International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 10, October- 2014

Dimensionality Reduction of Weighted Word Affinity Graph using Firefly Optimization

Dr. Poonam Yadhav
D.A.V. College of Engineering and Technology
India

***Abstract*— An information retrieval system highly relies on document analysis/ retrieval system. It includes numerous processing stages such as feature extraction, semantic representation, dimensionality reduction and similarity measure. Semantic representation aids for providing a better description to the documents. However, the probability of getting increased dimension for semantic descriptors is high. Hence, dimensionality reduction method plays crucial role. Conventional dimensionality reduction methods such as Principle Component Analysis (PCA), Independent Component Analysis (ICA), etc entertain complex means of dimensionality reduction. In the literature, numerous classical optimization algorithms such as Genetic Algorithm (GA), Particle Swarm Optimization (PSO), etc. have been reported to solve the similar problem. However, valiant attempts have been made on deriving robust optimization over the traditional algorithms. Hence, we exploited Firefly Algorithm (FA) to solve the dimensionality reduction problem. In this paper, we first present a theoretical overview of mapping a dimensionality reduction problem to an optimization problem. Subsequently, we describe the procedural steps to solve the problem using FA. This article is believed to be a context behind the experimental investigation on the performance of FA, when attempting to reduce the dimension of weighted word affinity graph and to retrieve the information effectively.**

Keywords*—Dimensionality; reduction; firefly; information; retrieval; semantic

I. INTRODUCTION

Today's world emphasizes electronic documents to use in all the applications, because electronic documents offer simplicity, easy communication, etc. Hence, all the traditional way of using documents are being converted in electronic format [1]. In contrast, the usage of databases is heavily increased that poses serious challenges to data analysts [2]. Nevertheless, such huge informative contents such as documents, either in text or numerical or structured format, are the pre – requisites of statistical and computational analyses. These have become the bottleneck for any information retrieval system to enable fast processing of the documents [2]. The information retrieval system aims at retrieving relevant documents from massive databases based on the user requirements [7] [8].

In – depth document analysis is a significant component for an information retrieval system to perform effectively while handling diverse user requirements. The primary challenge ahead with the analysis is on representing a concept and diverse words almost using same sense. This challenge remains unsolved in the conventional retrieval system [10]. When the query becomes complex, the conventional retrieval system becomes supine [3]. Hence, semantic representation has become a promising solution [9]. This increases the dimension of the feature vector [16] that led to have a computationally slow similarity check. Here, dimensionality reduction methods plays prominent role [17].

II. PRELIMINARIES

An overview of a document analysis and information retrieval system can be illustrated in Fig 1 [21]. Training phase includes generating feature library by well – constructing the extracted features.

In the first processing stage, local or global or even both the type of features are extracted from the given documents. Thus extracted features are given semantic descriptions. The dimension of the semantic representation may be high dimensional or even worse by attaining multi – dimensional representation [18] [19]. Hence, dimensionality reduction methods play a crucial role. It transforms the semantic descriptions from high dimensional space to a low dimensional space. Hence, converted features occupy feature library, which is used for similarity check, when a test document is given to the system.

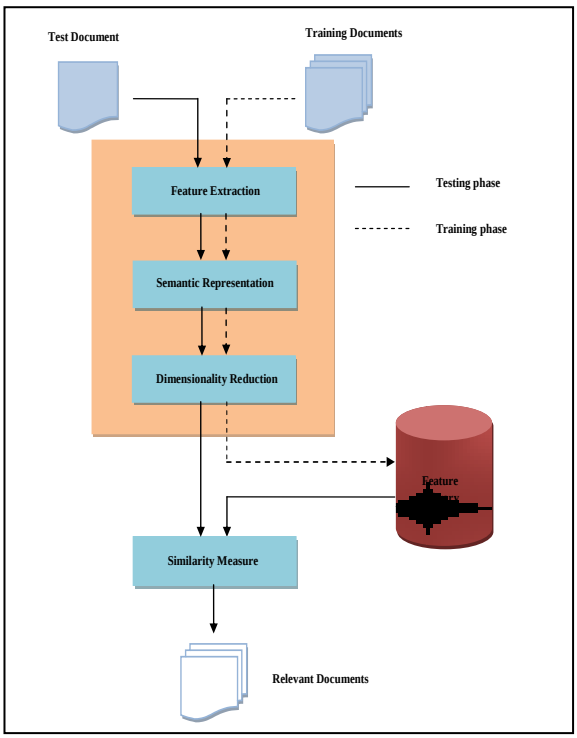

Fig. 1. Overview of document retrieval system [21]

III. MOTIVATION

It is well known about the significance of a dimensionality reduction, when we attempt to give a semantic description for low level features. In our previous work [21], we generated weighted word affinity graph for providing semantic description for the feature descriptors. Despite they are aimed to be computationally intensive, they can be high dimensional. Numerous research works have been carried out in the literature to handle, if any such dimensionality reduction persists. In [11] [12], Latent Semantic Indexing (LSI) has been used prominently. However, the trend has replaced LSI by PCA [14] so that the dimensionality reduction problem has been solved as an eigenvalue problem [5]. ICA and its variants can be considered as promising alternatives to replace PCA [20]. However, all these methods are statistics oriented and hence reliability and computational simplicity cannot be expected much.

This paper aims to solve the dimensionality reduction problem using optimization algorithms. First, we map the dimensionality reduction problem as a maximization problem. Subsequently, we use FA, which is a recent robust optimization algorithm, to solve the maximization problem [15].

IV. PROPOSED DIMENSIONALITY REDUCTION

The proposed dimensionality reduction methods employ FA to reduce the dimension of the global features associated with the document. It is to be noted that the global features are the semantic representation of all low level features extracted from the subjected documents. The proposed dimensionality reduction includes the processing steps as illustrated in Fig 2.

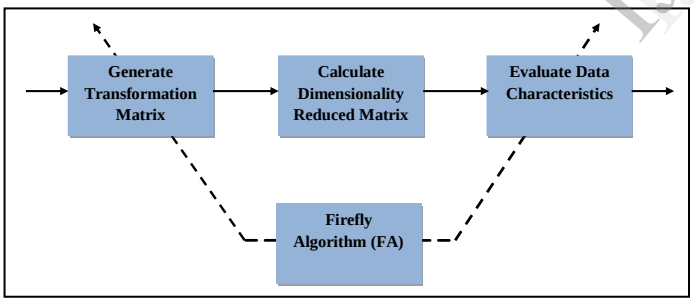

Fig. 2. Proposed dimensionality reduction method

The dimensionality reduction procedure illustrated in Fig 2, replaces the dimensionality reduction method block given in Fig 1. Here, first we describe about the problem formulation, followed by mapping towards optimization problem and solving the problem using FA.

A. Problem Formulation

Let us a consider a document set to be represented as

$$D = \{d_1, d_2, \ldots, d_n\} \in \Re^{m \times n} \quad (1)$$

where, D is a rectangular matrix with documents and terms. The need to apply dimensionality reduction is to reduce m, because it defines the volume of D, which is very big. The objective of dimensionality reduction is to determine D', with size $p \times n$, where $p \ll n$ and D' is the dimension reduced document matrix. This can be represented as

$$D' = V^T D \quad (2)$$

where, V is the transformation matrix to transform the high dimensional matrix to a lower dimensional space. The size of D' should be $m \times p$.

B. Optimization Model

The problem of identifying D' can be mapped to an optimization problem in which the objective is to determine optimal V so that the attributes of D' has to be achieved. Hence, the objective function of the optimization problem can be defined as

$$V^* = \arg\max_V f(V, D') \quad (3)$$

where, V^* is the optimal transformation matrix to be applied for dimensionality reduction and $f(V, D')$ is the function to be maximized.

The maximization function can be written as

$$f(V, D') = \frac{1}{p} \sum_{i=1}^{p} \sqrt{\sum_{j=1}^{m} \left(d'_{ij} - \bar{d}'_j \right)} \quad (4)$$

where, $d'_{ij} \in D'$ obtained after applying Eq. (2) and $\bar{d}'_j$ is the mean vector to be calculated as

$$\bar{d}'_j = \frac{1}{p} \sum_{i=1}^{p} d'_{ij} \quad (5)$$

C. Dimensionaltiy Reduction using FA

FA has been developed by Xin-She Yang based on the flashing behaviour of fireflies [15]. FA is used in our methodology to reduce dimensionality of global feature extracted from document. The pseudo code of FA used for dimensionality reduction method is given in Fig 3.

According to Fig 3, the fireflies V refer to the initial transformation matrix to be optimized. They are generated randomly. The light intensity I, here refers to the maximization function, often termed as fitness function. "Move current firefly in V towards current firefly in V^* ", in Fig 3 refers to update the firefly as per the following equation.

$$V^{new} = V + \beta \left(V^* - V \right) e^{-\gamma r^2} + \alpha \varepsilon \quad (6)$$

where, V^{new} is the updated firefly, V is the old firefly, β is a scaling factor, which should be set as 1 usually., γ is the absorption coefficient, which is a constant related with the problem scale, r^2 is the distance between V and V^*, α is the step size, which should be often related with the improvement of generations, ε is the Gaussian distributed random number generated within the interval $[0,1]$.

Initialize arbitrary fireflies V
Initialize current generation as zero
Calculate Light intensity I
While current generation is lesser than maximum acceptable generation, **do**
Generate V^* as duplicate of V
Sort both V^* and V based on I
For every firefly in V
For every firefly in V^*
If I of current firefly in V^* is greater than I of current firefly in V
Move current firefly in V towards current firefly in V^*
End If
Update attractiveness
Calculate I for updated firefly and update
End for
End for
Save the best firefly based on I
Increase current generation by one
End While

Fig. 3. Pseuodo code of FA on reducing dimensionality reduction

Once the firefly processes have reached the maximum number of generations, we can obtain $V*$ as the optimal transformation matrix to apply with D so that the dimensionality reduced matrix D' can be obtained.

V. CONCLUSION

In our previous work, we introduced weighted word affinity graph for betterment of semantic description for documents. As the dimension of the semantic description has become higher, in this paper, we recommended using FA to reduce the dimension of the semantic representation. Firstly, we mapped dimensionality reduction problem to a maximization problem. Then, we asserted to solve the maximization problem using FA, which is a recent promising optimization problem. The theoretical description of using FA to solve the problem is described further. In the future works, we attempt to study the performance of FA on dimensionality reduction over other dimensionality reduction methods such as PCA, ICA, etc.

REFERENCES

[1] Song Mao, Azriel Rosenfeld, Tapas Kanungo, "Document structure analysis algorithms: a literature survey", DRR 2003, p.p. 197-207, 2003.
[2] Carsten Gorg, Zhicheng Liu, Jaeyeon Kihm, Jaegul Choo, Haesun Park, Member, and John Stasko, "Combining Computational Analyses and Interactive Visualization for Document Exploration and Sensemaking in Jigsaw", IEEE Transactions on Visualization and Computer Graphics, Vol. 19, No. 10, p.p. 1646 – 1663, 2013.
[3] Jinxi Xu Amherst, W. Bruce Croft, "Query expansion using local and global document analysis", Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, p.p. 4-11, 1996.
[4] G. Salton, M. McGill, Eds. "Introduction to Modern Information Retrieval", New York: McGraw-Hill, 1983.
[5] S. Deerwester and S. Dumais, "Indexing by latent semantic analysis," J. Amer. Soc. Inf. Sci., vol. 41, no. 6, pp. 391–407, 1990.
[6] Haijun Zhang, John K. L. Ho, Q. M. Jonathan Wu, and Yunming Ye, "Multidimensional Latent Semantic Analysis Using Term Spatial Information", IEEE Transactions on Cybernetics, Vol. 43, No. 6, p.p. 1625- 1640, 2013
[7] W. B. Frakes and R. Baeza-Yates, "Information Retrieval: Data Structures and Algorithms", Prentice-Hall, Englewood Cliffs, NJ, 1992.
[8] Antoniol, G. ; Canfora, G. ; Casazza, G. ; De Lucia, A; "Information retrieval models for recovering traceability links between code and documentation", Proceedings of International Conference on Software Maintenance, p.p. 40-49, 2000.
[9] Yu-Gang Jiang ; Yang, J. ; Chong-Wah Ngo ; Hauptmann, A.G.; "Representations of Keypoint-Based Semantic Concept Detection: A Comprehensive Study", IEEE Transactions on Multimedia, Vol. 12, No. 1, p.p. 42 – 53, Jan. 2010.
[10] Eaddy, M. ; Antoniol, G. ; Gueheneuc, Y.-G., "CERBERUS: Tracing Requirements to Source Code Using Information Retrieval, Dynamic Analysis, and Program Analysis", 16th IEEE International Conference on Program Comprehension (ICPC 2008), p.p. 53 - 62, 10-13 June 2008.
[11] G. Antoniol, G. Canfora, G. Casazza, A. De Lucia, E. Merlo, "Recovering Traceability Links between Code and Documentation," IEEE Transactions on Software Engineering, Vol .28, No. 10, p.p.970–983, 2002.
[12] D. Poshyvanyk, Y.-G. Guéhéneuc, A. Marcus, G. Antoniol, V. Rajlich, "Feature Location Using Probabilistic Ranking of Methods Based on Execution Scenarios and Information Retrieval," IEEE Transactions on Software Engineering, Vol. 33, No. 6, p.p.420–432, 2007.
[13] Akiko Aizawa, "An information-theoretic perspective of tf–idf measures", Information Processing and Management, Vol. 39, p.p. 45–65, 2003.
[14] Wray Buntine and Aleks Jakulin, "Applying discrete PCA in data analysis", Proceedings of the 20th conference on Uncertainty in artificial intelligence, p.p. 59-66, 2004.
[15] Yang, X. S. (2008). Nature-Inspired Metaheuristic Algorithms. Frome: Luniver Press. ISBN 1-905986-10-6.
[16] Taiping Zhang; Yuan Yan Tang; Bin Fang; Yong Xiang, "Document Clustering in Correlation Similarity Measure Space", IEEE Transactions on Knowledge and Data Engineering, Vol. 24, No. 6, p.p. 1002 – 1013, 2012.
[17] Zhang, L. ; Zhao, Y. ; Zhu, Z. ; Wei, S. ; Wu, X. "Mining Semantically Consistent Patterns for Cross-View Data", IEEE Transactions on Knowledge and Data Engineering, Vol: 26, No. 11, p.p. 2745- 2758, 2014

[18] Chen, B. ; Kuan-Yu Chen ; Pei-Ning Chen ; Yi-Wen Chen, "Spoken Document Retrieval With Unsupervised Query Modeling Techniques", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 20, No. 9, p.p. 2602 – 2612, 2012.

[19] Hanhua Chen ; Hai Jin ; Xucheng Luo ; Yunhao Liu ; Tao Gu ; Chen, K. ; Ni, L.M., "BloomCast: Efficient and Effective Full-Text Retrieval in Unstructured P2P Networks", IEEE Transactions on Parallel and Distributed Systems, Vol 23, No. 2, p.p. 232 - 241, 2012.

[20] Sangwoo Moon ; Hairong Qi, "Hybrid Dimensionality Reduction Method Based on Support Vector Machine and Independent Component Analysis", IEEE Transactions on Neural Networks and Learning Systems, Vol. 23, No. 5, p.p. 749 – 761, 2012

[21] Poonam Yadav, "Weighted Word Affinity Graph for Betterment of Spatial Information Descriptors", Volume-02 , Issue-08, Page No : 117-120, 2014