ISSN:2320-0790

An International Journal of Advanced Computer Technology

Adaptive Firefly Optimization on Reducing High Dimensional Weighted Word Affinity Graph

Dr. Poonam Yadhav[1]

[1]D.A.V. College of Engineering and Technology, India

Abstract: Document analysis and retrieval system can best define an efficient information retrieval system. Among various processing stages in a document analysis and retrieval system, feature descriptors at processing volume limit require more importance while developing the system. This is mainly because of the increase in probability of getting high dimensional semantic description. This increases the vitality of opting a robust dimensionality reduction method for our retrieval system. Principle Component Analysis (PCA), Independent Component Analysis (ICA), etc are the most popular dimensionality reduction methods. However, they are highly complex while handling nonlinear data with multiple characteristics. Optimization algorithms can be a good alternative for the traditional methods. In fact, classical optimization algorithms such as Genetic Algorithm (GA), Particle Swarm Optimization (PSO), etc have been widely applied. However, the data handling remains inefficient under current data exploding scenario. In our previous work, we have exploited Firefly Algorithm (FA) to solve the optimization problem. Due to parameter selection dilemma in traditional FA, this paper concentrates on using Adaptive Firefly Algorithm (AFA). AFA adaptively varies step search of solutions and hence improves the convergence rate of the algorithm. As a result, near – optimal solution can be obtained qualitatively. We further recommend the dimensionality reduction method to handle weighted word affinity graph to improve the retrieval efficiency.

Keywords: firefly; adaptive; dimensionality; semantic; information; retrieval

I. INTRODUCTION

The modern era forces the utilization of electronic documents in all applications due to the simplicity, effortless communication and several other benefits presented by the electronic documents. So, the conventional means of exploiting the documents are being transformed into electronic format [1]. Massive database exploitation enables the data analysts to face severe problems [2], while the documents bearing bulk information in terms of text, number or structure are very much essential in performing statistical and computational analysis. Therefore, the information retrieval system is obstructed from offering a rapid document processing [2]. The information retrieval system intends to retrieve the related documents from a considerably large database depending on the needs of the user [7] [8].

An information retrieval system can operate effectively, if the documents are examined in detail and the various user requirements are met. The chief problem associated with this analysis is the representation of an idea or a group of different words that have similar meaning. The conventional retrieval systems lack the ability to provide a solution to the aforementioned problem [10]. A complex query can cause the conventional retrieval systems to become weaker [3]. Therefore, semantic representation can only be a hopeful solution [9]. But, semantic representation would raise the dimension of the feature vector [16] and this in turn causes the similarity check to be performed in a relatively time-consuming manner. So, dimensionality reduction methods have become more essential [17].

II. PRELIMINARIES

The general idea behind a document analysis and information retrieval system is portrayed in Fig 1 [21]. The training phase consists of creating the feature library with the help of the extracted features.

During the initial stage of processing, the local features or the global features and sometimes both the features are obtained from the documents provided. In this way, the extracted features are semantically described. The dimension of the semantic description may be very large and the situation may become even poorer, if multi-dimensional representation is achieved [18] [19]. So, the means of decreasing the dimensionality is of utmost

importance. With the dimensionality reduction techniques, the semantic descriptions can be converted from a high dimensional space to a low dimensional space. These transformed features get resided in the feature library and can be later utilized to check the degree of similarity of the test document that is provided to the system.

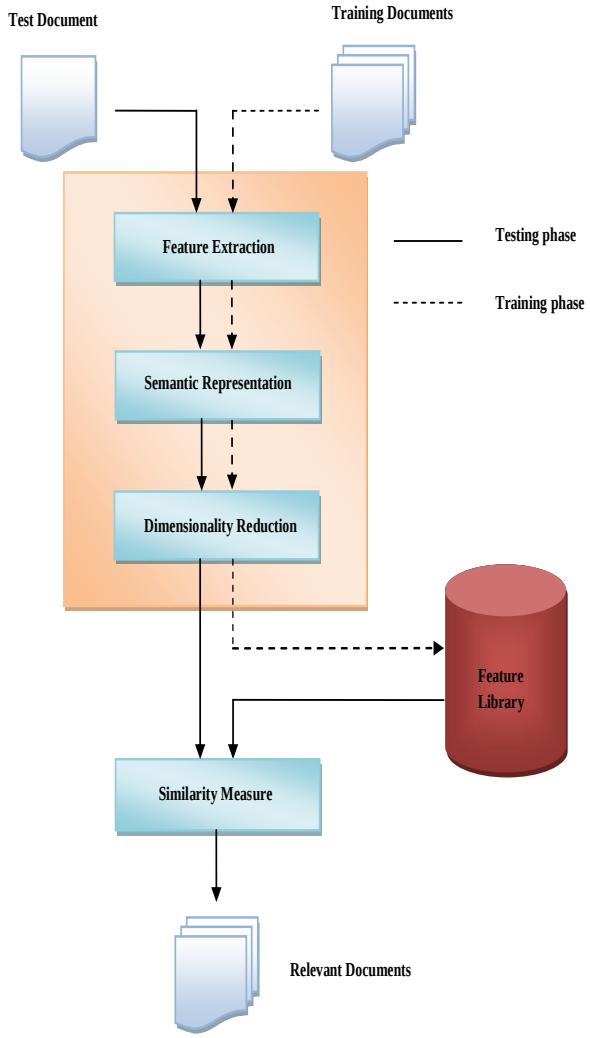

Figure 1: Overview of document retrieval system [21]

III. MOTIVATION

Dimensionality reduction and its importance become much familiar, while the low level features are to be described semantically. In our earlier work [21], the weighted word affinity graph is created to semantically describe the feature descriptors. Though the work was planned to be computationally intensive, high dimensionality still appears. Several attempts were made in the literature to obtain an approach for reducing the dimensionality. [11] and [12] has employed Latent Semantic Indexing (LSI). Then, PCA has been used in the place of LSI [14] and the solution to the dimensionality reduction problem is viewed as an Eigen value problem [5]. A better solution can be thought to be provided by ICA and its variants by replacing PCA [20]. Since all the above mentioned methods are based on statistics, reliability and computational simplicity cannot be greatly anticipated.

In this paper, optimization algorithms are exploited for providing a solution to the dimensionality reduction problem. For this, mapping of dimensionality reduction problem as a maximization problem is carried out initially. Subsequently, we describe proposed AFA based dimensionality reduction method on solving the maximization problem.

IV. DIMENSIONALITY REDUCTION PROBLEM

A. *Problem Formulation*

Let us represent a document set as follows

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

Here, D refers to a rectangular matrix with documents and terms. The primary objective of reducing dimension can be accomplished by minimizing m and hence the huge of size of D can be reduced.

Thus the dimensionality reduction can be a problem of determining D' of size $p \times n : p \ll m$. The obtained D' is the new document matrix with dimensionality reduced that can be represented as

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

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

B. *Optimization Model*

Determining D' can be mapped as an optimization problem by having the intent of optimizing V in such a way that the attributes of D' have to be preserved. The mapped optimization problem takes the objective function as follows.

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

Where, $f\left(V, D'\right)$ is a maximization function and V^* is the optimal transformation matrix to be obtained for performing dimensionality reduction. In order to preserve the attributes of D', the maximization function can be formulated as

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

Here, $d'_{ij} \in D'$ is got by the application of Eq. (2) and $\bar{d}'_j$ is the mean vector that can be computed by Eq. (5).

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

V. PROPOSED METHODOLOGY

A. Dimensionality Reduction using AFA

In this paper, we employ AFA to solve the objective function given in equation (3) and hence the dimensionality reduction can be accomplished. As stated in our previous works [21] [23], the document matrix refers to global features, which are semantic representation of features extracted from the documents at low level. The proposed dimensionality reduction method is illustrated in Figure 2.

The proposed dimensionality reduction method tunes transformation matrix in such a way that the resultant dimensionality reduced matrix can preserve the data attributes. The tuning process is accomplished by AFA. AFA is an improved version of original FA by introducing adaptive solution movement parameters. Subsequently, we explain original FA and AFA to be used for dimensionality reduction.

B. Traditional FA

Firefly algorithm (FA) has been developed by Xin-She Yang and this algorithm is inspired by the flashing behaviour of fireflies [15]. Figure 4 portrays the pseudo code of FA that is adapted for decreasing the dimensionality.

The fireflies V indicate the initial transformation matrix that should be optimized and they are randomly generated. The light intensity I denotes the maximization function that is frequently called as fitness function. "Move current firefly in V towards current firefly in V^* " represents that the firefly has to be updated with the expression stated below.

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

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

Immediately after the firefly processes have attained the maximum number of generations, $V*$, which is the optimal transformation matrix, can be got. This optimal transformation matrix along with D is then used to produce the dimensionality reduced matrix D'.

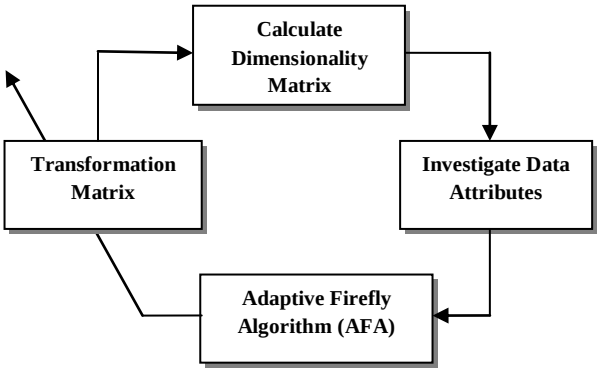

Figure 2: Proposed dimensionality reduction method

C. Adaptiveness in AFA

According to Equation (6), α is set as a constant. This makes FA to take constant step to converge towards the optimal transformation matrix. As a result, the convergence rate has become slow. In AFA, α is made adaptive. In other words, α tends to vary at every iteration according to the convergence of the algorithm. The new adaptive α can be given as

$$\alpha^{new} = \left(\frac{1}{2I^{\max}}\right)^{1/I^{\max}} \alpha^{old} \quad (7)$$

Where, $I^{\max}$ is the maximum number of generations, α^{new} is the new step size and α^{old} is the old step size. The adaptive variation decreases rapidly and exhibits transient variation over the growing number of iterations. This can be visualized from the Figure 3.

VI. CONCLUSION

We have developed FA based dimensionality reduction method to handle weighted word affinity graph. However, FA has taken constant step towards searching near-optimal solution. Due to the drawbacks, the original data characteristics may not be preserved in the dimensionality reduced data. Hence, this paper has introduced AFA in the position of FA. AFA takes the step value in an adaptive fashion so that the searching region has been increased/ decreased according to the iteration count. In our work, the searching ability has been made increased, when the iteration count is towards maximum number of iteration. This can help the solution to evade from local optima and hence to find the global optima. Our future works are based on investigating various recent advancements on optimization algorithm and to use them for dimensionality reduction.

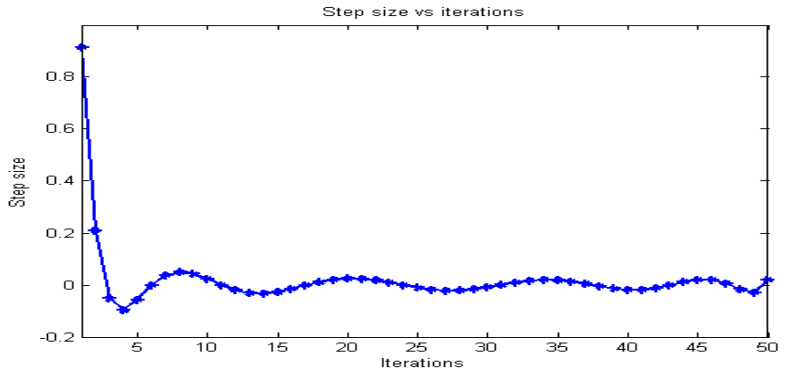

Figure 3: Response of adaptive step size over iterations

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

Figure 4: Pseuodo code of FA on reducing dimensionality reduction

VII. REFERENCES

[1] Song Mao, Azriel Rosenfeld, Tapas Kanungo, "Document structure analysis algorithms: a literature survey", DRR 2003, 2003, p.p. 197-207.

[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, 2013, p.p. 1646 – 1663.

[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, 1996, p.p. 4-11.

[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, 1990, pp. 391–407.

[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, 2013, p.p. 1625- 1640.

[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, 2000, p.p. 40-49.

[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, Jan. 2010, p.p. 42 – 53.

[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), 10-13 June 2008, p.p. 53 – 62.

[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, 2002, p.p.970–983.

[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, 2007, p.p.420–432.

[13] Akiko Aizawa, "An information-theoretic perspective of tf–idf measures", Information Processing and Management, Vol. 39, 2003, p.p. 45–65.

[14] Wray Buntine and Aleks Jakulin, "Applying discrete PCA in data analysis", Proceedings of the 20th conference on Uncertainty in artificial intelligence, 2004, p.p. 59-66.

[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, 2012 , p.p. 2602 – 2612.

[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, 2012 , p.p. 232 – 241.

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

[22] Niknam, T. ; Azizipanah-Abarghooee, R. ; Roosta, A., "Reserve Constrained Dynamic Economic Dispatch: A New Fast Self-Adaptive Modified Firefly Algorithm", IEEE Systems Journal, Vol. 6, No. 4, p.p. 635 – 646, 2012.

[23] Poonam Yadav, 'Dimensionality Reduction of Weighted Word Affinity Graph using Firefly Optimization", International Journal of Engineering Research & Technology, Vol. 3 - Issue 10, October – 2014. Ding, W. and Marchionini, G. 1997 A Study on Video Browsing Strategies. Technical Report. University of Maryland at College Park.