基于形式概念分析的 E-mail 分类本体模型

朱宗梅

(大连海事大学，计算机技术 ，大连 116026)

摘 要：对于大量的 E-mail，传统的分类方式是使用基于关键字比较的分类系统的过滤器对 E-mail 进行分类， 这种方式存在缺乏灵活性、分类粗糙等缺点。本文提出了一种基于形式概念分析(FCA)的 E-mail 分类模型，在传统规则分类方法的基础上引入 FCA 分类机制，即利用 FCA 来抽取隐含在 E-mail 文档中的潜在的概念关系，分析文档之间的相关性。用户对 E-mail 文档的浏览是在一个根据概念划分的概念格结构上进行的。试验验证了该模型的可行性，试验结果表明该模型较好地解决了概括性的分类问题。

关键词：FCA，属性抽取，E-mail 分类模型，E-mail 分类系统

中图法分类号：TP18 **文献标识码：**A

E-mail Classification System Based on Formal Concept Analysis

ZHU Zongmei

(Department of Computer Technology, Dalian maritime university, Dalian 116026, China)

Abstract: For a large number of E-mails, traditional sorting fashion is to sort them by the filter of the sorting system which is based on key word comparison. Aiming at actual fashion`s problem of returning rough sorts, this paper puts forward an E-mail sorting model based on Formal Concept Analysis(FCA), which adopts potential concept relation between the mails and analyzes relativity between them by FCA. The user browses mails organized on Ga lo is(concept) lattices. The paper emphasizes the implementation idea of E-mail sorting model. The test validates its feasibility, and test result indicates this model can make up the limitation of actual E-mail sorting system well.

Key words: FCA, attribute-abstract, E-mail sorting model, E-mail sorting system

0 引言

E-mail 是一种半结构化的文本文件，包括 E-mail 的头和正文。传统的客户端邮件管理系统如 Foxmail 或 Outlook 等本质上是一个文档管理系统，该系统把 Web 邮箱里的 E-mail 收取到本地计算机，并组织为与其物理目录相对应的树形结构供用户查阅。其优点在于树形结构直观地反映了 E-mails 与文档系统结构间的映射关系，缺点是在存储 E-mail 时，用户必须记住在哪个目录下能够再次找到这封 E-mail。尤其是对企业级的 E-mail，为了检索到一封 E-mail，用户必须预先知道该邮件存储在哪个文件夹下，并阅读该文件夹下的所有 E-mail 的列表。

概念格(Galois)是形式概念分析[1](Formal Concept Analysis，缩写为 FCA)的核心数据结构，概念格的每个节点是一个形式概念，由两部分组成:外延(对象)，即概念所涵盖的实例;内涵(属性)，即概念的内在描述，也是所有被涵盖的实例的共同特征。概念格对应的 Hasse 图体现了概念间的泛化和特化关系，它表现了数据的知识视图，目前

已经在知识发现[2]、机器学习[3]、软件工程[4]等诸多领域得到了一定的应用。

领域本体描述是一切基于领域本体的知识工程活动的前提，提高领域本体描述的形式化与规范化程度、语义表达能力和本体知识推理能力就成为领域本体描述所一直追求的目标。形式概念分析(FCA)是应用数学的一个分支，它是建立在概念和概念层次的数学化基础之上的一种新的知识描述手段和数据分析工具，本文提出了一种基于 FCA 的 E-mail 管理系统模型，该模型结合概念格的性质，根据邮件的内容或其它特征赋予给每个邮件一个唯一的 Key 对邮件进行聚类、管理，把存储在本地计算机的邮件作为对象，邮件的特征项作为属性，如是否有附件、主题、邮件来源等，从而使得对邮件按照其潜在概念上的关系进行分类，并在概念格结构上进行浏览，把概念格作为聚类技术[5]来组织结果，根据用户的观点来分类文档集，给用户提供多角度的分类、浏览方式。

1 关于形式概念的分析

根据基于形式概念分析（FCA）的领域本体描述原理，本文将 FCA 描述领域本体的过程分为三个阶段：准备阶段、分析阶段、描述阶段。从实际操作的层来看，上述三个阶段每阶段都包涵着许多错综复杂的相互作用的要素和内容，这给理解和掌握基于 FCA 域本体进行描述这一过程的本质造成了一定的困难。因此，本文采用模型化的思路，抓住这四个阶段中的主要要素并摒弃次要要素，进而深入研究各主要的关系，对基于 FCA 的领域本体描述过程进行构建了基于 FCA 的领域本体描述。

1.1 准备模块

该模块主要解决领域本体描述的前期准备问题。在知识工程专家、领域专家和领域本体用户三方面对所要建设的领域本体进行深入需求分析的基础上，搜集领域数据，并将其分为三类:结构化数据、半结构化数据、非结构化数据。随后使用相关技术(映射技术、NLP 技术等)从各类数据中抽取出领域核心术语集，并将术语集的格式统一为“对象-属性”集，文献[7]阐述了具体的方法:①对结构化数据(一般为关系数据库表)，利用逆向工程或映射技术将关系模型转换为 E-R 图，用数据库表的元组作为对象，而数据库表的属性作为属性，E-R 模型的关系表述概念间的关系；②对非结构化数据(一般是领域纯文本)的处理比较复杂，一般是通过自然语言的解析器，将领域文本中的每一个句子转换成一棵语法树，由语法树来分析，将词汇关系分为动宾关系、并列关系、偏正关系、主谓关系等，进而将这些关系转换成“对象-属性”关系；③半结构化数据一般是大量的 XML 格式的网页以及它们遵循的文档类型定义(XML Schema 或 DTD)等具有隐含结构的数据。半结构化数据具有结构化数据和非结构化数据的特征，从半结构化数据中抽取需要运用映射技术和自然语言分析技术相结合的办法来获取领域中的”对象-属性”关系。

1.2 分析模块

该模块是整个过程的核心，主要完成四项任务：

将准备模块得出的结果(即领域核心术语的”对象-属性”二元关系)纳入统一的形式背景下，并判断所形成的形式背景是否为标准形式背景，若不是，则分析原因(如多值背景、非净化背景等)，并采取对应措施(如多值形式背景单值化，背景净化)，将形式背景标准化。

通过造格算法，将标准形式背景转换成概念格，并将所得概念格通过 Hasse 图的形式显化出来，由领域专家和知识工程专家在可视化基础上判断概念格是否合理，对不合理的概念格通过一定的规则进行对象、属性编辑，循环操作，直至出现较为满意完备的概念格为止。对概念格的编辑处理的基本操作包括：添加或移除对象；添加或移除属性；当两个对象有相同的属性时，要么合并成一个对象，要么给对象添加属性，以区别对象。概念格可以产生新的对象，它们不在概念表中，可以增加这些对象；整个过程不断循环重复，直到合理完善为止。

将编辑后的完备概念格进行转换，主要包括节点转换(命名顶端节点，标示中间节点，删除底端节点)和节点关系转换(转换为概念及概念间的关系)两部分，转换的结果是得出领域本体原模型。

在领域专家的参与下，将领域本体原模型进行属性扩充、实例扩充、公理扩充及关系扩充，对领域本体原型进行完善，最终形成扩充后的领域本体原型。其中，属性扩充和实例扩充分别用于完善本体概念的内涵和外延的两个方面，关系扩充的目的在于完善领域本体概念除分类关系外的其余关系，而对公理和推理规则的扩充可以帮助实现本体推理。

1.3 本体描述模块

该模块的主要任务是选择合适的本体描述工具和本体描述语言，对扩充后的领域本体模型进行形式化描述，即完成本体的编码过程，最终得到领域本体。本体描述包括对领域概念、概念间关系、属性、实例、公理和推理规则等各个方面的描述。

本体描述的过程相当复杂，为方便和简化领域本体描述的具体过程，相关研究机构开发了一些有代表性的本体描述工具：JOE、OILed、OntoEdit、Protégé、WebOnto 等。这些工具在描述领域本体的能力上各有特点和优势，因此要结合具体的情况来选择使用。

本体描述语言近年来也呈现出多样化(如 OWL、DAML、RDF 等)的趋势，在此背景下，本体描述语言的选择就成为一个需要关注的问题。本文的观点是，本体描述语言的选择并非是唯一的，而是需要与具体的项目结合起来，与选择的本体描述工具结合起来，综合考虑各方面的因素，然后做出选择。一般情况下，选择 OWL 描述语言对本体进行描述。

2 E-mail 分类模型的设计

形式背景(formal context)是形式概念分析的基础数据，形式概念分析是由形式背景利用一定的方法得出形式概念，所有概念及其之间的偏序关系便构成了概念格。属性有单值和多值之分，所以形式背景也分为单值形式背景和多值形式背景。

定义 1 单值形式背景 $K=(G, M, I)$中的一个形式概念是一个对(A, B)，其中 $A\subseteq G$，$B\subseteq M$，满足:$A'=B$ 且 $B'=A$。A、B 分别称为形式概念(A, B)的外延(extent)和内涵(intent)。$B(G, M, I)$表示形式背景(G, M, I)所有形式概念的集合。

定义 2 一个多值形式背景(G, M, W, I)由集合 G、M 和 W 以及这三者之间的一个三元关系 I 组成，$I\subseteq G\times M\times W$ 且下式成立$(g, m, w)\in I$ 且$(g, m, v)\in I$ 总蕴含有 $w=v$。G 的元素称为对象，M 的元素称为(多值)属性，W 的元素称为属性值。$(g, m, w)\in I$ 读为"对象 g 的属性 m 有值 w"。称(G, M, W, I)为 n-值形式背景。

形式概念分析中使用的是单值形式背景，所以在处理多值形式背景时，需要将其转化为单值形式背景。将多值背景转化为单值背景，

可以通过概念定标(conceptual scaling)来完成。概念定标的基本思想是通过概念标尺(scale)从多值背景中导出形式背景。每个多值属性 $m\in M$ 可以被赋以一个概念标尺 S_m，它也是一个形式背景 $S_m:=(G_m, M_m, \in m)$，其中 $m(G)\subseteq G_m$。表 1 是以邮件为对象的多值形式背景示例，转换后的单值形式背景如表 2 所示。

表 1 关于邮件的多只形式背景

	附件类型	来源
D1	PIC DOC	163
D2	TXT	sina
D3	PPT EXE DOC	yahoo
D4	PPT TXT	163

表 2 转换后的单值形式背景

	PPT	PIC	TXT	EXE	其他	DOC	sina	163	yahoo
D1	Y	Y			Y		Y		
D2							Y		
D3	Y			Y		Y			Y
D4	Y		Y				Y		

FCA 的应用与文件管理的优势在于标准文档聚类的算法方面：任意一个文本都可以通过它的关键词(术语)来描述它的内容特征，特征是概念的外在表现形式，特征(关键词或术语)间存在很大的相关关系，即存在潜在的概念结构，如词汇之间的共现关系、同义关系等，分析这种相关关系对文本挖掘提供有用的帮助。把关键词结合 FCA 来对每一个邮件文档进行描述，能充分体现出文档间的关系，使得文档分类更明确。而且，聚类机制是一个格，而不是层次结构。概念格及其 Hasse 图体现了概念内涵和外延的统一，反映了对象和属性(特征)间的联系以及概念间的泛化和特化关系，可以用来为用户浏览导航。格结构在对文档进行多角度分类的时候使用起来更自然。

FCA 应用于文本检索在国内外已经有成功的例子，与项目组前期所做的 FCA 搜索引擎工作有所区别的是电子邮件与网页文档相比有自己的特点，需要在预处理模块做部分调整。电子邮件是一种半结构化的文本文件，包括邮件头(Head)和正文(即单一信息体，Body)。E-mail 头包括发送者的邮件地址(Sender Address)及邮件标题(Subject)等信息，邮件的标题通常表明了该 E-mail 的主要内容。E-mail 正文则包括该邮件的真实内容。对 E-mail 进行自动分类主要集中于对 E-mail 头的结构信息及正文内容进行分类。

3 E-mail 分类模型的建模

3.1 构造形式背景

为了更好地应用概念格，我们为 FCA 的 E-mail 分类模型提出新的定义。设定 E-mail 文档集合为对象集，能够分别代表 E-mail 的特征的关键词(Term)的集合构成属性集。形式背景定义如下:

定义 3 形式背景 $C=(D，T，I)$，其中 $G=\{d1，d2，d3，d4\}$，$T=\{t1，t2，t3，t4，t5，t6\}$，$I=D\times T$，表中的"Y"表明文档中有该关键词，"N"表示文档中没有该关键词。其二元关系见表 3，我们把已经取到本地的 E-mail 集合作为对象集 G，属性集 M 的大小决定了形式背景的规模。

表 3 形式背景

	t1	t2	t3	t4
D1	Y	N	Y	N
D2	Y	N	N	Y
D3	N	Y	N	Y
D4	Y	N	N	N

3.2 E-mail 属性抽取

E-mail 属性抽取最关键的一步是摘词，它决定 M 的大小和 E-mail 检索模块的精确度。E-mail 属性抽取的操作步骤如下:

①根据 POP3（或 MIME 协议），提取电子邮件文本信息；

②E-mail 文本信息预处理，将 html 等文件格式转换为 Txt 文本格式进行处理，包括通过用户或系统自主学习方式制定的“恶意地址规则库”采用简单的地址过滤或地址过滤加简单的关键词匹配过滤方法将一部分邮件过滤掉。对未过滤掉的 E-mail 去掉对 E-mail 分类无用的结构信息，只提取 E-mail 的标题和正文组成的文本文件；

③利用分词组件将文本文件分词，对要处理的文本进行词组分解的操作，成为一个词的序列，词组分解的同时已经标注出词性，所以统计词频之前可以首先将文本的词组进行过滤，比如可以根据需要只取或者兼取名词、动词、形容词和副词；

结束语 本文提出了基于 FCA 的 E-mail 分类方法，概念格作为 E-mail 检索浏览机制是有效的，基于 FCA 的 E-mail 分类模型解决了用户在浏览时邮件分类粗糙、每一类文件查询繁琐的问题。如何更精确地抽取概括性强、特征性强的属性集是我们将要进行的下一步工作。

参考文献

[1] 陈小莉.基于形式概念分析构建本体的方法研究[J].重庆广播电视大学远程教育技术中心，2009.

[2] 闫晶，王燕涛.基于形式概念分析的本体研究[J].东北电力大学，2012.

[3] 黄伟，金远平.形式概念分析在本体构建中的应用[J].东南大学计算机科学与工程系，2005.

[4] 曲开社，翟岩慧.偏序集-包含度与形式概念分析[J].山西大学计算机与信息技术学院，2006.

[5] 李云，刘宗田，吴强，沈夏炯，强宇.概念格的分布处理研究[J].上海大学计算机学院，2005.

[6] 杨帆，翟岩慧，曲开社，李德玉.基于形式概念分析的词义理解研究[J].山西大学计算机与信息技术学院，2011.

[7] 刘萍，高慧琴，胡月红.基于形式概念分析的情报学领域本体构建[J].武汉大学信息管理学院，2012.