

概念的矩阵化描述及相关定理的矩阵化证明

董辉 马垣 宫玺

(辽宁科技大学计算机科学与工程 鞍山 114051)

摘要 概念格作为形式概念分析理论中的核心数据结构,已在知识工程和软件工程等领域得到了广泛的应用。然而,对概念格理论的研究仅存在于纯数学领域^[1],虽然已有的 Hasse 图能体现概念间的泛化与特化关系,但对概念本身的理解帮助不大,用于定理的证明更为少见。从可视化着手,对概念格的图形化研究做了一点尝试,把概念信息理解成概念矩形,并用之于概念格定理的证明,取得满意的效果。

关键词 概念格,概念矩形,概念格生成算法,批处理,渐进式

Rectangle Description of Concept and the Rectangle Proof of Correlative Propositions

DONG Hui MA Yuan GONG Xi

(School of Computer Science and Engineering, Liaoning University of Science and Technology, Anshan 114051, China)

Abstract Concept lattice, the core data structure of formal concept analysis, is used widely in knowledge engineering and software engineering. However, the scientific researches on the concept lattice is confined in the domain of mathematics^[1]. The Hasse diagram of a concept lattice can describe and perform the connections between generalization and specialization, but it is helpless to the comprehension of the concept essentially and rareness to be used in the theorem proof. Based on the visualization, this article made an attempt to the visualization of the concept lattice, regarded the concept as concept rectangle, and used in the concept lattice theorem proof, obtained the satisfactory effect.

Keywords Concept lattice, Concept rectangle, Constructing algorithm, Batch algorithm, Incremental algorithm

数据库知识发现是当前人工智能研究中的一个重要领域,由 Wille 在 1982 年提出的形式概念分析是近来获得飞速发展的数据分析的有力工具,在数据挖掘和规则提取领域已有很好的应用。形式概念分析的核心数据结构是概念格,其每个节点是一个形式概念,由两部分组成:外延和内涵。概念格所对应的 Hasse 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,实现了对数据的可视化。

概念的图形化就是用直观的方式描述概念格中的概念。目前,对概念格的研究主要集中在以数学为基础的格理论、集合理论和逻辑理论等领域,从一个称为形式背景的二维表格中生成所有的概念,进而研究概念间的关系。然而,在生成每个概念前,形式背景中对象集中的每个元素或属性集中的每个元素,在未规定字典序前,是不可比的。本文从这一点着手,基于概念的图形化特点,把概念所包含的信息理解为一个概念矩形,对概念格的图形化研究做了一点尝试,并用得出的结论对建格算法中相关定理进行了证明。

1 基本概念

定义 1 形式概念中称给定的信息表 $K=(G, M, I)$ 为形式背景,其中 G 是对象的集合, M 是属性集合, I 是 G 和 M 之间的一个二元关系,且 $I \subseteq G \times M$, gIm 表示 $g \in G$ 与 $m \in M$ 之间存在关系 I 。对于一个对象 $g \in G$,那么 gIm 就表示对象 g

具有属性 m ^[2]。

定义 2 形式背景 $K=(G, M, I)$ 中的对象集合 $A \in P(G)$, 属性集合 $B \in P(M)$, 满足 $A \times B \subseteq I$, 使得 $\forall x \in G - A$, 有 $\exists y \in B | (x, y) \notin I$ 和 $\forall y \in M - B$, 有 $\exists x \in A | (x, y) \notin I$ 的二元组 (A, B) 称为关系 I 的一个形式概念或一个概念矩形,称 A 为该概念矩形的外延, B 为该概念矩形的内涵,其中 $P(G)$ 是对象集合 G 的幂集, $P(M)$ 是属性集合 M 的幂集,称 $\{m\}' | m \in M$ 为 m 的对象矩形条, $\{g\}' | g \in G$ 为 g 的属性矩形条。且定义运算: $A' = B, A = B'$ 。

I	内涵					
	a	b	c	d	e	f
外延	1	×	×			×
	2		×	×		×
	3			×	×	×
	4	×	×	×	×	×
	5			×	×	

图1 一个形式背景 K

I	内涵					
	a	b	c	f	e	d
外延	1	×	×		×	×
	2		×			
	3					×
	4	×	×			×
	5		×			×

图2 背景 K(属性 f 与 d 互换)

例 1 图 1 给出了一个形式背景 K , $\times$ 表示所对应的对象和属性具有关系 I , 即 gIm 。该背景下的概念矩形为 $(1, abef)$, $(14, abf)$, $(124, bf)$, $(1234, f)$, $(12345, \emptyset)$, $(24, bc f)$, $(234, cf)$, $(2345, c)$, $(4, abcd f)$, $(34, cd f)$, $(345, cd)$ 。

未规定字典排序时,属性间或对象间没有可比性,可以互

到稿日期:2008-05-20 本文受国家自然科学基金(No. 60775036)资助。

董辉(1975—),男,硕士研究生,工程师,主要研究方向为概念格、数据挖掘等;马垣(1941—),男,博导,主要研究方向为粗糙集、概念格、人工智能等, E-mail: mayuanas@sina.com;宫玺(1981—),女,硕士研究生,主要研究方向为概念格、数据挖掘等。

换对象或属性的位置来突出显示概念矩形,如图2所示中的概念矩形(234,cf)。

一个概念矩形可简化为如图3所示,对图2而言,图3的边A表示对象集合{234},边B表示属性集合{cf}。

属性f的对象矩形条如图4所示,窄边表示单个对象,长边表示阴影部分的对象集合。属性矩形条同理可得。

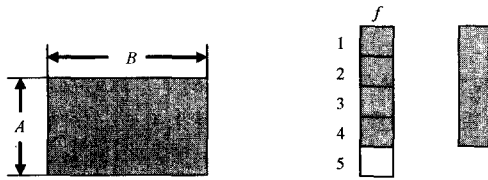


图3 概念矩形的简化

图4 属性f对象矩形条及其简化

定义3 (A_1, B_1) 和 (A_2, B_2) 是某个背景上的两个不同概念矩形,且 $A_1 \subset A_2 \Leftrightarrow B_2 \subset B_1$,如果不存在概念矩形 (A_3, B_3) 有 $A_1 \subset A_3 \subset A_2 \Leftrightarrow B_2 \subset B_3 \subset B_1$,则称 (A_1, B_1) 是 (A_2, B_2) 的直接例化, (A_2, B_2) 是 (A_1, B_1) 的直接泛化^[2], (A_1, B_1) 是 (A_2, B_2) 的子节点(子概念矩形);称 (A_2, B_2) 是 (A_1, B_1) 的父节点(父概念矩形)。

例2 例1中,父概念矩形(234,cf)及子概念矩形(24,bcf),如图5所示。其中, R_1 表示概念矩形(234,cf)和 R_2 表示概念矩形(24,bcf)。

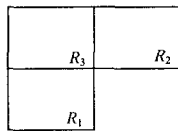


图5 父子概念间的关系

2 基本定理

定理1 对于概念矩形 (A, A') 和 (B, B') ,满足 $A \cap B \neq \emptyset$,称 (A, A') 和 (B, B') 的对象矩形边是可以部分重合的。同理,适用于属性边。如果 (A, A') 和 (B, B') 的对象边和属性边都可以部分重合,则称这两个对象矩形是可以部分重合的。

未规定字典序时,可以通过调整对象的顺序或属性的顺序来达到重合的目的,把 $A \cap B$ 中的元素和 $A' \cap B'$ 中的元素分别连续排列,设 $A \cap B = C$,则这3个概念矩形的图形关系如图6所示。

定理2 未规定字典序时,互换对象或属性的位置,父概念矩形 R_1 和子概念矩形 R_2 的图示关系唯一,如图6所示,即不存在这样概念矩形 R_3 和对象矩形条 m_1 和 m_2 (或属性矩形条)存在,如图6和图7所示。

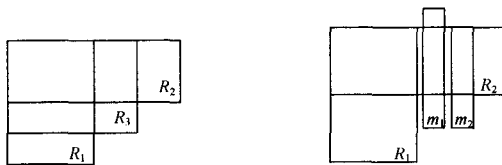


图6 R_3 与 R_1 和 R_2 关系

图7 m 与 R_1 和 R_2 关系

证明:设矩形 R_1, R_2 和 R_3 分别对应于概念矩形 $(A_1, B_1), (A_2, B_2)$ 和 (A_3, B_3) ,满足关系 $A_1 \subset A_2 (\Leftrightarrow B_2 \subset B_1)$,因为概念矩形 (A_2, B_2) 是概念矩形 (A_1, B_1) 的父概念矩形,即不存在概念矩形 (A_3, B_3) 存在满足关系 $A_1 \subset A_3 \subset A_2 (\Leftrightarrow B_2 \Leftrightarrow$

$B_3 \subset B_1)$,即不存在图中所示 R_3 的存在,同理可证概念矩形条和属性矩形条,得证。

3 建格算法中相关定理的矩阵化证明

目前概念格的构造算法可以分为两大类:渐进式算法和批处理算法。批处理算法的基本思想是先生成所有概念,然后根据它们之间的泛化-例化关系,生成边,从而形成概念格。渐进式构造算法的基本思想是先初始化为空概念,然后根据当前要插入对象的属性与概念格节点的内涵交集的不同,而采取不同的操作,逐渐形成概念格。

3.1 批处理算法

在概念格的批处理算法中,最具有代表性的就是Next-Closure算法。该算法假定在属性集上存在一个线性序,根据已得到的结果作为下一次的计算的属性集,采用正规测试,只生成满足要求的属性集里的某些闭包^[2]。

定义4 设集合 M 的两个不同子集 A 和 B ,规定属性字典序为 $m_i < m_j (i < j, m_i \in M, m_j \in M)$,若 A 和 B 中不同的最小元素 m_i 属于 B ,则称 A 按字典排序小于 B ,记为 $A <_i B$ 。即对于 $A, B \subseteq M$,且 $A \neq B \Leftrightarrow \exists i (m_i \in B, m_i \notin A, \forall j < i (m_j \in A \Leftrightarrow m_j \in B))$ 。

定义5 设 $A \subseteq M, m_i \in M$,则定义 $A \oplus m_i = (f(A))'$,其中 $f(A) = (A \cap \{m_i, \dots, m_{i-1}\}) \cup \{m_i\}$ 。

设 $R_{f(A)}, R_A$ 和 $R_{A \oplus m_i}$ 表示 $f(A), A$ 和 $A \oplus m_i$,根据边的部分可重合性,三者之间的关系如图8所示。

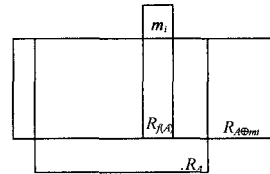


图8 $R_{f(A)}, R_A$ 和 $R_{A \oplus m_i}$ 的关系图

定理3 给定集合 $A, A \subset M$,在某字典排序下,比 A 大的最小封闭集是 $A \oplus m_i$,其中, m_i 是 M 中满足 $A <_i A \oplus m_i$ 的最大元素^[2]。

证明:用 R_A 和 R_{A^*} 分别表示概念矩形 (A, A') 和 $(A \oplus m_i, (A \oplus m_i)')$,比 A 大的封闭集存在于 (A, A') 的子概念矩形的内涵的集合中,若与概念矩形 (A, A') 作运算的矩形为对象矩形条,如图9和图10所示,则可使运算后所得的概念矩形的外延的势最大,内涵的势最小。如果 R_{A^*} 的左边界不超出 R_A 的左边界,则 m_i 就是满足要求的最大元素,得证。

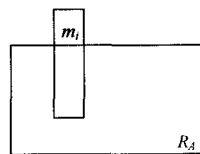


图9 R_A 与 m_i 的关系图

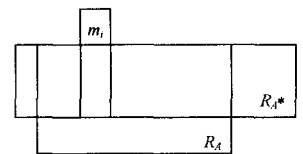


图10 R_A, R_{A^*} 与 m_i 的关系图

3.2 渐进式算法

渐进式算法又称为增量式算法,基本思想是将当前要插入的对象和格中所有的概念交,根据交的结果采取不同的行动。

基于属性的渐进式构造概念格算法就是在原始背景 $K = (G, M, I)$ 所对应的原始概念格 $L(K)$ 和新增属性 m^* 的情况

下,来求解形式背景 $K^* = (G, M \cup \{m^*\}, I^*)$ 所对应的概念格 $L(K^*)$ 。可以把概念格 $L(K^*)$ 中概念分为 3 类:①不变概念;②更新概念;③新增概念^[3]。

设新增的属性为 m^* , 具有该属性的对象集为 $\{m^*\}'$, 这样就可以用属性概念的形式表示为 $C^* = (\{m^*\}', m^*)$, 对这 3 种类型的概念进行如下定义。

定义 6 如果一个概念矩形 $C = (A, A')$ 满足 $A \cap \{m^*\}' = \emptyset$, 则称该概念矩形为不变概念矩形。显然地, 对于一个不变概念矩形来说, 它在新概念格 $L(K^*)$ 中保持不变, 仍然是 (A, B) , 如图 11 所示。

定义 7 如果一个概念矩形 $C = (A, A')$ 满足 $A \subset \{m^*\}'$, 则称该概念矩形为更新概念矩形。显然地, 对于一个更新概念矩形来说, 它将被更新为 $(A, A' \cup \{m^*\})$, 如图 12 所示。

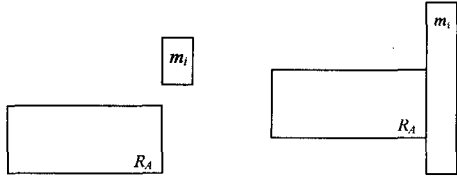


图 11 不变概念矩形

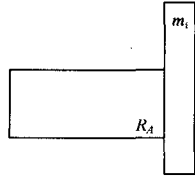


图 12 更新概念矩形

定义 8 如果存在某个概念矩形 $C_1 = (A_1, B_1) \in L(K)$, 并且满足条件:①在格中不存在任意概念矩形 $C_2 = (A_2, B_2)$ 的外延是 $A_2 = A_1 \cap \{m^*\}'$;②对于 C_1 的任意孩子概念矩形 $C_3 = (A_3, B_3)$, 都没有 $A_3 \cap A_1 = A_1 \cap \{m^*\}'$, 则称 C_1 为新增概念矩形的产生子概念矩形, 如图 13 所示。

定理 4 若 (A_{N+1}, B_{N+1}) 是 (A_N, B_N) 的父概念, 则 $\forall m \in B_N - B_{N+1}$, 有 $\{m\}' \cap B_{N+1}' = A_N$ 。

证明: 设概念矩形 (A_{N+1}, B_{N+1}) 和 (A_N, B_N) 分别对应图中的 R_{N+1} 和 R_N , 根据定理 1 知, 仅有 $\{m\}'$ 这样的对象矩形条存在, 如图 14 所示可知 $\{m\}' \cap B_{N+1}' = A_N$, 得证。

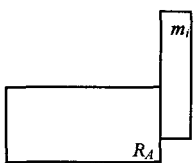


图 13 新增概念矩形

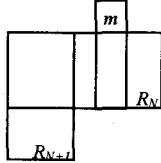


图 14 父子概念生成关系

定义 9 如果原概念格中概念矩形 (A, A') 和 (B, B') 满足 $A \subseteq B$, 添加新的对象矩形条 $\{m\}'$ 后, 都能生成新的概念矩形 (C, C') , 称 $R_A = (A, A')$ 是 $R_C = (C, C')$ 的典型生成元, $R_B = (B, B')$ 是 (C, C') 的非典型生成元, 如图 15 所示。

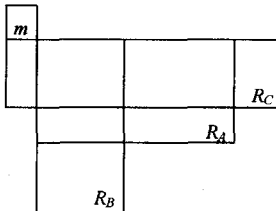


图 15 典型生成元与非典型生成元

定理 5 如果 (A, A') 是一个新概念 (C, C') 的典型生成元, 而 (B, B') 是 (C, C') 的非典型生成元, 有 $A \subseteq B$, 则任意概念 (E, E') 满足 $E \subset A'$, $E \not\subset A$, 即不是更新概念也不是任何新

概念的典型生成元^[4]。

证明: 未定义字典序时, 设 R_A, R_B, R_C 和 R_A 分别表示概念矩形 $(A, A'), (B, B'), (C, C')$ 和 (E, E') , $\{m\}'$ 为 m 的对象矩形条, R 的长和宽分别代表该概念矩形的外延和内涵, 则这些矩形的图示分布可分为下列 3 种情况:

①概念矩形 (E, E') 满足 $E \subset A'$, $A \subseteq E$, $A \cap E = \emptyset$ 则 (E, E') 为非典型生成元, 非更新概念, 如图 16 所示。

②概念矩形 (E, E') 满足 $E \subset A'$, $A \cap E \neq \emptyset$, $A - E \neq \emptyset$, $E - A \neq \emptyset$, 则 (E, E') 为非更新概念, R_A 与 R_A 重合部分一定是一个伪概念矩形, 则 (E, E') 是非典型生成元, 如图 17 所示。

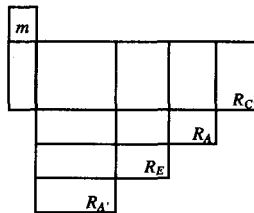


图 16 典型生成元情况 1

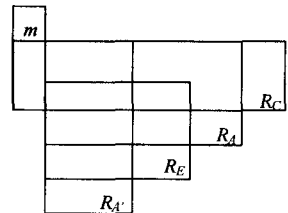


图 17 典型生成元情况 2

③概念矩形 (E, E') 满足 $E \subset A'$, $A \cap \{m\}' = \emptyset$, 则 (E, E') 为不变概念, 如图 18 和图 19 所示。得证。

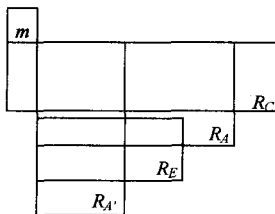


图 18 典型生成元情况 3

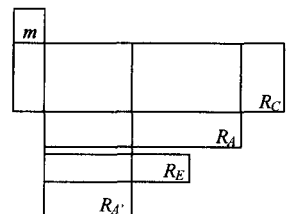


图 19 典型生成元情况 4

定理 6 如果 (A, A') 是一个原概念且 $A \cap \{m\}' = C$, $(C, C') \in L$ 是更新概念, 则任意概念 (E, E') 满足 $E \subset A$, $E \not\subset C$, 即不是更新概念也不是任何新概念的典型生成元^[4]。

证明: 未定义字典序时, 设 R_A, R_C 和 R_E 分别表示概念矩形 $(A, A'), (C, C')$ 和 (E, E') , $\{m\}'$ 为 m 的对象矩形条, R 的长和宽分别代表该概念矩形的外延和内涵, 则这些矩形的图示分布可分为下列 3 种情况:

①概念矩形 (E, E') 满足 $E \subset A$, $C \subseteq E$, 则 (E, E') 为非典型生成元, 非更新概念, 如图 20 所示。

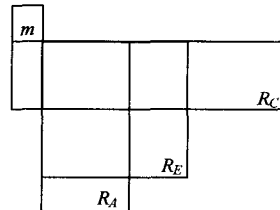


图 20 原概念情况 1

②概念矩形 (E, E') 满足 $E \subset A$, $C \cap E \neq \emptyset$, $C - E \neq \emptyset$, $C - A \neq \emptyset$, 则 (E, E') 为非更新概念; R_C 与 R_E 重合部分一定是一个伪概念矩形, 则 (E, E') 是非典型生成元, 如图 21 所示。

③概念矩形 (E, E') 满足 $E \subset A$, $C \cap E = \emptyset$, 则 (E, E') 为不变概念, 如图 22 所示。得证。

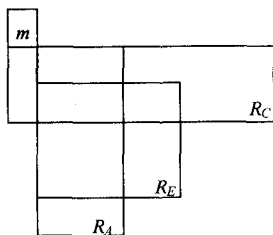


图 21 原概念情况 2

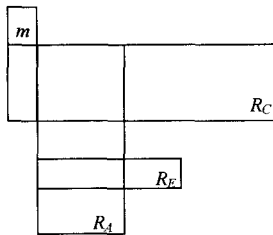


图 22 原概念情况 3

4 并行建格算法中相关定理的矩阵化证明

概念格的分布处理思想就是根据不同的原理,采用分治策略,通过形式背景的拆分,形成分布存储的多个子背景,然后构造相应的子概念格,再由子概念格的合并得到所需的概念格。

4.1 现有并行建格算法简介

4.1.1 形式背景的并置和叠置

并置或叠置可理解为两个对象域相同、属性项不同或两个对象域不同、属性项相同的形式背景间的合并。该算法在内涵或外延一致的形式背景、基于横向合并的子形式背景的概念格和子背景所对应的子概念格的横向并是同构的基础上,通过概念的横向(或纵向)加运算和概念格的横向(或纵向)并运算,引入了子格和完整格之间的两个映射函数来通过部分格概念计算原始背景的所有概念^[2,5]。

4.1.2 形式背景的折叠搜索子空间划分

此算法基于对生成概念的搜索空间的分解,使用了一个新的方法将初始背景分解为不交叠的子空间,在每个子空间内计算并生成概念。先将初始背景按属性的权值排序,按给定的划分系数来分解初始背景。经过划分,概念的求解的时间复杂度和空间复杂度有明显的改善^[6,7]。

4.2 基于闭包划分的概念格并行生成算法相关定理的矩阵化证明

概念格是其形式背景中概念间关系的表现形式,它和对应的形式背景是一一对应的。把一个形式背景拆分成了多个子背景,这样形式背景所对应的概念格就可由各子格进行合并来实现。

4.3 Para_Prun 建格算法的相关定义

定义 10 设 $F(I)$ 是形式背景 $K=(G, M, I)$ 对象 G 上的闭包系统,定义

$$F_g(I) = \{F \in F(I) \mid g \in F, F \subseteq 2^G\}$$

$$F_g^-(I) = \{F \in F(I) \mid g \notin F, F \subseteq 2^G\}$$

$$F_g^+(I) = \{F \in F(I) \mid g \notin F, F \subseteq 2^G\},$$

并且是背景 K 中不包含 g 的闭包系统,满足 $F(I) = F_g(I) \cup F_g^-(I), F_g^-(I) \subseteq F_g^+(I)$, 其关系如图 23 所示。

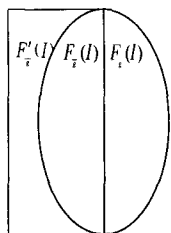


图 23 三者关系

定义 12 设 $K=(G, M, I)$ 为一形式背景,定义运算 $\uparrow g$, 满足 $\uparrow g = \{g_i \in G \mid g'_i \subseteq g'\}$ 。

定义 13 设 $K=(G, M, I)$ 为一形式背景, K_g 和 K_g^- 是由 K 分解包含 g 与不包含 g 的两个子背景,满足 $K_g = (G_1, g', G_1 \times g' \cap I), K_g^- = (G \setminus \uparrow g, M_2, G \setminus \uparrow g \times M_2 \cap I)$, 其中 $\forall g \in G_1 \mid g' \neq \emptyset, \forall m \in M_2 \mid m' \neq \emptyset$ 。

设对象 g 及形式背景 K , 经重新排序对象元素和属性元素, $\uparrow g$ 集合及形式背景如图 24 所示, K_g 为矩形的上半部分, K_g^- 为矩形的下半部分。

定义 14 设 $K=(G, M, I)$ 为一形式背景,定义 $F(I)$ 背景 K 的外延的闭包系统, g 为 $\vee$ -Prime 元素, 满足

$$\textcircled{1} |F(I)| \leq 1;$$

$\textcircled{2} F(I) = F_g(I) \cup F_g^-(I), F_g(I)$ 和 $F_g^-(I)$ 分别为背景 K_g 和 K_g^- 的外延的闭包系统。

定理 7 设 $K=(G, M, I)$ 为一形式背景, $g \in G$ 是 $\vee$ -Prime 元素, 则 $\forall A \in F_g(I), \exists m \in M \setminus \uparrow g'$, 满足 $A \setminus \uparrow g \subseteq m'$ 。

证明: 设 $g \in G$ 是 $\vee$ -Prime 元素, 实际上是要求在背景 $K' = (G \setminus \uparrow g, g', G \setminus \uparrow g \times g' \cap I)$ 不存在这样的属性 m , 满足 $m' \not\subseteq m'_i$; 在背景 $K'' = (G \setminus \uparrow g, M, G \setminus \uparrow g \times M \cap I)$ 中求概念就不会有伪概念出现, 满足 $K_g = (G_1, g', G_1 \times g' \cap I)$ 和 $K_g^- = (G \setminus \uparrow g, M_2, G \setminus \uparrow g \times M_2 \cap I)$ 是背景 $K=(G, M, I)$ 的两个划分, 如图 25 所示。得证。

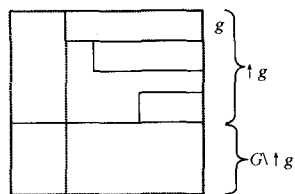


图 24 $g, \uparrow g$ 及 $G \setminus \uparrow g$ 关系图

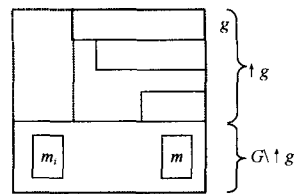


图 25 $\vee$ -Prime 判断证明

由图 25 还可得出以下结论:

定理 8 设 $K=(G, M, I)$ 为一形式背景, $g \in G$ 是 $\vee$ -Prime 元素, 则 $\forall A \in F_g(I), \exists B \in F_g^-(I), (A \setminus \uparrow g)'' \subseteq B$ 。

定理 9 设 $K=(G, M, I)$ 为一形式背景, $g \in G$ 是 $\vee$ -Prime 元素, 则 $\forall m \in g'$, 满足 $(m' \setminus \uparrow g) \subseteq g'$ 。

定理 10 设 $K=(G, M, I)$ 为一形式背景, $g \in G$ 是 $\vee$ -Prime, $\forall m \in g'$, 则 $\exists m_i \in M \setminus \uparrow g', m' \setminus \uparrow g \subseteq m_i$ 。

定理 11 设 $K=(G, M, I)$ 为一形式背景, $g \in G$ 是非 $\vee$ -Prime 元素, 则 $\exists m \in g', m' \setminus \uparrow g \setminus m'_i \neq \emptyset$, 其中 $m_i \in M \setminus \uparrow g$ 。

结束语 本文从一个称为形式背景的二维表格出发, 基于集合中未给定排序元素间的无可比性, 把概念及概念间的关系信息以概念矩形的图形形式描述出来, 并用之于概念格构造算法定理的证明, 取得初步成效。若规定元素的排序, 将会增加概念矩形图形化的复杂度, 如定理 3 的证明。如何简单明了地用图形描述概念及概念间的关系将是下一步的研究工作。

参考文献

- [1] 马垣. 粗糙几何学//中国第6届粗糙集及软计算学术会议报告. 2006
- [2] Ganter B, Wille R. Formal Concept Analysis; Mathematical Foundations. Berlin-Heidelberg, Springer, 1999

[3] Godin R. Incremental Concept formation algorithm base on Galois (concept) lattices. Computational Intelligence,1995,11(2): 246-267

[4] van der Merwel D. AddIntent : A New Incremental Algorithm for Constructing Concept Lattices// ICFCA 2004, LNCS 3403. 2004;372-385

[5] 李云,刘宗田,陈峻,等. 多概念格的横向合并算法. 电子学报, 2004,32 (11):1849-1854

[6] Fu H,Nguifo E M. A parallel algorithm to generate formal concepts for large data. ICFCA,2004;394-401

[7] 齐红,刘大有,等. 基于搜索空间划分的概念生成算法. 软件学报,2005,16

(上接第 154 页)

系统能有效地检索出与用户输入相关的类别,并给出输入词到它们的语义距离,且这个距离是满足图的最优选择性的。例如:fish 到下位词 dolphin 的语义距离为 5,到上位词 animal 的语义距离为 4。因此,dolphin 到上位词 fish 的语义距离也为 5,而到 animal 的语义距离则为 5(与 fish 的距离)+4(fish 与 animal 的语义距离)=9。

表 1 检索结果

输入	同义词		下位词		上位词	
	语义距离	关键词	语义距离	关键词	语义距离	关键词
fish	0	fish	3 5	shark dolphin	4	animal
shark	0	shark	\	\	3 7	fish animal
dolphin	0	dolphin	\	\	5 9	fish animal
pig	0	pig	\	\	8	animal
dog	0	dog	\	\	2	animal
horse	0	horse	\	\	8	animal

7 语义网络的扩展

WordNet 并不一定在所有领域里都是很完备的,因此利用它构建的语义网络也可能存在不合事实的情况,所以,为了进一步完善我们的语义网络,要求它必须是可修改和可扩展的。我们建立的语义网络是用本体描述语言 OWL 来描述的,OWL 语言具有很强的描述事物的表达能力,它是 W3C 组织推出的一种语言,得到大多数人的认可。用它来描述语义网络不仅具有很好的扩展性,而且所得的语义网络是可共享的。

结束语 我们系统的原始词汇只有 197 个,即系统只有 197 个关键节点,经过扩展后得到 13027 个概念。这大大增加了系统的词汇量,提高了系统词汇的覆盖面,从而提高了召回率。而我们建立的语义网络,是对原始词汇的语义扩展,检索时是根据语义关系来做推理的,因此,检索结果和用户输入的检索词在语义上具有较大的相似度,这就较好地保持了检索的准确率。

然而,我们还没有对 WordNet 的语义关系作充分的挖掘,所构建的语义网络还是一个较为简单的模型,还处于初期阶段。今后,我们将会不断地对系统做扩展,并利用其它本体对 WordNet 作补充,使所构建的系统更加完善。

参 考 文 献

[1] Fellbaum C. WordNet : An Electronic Lexical Database [M].

Cambridge, Mass; Bradford Books,1998

[2] Khan L,Luo F. Ontology Construction for Information Selection [A]// ICTAI'02: Proceedings of the 14th IEEE International Conference on Tools with Artificial Intelligence[C]. Washington DC:IEEE,2002;122-127

[3] Baziz M,Boughanem M, Aussenac-Gilles N,et al. Semantic cores for representing documents in IR[A]//SAC '05:Proceedings of the 2005 ACM symposium on Applied computing [C]. New York,ACM,2005;1011-1017

[4] Guarino N, Masolo C, Vetere G. OntoSeek: Content-Based Access to the Web[J]. IEEE Intelligent Systems,1999,14(3):70-80

[5] Hovy E,Knight K, Marcu D,et al. Large Resources Ontologies (SENSUS) and Lexicons[EB/OL]. [http://www. isi. edu/natural-lang uage/projects/ONTOLOGIES. html](http://www.isi.edu/natural-lang uage/projects/ONTOLOGIES. html),2008-05-14

[6] 宋炜,张铭. 语义网简明教程[M]. 北京:高等教育出版社,2004: 136-151

[7] Bwalenz,Didion J. JWNL - Java WordNet Library[EB/OL]. <http://jwordnet. sourceforge. net>,2008-05-14

[8] HP Labs Semantic Web Research. Jena-A Semantic Web Framework for Java[EB/OL]. <http://jena. sourceforge. net/index. html>,2008-05-14

[9] digital craftsmen. WordNet-Similarity[EB/OL]. <http://search. cpan. org/ dist/WordNet-Similarity>,2008-05-14

[10] Rada R,Mili H,Bichnell E,et al. Development and application of a metric on semantic nets[J]. IEEE Transactions on Systems, 1989,19(1):17-30

[11] Banek M,Vrdoljak B,Tjoa A M. Using Ontologies for Measuring Semantic Similarity in Data Warehouse Schema Matching Process[A]// ConTEL 2007: Proceedings of the 9th International Conferenceon Telecommunications[C]. Washington DC: IEEE, 2007;227-234

[12] Liu X , Zhou Y , Zheng R . Measuring Semantic Similarity in Wordnet[A] // Proceedings of the Sixth International Conference on Machine Learning and Cybernetics [C]. Washington DC:IEEE,2007;3431-3435

[13] Yang D,Powers D M. Measuring semantic similarity in the taxonomy of WordNet[A]// Proceedings of the Twenty-Eighth Australasian Conference on Computer Science [C]. Darlinghurst, Australia, Australia: Australian Computer Society, 2005: 315-322