第34卷 第16期 计 算 机 工 程 2008年8月
Vol.34 No.16 Computer Engineering August 2008
·开发研究与设计技术· 文章编号：1000—3428(2008)16—0268—03 文献标识码：A 中图分类号：TP391

基于FCA的产品特征信息分类

王 芳[1]，马 骏[1,2]，陈 明[1]
(1. 河南大学计算机与信息工程学院，郑州 475004；2. 河南大学数据与知识工程研究所，郑州 475004)

摘 要：根据从产品信息中提取的形式概念，建立形式背景，提出利用概念格结构表示特征信息的设计思想，分析关键格结构和概念聚类2种优化显示策略。通过与传统信息浏览方式的比较，说明该策略具有更高的灵活性，能够更清晰地反映各类产品间的联系与差异。
关键词：形式概念分析；概念格；产品信息inder；特征分类

Product Characteristic Information Categorization Based on FCA

WANG Fang[1], MA Jun[1,2], CHEN Ming[1]
(1. College of Computer and Information Engineering, Henan University, Zhengzhou 475004;
2. Institute of Data and Knowledge Engineering, Henan University, Zhengzhou 475004)

【Abstract】 This paper presents a method to extract the formal concept from the product information, and constructs the formal context. By using concept lattices, the idea to describe the product information is proposed. Two optimum strategies for browsing concept lattices are given. One is key-lattices and the other is cluster concept. Compared with the conventional imformation browse, test results show that it has more flexibility and can display the relationship and difference between products clearly.
【Key words】 Formal Concept Analysis(FCA); concept lattices; product information inder; characteristic categorization

1 概述

传统的基于关键词的搜索引擎一般都是以列表或树型结构的形式将搜索结果呈现给用户，没有进一步分类的综合处理过程。虽然列表方式呈现的内容清晰明了，但是无法体现商品信息之间的联系与差异。树型结构能较好地体现层次关系，但查找商品信息时，往往需要按固定的路径查询，缺乏灵活性，由于没有进一步的分类处理，因此用户要得到所需信息，常需在不同的站点间切换浏览，并人工筛选搜索结果。为解决这些问题，笔者提出一种基于FCA的产品特征信息分类处理方式，即利用形式概念分析[1](Formal Concept Analysis, FCA)理论提取、综合并结构化搜索的产品信息，并以文字与概念格相结合的形式展现给用户，从而简化了人工筛选的难度，提高产品特征信息检索的效率。

形式概念分析是概念知识形式化的一种方法，其主要思想是从由二元关系构成的形式背景中提取概念层次结构，并从数据集中生成概念格。为了适合实际的产品特征搜索需求，首先根据搜索关键词从Web中提取相关的信息，并将提取后的结果转化为标准形式背景的形式，然后在此基础上构造概念格。概念格是建立在偏序关系上的一种结构化特征描述，它在表示概念之间分类规则方面有其独特的优势，能够清晰、直观地体现概念间的偏序关系，这种结构非常适合于综合相似信息，并进行浏览和比较。

2 Web信息提取及显示

基于概念格结构的Web信息分类处理及优化显示的首要任务是从网页中提取形式概念。先利用网络中已有的多个搜索引擎同时检索相关信息，然后采用模式匹配的方法提取需要的特征信息，并根据提取的信息构造标准形式背景，最后利用概念格结构综合、描述这些信息。其过程如图1所示。

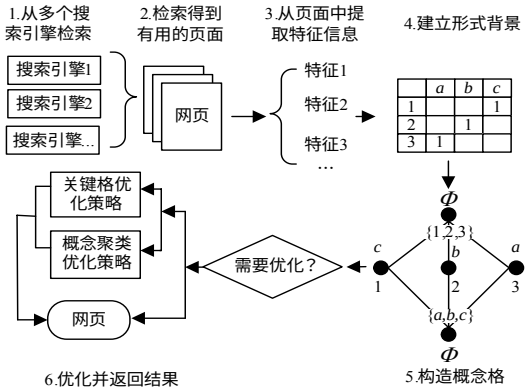

图1 基于FCA的产品信息处理过程

2.1 从网页中提取形式概念

从网页中提取形式概念的过程分2步完成。先利用目前已有的多个搜索引擎同时查询信息。这种方式可以充分利用现有的资源，使搜索的范围更广、内容更全面；然后分析搜索结果，并使用模式匹配的方法从中提取出需要的信息，这部分工作主要是对反馈的HTML页面进行分析，通过正则表达式确定并抽取需要的内容，并将这些内容以某种形式存入库中以备构造形式背景使用。

2.2 信息特征的形式背景表示

利用FCA分析的数据一般是用形式背景来表示的。在标

基金项目：河南省高校杰出科研人才创新工程基金资助项目(2007KYCX018)；河南大学自然科学基金资助项目(05YBZR008)
作者简介：王 芳(1980－)，女，硕士，主研方向：软件工程，知识发现，网络应用；马 骏，副教授；陈 明，硕士
收稿日期：2007-09-30 **E-mail：**hedawangfang@126.com

准的形式背景中，若某个对象 g 与某个属性 m 存在二元关系，就表示对象 g 具有属性 m。但在现实世界中，一个对象不仅仅是具有或不具有某一个“属性”，诸如“颜色”、“重量”等属性都具有很多的值，因此，需要采用多值形式背景的方式描述这些信息。由于不同用户对所关注对象的属性划分的精确程度要求不同，为了使形式背景表现的信息更符合实际，在对属性进行处理时，不采用固定的划分细度，而是将由具体数值表示的属性粗糙地划分为 n 个子属性段，并为每个子属性段设置适当的取值范围，若具体的属性值在某一范围内，就将此子属性段取值为“1”，其他子属性段取值为“0”。如图 2 所示，将由具体数值表示的属性转化为多值形式背景。

	价格/元
1	480
2	670
3	1 100
4	2 500
5	3 600
6	870
7	1 350
8	2 100
9	2 800

⇨

	＜500	500~1 499	1 500~2 500	＞2 500
1	1	0	0	0
2	0	1	0	0
3	0	1	0	0
4	0	0	1	0
5	0	0	0	1
6	0	1	0	0
7	0	1	0	0
8	0	0	1	0
9	0	0	0	1

图 2　将价格信息转化为形式背景

这种方式构造的形式背景由三元组(G,M,R)构成，其中，$G=\{g_1,g_2,\cdots,g_n\}$是对象集，即具体的某一个或某类产品集；$M=\{m_1,m_2,\cdots,m_n\}$是属性集，即产品所具有的特征集；R是G和M之间所具有的二元关系。确定了形式背景，接下来就可以构造概念格。

2.3　产品信息的概念格描述

在构造概念格过程中，利用Godin算法[2]建立一种基于层次图的概念格模型[3]。在此概念格模型中，整个格结构是信息抽象分类的一种结构化形式，其中每个格节点代表一个或一类信息，即一个节点是一个或一类对象；节点的内涵是此信息或此类信息的特征描述，即某个或某类对象所具有的属性集。在格图中，当且仅当从节点c_2到节点c_1有一个上升的路径连线，那么c_1所代表的一个或一类信息是c_2所代表的一个或一类信息的子集，其中c_2称为c_1的下邻，c_1称为c_2的上邻。若一个节点有多个上邻，则说明它所具有的多个属性来自不同的上邻点。以用户输入搜索关键词“诺基亚”，搜索类别选择“手机”，搜索目的选择“特征比较”为例，来构建相应的概念格，如图 3 所示。

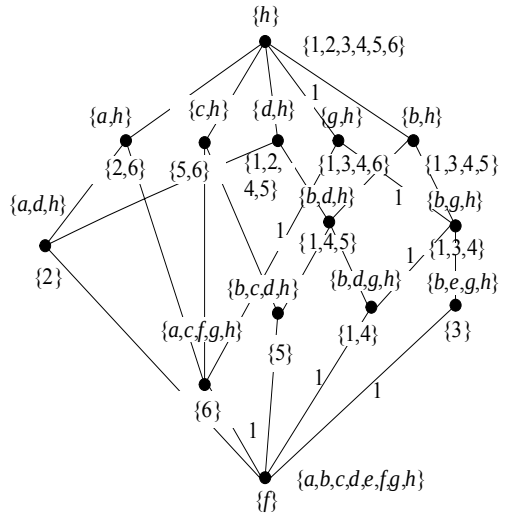

图 3　概念格

图 3 所示的概念格中，节点上方用字母表示的集合为内涵集，节点下方用数字表示的集合为外延集。

3　信息格的显示优化策略

良好的格结构不仅要能准确地展示信息间的关联，而且要能剔除不相关的信息，以降低显示复杂度，在最小的浏览范围内尽量提高信息的可用性[4]。根据这些需求，提出 2 种优化概念格结构的方法，分别是关键格结构和概念聚类，前者用于优化综合后的信息，后者用于提高浏览清晰度。

3.1　用关键格结构优化综合后的信息

利用概念格结构表示信息间的关联时，若格结构规模适中，其浏览优势显而易见，但如果其结构规模庞大而复杂，就完全丧失了用于实际浏览的意义。因此，如何降低概念格结构的规模就成为显示产品特征搜索结果的关键问题。在实际情况中，用户在信息浏览时往往只对其中的某一类或某几类对象或者属性感兴趣，即他们关注的只是局部信息。因此，提出了利用关键格结构来优化信息之间关系的策略。确定关键格算法的主要思想如下：

(1)从初始节点(格结构中最顶层的节点，即外延最大的节点)出发，依次访问初始节点的各个未曾访问过的下邻节点，为初始节点与包含所关注属性的节点之间的连线标注其权值为 1，为初始节点与不包含关注属性的节点之间的连线标注其权值为 0。

(2)分别从那些标注权值为 1 的连线所连接的节点出发，依次访问它们的下邻节点，将此节点与其下邻节点中包含所关注属性的节点之间的连线标注权值为 1，将此节点与其下邻节点中不包含关注属性的节点之间的连线标注权值为 0。

(3)重复(2)，直至找到包含关注属性的所有节点为止，最后得到的所有标注为 1 的边与其所连接的节点所构成的图就是所求的关键格结构。

图 4 为用户右键单击“彩信、和弦”节点，在快捷菜单中选择“只显示相关信息”时，显示的关键格结构。确定了关键格结构，就可以为用户呈现所有那些具有用户所感兴趣属性的对象集，并可在此基础上进行详细查询。

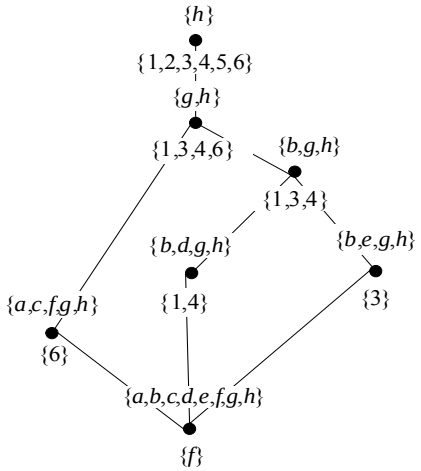

图 4　关键格

3.2　用属性相似性聚类的结果

利用关键格结构策略优化格结构化后的产品特征信息适用于根据部分特征信息查找具体的某个或某类产品的情况，但在实际应用中，某个产品特征信息是所有产品共有的属性，此时，关键格结构与原始格结构就会完全相同。显然在这种

情况下，关键格结构策略无法达到显示优化格结构化后的产品特征信息的目的。因此，又提出另外一种策略——概念聚类。其基本思想是用一种相似性的尺度来度量概念之间的相似程度，并根据一定的属性隶属程度完成对概念的分类。

由于概念格结构中，对象与属性之间具有偏序关系的特性，因此在讨论概念相似性时，必然要考虑到属性的相似度。由于所有的对象都是由相同域中的属性构成，因此对象间的相似性可映射为属性间的相似性度量。属性相似性越高，说明 2 个对象的相似度越高。

根据属性确定概念对象间的相似性，而后根据相似程度分清对象之间的疏近程度，对概念进行同类分组，将相似度高的概念集作为一个簇，并将这些簇作为节点进行造格。这样，浏览者只需找到那些具有所感兴趣的属性簇即可。

以下定义是根据概述得到的用属性相似度进行对象相似度比较的方法[5]。

$$Sim(m,n)=\frac{f(M\cap N)}{f(M\cap N)+\alpha\cdot f(M-N)+\beta\cdot f(N-M)}$$

其中，M 和 N 分别是对象 m 和 n 的属性集；$f(M-N)$ 表示包含在属性集 M 中而不包含在属性集 N 中的属性集合；$f(M\cap N)$ 为属性集 M 与属性集 N 的共有属性的集合；α 和 β 是设定的参数。

根据以上定义，利用属性相似度计算 2 个对象相似性可定义为

$$Sim(ob_1,ob_2)=\frac{\left|(m_1\vee m_2)_{LA}\right|}{\left|(m_1\vee m_2)_{LA}\right|+\alpha\left|m_{1LA}-m_{2LA}\right|+(1-\alpha)\left|m_{2LA}-m_{1LA}\right|}$$

其中，ob_1 和 ob_2 是所要进行相似度比较的 2 个对象；$m_{1LA}\vee m_{2LA}$ 表示对象 ob_1 和 ob_2 所具有属性的最小上界；$m_{1LA}-m_{2LA}$ 表示属于 m_{1LA} 而不属于 m_{2LA} 的属性集；同理，$m_{2LA}-m_{1LA}$ 表示属于 m_{2LA} 而不属于 m_{1LA} 的属性集；α 是度量 2 属性相似度的标准，可以设置一些枚举值以便使用者选择。

在软件实现中，可以将对象相似度大于此标准值的相互比较的 2 个对象归类为一个簇。同理，也可以将所有的对象分别进行比较、分簇。最后所得即是由属性相似性确定出的概念聚类。

算法的主要思想如下：

输入 一组将要相互比较相似性的对象：ob_1和ob_2以及相似度系数 Φ

输出 产生的概念聚类集合S_C

```
Sc:= ϕ
for(p=1;p<n; p++)
{
    for(q=p+1;q<n; q++)
     {if
       then      Sc←Sc∪{p, q}
     else        Sc←Sc∪{ ϕ }
    }
}
```

通过度量两两概念间的属性相似度得到对象相似性，然后根据概念间的相似性强弱关系进行概念的聚类处理。

其现实意义在于：功能相似性系数高的手机归类为一个概念簇，方便用户对手机选择的整体把握，之后再详细分析、比较，最后选出满意的手机款型。

利用概念聚类的策略能够达到控制格结构规模的目的，并且聚类结果与实际相符，聚类效果良好。

值得注意的是，当格节点数目取值范围为 10～15 时，利用关键格结构效果较好；当格节点数目超过 20 时，利用概念聚类效果较好。

4 软件实现

在 Visual Studio2005 环境下，采用组件、GDI+等可视化技术以及 C#语言实现了利用概念格结构二维显示处理的产品信息处理结果。功能界面由 3 部分组成：

(1)左侧以树状图文字的方式显示产品信息内容，用户可根据不同的划分规则进行选择。

(2)中间部分为概念格结构描述的产品信息，用户可以通过格结构选择查看具有某些特征的手机型号。

(3)右侧为格节点操作区域，用户可通过选择控件对格结构进行缩放、平移以及节点设置等操作。概念格结构与树状结构结合使用，能够更清晰地反映各类产品间的关系与区别，对用户的选择有很好的导向作用。

5 结束语

本文从增强传统搜索引擎的功能入手，利用形式概念分析理论，讨论了产品特征信息搜索、综合与浏览问题，并用关键格来简化浏览复杂度，利用概念相似性度量方法以及概念聚类思想完成对象的分类。

从结构观察角度来讲，利用概念聚类得到的格结构比单纯的文字表示以及树状结构更形象、更容易理解和便于观察。同时，由于对产品信息进行了分类和综合处理，更利于用户快速找到所需的信息。

参考文献

[1] Ganter B, Wille R. Formal Concept Analysis: Mathematical Foundations[M]. Berlin, Germany: Springer-Verlag, 1999.

[2] Godin R, Missaoui R, Alaoui H. Incremental Concept Formation Algorithms Based on Galois[J]. Computational Intelligence, 1995, 11(2): 246-267.

[3] 马　骏, 沈夏炯, 刘宗田. 基于三维空间的概念格自动布局[J]. 计算机科学, 2006, 33(5): 244-246.

[4] Cigarrán M, Gonzalo J. Browsing Search Results via Formal Concept Analysis: Automatic Selection of Attributes[M]. [S. l.]: Springer Press, 2004.

[5] Zhao Yi, Wang Xia. Ontology Mapping Based on Rough Formal Concept Analysis[M]. [S. l.]: IEEE Press, 2006.