ISOMAP 算法参数的递增式选取

邵 超
SHAO Chao

河南财经学院 信息学院,郑州 450002
School of Information,Henan University of Finance and Economics,Zhengzhou 450002,China
E-mail:sc_flying@163.com

SHAO Chao.Incremental parameter selection for ISOMAP algorithm.Computer Engineering and Applications,2008,44(21):119-122.

Abstract: The success of the ISOMAP algorithm depends greatly on the suitable selection of its only one parameter,i.e.the neighborhood size;however,it's an open problem how to do this efficiently.When the neighborhood size becomes unsuitable,short-cut edge can be introduced into the neighborhood graph and destroy the approximation ability of the involved shortest-path distances to the corresponding geodesic distances greatly.Unlike non-shortcut edge,shortcut edge links two endpoints lying close in Euclidean space but far away on the manifold,which can be measured approximately by its order presented in this paper.Based on the observation,this paper presents an efficient method to find a suitable neighborhood size,which only requires running the breadth-first search algorithm incrementally,but doesn't need to run the whole ISOMAP algorithm for every possible neighborhood size as those methods based on residual variance do.Finally,the feasibility of this method can be verified by experimental results.
Key words: data visualization;ISOMAP;neighborhood size;residual variance;shortcut edge;order;breadth-first search

摘 要:ISOMAP 算法能否被成功应用依赖于其唯一参数——邻域大小的选取是否合适,然而,如何高效地选取一个合适的邻域大小目前还是一个难题。当邻域大小变得不合适时,短路边将会出现在邻域图中,从而严重破坏与之相关的最短路径距离对测地距离的逼近能力。和非短路边不同,短路边的两个端点虽然在欧氏空间中相距较近,但在流形上却相距甚远。基于短路边的这一特点,采用序来近似度量一条边的两个端点在流形上的远近程度,因而能够递增式地对邻域大小进行合适的选取。和基于残差的参数选取方法不同,该方法只需递增式地运行广度优先搜索算法,而无需就每一个可能的邻域大小分别运行整个 ISOMAP 算法,从而具有比较高的运行效率。最终的实验结果证实了该方法的可行性。
关键词:数据可视化;ISOMAP;邻域大小;残差;短路边;序;广度优先搜索
DOI:10.3778/j.issn.1002-8331.2008.21.033 **文章编号:**1002-8331(2008)21-0119-04 **文献标识码:**A **中图分类号:**TP181

1 引言

在当今时代,数据量及其维数的急剧膨胀使得数据可视化显得越来越重要。在过去数年内,人们提出了大量的数据可视化技术,大致可归为以下五类:(1)将整个可视窗口划分为多个子窗口,分别用来表示数据维的不同组合,如散列图矩阵[1](scatterplot matrices)和面向像素技术[2](pixel-oriented techniques);(2)在低维可视空间中对所有数据维进行重排,如平行坐标系[3](parallel coordinates)和星型坐标系[4](star coordinates);(3)按照所有数据维对低维可视空间进行层次划分,如dimensional stacking[5]和 Treemap[6];(4)采用某些具有多个可视特征的图标,每一个可视特征都可用来表示数据的一维,如 Chernoff-faces[7]和 stick figures[8];(5)利用降维算法将数据映射在一个低维可视空间中,如主成分分析[9](Principal Component Analysis,PCA)、多维尺度变换[9](Multidimensional Scaling,MDS)、自组织映射[10](Self-Organizing Map,SOM)、等距映射[11](Isometric Mapping,ISOMAP)、局部线性嵌入[12](Locally Linear Embedding,LLE)、拉普拉斯特征映射[13](Laplacian Eigenmap)、海森特征映射[14](Hessian Eigenmap)等。不同于其它数据可视化技术,降维算法不仅能尽可能真实地展现数据的内在结构,而且还能对数据进行预处理,从而得到了十分广泛的应用。

作为一种有效的非线性降维算法,ISOMAP 算法[11]采用能有效表征数据全局几何结构的测地距离对古典 MDS 算法进行了非线性扩展,从而能较好地对嵌入在高维欧氏空间中的低维非线性流形如 swiss roll 数据集[11]等进行可视化。

ISOMAP 算法有很多优点,其中之一就是它只有一个参数——邻域大小。ISOMAP 算法能否被成功应用严重依赖于该参数的选取是否合适[15],此外,该算法所谓的拓扑不稳定性[15]也与此相关。然而,目前大多数的参数选取方法[11,16]都基于残差(residual variance),需要就每一个可能的邻域大小分别运行整个 ISOMAP 算法,因而都非常耗时。众所周知,不合适的邻域大

基金项目:河南省基础与前沿技术研究项目(No.082300410110)。
作者简介:邵超(1977-),男,工学博士,副教授,主要研究方向包括:机器学习、数据可视化等。
收稿日期:2008-04-30 **修回日期:**2008-05-29

小将在邻域图中引入短路边,从而严重破坏与之相关的最短路径距离对测地距离的逼近能力。本文根据短路边的特点提出了一种递增式的参数选取方法。该方法无需多次运行整个 I-SOMAP 算法,从而具有比较高的运行效率。

2 ISOMAP 算法和残差

在数据的全局几何结构未知(通常呈非线性)的情况下,欧氏距离只在很小的邻域内才有意义[17]。因此,需要用这些已知的局部欧氏距离来逼近能有效表征数据全局几何结构的测地距离(其合理性证明请参见文献[18]),ISOMAP 算法便是这么做的[11]:

(1)对于大数据集,为降低计算量,用矢量量化方法[19]从中选取 n 个代表点以执行下面的操作(可选);

(2)用 k-近邻法创建能正确表达数据邻域结构的邻域图,这需要一个合适的邻域大小 k;

(3)在该邻域图上运行最短路径算法得到所有数据点间的最短路径距离,以此作为测地距离的近似;

(4)以这些最短路径距离作为输入运行古典 MDS 算法,将数据映射在一个低维可视空间中。

当数据具有良好抽样且位于内在扁平的单一流形之上时(如 swiss roll 数据集),邻域图中的最短路径距离能否对测地距离进行良好逼近,进而 ISOMAP 算法能否被成功应用就完全依赖于其唯一参数——邻域大小的合适与否了。目前大多数的参数选取方法都是通过估计最终映射"质量"的好坏反过来对邻域大小进行选取的,而衡量映射"质量"好坏的标准通常为残差[11]:$1-\rho^2_{\hat{D}_X(k)D_Y}$,其中,$\hat{D}_X(k)$和 D_Y 分别为数据 X 在原数据空间中的测地距离矩阵(由最短路径距离进行逼近,在给定数据 X 的情况下为邻域大小 k 的函数)和在低维可视空间中的欧氏距离矩阵(数据 X 在低维可视空间中的映射为 Y),而 $\rho_{\hat{D}_X(k)D_Y}$ 则为它们之间的线性相关系数。残差越小,映射的"质量"就越好,邻域大小也就越合适。因此,最优邻域大小可定义如下[16]:

$$k_{opt}=\arg\min_k\left(1-\rho^2_{\hat{D}_X(k)D_Y}\right) \tag{1}$$

由于残差用到了 ISOMAP 算法的运行结果 Y,因此,为计算残差需要运行整个 ISOMAP 算法。此外,如上所述,残差只能用来比较两个邻域大小的相对合适程度,而不能用来判断某一个邻域大小的合适与否,因此,基于残差的参数选取方法需要就每一个可能的邻域大小分别运行整个 ISOMAP 算法并求出相应的残差,这是极其耗时的,从而也极大地限制了 ISOMAP 算法的进一步应用。

3 递增式的参数选取方法

众所周知,一个合适的邻域大小不应太大,以避免邻域图中出现短路边,同时也不应太小,以防止邻域图过于稀疏[15]。这就意味着邻域大小的选取仅仅依赖于数据本身的分布,而与特定算法如 ISOMAP 算法无关。因此,邻域大小的选取可以不必象残差那样需要运行整个 ISOMAP 算法。简言之,一个合适的邻域大小应能使测地距离得到良好逼近,要做到这一点,邻域图应能正确表达数据的邻域结构,也就是说,邻域图应足够稠密,但不能包含任何短路边。

一旦邻域大小变得不合适,短路边就会出现在邻域图中,使得邻域图不能正确表达数据的邻域结构,从而严重破坏与之相关的最短路径距离对测地距离的逼近能力。因此,如果能够监测到短路边何时将第一次出现在邻域图中,那么在此之前最大的邻域大小就是一个合适的邻域大小,因为此时的邻域图在不包含任何短路边的前提下最为稠密。

和非短路边不同,短路边的两个端点虽然在欧氏空间中相距较近,但在流形上却相距甚远。不幸的是,数据的内在流形结构事先是未知的,这也正是 ISOMAP 算法要去发现的。然而,正象 ISOMAP 算法所做的那样,数据的内在流形结构可以用一个不包含任何短路边的邻域图来近似表达。因此,可以用序(order)来近似度量一条边的两个端点在流形上的远近程度,定义如下:

定义 1 数据点 i 的第 $k+1$ 条边(连接到该点的第 $k+1$ 个近邻数据点)的序(记作 O_{ik})定义为在邻域大小为 k 的邻域图(k-阶邻域图)上用以连接该边两个端点所需的最少边数(如图 1 中的数字所示)。

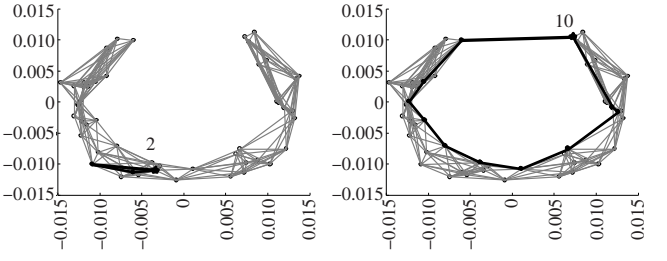
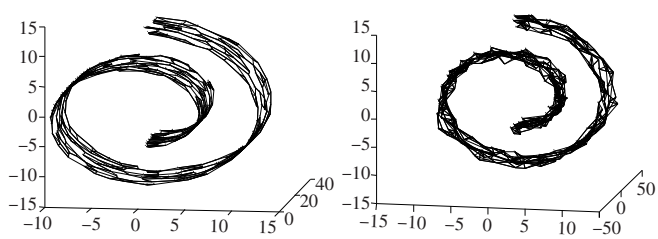


图 1 某个二维数据集上不同数据点(用五星表示)的序 $O_{ik}(k=9)$

如果 k-阶邻域图不包含任何短路边,也就是说,该邻域图能近似表达数据的内在流形结构,那么就可以根据序 O_{ik} 的大小来判断数据点 i 的第 $k+1$ 条边是否为短路边:如果序 O_{ik} 依然保持在很小的水平(和 $O_{i(k-1)}$ 相比),那么该边将依旧是一条正常的非短路边;而如果序 O_{ik} 突然变得很大(和 $O_{i(k-1)}$ 相比),那么该边就是一条短路边。

除序之外,最短路径距离同样可用来近似度量一条边的两个端点在流形上的远近程度,但其时间复杂度明显要大得多,因为序可以通过运行广度优先搜索算法来获得,其时间复杂度仅为线性。

一般来说,当邻域大小为能使邻域图连通的最小 k 值(记作 $k_{\min}$)时,相应的邻域图(也就是 $k_{\min}$-阶连通图)将不会包含任何短路边,因而也就能近似表达数据的内在流形结构(事实上,这也是具有单一邻域大小 k 的 ISOMAP 算法得以成功应用的一个前提),如图 2 所示。为获得 $k_{\min}$ 的值,需要从 $k=1$ 开始递增式地判断相应的邻域图是否连通,该操作耗时相对较少,因为 $k_{\min}$ 通常都比较小(见下一章),另外,判断一个图是否连通只需遍历一次即可,其时间复杂度也仅为线性。



(a)swiss roll 数据集 (b)加入了噪音的 swiss roll 数据集

图 2 不同数据集(见下一章)的 $k_{\min}$-阶连通图

因此,可以从 $k=k_{\min}$ 开始递增式地计算所有数据点的序 O_{ik}

(i=1,2,…,n)。在 k 成为最大的合适邻域大小之前,也就是说,在(k+1)-阶邻域图依旧没有包含任何短路边之前,所有序都将保持在很小的水平。另外,随着 k 的增加,数据的无序性会越来越小,k-阶邻域图会越来越稠密,越来越逼近于数据的内在流形结构;显然,作为正常的非短路边,所有数据点的第 k+1 条边也都将越来越贴近于此前的 k-阶邻域图,其结果是最大序值 $\max_{1\leqslant i\leqslant n}(O_{ik})$ 将会单调减小,即有 $\max_{1\leqslant i\leqslant n}(O_{ik_{\min}})\geqslant\max_{1\leqslant i\leqslant n}(O_{i(k_{\min}+1)})\geqslant\cdots\geqslant\max_{1\leqslant i\leqslant n}(O_{ik})$。而当最大序值 $\max_{1\leqslant i\leqslant n}(O_{ik})$ 第一次陡然增加时,即有 $\max_{1\leqslant i\leqslant n}(O_{ik_{\min}})\geqslant\max_{1\leqslant i\leqslant n}(O_{i(k_{\min}+1)})\geqslant\cdots\geqslant\max_{1\leqslant i\leqslant n}(O_{i(k-1)})<\max_{1\leqslant i\leqslant n}(O_{ik})$,这意味着存在某个数据点,它的第 k+1 条边将作为第一条短路边出现在(k+1)-阶邻域图中,此时的 k 就是所要选取的合适的邻域大小,更确切地说应该是最大的合适邻域大小。

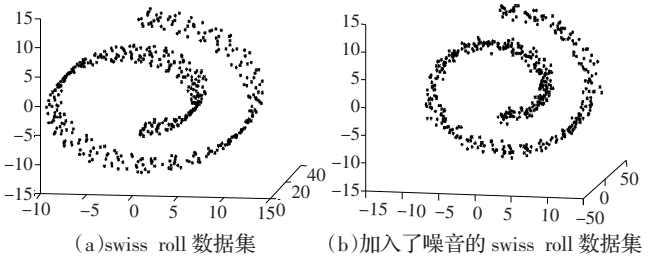
该参数选取方法的时间复杂度相对较小,因为该方法无需运行比较耗时的最短路径算法和古典 MDS 算法(ISOMAP 算法的第 3 步和第 4 步)即可选取一个合适的邻域大小,它只需递增式地计算所有数据点的序,而计算所有数据点的序只需运行 n 次广度优先搜索算法即可,其时间复杂度仅为 $O(n^2)$。

4　实验结果

在这一章,将使用以下两个数据集来对该参数选取方法进行验证:

(1)具有 2 000 个随机样本点的 swiss roll 数据集[11]。为降低计算量,在 ISOMAP 算法的第 1 步,采用 Matlab v6.5 工具箱中的 k-均值算法从中选取了 n=500 个代表点,如图 3(a)所示。

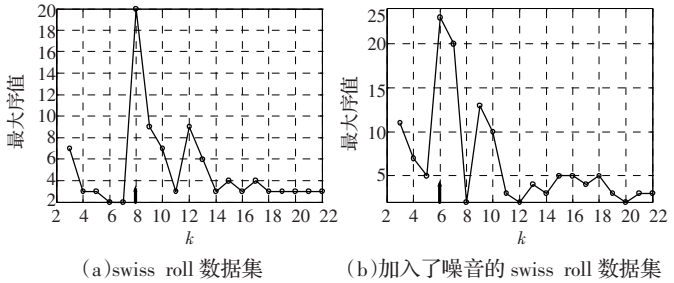
(2)加入了噪音的 swiss roll 数据集[15]。该数据集是在如上所述的 swiss roll 数据集的每一个数据点上都加入了一服从正态分布的噪音,该正态分布的期望为 0,标准差为各维跨度最小值的 2%。为降低计算量,采用同样的方法从中选取了 n=500 个代表点,如图 3(b)所示。



(a)swiss roll 数据集　(b)加入了噪音的 swiss roll 数据集

图 3　实验中用到的两个数据集

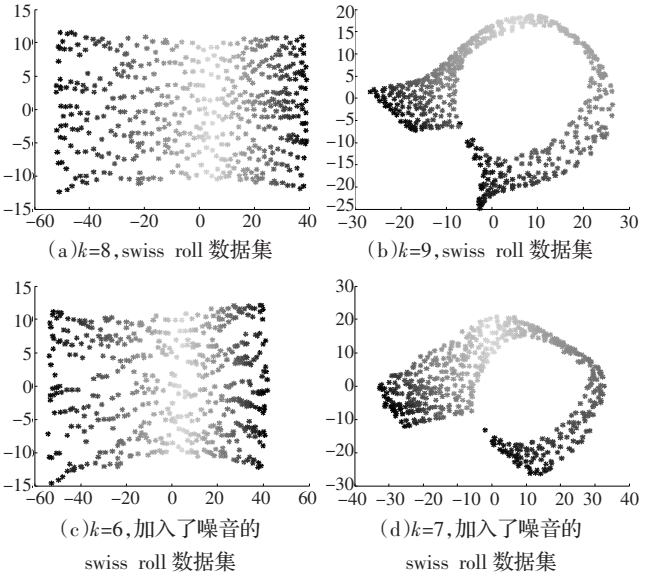
为验证该参数选取方法的可行性,就每一个可能的邻域大小 $k\in[k_{\min},k_{\max}]$($k_{\max}$ 是预先设定的一个最大可能 k 值,将其设定为 $k_{\max}=k_{\min}+20$,事实上,该参数在递增式的参数选取方法中是不必要的)分别计算了所有数据点的序,其最大值随邻域大小的变化情况如图 4 所示。



(a)swiss roll 数据集　(b)加入了噪音的 swiss roll 数据集

图 4　不同数据集的最大序值随邻域大小的变化情况

从图 4 可以看出,在最大序值第一次陡然增加之前,最大序值随邻域大小的增加都呈单调减小的趋势,这意味着递增式地对邻域大小进行选取是可行的;而第一次陡然增加的那个最大序值对应的就是我们所要选取的邻域大小,在 swiss roll 数据集中为 k=8,在加入了噪音的 swiss roll 数据集中为 k=6,I-SOMAP 算法的运行结果分别如图 5(a)和图 5(c)所示。为了进一步说明选取的邻域大小是最大的合适邻域大小,将相应的邻域大小分别加 1,得到的运行结果分别如图 5(b)和图 5(d)所示。

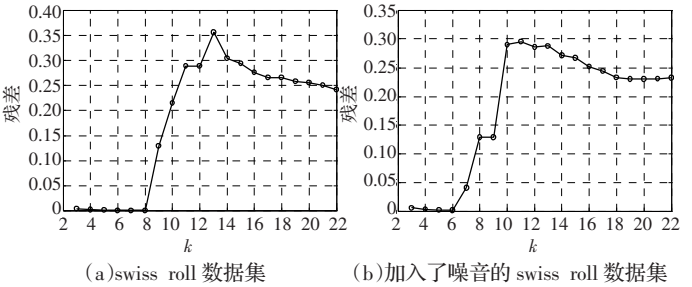


(a)k=8,swiss roll 数据集　(b)k=9,swiss roll 数据集

(c)k=6,加入了噪音的 swiss roll 数据集　(d)k=7,加入了噪音的 swiss roll 数据集

图 5　ISOMAP 算法的运行结果

此外,通过计算相应的残差(如图 6 所示),同样可以证明:用该参数选取方法选取出来的这些邻域大小都是合适的,而且还都是最大的合适邻域大小。



(a)swiss roll 数据集　(b)加入了噪音的 swiss roll 数据集

图 6　不同数据集上的残差随邻域大小的变化情况

从图 4~图 6 还可以看出,噪音对邻域大小的合适与否影响很大,这也是造成 ISOMAP 算法拓扑不稳定的重要原因。本文提出的参数选取方法在有噪音的情况下依然能够选取合适的邻域大小(如图 4(b)所示),这说明该参数选取方法具有一定的鲁棒性。

5　结论

众所周知,ISOMAP 算法能否被成功应用依赖于其唯一参数——邻域大小的选取是否合适,然而,如何高效地选取一个合适的邻域大小目前还是一个难题。目前大多数的参数选取方法都基于残差,因而都非常耗时。

和非短路边相比,短路边的两个端点在流形上相距明显要远得多,本文用序对此进行近似度量,从而能够递增式地对邻

域大小进行合适的选取。和基于残差的参数选取方法不同,该方法只需递增式地运行广度优先搜索算法,而无需就每一个可能的邻域大小分别运行整个 ISOMAP 算法,从而具有比较高的运行效率。在实验中也发现,该参数选取方法具有一定的鲁棒性。

参考文献:

[1] Cleveland W S.Visualizing Data[M].Summit,NJ,USA:Hobart Press,1993.

[2] Keim D A.Designing pixel-oriented visualization techniques:Theory and applications[J].IEEE Transactions on Visualization and Computer Graphics,2000,6(1):59-78.

[3] Inselberg A,Dimsdale B.Parallel coordinates:a tool for visualizing multi-dimensional geometry[C]//Proceedings of IEEE Conference on Visualization'90,San Francisco,CA,USA,1990:361-378.

[4] Kandogan E.Visualizing multi-dimensional clusters,trends,and outliers using star coordinates[C]//Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,San Francisco,CA,USA,August 2001:107-116.

[5] LeBlanc J,Ward M O,Wittels N.Exploring n-dimensional databases[C]//Proceedings of IEEE Conference on Visualization'90,San Francisco,CA,USA,October 1990:230-237.

[6] Shneiderman B.Tree visualization with treemaps:a 2d space-filling approach[J].ACM Transactions on Graphics,1992,11(1):92-99.

[7] Chernoff H.The use of faces to represent points in k-dimensional space graphically[J].Journal of the American Statistical Association,1973,68(342):36-368.

[8] Pickett R M,Grinstein G G.Iconographic displays for visualizing multidimensional data[C]//Proceedings of the 1988 IEEE Conference on Systems,Man and Cybernetics,Beijing and Shenyang,China,August 1998:514-519.

[9] Duda R O,Hart P E,Stork D G.Pattern classification[M].2nd ed.New York,NY,USA:John Wiley & Sons Inc.,2000.

[10] Kohonen T.Self-organized formation of topologically correct feature maps[J].Biological Cybernetics,1982,43(1):59-69.

[11] Tenenbaum J B,de Silva V,Langford J C.A global geometric framework for nonlinear dimensionality reduction[J].Science,2000,290(5500):2319-2323.

[12] Roweis S T,Saul L K.Nonlinear dimensionality reduction by locally linear embedding[J].Science,2000,290(5500):2323-2326.

[13] Belkin M,Niyogi P.Laplacian eigenmaps for dimensionality reduction and data representation[J].Neural Computation,2003,15(6):1373-1396.

[14] Donoho D L,Grimes C.Hessian eigenmaps:locally linear embedding techniques for high dimensional data[C]//Proceedings of the National Academy of Sciences,2003,100(10):5591-5596.

[15] Balasubramanian M,Shwartz E L,Tenenbaum J B,et al.The isomap algorithm and topological stability[J].Science,2002,295(5552):7a-7.

[16] Kouropteva O,Okun O,Pietikäinen M.Selection of the optimal parameter value for locally linear embedding algorithm[C]//Proceedings of the 1st International Conference on Fuzzy Systems and Knowledge Discovery,Singapore,2002:359-363.

[17] Saxena A,Gupta A,Mukerjee A.Non-linear dimensionality reduction by locally linear isomaps[C]//Proceedings of the 11th International Conference on Neural Information Processing,Calcutta,India,2004:1038-1043.

[18] Bernstein M,de Silva V,Langford J C,et al.Graph approximations to geodesics on embedded manifolds[R].Technical Report,Department of Psychology,Stanford University,2000.

[19] Lee J A,Lendasse A,Verleysen M.Nonlinear projection with curvilinear distances:isomap versus curvilinear distance analysis[J].Neurocomputing,2004,57(1):49-76.

(上接 118 页)

$L(K_1)\times L(K_2)$的映射 $\varphi:\varphi([A_1\times A_2,F(B_1\times B_2)])=([A_1,B_1],[A_2,B_2])$是一个 $\wedge$-维持的序嵌入。

证明 由命题 10 以及定义 4 和定义 5 容易证明。

4 总结和进一步的问题

本文主要是把概念格理论引入文本知识挖掘领域,用于描述较为规则的文本知识的数据挖掘。为更简洁地从文本知识中抽取概念格,适当修改了 Galois 联通,从而避免了在概念换算中概念格结构因使用属性标尺不同可能差别很大的问题。由于创建了一些概念本体,所以基本上可以自动地把描述规则的文本知识翻译为多值上下文。例如基于人物概念本体(战役本体)就可以自动地从描述人物(著名战役)的文本知识中抽取出多值上下文。本文假设已经把文本知识转化为多值上下文,然后分析了两类情况:一是当在一个多值上下文中增加/删除一些对象或属性时,即当多值上下文动态变化时概念的变化情况;二是多值上下文乘积的概念与因子上下文的概念间的关联。

本文得到一些重要命题,这些命题可用来判定执行一些操作后哪些概念保持不变以及概念格之间的结构关联。本文主要从理论上分析了多值上下文的一些操作对概念以及概念格结构的影响。实践证明,概念格理论更适合处理较规则的文本知识。依据一定的原则,增加一些对象或一些属性并不会改变原有的概念。反过来,依据一定的策略,也可由多值上下文中的概念确定子多值上下文的概念。另外,多值上下文乘积的概念格可嵌入到两个因子概念格的乘积中。目前本文分析的只是描述较为规则的文本知识的形式处理方法,今后将深入研究一般性文本知识的概念格挖掘以及分析。

参考文献:

[1] Wille R.Formal concept analysis as mathematical theory of concepts and concept hierarchies[C]//Ganter B.LNAI 3626:Formal Concept Analysis,2005:1-33.

[2] Lei Yuxia,Wang Yan,Cao Baoxiang,et al.Concept interconnection based on many-valued context analysis[C]//Zhou Z H,Li H,Yang Q.LNAI 4426:PAKDD 2007,Nanjing China,2007:623-630.

[3] Kim M,Compton P.Formal concept analysis for domain-specific document retrieval systems[C]//Brooks M,Corbett D,Stumptner M.LNAI 2256:2001:237-248.

[4] Lei Y,Cao C,Sui Y.Acquiring military knowledge from texts in the electronic encyclopedia of China[C]//ICYCS'2001,Hangzhou,China,2001,1:367-371.

[5] Hahn U,Schnattinger K.Knowledge mining from textual sources[C]//Proceedings of the Sixth International Conference on Information and Knowledge Management,1997:83-90.

[6] Mooney R J,Bunescu R.Mining knowledge from text using information extraction[J].SIGKDD Explorations,2005,7(1):3-10.

[7] Pechsiri C,Kawtrakul A.Mining causality for explanation knowledge from text[J].Journal of Computer Science and Technology,2007,22(6):877-889.

[8] Ganter B,Wille R.Formal concept analysis:mathematical foundations[M].[S.l.]:Springer,1999.