WEIGHTED PAIR GROUP METHOD ARITHMETIC TECHNIQUE FOR DOCUMENT CLUSTERING

[1]G.Vijayalakshmi, [2]Dr.M.karthikeyan
[1]M.phil Research Scholar, [2]Assistant Professor
[1]Department of Computer Science and Information Science
[1]Annamalai University, Tamilnadu, India

ABSTRACT: Document clustering (or) text clusters in the application of cluster analysis to textual documents. It has application in automatic document organization fast information and retrieval or filtering. One of the most commonly used variants. It defines as the average dissimilarity between dissimilarity clusters (its name). The WPGMA produces in a simple averaging weighted result and proportional averaging document clustering produces un-weighted result. Numerical tests data are used carried out on synthetic for comparing different new algorithm generated approximation the certain approximations obtained by classical method. In this paper we proposed method called concept factorization adaptive neighbors, (concept factorization with adaptive neighbors). The idea of concept factorization with adaptive neighbors I to integrate an adaptive neighbors regularization constraint into the concept factorization decomposition. The goal of concept factorization adaptive neighbors is to extract the representation space that maintains geometrical neighborhood structure of the data. It is similar to the existing graph-regularized concept factorization, concept factorization with adaptive neighbors builds a neighbors graph weights matrix. The key difference is that the concept factorization adaptive neighbors performs dimensionality reduction and finds the neighbor graph weights matrix simultaneously. An efficient algorithm derived to solve the proposed problem. We apply the proposed method to solve the problem of document clustering on the 20 newsgroups, reuters-21578, and TDT2 document data sets. Our experiments demonstrate the effectiveness of the method.

Index Terms- Document Clustering, Weighted Matrix, Concept Factorization, WPGMA

I.INTRODUCATION

Concept factorization and non negative matrix factorization more important methods of dimensionality reduction. The data representation with clustering and classification. The two non negative matrix U and H the non negative matrix factorization decomposes a data matrix X, that X~~UHT over the past decade. Non negative matrix factorization many algorithms extended have been presented. The three-factor of non negative matrix factorization model was presented by ding et al. the non negative matrix factorization is equivalent to spectral clustering. The encode geometrical information the graph regularized, non negative matrix factorization (GNMF) approach was proposed. The proposed a low-rank matrix factorization on integrating regularization in to the matrix factorization non negative local coordinates factorization with proposed by chan for feature extraction. Non negative matrix factorization with presented on zniet al Graph-preserving sparse with facial expression recognition

1.1 SCOPE

The major advantage of performed on concept factorization over non negative factorization is the concept factorization with on either the original feature space or the reproducing kernel Hilbert space. The application of concept factorization has been drawn in data clustering. To determine the number of clusters is a problem in clustering algorithms the number of methods have been proposed in determining to number of clusters. Many methods of popular in clusters such as the k-means algorithm require on the number of cluster in exact preassignment the method chosen in the concept factorization and non negative clustering methods in number of cluster to pressing for presented was concept factorization in document clustering. The proposed a local coordinate concept factorization by incorporating into constraint the

traditional concepts to be as close the original samples as possible. The most GNMF or concept factorization algorithms consist of two steps: graph weights matrix from the constructs a neighborhood to original data sample, second step: the factorization a data matrix into two or three non negative matrices. The basic methods for non negative matrix factorization and concept factorization:

Let X =[X1, X2……, X n] € R m ×n be a given nonnegative data matrix, where n is the total number of samples, m is the feature dimension. The goal of NMF is to decompose data matrix X into two nonnegative matrices, the encoding matrix H and the basis matrix U, X~~ UH t the NMF.

1.2 The Java Programming Language

The Java programming language is a high-level language that can be characterized by all of the following buzzwords:

- Simple
- Architecture neutral
- Object oriented
- Portable
- Distributed
- High performance
- Interpreted
- Multithreaded
- Robust
- Dynamic
- Secure

With most programming languages, you either compile or interpret a program so that you can run it on your computer. The Java programming language is unusual in that a program is both compiled and interpreted. With the compiler, first you translate a program into an intermediate language called *Java byte codes* —the platform-independent codes interpreted by the interpreter on the Java platform. The interpreter parses and runs each Java byte code instruction on the computer. Compilation happens just once; interpretation occurs each time the program is executed. The following figure illustrates how this works.

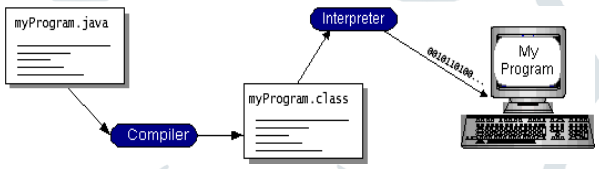

Figure 1: Java virtual machine

You can think of Java byte codes as the machine code instructions for the *Java Virtual Machine* (Java VM). Every Java interpreter, whether it's a development tool or a Web browser that can run applets, is an implementation of the Java VM. Java byte codes help make "write once, run anywhere" possible. You can compile your program into byte codes on any platform that has a Java compiler. The byte codes can then be run on any implementation of the Java VM. That means that as long as a computer has a Java VM, the same program written in the Java programming language can run on Windows 2000, a Solaris workstation, or on an iMac.

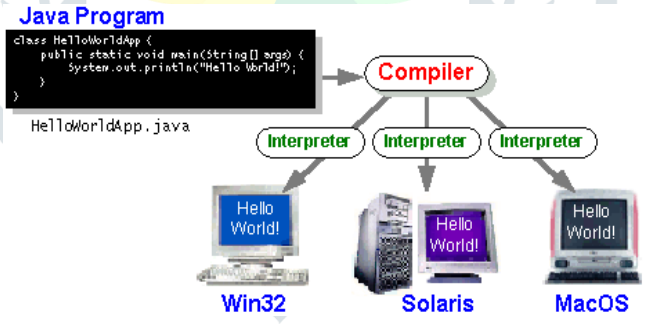

Figure 2: Solaris Workstation

II. LITERATURE SURVEY

Co-clustering different data has attracted with extensive attention recently due to its high impact on various important applications used to text mining, image retrieval, and bioinformatics. The data co-clustering is background information any prior knowledge of still a challenging problem. The propose in a Semi supervised non-negative matrix factorization framework for the data co-clustering. the computes our method a new relational matrices by incorporating client provided constraints through simultaneous distance metric algorithm, then perform trifactorizations of the new matrices into infer the clusters of different data types and their correspondence. We prove the convergence and correctness of semi supervised non -negative matrix factorization co-clustering the relationship between semi supervised non- negative matrix factorization with other well-know co-clustering models. Our extensive experiments on publicly available text, gene expression, and image data sets, the different data co-clustering we demonstrate the superior performance of semi supervised non negative matrix factorization.

The non negative matrix factorization has been successfully applied in to document clustering, image representation, or other domains, non negative matrix factorization have received considerable into interest from the data information retrieval fields. This proposes an online non- negative matrix factorization algorithm for efficiently handle very large-scale and/or streaming data sets. Non negative matrix factorization solutions with require the entire data matrix to reside in the memory, online non negative matrix factorization with algorithm proceeds in one data into or one chunk of the data points at a time. Our experiments with other one pass and multi pass online non negative matrix factorization on real datasets in presented.

The classification methods which employ non negative matrix factorization employ of two consecutive independent steps for the first one performs data transformation (dimensionality reduction) and second one classifies the transformed data using for classification methods, and nearest neighbor/centroid or support vector machines. The focus of using non negative matrix factorization followed by support vector machines classification. The parameters of two steps, e.g., non negative matrix factorization base coefficients and the support vectors, in optimized independently. The leading too suboptimal classification performance. We merge two steps into one by incorporating maximum margin classification of standard into constraints non negative matrix factorization optimization the support vector optimization and non negative matrix factorization are performed a set of multiplicative update rules.

The maximum margin and same context, classification constraints are imposed on the non negative matrix factorization problem with additional of discriminate and respective multiplicative update rules are extracted. Experimental results in several databases indicate the incorporation of the maximum margin classification into the constraints of non negative matrix factorization and discriminate of non negative matrix factorization objective functions improves the accuracy of the classification. This paper proposes a new model in low-rank matrix that incorporates manifold regularization of matrix factorization. The new regularization model has globally optimal and closed form solutions. Superior to the graph-regularized non negative matrix factorization, the direct algorithm (data with small number of points) and alternate iterative algorithm with inexact inner iteration for large scale data in proposed to solve the new model. A establishes of convergence analysis the global convergence in the iterative algorithm. The efficiency and precision of the algorithm are demonstrated numerically through applications of six-real-world datasets on clustering and classification. Comparison with existing of algorithms shows the effectiveness of the proposed method for low-rank factorization in general.

III. PROPOSED SYSTEM

A new algorithm is proposed for generating in given similarity matrix for min-transitive approximations (symmetric matrix with elements in the unit interval and diagonal elements equal to one) an aggregation operator in plays a central role in the algorithm for different approximations are generated depending on the choice. The maximum operator is chosen, for the approximation coincides another for the min-transitive closer of the given similarity matrix. The arithmetic mean, a transitive approximation is generated by, on the average, as close to the given similarity matrix as the approximation generated by the concept factorization hierarchical clustering algorithm. The new algorithm for allows in generated approximations in a purely ordinal setting. A new approach is weight-driven, min-transitive for partition free associated of corresponding approximation can be built layer by layer.

We proposed scheme can be support to complicated logic search the mixed “AND”,”OR”,”NOT” operations of keywords. The develop a practical and very efficient multi-keyword search scheme. To matching search scheme coordinate (MRSE) which can be regarded to a searchable encryption scheme with “OR” operation. The returned documents matching all keywords “AND” operation which can be regarded as a searchable encryption scheme. The data access by multi users we are going to make at same time. We are using algorithm for producing key and encryption and decryption called identity based algorithm.

We introduce the relevance scores and preference factors of keywords for searchable encryption. The relevance scores of keywords an enable more precise returned result, and preference factors of keywords to represent the importance of keywords in the search keyword set by specified search users and enables to correspondingly personalized search to cater to specific users preferences. The operations of “AND,”OR”,”NOT” in the multi-keyword search for searchable encryption. The proposed scheme to more achieve in comprehensive functionality and lower query complexity. The classified sub-dictionaries technique to enhance the efficiency of the above two schemes. That the enhanced schemes can achieve better efficiency in extensive experiments demonstrate in terms of index building, trapdoor generating and query in the comparison with schemes.

IV. IMPLEMENTATION

Implementation is the most crucial stage in successful achieving a system and giving the users confidence that the new system in workable and effective, this type in conversation easy to handle to relatively, no major changes in the system. The modified application of implementation to replace an existing one .Each program is tested individually at the time of development the data has verified this program in the way specified linked together in the program specification, user satisfaction of tested to environment of computer system. The system is going to be implanted very soon. A simple operating procedure is include to user can understand the different functions clearly and quickly.

Implementation is the stage of the project when the theoretical design to turned out into a working system. The most critical stage in achieving to considered to a successful new system and it's giving the user, confidence that the new system will be work and be effective. The implementation stage involves careful planning, investigation of the existing system and constraints on implementation, designing of method to achieve changeover methods.

4.1 Software Risk Planning

1. Defining preventive measure that would lower down the likelihood or probability of various risks.
2. Define measures that would reduce the impact in case a risk happens.
3. Constant monitoring of processes to identify risks as early as possible.

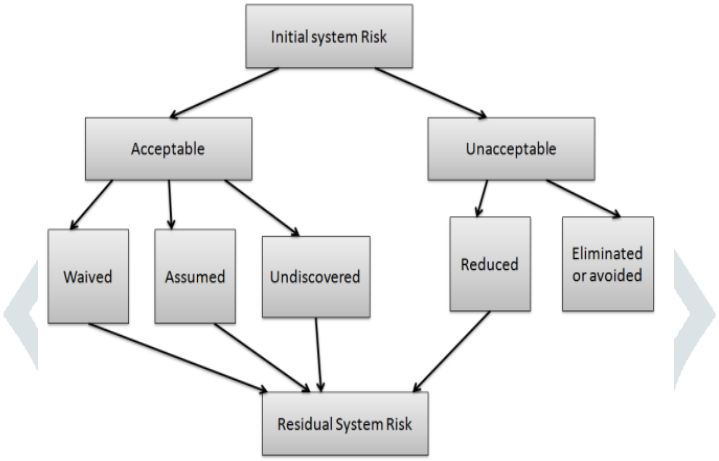

Figure3: Software Risk planning

V. Result

Figure 4: Security key generate

Figure 5: Database collection

VI. CONCLUSION

This paper focuses on the method for concept factorization with document clustering. We present a novel regularized method of concept factorization by incorporating an adaptive neighbors regularization constraint into the concept factorization model. The advantage of approach is that the concept factorization adaptive neighbors dimensionality reduction and finds to neighbors graph weights matrix simultaneously. To solve the proposed problem is efficient algorithm for presented. The proposed method applied to document clustering on document data sets routers-21578, 20 newsgroups, and TDT2. The developed technique some important issues need further investigation. For example, the proposed method the number of cluster to determining should be investigated. In this paper, for the proposed method cannot determine the number of cluster automatically. How to determine the number of cluster automatically in our future work. The parameters are chosen based on the results of experiments. It remains unclear how to select the parameters efficiently and theoretically.

REFERENCES

[1] W. Xu and Y. Gong, "Document clustering by concept factorization," in Proc. Int. Conf. Res. Develop. Inf. Retry. (SIGIR), Sheffield, U.K., Jul. 2004, pp. 202–209.
[2] P. D. O'Grady and B. A. Pearlmutter, "Convolutive non-negative matrix factorisation with a sparseness constraint," in Proc. 16th IEEE Signal Process. Soc. Workshop Mach. Learn. Signal Process. (MLSP), Sep. 2006, pp. 427–432
[3] D. D. Lee and H. S. Seung, "Learning the parts of objects by nonnegative matrix factorization," Nature, vol. 401, pp. 788–791, Oct. 1999.
[4] Z. He, S. Xie, R. Zdunek, G. Zhou, and A. Cichocki, "Symmetric nonnegative matrix factorization: Algorithms and applications to probabilistic clustering," IEEE Trans. Neural Netw., vol. 22, no. 12, pp. 2117–2131, Dec. 2011
[5] W. Xu, X. Liu, and Y. Gong, "Document clustering based on nonnegative matrix factorization," in Proc. Int. Conf. Res. Develop. Inf. Ret., Aug. 2003, pp. 267–273
[6] F. Wang, P. Li, and A. C. König, "Efficient document clustering via online nonnegative matrix factorizations," in Proc. SIAM Int. Conf. Data Mining (SDM), 2011, pp. 908–919.
[7] Y. Chen, J. Zhang, D. Cai, W. Liu, and X. He, "Nonnegative local coordinate factorization for image representation," IEEE Trans. Image Process., vol. 22, no. 3, pp. 969–979, Mar. 2013. [22] R. Zhi, M. Filer, Q. Q. Ruan, and W. B. Kleijn, "Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition," IEEE Trans. Syst. Man, Cybern. B, Cybern. vol. 41, no. 1, pp. 38–52, Feb. 2011.
[8] F. Sha, L. K. Saul, and D. D. Lee, "Multiplicative updates for large margin classifiers," in Lecture Notes in Computer Science, vol. 2777. 2003, pp. 188–202.
[9] Y.-M. Cheung and H. Jia, "Categorical-and-numerical-attribute data clustering based on a unified similarity metric without knowing cluster number," Pattern Recognit., vol. 46, no. 8, pp. 2228–2238, 2013
[10] Z. Yang and E. Oja, "Linear and nonlinear projective nonnegative matrix factorization," IEEE Trans. Neural Netw., vol. 21, no. 5, pp. 734–749, May 2010.
[11] S. Lloyd, "Least squares quantization in PCM," IEEE Trans. Inf. Theory, vol. 28, no. 2, pp. 129–137, Mar. 1982
[12] Y. Chen, L. Wang, and M. Dong, "Non-negative matrix factorization for semisupervised heterogeneous data is coclustering," IEEE Trans. Knowl. Data Eng., vol. 22, no. 10, pp. 1459–1474, Oct. 2010.
[13] C. Ding, T. Li, and M. I. Jordan, "Nonnegative matrix factorization for combinatorial optimization: Spectral clustering, graph matching, and clique finding," in Proc. 8th IEEE Int. Conf. Data Mining (ICDM), Dec. 2008, pp. 183–192
[14] O. Zoidi, A. Tefas, and I. Pitas, "Multiplicative update rules for concurrent nonnegative matrix factorization and maximum margin classification," IEEE Trans. Neural Netw. Learn. Syst., vol. 24, no. 3, pp. 422–434, Mar. 2013.

[15] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," J. Mach. Learn. Res., vol. 5, pp. 1457–1469, Dec. 2004.
[16] C. Ding, T. Li, W. Pang, and H. Park, "Orthogonal nonnegative matrix t-factorizations for clustering," in Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, Aug. 2006, pp. 126–135.
[17] S. Lloyd, "Least squares quantization in PCM," IEEE Trans. Inf. Theory, vol. 28, no. 2, pp. 129–137, Mar. 1982.
[18] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," Science, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000.
[19] Y. Chen, J. Zhang, D. Cai, W. Liu, and X. He, "Nonnegative local coordinate factorization for image representation," IEEE Trans. Image Process., vol. 22, no. 3, pp. 969–979, Mar. 2013. [22] R. Zhi, M. Filer, Q. Q. Ruan, and W. B. Kleijn, "Graph-preserving sparse nonnegative matrix factorization with
[20] Y.-M. Cheung and H. Jia, "Categorical-and-numerical-attribute data clustering based on a unified similarity metric without knowing cluster number," Pattern Recognit., vol. 46, no. 8, pp. 2228–2238, 2013.
[21] Y. Chen, M. Rege, M. Dong, and J. Hua, "Non-negative matrix factorization for semi-supervised data clustering," Knowl. Inf. Syst., vol. 17, no. 3, pp. 355–379, Dec. 2008
[22] D. Cai, X. He, X. Wu, and J. Han, "Non-negative matrix factorization on manifold," in Proc. Int. Conf. Data Mining, Dec. 2008, pp. 63–72.
[23] Z. He, S. Xie, R. Zdunek, G. Zhou, and A. Cichocki, "Symmetric nonnegative matrix factorization: Algorithms and applications to probabilistic clustering," IEEE Trans. Neural Netw., vol. 22, no. 12, pp. 2117–2131, Dec. 2011.
[24] S. Xiang, F. Nye, C. Zhang, and C. Zhang, "Nonlinear dimensionality reduction with local spine embeds," IEEE Trans. Knowl. Data Eng., vol. 21, no. 9, pp. 1285–1298, Sep. 2009.
[25] Z. Zhang and H. Zha, "Principal manifolds and nonlinear dimension reduction via local tangent space alignment," SIAM J. Sci. Comput., vol. 26, no. 1, pp. 313–338, 2002.