基于模糊概念格的 Web 搜索结果聚类算法

黄健斌[1,2]，姬红兵[1]
(1. 西安电子科技大学 电子工程学院,陕西 西安 710071;
2. 西安电子科技大学 计算机学院,陕西 西安 710071)

摘要:提出了一种模糊形式概念分析方法,给出了在对象和属性的模糊二元关系上生成模糊概念格的过程.提出了一种在格的拓扑序列上进行概念聚类的快速算法,并且定义了概念聚类间基于偏序的层次关系.该方法利用格理论解决了概念聚类中概念间的多重继承关系,应用在 Web 搜索结果聚类上,实验结果表明算法在聚类质量和检索性能上都有改进和提高.

关键词:模糊概念格;概念聚类;Web 搜索结果聚类
中图分类号:TP311 **文献标识码**:A **文章编号**:1001-2400(2005)06-0856-05

A Web search results clustering algorithm based on fuzzy concept lattices

HUANG Jian-bin[1,2], *JI Hong-bing*[1]
(1. School of Electronic Engineering, Xidian Univ., Xi'an 710071, China;
2. School of Computer Sceince and Technology, Xidian Univ., Xi'an 710071, China)

Abstract: Fuzzy logic is incorporated into the Formal Concept Analysis to tackle the vague data in the formal context. A fuzzy concept lattice can be generated in the fuzzy formal context. Next, a fast fuzzy conceptual clustering algorithm is proposed to cluster the fuzzy concept lattice into conceptual clusters. Then, partial order relations are defined between the clusters for construction of the concept hierarchy. This method is applied in the Web search results clustering. Experimental results show the algorithm's effectiveness and suitability for solving the multi-inheritance problems between conceptual clusters.

Key Words: fuzzy concept lattices; conceptual clustering; Web search results clustering

概念聚类是一种基于模型的聚类方法,它能够对输出的聚类确定其属性特征,从而对聚类结果给予一定的概念解释[1,2].根据概念属性的范化与特化关系,可得到不同概念间的层次关系.典型的概念聚类方法有 COBWEB[2]、基于神经网络的竞争学习和自组织特征映射等.但这些概念聚类算法产生的聚类层次大多是基于树状结构的,不支持概念间多重继承关系的表示.近年来,出现了一些基于形式概念分析(FCA)[3]的概念聚类系统,例如 TOSCANA,INCOSHAM 等.传统 FCA 定义的形式背景中对象与属性的关系是二值的,在知识完全的情况下产生的概念格是准确的.但是,多数情况下数据聚类的信息是不确定的.为了处理聚类过程中的模糊信息,给出了一种基于模糊概念格的概念聚类方法,并将这种概念聚类方法应用在 Web 搜索结果聚类上,实验结果表明该方法是有效的.

1 模糊形式概念分析

传统的形式概念分析是目前得到成功应用的一种概念知识结构形式化方法.它首先将领域中的对象和属性间的二元关系建模为形式背景,然后从固定的形式背景中抽取形式概念,所有概念间的包含关系构成一

收稿日期:2005-03-17
基金项目:"十五"国家部委预研基金资助项目(41101050108)
作者简介:黄健斌(1975-),男,讲师,西安电子科技大学博士研究生.

个完备格,这使得概念范化与特化关系十分清晰. 结合模糊逻辑 Pollandt 等[4] 提出了一种基于 L-Fuzzy 形式背景的模糊形式概念分析方法,它采用语言变量表达形式背景的不确定性,但从 L-Fuzzy 形式背景上生成模糊概念格的过程很容易导致组合爆炸,为了解决这个问题给出以下模糊形式概念分析方法.

1.1　模糊形式背景

定义 1　一个模糊形式背景是一个三元组 $K = < G, M, \tilde{I} >$,其中 G 是对象集,M 是属性集,$\tilde{I}$ 是 $G \times M$ 上的一个模糊集,$\mu_{\tilde{I}}(x): G \times M \to [0,1]$ 是模糊集 $\tilde{I}$ 的隶属函数[3,5].

定义 2　设有模糊形式背景 $K = < G, M, \tilde{I} >$ 和阈值 $\lambda \in [0,1]$. 给定对象集合 $A \subseteq G$,A 的内涵是 A 中所有对象都拥有且隶属度大于等于 λ 的属性所构成的集合,记为 A'.

$$A' = \{m \in M \mid \forall a \in A \to \mu_{\tilde{I}}(< a, m >) \geqslant \lambda\} \quad .$$

同样,给定属性集合 $B \subseteq M$,B 的外延是拥有 B 中所有属性的隶属度均大于等于 λ 的对象所构成的集合,记为 B'.

$$B' = \{g \in G \mid \forall b \in B \to \mu_{\tilde{I}}(< g, b >) \geqslant \lambda\} \quad .$$

定义 3　一个模糊形式概念是由对象集合 $A \subseteq G$ 上的一个模糊集和属性集 $B \subseteq M$ 组成的序偶 $< \tilde{A}, B >$,且满足 $A' = B$ 和 $B' = A$. 其中 $\tilde{A}$ 的隶属函数 $\mu_{\tilde{A}}(x) = \min\limits_{m \in B} \mu_{\tilde{I}}(< x, m >)$,$x \in A$.

定义 4　设有两个模糊形式概念 $C_1 = < \tilde{A}_1, B_1 >$ 和 $C_2 = < \tilde{A}_2, B_2 >$,称 C_1 是 C_2 的子概念,当且仅当 $\tilde{A}_1 \subseteq \tilde{A}_2$,记为 $C_1 \leqslant C_2$. 根据 Galois 联络的对偶性,同样有 $C_1 \leqslant C_2$ 当且仅当 $B_2 \subseteq B_1$.

1.2　模糊概念格

通过以上方式定义的模糊形式背景,可借助传统的形式概念分析工具,例如 Conexp,Galicia,ToskanaJ 等,在模糊形式背景上生成模糊概念格.

定义 5　设有模糊形式背景 $K = < G, M, \tilde{I} >$ 和阈值 $\lambda \in [0,1]$,K 的 λ 截集 $K_\lambda = < G, M, I_\lambda >$,其中 $I_\lambda = \{x \mid \mu_{\tilde{I}}(x) \geqslant \lambda, x \in I\}$. K_λ 是一个二值形式背景.

定义 6　设模糊形式背景 K 的 λ 截集 K_λ 上产生的形式概念构成的集合记为 C_{K_λ},则 $< C_{K_\lambda}, \leqslant >$ 是二值形式背景 K_λ 上的概念格. $< C_{K_\lambda}, \leqslant >$ 是一个完备格.

定义 7　与 $< C_{K_\lambda}, \leqslant >$ 同构的模糊形式背景 K 上的模糊概念格记为 $< C_K, \leqslant >$. 其中 $C_K = \{< \tilde{A}, B > \mid < A, B > \in C_{K_\lambda}, \mu_{\tilde{A}}(x) = \min\limits_{m \in B} \mu(< x, m >), x \in A\}$.

定义 8　设有模糊概念格 $< C_K, \leqslant >$. $F \subseteq C_K$,若 F 存在上确界 $\mathrm{lub}(F)$,则称 $\mathrm{lub}(F)$ 为 F 中的顶概念,记为 ∇_F,若 F 存在下确界 $\mathrm{glb}(F)$,则称 $\mathrm{glb}(F)$ 为 F 中的底概念,记为 $\triangle_F$. $< C_K, \leqslant >$ 中的上确界称为概念格的顶概念,记为 ∇. $< C_K, \leqslant >$ 中的下确界称为概念格的底概念,记为 $\triangle$. 模糊形式概念 $C_1 \in C_K$,C_1 的子概念集 $\mathrm{sub}(C_1) = \{C \mid C \leqslant C_1 \wedge C \neq C_1\}$. C_1 的超概念集 $\mathrm{sup}(C_1) = \{C \mid C_1 \leqslant C \wedge C \neq C_1\}$.

2　模糊概念格上的概念聚类

与传统形式概念分析一样,模糊概念格中产生的概念会随着形式背景中属性数量的增大而成指数级增长,同时很多对象会因为在属性值上的细微不同而被分离到不同的模糊形式概念中,而实际上这些对象应该属于同一类. 这样,太大的模糊概念格存在许多冗余的概念,这里采用以下概念聚类算法来约简概念的数量以及概念层次的规模.

2.1　概念聚类算法

在概念格上对格进行同态映射从而将一个规模很大的格缩放为一个小规模的格,可保持概念聚类间的格关系,但是同态映射函数的选取是一个困难的问题. 这里给出一个在格上通过模糊形式概念的距离度量进行概念聚合的快速算法.

定义 9　设有两个模糊形式概念 $C_1 = < \tilde{A}_1, B_1 >$ 和 $C_2 = < \tilde{A}_2, B_2 >$. 若 $C_1, C_2 \notin \{\nabla, \triangle\}$,则 $E(C_1, C_2) = |(\tilde{A}_1 \cap \tilde{A}_2)/(\tilde{A}_1 \cup \tilde{A}_2)|$,否则 $E(C_1, C_2) = 0$. 其中 $\cup$ 和 $\cap$ 是普通模糊集的并和交运算.

定义 10　设有模糊概念格 $< C_K, \leqslant >$ 和模糊概念相似度阈值 T_e,$F \subseteq C_K$,若 F 满足:

(a) F 存在顶概念 ∇_F 且 ∇_F 不存在任何超概念 $S \in \sup(\nabla_F)$ 使得 $E(\nabla_F, S) \geqslant T_e$.

(b) 若 $C \in F$ 且 $C \neq \nabla_F$,则 C 至少存在一个超概念 $S \in \sup(\nabla_F)$ 且 $S \in F$ 使得 $E(C, S) \geqslant T_e$.

则称 F 是模糊概念格 $< C_K, \leqslant >$ 上的一个概念聚类,并设概念聚类集 $F_K = \{F \mid F$ 是模糊概念格 $< C_K, \leqslant >$ 上的概念聚类$\}$.

定理 1 设有模糊概念格 $< C_K, \leqslant >$ 和模糊概念相似度阈值 T_e. F_K 是 $< C_K, \leqslant >$ 上的概念聚类集. 对于 $< C_K, \leqslant >$ 上的每一个模糊形式概念 $C \in C_K$,都存在 $F \in F_K$ 使得 $C \in F$. 分别由 $< C_K, \leqslant >$ 中的顶概念 ∇ 和底概念 $\triangle$ 为元素构成的集合 $\{\nabla\}$, $\{\triangle\} \in F_K$. (证明略)

定义 11 设 $F \in F_K$,F 的聚类中心是顶概念 ∇_F. $F[\nabla_F]$ 表示以 ∇_F 为中心的概念聚类.

下面给出在模糊概念格的拓扑序列上生成概念聚类的快速算法:

算法 基于相似度阈值的模糊概念格上的概念聚类生成算法.

输入 概念格 $< C_K, \leqslant >$,相似度阈值 T_e.

输出 概念聚类集 F_K.

处理过程如下:

```
for < C_K, ≤ > 中的每个模糊形式概念 C do F_K ← {C} endfor
for < C_K, ≤ > 中的每个极小元素 m do
    flag = flase;
    for m 的每一个超概念 s ∈ sup(m)
        if E(m,s) ≥ T_e then
            F[s] = F[s] ∪ F[m]
            flag = ture;
        endif
    endfor
    if flag then Delete F[m] from F_K endif
Delete m from C_K
endfor.
```

2.2 概念聚类层次的产生

通过模糊概念聚类产生了模糊概念格上的一组概念聚类,为了重新组建概念聚类层次,需要定义概念聚类间的层次关系[6].

定义 12 设 F_K 是模糊概念格 $< C_K, \leqslant >$ 上的一个概念聚类集,且有 $F_1, F_2 \in F_K$. 若 $\forall C(C \in F_1 \rightarrow C \leqslant \nabla_{F2})$ 称 F_1 是 F_2 的子概念聚类,记为 $F_1 \leqslant' F_2$.

定理 2 设 F_K 是模糊概念格 $< C_K, \leqslant >$ 上的一个概念聚类集,$F_1, F_2 \in F_K$. $F_1 \leqslant' F_2$ 当且仅当 $\nabla_{F1} \leqslant \nabla_{F2}$. (证明略)

定理 3 $< C_K, \leqslant >$ 上的一个概念聚类集 F_K 及其上的二元关系 $\leqslant'$ 构成的集合 $< F_K, \leqslant' >$ 是一个偏序集合,但不一定是格.

证明 F_K 上的二元关系 $\leqslant'$ 是自反,反对称和传递的,因此 $\leqslant'$ 是一个偏序.

任取 $F_1, F_2 \in F_K$,有 $\nabla_{F1}, \nabla_{F2} \in C_K$. 设 $\mathrm{glb}(\nabla_{F1}, \nabla_{F2}) = C_i, C_i \in C_K$.

若 C_i 是概念聚类 $F_m \in F_K$ 的聚类中心,则 $\mathrm{glb}(F_1, F_2) = F[C_i]$.

若 C_i 不是任何概念聚类 $F \in F_K$ 的聚类中心,则考虑 C_i 的子概念集 $\mathrm{sub}(C_i)$ 的极大元素. 如果 $\mathrm{sub}(C_i)$ 中没有聚类中心,则继续搜索 $\mathrm{sub}(C_i)$ 中元素的子概念集中的极大元素. 如果仅存在惟一 $C_j \in \mathrm{sub}(C_i)$ 且 C_j 是概念聚类 $F_n \in F_K$ 的聚类中心,那么 $\mathrm{glb}(F_1, F_2) = F[C_j]$. 如果存在 $C_j, C_k \in \mathrm{sub}(C_i)$,$C_j \neq C_k$ 且 C_j, C_k 分别是概念聚类 $F_u, F_v \in F_K$ 的聚类中心,那么 F_1, F_2 没有下确界.

同理,F_1, F_2 也不一定有上确界.

因此,$< F_K, \leqslant' >$ 是一个偏序,但不一定是格.

3　基本模糊概念格的 Web 搜索结果聚类

从 Web 这个异构的、分布的并且是动态的信息库中发现真正对用户有用的信息和知识是一个具有挑战性的课题[7]. 目前,用户在 Web 上搜索信息主要依赖于搜索引擎,而大多数的搜索引擎都是基于关键词匹配的,存在着对用户的查询意图不明确,返回结果过多等问题. 一种解决方法是对搜索引擎检索的结果自动聚类成若干个具有特定语义的组,以启发用户对搜索结果的访问[8].

将以上介绍的基于模糊概念格的概念聚类方法应用在 Web 文档搜索结果聚类上,具体的步骤如下:(1)搜索结果文档的预处理,包括语法标签的清除、过滤名词和形容词词组、删除停用词以及词干化等过程;(2)生成模糊形式背景. 采用改进的 C-value[9] 算法抽取检索结果中的巢状词条,分析每个词条在不同搜索结果上的相关度,从而在检索结果集合到多字词条集合上建立一个模糊二元关系;(3)模糊概念格的产生和聚类. 采用形式概念分析工具在模糊形式上下文上生成模糊概念格,随后在对模糊概念格上的模糊形式概念进行聚类分析;(4)搜索结果主题的标注;(5)聚合类间层次关系的产生. 整个处理过程如图 1 所示.

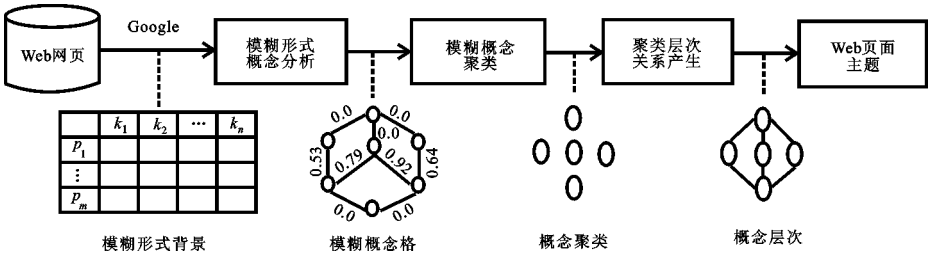

图 1　基于模糊概念格的 Web 文档聚类过程

将该算法与 COBWEB 概念聚类算法进行了比较,并且对概念格聚类与普通概念聚类的检索性能进行了对比分析. 实验中采用的 Web 文档是用 Google™ 提供的应用程序接口从 Internet 相关网站上搜索得到的. 整个实验在 PⅢ-1.7G(256MB 内存) PC 机的 Windows 2000 平台上进行.

首先,使用平均松弛误差(ARE)来测试算法产生概念的质量. 概念聚类集 F_K 的松弛误差定义为

$$A_{RE}(F_K) = \frac{1}{|F_K|}\sum_{F\in F_K}\sum_{a\in A}\sum_{i=1}^{n}\sum_{j=1}^{n} P(x_i)\ P(x_j)\ d^a(x_i,x_j) \quad , \tag{1}$$

其中 A 是聚类 F 中所有概念所拥有的属性构成的集合,$P(x_i)$ 是对象 x_i 属于 F 的概率,$d^a(x_i,x_j)$ 是对象 x_i 与对象 x_j 在属性 a 上的距离. 设 $m(x_i,a)$ 表示对象 $< x_i,a >$ 在模糊形式背景上的隶属度,$d^a(x_i,x_j)$ 定义如下:

$$d^a(x_i,x_j) = |m(x_i,a) - m(x_j,a)| \quad . \tag{2}$$

从图 2 中可看出,在属性数量 N 相同的情况下模糊概念聚类的产生概念的平均质量要优于 COBWEB.

其次,使用平均无内插精度(AUP)来评价概念层次的检索性能. 平均无内插精度定义为每个词条在其所出现的每个概念中精度的均值. 为了评价分类结果的 AUP,手工将搜索得到的结果按其主题进行了分类. 对于每一类,选取相应 5 个最频繁使用的词条. 随后,用这些词条作为输入形成检索查询,进而用 AUP 来评价检索结果的性能. 对于每一个文档,产生一组文档关键词. 这里有两种方法产生文档关键词:一是用概念格上每一个概念聚类所关联的一组属性关键词;二是用每个文档关联的一组关键词. 然后,将文档关键词和查询输入矢量化,并且通过计算矢量之间的欧氏距离来度量检索的性能. 第一种方法每个概念继承了其超概念的所有属性,计算其层次平均无内插精度记为 AUP(H). 第二种方法每个概念间无继承关系,计算其平均无内插精度记为 AUP(U).

图 3 给出了 AUP(H)和 AUP(U)在抽取不同属性个数 N 情况下的性能结果. 从图 3 中发现 N 越大 AUP(H)和AUP(U)的性能越好. 当 N 大于 5 时,AUP(H)和AUP(U)都取得了良好的性能,这说明抽取关键词的个数将会影响概念聚类的检索性能. 另外,AUP(H)的性能要优于AUP(U),这说明从概念格聚类产生的属性关键词更适合表示概念间的层次结构.

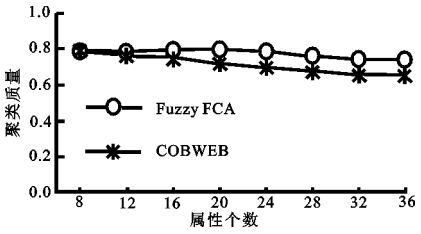

图 2 概念聚类质量测试

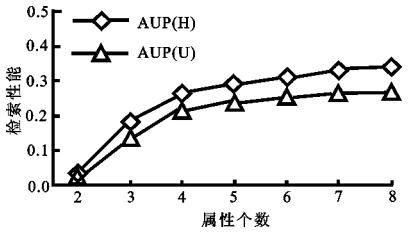

图 3 概念层次的检索性能评价

4 结 束 语

模糊概念格是一种将模糊逻辑与形式概念分析相结合的一种概念分析方法. 给出了在对象和属性的模糊二元关系上生成模糊概念格的过程. 提出了一种在格的拓扑序列上进行概念聚类的快速算法. 最后定义了概念聚类间基于偏序的层次关系. 将该方法应用在 Web 搜索聚类上,实验结果表明,该算法是有效的,能处理概念聚类间的多重继承关系. 这种基于概念格的概念聚类方法还可应用到许多需要自动分类的系统中.

参考文献:

[1] Hotho A, Staab S, Stumme G. Explaining of Text Clustering Result Using Semantic Structure[A]. Principles of Data Mining and Knowledge Discovery, 7th European Conference, PKDD 2003[C]. Dubrovnik: PKDD, 2003. 22-26.

[2] Fisher D H. Knowledge Acquisition Via Incremental Conceptual Clustering[J]. Machine Learning, 1987, 2(2): 139-172.

[3] Ganter B, Wille R. Formal Concept Analysis: Mathematical Foundations[M]. Heidelberg: Springer-Verlag, 1999.

[4] Polland S. Fuzzy-Concepts[A]. Formal Concept Analysis for Imprecise Data[C]. Berlin: Springer-Verlag, 1996.

[5] Zadeh L. Fuzzy Sets[J]. Information and Control, 1965, 69(3): 338-353.

[6] Reich Y, Fenves S J. The Formation and Use of Abstract Concepts in Design [A]. Concept Formation: Knowledge and Experience in Unsupervised Learning[C]. [s. l.]: Morgan Kaufmann, 1991. 323-353.

[7] Chen Li, Jiao Licheng. International Study on Internet/Web Data Mining with the State of Art and Advances[J]. Journal of Xidian University, 2001, 28(1): 114-119.

[8] Zamir O, Etzioni O. Grouper: a Dynamic Clustering Interface to Web Search Results[J]. Computer Networks, 1999, 31(1): 1 361-1 374.

[9] Frantzi K, Ananiadou S, Mima H. Automatic Recognition of Multiword Terms[J]. International Journal of Digital Libraries, 2003, 32(2): 117-132.

(编辑:齐淑娟)