第16卷 第1期
2006年1月

计算机技术与发展
COMPUTER TECHNOLOGY AND DEVELOPMENT

Vol.16 No.1
Jan. 2006

基于概念格的文本挖掘

王 娜[1,2],李云松[2]

(1.郑州航空工业管理学院,河南 郑州 450015;
2.兰州理工大学 计算机与通信学院,甘肃 兰州 730050)

摘 要:文本挖掘是从非结构化的文本中发现潜在的概念以及概念间的相互关系。作为从浩瀚的Web信息资源中发现潜在的、有价值知识的有效技术,Web文本挖掘已倍受关注。文中提出了利用概念格来抽取隐含在文本中潜在的概念关系,将文本挖掘中文档与关键词之间的关系通过概念格结构呈现出来。

关键词:文本挖掘;概念格;特征抽取
中图分类号:TP311 **文献标识码**:A **文章编号**:1005-3751(2006)01-0114-03

Text Mining Based on Concept Lattice

WANG Na[1,2], LI Yun-song[2]

(1. Zhengzhou Institute of Aeronautical Industry Management, Zhengzhou 450015, China;
2. College of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, China)

Abstract: Text mining depends on finding out the latent concepts and relationships between themselves from the original texts. As an effective technology to find potential, valuable knowledge from Web information resources, Web text mining has already been paid close attention to. The paper presents using the concept lattice to extract the latent concept relationships in the text. The relationships of the documents and the terms in the text mining are shown by the concept lattice.

Key words: text mining; concept lattice; feature extraction

0 引 言

面对今天浩如烟海的文本信息,如何帮助人们有效地收集和选择所感兴趣的信息,关键是如何帮助用户在日益增多的信息中发现新的概念,并自动分析它们之间的关系,使之能够真正做到信息处理的自动化,这已经成为信息技术领域的热点问题。在这样的需求驱动下,文本挖掘的概念产生了。

文本挖掘是从文本文件中抽取有效、新颖、有用、可理解的、散布的有价值知识[1],并且利用这些知识更好地组织信息的过程。在文本挖掘过程中,需要从信息中抽取有用的知识,即:通过为文本建立一个有意义的概念集合来看清概念的层次结构[2],从而在文本和概念之间挖掘它们的关系。而概念格,也称为Galois格,又叫做形式概念分析,由R. Wille等根据二元关系提出的一种概念层次结构[3],是数据分析与规则提取的一种有效工具。从数据集中生成概念格的过程实际上是一种概念聚类的过程,它的每个节点被称为一个概念,概念的外延表示为属于这个概念的所有对象的集合,而内涵则表示为所有这些对象所共有的属性的集合。概念格在本质上描述了对象和属性之间的联系,表明了概念之间的泛化和特化关系,而它的Hasse图则实现了对数据的可视化,作为数据分析和知识处理的形式化工具,概念格理论已被广泛地应用于软件工程、知识工程、数据挖掘、信息检索等领域。

文中利用概念格来揭示文本中潜在的概念间关系并从中抽取有用的知识。

1 概念格(concept lattice)基本理论

概念是由外延和内涵组成的。基于概念的这一哲学思想,德国Wille在20世纪80年代初期提出了一种形式化概念分析方法,用于概念的发现、排序和显示[4,5]。概念的外延表示为属于这个概念的所有事物的集合,而概念的内涵表示为所有这些事物所共同具有的特有属性的集合,即概念的描述。

概念之间是有序关系的,同时概念的内涵和外延间存在着反变关系,基于这种序关系所构造的格结构,就称为概念格。它是一种二元关系,描述了对象和特征之间的联系,表明了概念之间的泛化和特化关系,同时体现了概念内涵和外延的统一。而概念格所对应的哈斯图则形象地揭示了概念间的泛化和特化关系,反映出一种概念层次结构(Concept Hierarchy),实现了对数据的可视化,非常适用于从数据库中进行知识挖掘,从而成为数据分析和规则提

收稿日期:2005-04-11
作者简介:王 娜(1977—),女,河南南阳人,硕士研究生,研究方向为信息挖掘、知识发现。

取的一种有效工具。

定义 1:一个形式背景(context)是一个三元组 (U, D, R),其中,U 是对象的集合,D 是属性的集合,R 是 U 和 D 之间的二元关系,对于 $\forall x \in U, y \in D$,若 x 具有属性 y,则说 x 与 y 是有关的,记为 xRy 或者 $(x, y) \in R$。

定义 2:形式背景 (U, D, R) 的一个形式概念(简称概念)是一个二元组 (X, Y),它满足 $X' = Y$ 且 $Y' = X$,其中 $X \subseteq U, Y \subseteq D, X \to X' := \{d \in D \mid \forall u \in X: (u, d) \in R\}, Y \to Y' := \{u \in U \mid \forall d \in Y: (u, d) \in R\}$。$X$ 是概念 (X, Y) 的外延,Y 是概念 (X, Y) 的内涵。

定义 3:在概念节点之间能够建立起一种偏序关系。对于给定 (X_1, Y_1) 和 (X_2, Y_2),若 $(X_1, Y_1) \leqslant (X_2, Y_2) :\Leftrightarrow X_1 \subseteq X_2 (\Leftrightarrow Y_2 \subseteq Y_1)$ 成立,则称 (X_1, Y_1) 是 (X_2, Y_2) 的子概念,(X_2, Y_2) 是 (X_1, Y_1) 的超概念。关系"$\leqslant$"是概念的一个偏序(partial order)。

根据偏序关系可生成概念格的 Hasse 图。如果有 $C_1 > C_2$,并且不存在另一个元素 C_3 使得 $C_1 > C_3 > C_2$,则从 C_1 到 C_2 就存在一条边,即 C_1 是 C_2 的直接超概念,C_2 是 C_1 的直接子概念,形式背景 (U, D, R) 中,满足直接子概念—超概念关系的所有概念节点的集合是一个完全格,称之为 Galois 概念格,简称概念格。

2　概念格在文本挖掘中的应用

随着互联网飞速发展,互联网上的信息呈爆炸式增长。面对信息的海洋,如何准确有效地检索 Web 信息,帮助用户从大量文档信息集合中找到与给定查询请求相关的文档子集,也就成为一项重要的研究课题。搜索引擎是一种最为常见的 Web 信息检索系统,它虽然部分地解决了 Web 上资源发现问题,但是它往往会返回给用户成千上万个检索到的网页,而其中很大一部分与用户的检索要求无关,用户不能快速、准确地得到所需的有价值的信息。为此,需要开发比搜索引擎检索技术更高的新技术,这就是 Web 文本挖掘技术。

在文本挖掘过程中,需要从信息中抽取有用的知识,即:通过为文本建立一个有意义的概念集合来看清概念的层次结构,从而在文本和概念之间挖掘它们的关系。任一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[6],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,分析这种相关关系会对文本挖掘提供有用的帮助。如何在文本流中挖掘到潜在的概念以及概念间的相互关系,是十分必要的。而概念格及其 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。基于这种情况,文中提出了利用概念格来挖掘概念间的潜在关系。文中的主要思想是在自动索引和特定领域知识的辅助下,用关键词和文档作为文本特征,通过概念格找出文本中潜在的概念结构及文档和关键词之间的内在联系。

为了应用概念格,文中设定文档对应对象,依附于文档的关键词或术语(Term)构成属性集。这样,形式背景定义如下。

定义 4:一个基于文档的形式背景是一个三元组 $C = (D, T, I)$,其中 D 是文档(对象)集,T 是关键词(属性)集,I 是一个二元关系,它表明文档 d 中是否有关键词 t。如果 t 是 d 的关键词,则记为:dIt 或者 $(d, t) \in I$。

形式背景可以很容易地用交叉表来表示,交叉表的行表示文档,列表示关键词。如表 1 所示的一个形式背景,文档集 $D = \{d1, d2, d3, d4\}$,关键词集 $T = \{t1, t2, t3, t4, t5, t6\}$,$I = D \times T$,表中的"1"表明文档中有该关键词,"0"表示文档中没有该关键词。

表 1　形式背景

I	$t1$	$t2$	$t3$	$t4$	$t5$	$t6$
$d1$	1	0	1	0	1	1
$d2$	1	0	1	0	1	1
$d3$	0	1	0	1	0	0
$d4$	1	0	0	1	0	1

为了构建一个概念层次结构,必须找到形式背景 C 的所有概念。公式(1)和公式(2)可以用来计算所有的概念。首先,获得所有的行内涵 $\{d\}', d \in D$(公式 1)或者所有的列外延 $\{t\}', t \in T$(公式 2)。然后找出它们的交集以便 C 的概念的所有外延 X' 或者内涵 Y' 得以确定。依次类推,计算出所有确定外延的内涵。C 的所有概念的集合表示为 $B(D, T, I)$。

$$X' = \bigcap_{d \in X} \{d\}' \tag{1}$$

$$Y' = \bigcap_{t \in Y} \{t\}' \tag{2}$$

基于公式(2),表 2 给出了从表 1 的形式背景中获得所有概念的一个实例。详细算法过程如下:

步骤 1:明确表示出包含所有对象的概念的外延,即所有文档的集合 D。然后,对每个关键词(属性) t 执行下列步骤 m。

步骤 m:找出包含关键词 t 的文档集 X。随后,检查在列表中的任一外延是否等于 X。如果 X 的一个等价外延不在列表中,那么集合 X 就被作为一个外延。然后,计算 X 和前边几步计算的外延的交集。当这个交集不在列表中时,那么这个集合也添为一个外延。最后,根据形式背景 C(Context),可以找出对应每个概念外延的概念的内涵。这样,就获得了形式背景的所有概念的列表。由表 2 可以看出,针对表 1 中的形式背景,得出了 7 个形式概念。

最后,根据定义 3 找出概念集合 $B(D, T, I)$ 中的所有子概念—超概念关系(又称泛化-特化关系),依据它们之间的偏序关系就可以构建出概念格了。从表 1 形式背景 C 中得到的概念格如图 1 实线部分所示。图中每个节点代表一个形式概念 (X, Y),其中,X 是文档的集合,Y 是关键词的集合。

由于概念格的每个节点就是一个概念,概念的外延表示为属于这个概念的所有对象的集合,而内涵则表示为所

有这些对象所共有的属性的集合，所以概念格的每个概念就是具有最大共同属性的对象的集合，在形式背景中，外延即是由内涵所确定的等价类。由图 1 中的概念格就可以清楚地看出文档集和关键词集之间的这种内在关系。

表 2 由表 1 的形式背景获得的所有形式概念

步骤	关键词	外延	内涵
1		{d1,d2,d3,d4}	{}
2	t1,t6	{d1,d2,d4}	{t1,t6}
3	t2	{d3} {}	{t2,t4} {t1,t2,t3,t4,t5,t6}
4	t3,t5	{d1,d2}	{t1,t3,t5,t6}
5	t4	{d3,d4} {d4}	{t4} {t1,t4,t6}

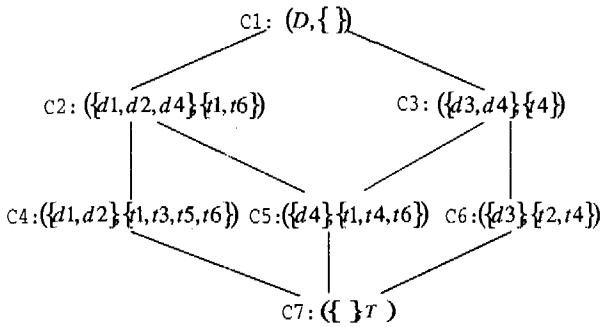

图 1 表 1 中形式背景 C 的概念格(Hasse 图)

另外，概念格节点间关系体现了概念之间的泛化和特化关系，因此非常适合作为规则发现的基础性数据结构。从构建好的概念格中可以提取蕴含规则，Godin 等[7] 提出了由概念格来提取蕴涵规则的算法，文献[8] 则提出了近似蕴涵规则的提取算法。从图 1 所示的概念格中提取蕴含规则示例如下：对于形式概念 C4 有两个自己的术语(关键词){t3,t5}，还继承了 C2 的两个术语{t1,t6}，如图中虚线所指。由形式概念 C4 可得到以下规则：

R1: $t3 \Rightarrow t1 \wedge t6$

R2: $t5 \Rightarrow t1 \wedge t6$

R3: $t3 \Leftrightarrow t5$

即规则 R1 和 R2 表示：关键词(术语)t3 或者 t5 的应用总是和关键词 t1 和 t6 的应用联系在一起的；规则 R3 表示：关键词 t3 和 t5 是相互等价的(或共现的)，也即：含有关键词 t3 的所有文档也包含有关键词 t5。

3 小 结

介绍了概念格和文本挖掘的有关理论，将文本中关键词和文档集之间的关系及文本中潜在的概念结构通过概念格清晰地表现出来，也得出一些有用的规则，如用户可以通过关键词搜索找到其中感兴趣的文档及关键词(术语)共现等。大部分搜索系统存在的明显问题是：当用户不能得到一个合适的搜索结果时，搜索引擎很难根据用户输入的关键词给出或设置合适的相关关键词，以启发用户进一步明确自己的查询问题。下一步的研究就是利用概念格来进行文本检索以期提高检索性能。

参考文献：

[1] 王继成，潘金贵，张福炎. Web 文本挖掘技术研究[J]. 计算机研究与发展，2002,37(5):513－520.

[2] 梅 馨，邢桂芬. 文本挖掘技术综述[J]. 江苏大学学报，2003,24(5):72－76.

[3] Zupa B, Bohance M. Learning by discovering concept hierarchies[J]. Artificial Intelligence, 1999, 109: 211－242.

[4] 朱福喜，汤怡群，傅建明. 人工智能原理[M]. 武汉：武汉大学出版社，2002.

[5] Wille R. Concept lattices and conceptual knowledge systems [J]. Computers and Math Applications, 1992, 23:5－9.

[6] 林鸿飞，战学刚，姚天顺. 中文文本挖掘的特征导航机制[J]. 东北大学学报，2000,21(3):240－243.

[7] Godin R, Missaoui R. An incremental concept formation approach for learning from databases[J]. Theoretical Computer Science, 1994, 133: 387－419.

[8] Missaoui R, Godin R. Extracting exact and approximate rules from databases[A]. In: Alagar V S, Bergler S, Dong F Q. Incompleteness and Uncertainty in Information Systems[C]. London: Springer－Verlag, 1994. 209－222.

(上接第 113 页)

4 结论及展望

在线健康保险系统是一套具有先进开发模式和先进管理理念的在线保险系统，将工作流思想引入系统的开发流程，能够大大提高系统的自动化程度，不仅规范了系统开发流程，同时加快了系统信息的更新周期，符合保险业的行业特点。

在下一步的工作中，将进一步深入研究工作流技术在系统业务管理[5] 中的应用。随着工作流技术在提高自动化流程控制应用上的不断发展与完善，保险业在信息化及网络化服务水平也将不断得到提高，健康保险业实现跨越式的发展就在眼前。

参考文献：

[1] 江 平，左 春，陈宝兵. 基于 J2EE 体系结构的保险电子商务系统的设计研究[J]. 计算机应用研究，2004(3):18－20.

[2] 杨一平. 软件能力成熟度模型 CMM 方法及其应用[M]. 北京：人民邮电出版社，2001. 10－17.

[3] 范玉顺. 工作流管理基础[M]. 北京：清华大学出版社，2001. 27－32.

[4] van der Aalst W, van Hee K. Workflow Management: Models, Methods, and Systems (Cooperative Information Systems)[M]. Cambridge, London, England: MIT Press, 2002. 148－150.

[5] Fischer L. The Workflow Handbook 2002[M]. FL, USA: Future Strategies Inc., 2002. 133－138.