策略。在协同信息检索中，针对多个子文本数据库，采用增量更新算法构建概念格。然后以形式概念相似度作为度量方法，对已有的形式概念进行匹配，找到暂存本体后，合并获取新概念，再对新概念进行相似度匹配，得到满足用户需求的集合。整个系统充分体现了协作的思想，可以在分布式环境中方便地部署和实现。同时，在查询关键字匹配中使用不精确的方式，从结构和语义两个层面衡量，体现了人性化的需求。

2 形式概念分析

1982 年，德国的 Wille R 提出了形式概念分析。它不仅是一种有效的数学工具，也是数据分析和规则提取的工具，同时也是知识处理的一种新方法，它能够帮助人们认识集合中各个元素之间的关系，用数学方法表达概念和概念层次。

其内容是形式上下文、形式概念以及形式概念之间的关系。形式上下文定义为三个 K=(G, M, I)，其中 G 是对象集，M 是属性集，I 是 G 和 M 之间的二进制关系，即 I 基于三角形统计图的数据集，基于三角形统计图的数据集。在形式环境 K 中，两个映射函数 f 和 g 定义为：

$$A^{'}=\{m\in M|gR_{m},\ \forall_{g}\in A\}$$
$$B^{'}=\{g\in G\ |gR_{m},\ \forall_{m}\in B\}$$

格（lattice）的意义是任两个元素的上确界和下确界都存在的偏序集。完备格为任一子集的上确界和下确界存在的偏序集，其特点是只有一个最高点，且只有一个最低点，且图中任何两点连通。概念格是元素为概念的完备格。概念格，也称为 Cralois 格，它提供了一种支持数据分析的有效工具。概念格的每个节点是一个形式概念，每一个形式概念都是由外延和内涵两部分组成。概念格通过 Hasse 图生动和简洁地体现了这些概念之间的泛化和特化关系。从形式背景中生成概念格的过程实质上是一种概念聚类过程。

概念格的表示形式是 Hasse 图，概念格的构建的基础是形式背景，形式背景描述了多个形式概念之间的关系，单个形式概念描述了形式对象以及形式对象所具有的形式属性之间的关系。所以，概念格的构建必须明确不同形式对象以及不同形式对象所具有的形式属性。

概念格的构建包含以下几个步骤：生成形式背景，约简形式背景，生成单值形式背景，确定父子关系，绘制 Hasse 图，补充各形式概念的上确界和下确界，最后获得概念格。

3 基于形式概念分析的协同信息检索

通过使用形式概念格中包含的丰富信息，可以更好地实现多个数据库。协同概念信息检索系统，在形式化概念分析的基础上，提出了三种不同的检索系统。主要是对每个子上下文进行约简，同时获得等价的对象集合。然后，通过检索每个子上下文来实现一个临时结果集。在此基础上，使用合并算法对这些临时结果进行合并，并将基于等效对象集合的最终检索结果返回给用户。然而，也有一些缺点。例如，在进行子上下文匹配时，它使用精确的对象集来寻找满足条件，而在实际情况下需要一个模糊的、不精确的对象匹配过程。此外，在合并临时结果集时，匹配过程也很精确。将形式概念格用于对象约简的另一个缺点是，所构造的形式概念格是一个完整的格，可能会消耗大量的时间和空间资源。

3.1 协同信息检索的结构框架

这里提出一种改进的协同概念信息检索系统，分别构造了多个形式子上下文对应的概念子格，同时生成了等价的对象集。针对用户查询的特定关键字，在子格中使用相似度计算公式得到一个满足概念相似度阈值的临时形式概念集，然后利用合并算法得到临时形式概念集。合并后的形式概念度量仍然可以使用概念相似度来完成。概念相似度阈值中包含的形式概念将被添加到最终的结果集中。最后，最终结果集中的每个形式概念对象，包括与它们具有等价关系的对象，都将作为最终的搜索结果返回给用户。

每个分离的子数据库都被构造为分离的子上下文。利用相关构造算法和构造过程中得到的等价对象集构造子概念格，将相关等价对象约简，形成基本结构。当用户输入各种关键字进行查询时，系统对其进行分析，得到相应的相关属性集，并对各种次形式概念格进行对合，对各自格中的形式概念集进行匹配，使其满足一定的相似性，并将其放入

临时结果数组中。在此过程中，由于每个子格上的概念匹配操作是独立的，因此可以分布式进行。每个子流程完成后，将存储在临时结果数组中的概念进行合并，再进行一次相似度匹配，将得到的最终结果返回给用户。基于上述思想，本文提出了相关策略，并设计了分布式算法来实现细节。

3.2 等价对象集的约简

定义 1. 闭包：假设一个对象来自 x∈G，A∈G 的对象集，I 是来自 G×M 的一个关系，那么闭包(x)=g(f(x))，闭包(A)=g(f(A)).

定义 1. 等价对象集：假设一个对象来自 x∈G，A∈G 的对象集，I 是来自 G×M 的一个关系，对象 x 等价于对象集 A,

表 1 背景说明

	A	B	C
o1	x	x	x
o2	x		x
o3	x		
o4		x	x
o5			x

其中，对象 o5 等价于对象集(o1, o2, o4)，因为包含对象 o5 的形式概念是 CP ((o1, o2, o4, o5), (c))，而相反，包含对象集(o1, o2, o4)的形式概念也是 CP，与上面等价对象集的定义相匹配。对象 o5 等价于对象集(o1, o2, o4)。

由于对象 o5 对象可以用 set (o1, o2, o4)等价表示，它可以在原始上下文中被约简，这意味着对上下文的对象级约简。整个约简过程可以用算法 1 来描述。

算法 1 是针对概念对象的约简算法，不进行属性层面的约简。该算法在属性抽取过程中进行属性约简，在形成基本上下文后进行对象级约简。然后会得到整个流线型 f 上下文。

算法 1. 等价对象约简算法(FC)

INPUT : FC, 原始上下文

OUTPUT : 对象缩减后的上下文

STEP : 对 FC 中的每个对象 o :

P 是一个属性集合，对应于 o.

搜索除 o 以外的对象子集的上下文，得到的属性集合为 sub_p

IF p == sub_p：

将 o 从上下文或上下文 FC 中删除

Return FC

3.3 协同信息检索的过程

基于协同信息检索的基本思想，本文提出了改进的协同信息检索系统，其过程可分为三个步骤:

1)对于每个子数据库 DBi,经过相应的预处理，形成一个基本上下文 FCi，使用算法 1 对基本上下文进行对象约简，得到等价对象的集合。然后，使用增量更新算法生成格(Godin Algorithm)到简化上下文和生成器的子形式概念格 Li。

2）对特定查询词进行提取，得到相关查询属性集 T1、T2、…， Tn，其形式概念可记为((), (T1, T2，…， Tn))，范围为空，计算形式概念相似度时，可记录为 0 级。然后在每个次形式概念格 Li 上进行形式概念相似度匹配，将每个次临时结果集存储到临时概念结果集中。具体过程将在下面的算法 2 中描述。

3）临时结果集概念,使用合并算法 3 结合这些临时结果集概念,相似性匹配将获得最终的概念集,做过的对象,作为最终的对象集,在区段的正式的概念,将结果返回给用户。

在这些步骤中，在每个子形式概念格中搜索目标概念，得到子临时结果集，作为子概念集，该算法是典型的分布式，在时间和空间上都具有较高的效率。

如果子任务上下文的数量是 n, n 构成的子正式概念格,然后使用算法 2 得到的 n 次临时结果集,SubConceptSeti,该算法易于理解和分布式实现,那么将会有更多的时间和空间效率。

算法 2. 子格匹配算法(L_i, (T_1, T_2, …T_n), sim1)

INPUT : Li 形式概念子格，(T1, T2，…， Tn) 查询属性集，sim1 形式概念相似度阈值

OUTPUT：子形式概念结果集，如 SubConceptSet

STEP : 子临时结果集，如 SubConceptSet 不为空.

基于查询属性集(T1, T2，…)， Tn)，找到与形式概念相似度最大的形式概念 Ci， (T1, T2，…)Tn)).

对于所有形式概念 Cj，除 Ci 外，在每个次形式概念格 Li 中.

使用公式计算 Ci 与 Cj 的概念相似度，得到 Sim (Ci, Cj)
If Sim (Ci, Cj)> sim:

形式概念 Cj 被添加到子临时结果集，即 SubConceptSeti.

Return SubConceptSeti 的次临时结果.

算法 3. 概念合并算法 (SubConceptSeti, (T_1, T_2, …,T_n), sim2)

INPUT : 子临时结果集为 SubConceptSeti，子形式概念格的数量为 i，查询属性集 (T1, T2, …,Tn)，形式概念相似度阈值为 sim2.

OUTPUT : 最终的概念对象，设为 FinalObjectSet

STEP : 所有概念结果集，如 AllConceptSet 均不为空

对于每个次级临时结果集，SubConceptSeti：

#将每个子临时结果集添加到所有概念结果集

AllConceptSet = AllConceptSet SubConceptSeti

对于所有概念结果集中的每个概念 Ci，如 AllConceptSet：

#将每一对组合起来得到新的形式概念

对于所有作为 AllConceptSet 的概念结果集中的每个形式概念 Cj，除 Ci 外：

NewConcept(i,j)=

$$\left(\left(\text{extent}(C_i) \text{I } \text{extent}(C_j),\ \text{intent}(C_i) \cup \text{intent}(C_j)\right)\right)$$

计算得到新的形式化概念如下

NewConcept and the new similarity of formal concept T (),(T1, T2, ..., Tn)) as Sim (NewConcept, T).

IF Sim (NewConcept, T) 大于 sim2 :

将 NewConcept 的范围添加到最终的概念对象集 FinalObjectSet.

Return 最终的概念对象集 FinalObjectSet

通过算法 3，可以得到最终满足一定相似性阈值的概念集，这些集合中的每一个程度的形式概念，都是满足用户查询条件的对象集。其中，算法 2 和算法 3 分别涉及两个相似度阈值，如 sim1 和 sim2，在子形式概念格的相似度匹配中，由于形式概念格中的内涵总数可能较小，可能不完整，所以小正式概念应设置 sim1 害怕错过,而在合并过程中,合并正式的概念是通过其内涵的合并,其信息比较丰富,所以大正式的概念应该设置 sim2 过滤不匹配的正式的概念。

4 复杂度分析

假设子上下文个数 n，用 Godin 算法得到子形式概念格，其时间复杂度为 $O(2^{2K}|G|)$。G 是物体的数量，K 是正常情况下的常数。等价对象约简算法作为算法 1，其时间复杂度为 $O(|C|2^M)$，C 是形式概念格中所有的形式概念，M 是属性的个数。子概念格上的形式概念匹配算法，即算法 2，只需要遍历整个形式概念集就可以得到满足一定相似性阈值的形式概念，所以时间复杂度为 O(|C|)。对于合并算法 3，由于需要合并每对形式概念，如果临时概念结果集中有 N 个形式概念，则合并的时间复杂度为 $O(N^2/2)$。综合考虑，整个过程的时间复杂度是 $O(2^{2K}|G| + |C|2^M + N^2/2)$。

但是，从整体上考虑，那么本文提出的系统可以很容易地部署在分布式计算环境中，那么整个系统的性能在时间和空间上都会有较大的提升。

5 总结与展望

在协同信息检索中，利用形式化概念相似度度量方法对各子上下文下给定的形式化概念进行匹配，在找到临时概念集后，合并得到新概念，然后将新概念与相似度进行匹配。最后得到满足用户需求的结果集。整个系统充分体现了协作的思想，可以很容易地部署在分布式环境中。同时，在查询关键字匹配中使用不精确的方式，从结构和语义两个层面衡量，体现了人性化的需求。但是虽然本文给出了整个系统的框架和具体的算法，但是由于硬件的限制，整个系统无法真正的在分布式环境下进行，所以在最终的分析中，并没有真实的数据来分析性能，只从理论上分析时间和空间的复杂性。这将在今后的工作中得到改进，报告更精确的数据，以便进一步改进。

参 考 文 献

[1] 魏玲，祁建军，张文修. 概念格与粗糙集的关系研究［J］. 计算机科学，2006，33(3) : 18-21.
WEI Ling，QI Jianjun，ZHANG Wenxiu. Study on relationships betw een concept lattice and rough set ［J］. Computer Science，2006，33(3) : 18-21.

[2] 仇国芳，张志霞，张炜. 基于粗糙集方法的概念格理论研究综述［J］. 模糊系统与数学，2014，28(1) : 168-177.
QIU Guofang，ZHANG Zhixia，ZHANG Wei. A survey for study on concept lattice theory via rough set［J］. Fuzzy systems and M athematics，2014，28 (1) : 168-177.

[3] 张文修，魏玲，祁建军. 概念格的属性约简理论与方法［J］. 中国科学(信息科学) ，2005，35 (6) : 628-639.
ZHANG Wenxiu，WEI Ling，QI Jianjun. Attribute reduction theory and approach to concept lattice［J］. Science China(Information Sciences) ，2005，35(6) : 628-639.

[4] 智慧来，智东杰，刘宗田. 概念格合并原理与算法［J］. 电子学报，2010，38(2) : 455-459.
ZHI Huilai，ZHI Dongjie，LIU Zongtian. Theory and algorithm of concept lattice union［J］. Acta Electronica Sinica，2010，38(2) : 455-459.

[5] 胡可云，陆玉昌，石纯一. 概念格及其应用进展［J］. 清华大学学报(自然科学版) ，2000，40(9) : 77-81.
HU Keyun，LU Yuchang，SHI Chunyi. Advances in concept lattice and its application［J］. Journal of Tsinghua University (Science ＆ Technology) ，2000，40(9) : 77-81.