SIMPLIFYING WEIGHTED WORD AFFINITY GRAPH: AN APPROACH USING ARTIFICIAL BEE COLONY ALGORITHM

Poonam Yadav
D.A.V. College of Engineering and Technology, Kanina, India,
Email: poonam.y2002@gmail.com

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 entertains 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 Artificial Bee Colony Algorithm 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 artificial bee colony algorithm. This article is believed to be a context behind the experimental investigation on the performance of artificial bee colony algorithm, when attempting to reduce the dimension of weighted word affinity graph and to retrieve the information effectively.

Keywords: semantic; dimensionality reduction; artificial bee colony (ABC); information retrieval; word affinity

1. INTRODUCTION

Information retrieval (IR) has found as a promising technology in computer based search systems and has been commercialized since 1960s. Over the period, the need and the capability of the IR system in terms of processing speed and storage capacity has increased due to the increased computer technologies [7]. An IR system can be defined as a system that localizes the information relevant to the given user query [5] [6]. The information searching can be done on a structured, semi-structured and even on unstructured data such as web pages, video, etc [1]. However, the necessity of having an effective IR system increases, when the volume of searching database increases rapidly [4].

According to Moore’s law of continual processor speed increase, the digital storage capacity has predicted to be doubled in every two years. For instance, a hard drive has stored 2000 bits in 1956 that has been increased to 100 billion bits in 2005 [2]. Moreover, the traditional way of using documents are being converted into electronic format [3]. Due to handle such rapidly increased database in searching for relevant contents, IR system plays crucial role [1].

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].

In an IR system, in-depth document analysis plays a crucial role in understanding the requirements of the user. However, it faces great challenge when a same intention is presented in different form of words. The traditional IR system did not focus on this system [8]. Moreover, a complex set of query affects the performance of the conventional IR system [9]. This can be solved by introducing more descriptive feature set, but it may degrade the efficiency of similarity measurement process [10].

2. PRELIMINARIES

A prefatory note on document analysis and information retrieval system can be illustrated in Fig 1. The system includes two databases, namely, offline docs and feature library. The offline docs refer to the documents used for training the system, whereas the feature library refers to storing the feature extracted from the documents of offline docs. The elements of feature library are used to determine the similarity between the query document and the database documents.

The system is comprised of three major components such as feature extraction, semantic representation and dimensionality reduction. The feature extraction stage extracts both local and global features from the subjected documents. These representations are given semantic description in the subsequent stage followed by reducing the dimensionality of the extracted features. It is obviously known that semantic representation of extracted features often results in high and multidimensional dataset [11]

[12]. Under these circumstances, dimensionality reduction plays crucial role. Dimensionality reduction is a process of transforming a high dimensional data into a low dimensional data. Hence obtained features are subjected to similarity measure, where the features of the database documents that are similar to the user document are determined.

3. MOTIVATION

The significance of dimensionality reduction in an IR system is highly crucial. Earlier, we have provided an efficient semantic description for the documents [13]. However, the description is in high dimension. Numerous methods have been reported in the literature to perform dimensionality reduction. Latent Semantic Indexing (LSI) has been used as a promising dimensionality reduction method [14] [15]. However, PCA was proved to be better than LSI, over the period [16]. In PCA, the dimensionality reduction problem has been considered as eigenvalue problem [17]. Subsequently, ICA has replaced PCA [18]. Since, these methods are based on the statistics, the reliability and computational efficiency remains questionable.

This paper considers the dimensionality reduction problem as an optimization algorithm. Further, we introduce ABC algorithm to solve the optimization problem. Here, the optimization problem is mapped as a maximization function.

4. PROPOSED DIMENSIONALITY REDUCTION

This paper utilizes Artificial Bee Colony (ABC) algorithm as the dimensionality reduction approach and hence, attempts to achieve a reduction in the dimension of the global features pertaining to the document. The semantic description of the entire number of low level features, which are extorted from the documents under consideration, serves as the global features. Fig 2 elucidates the various processes involved in the proposed dimensionality reduction scheme.

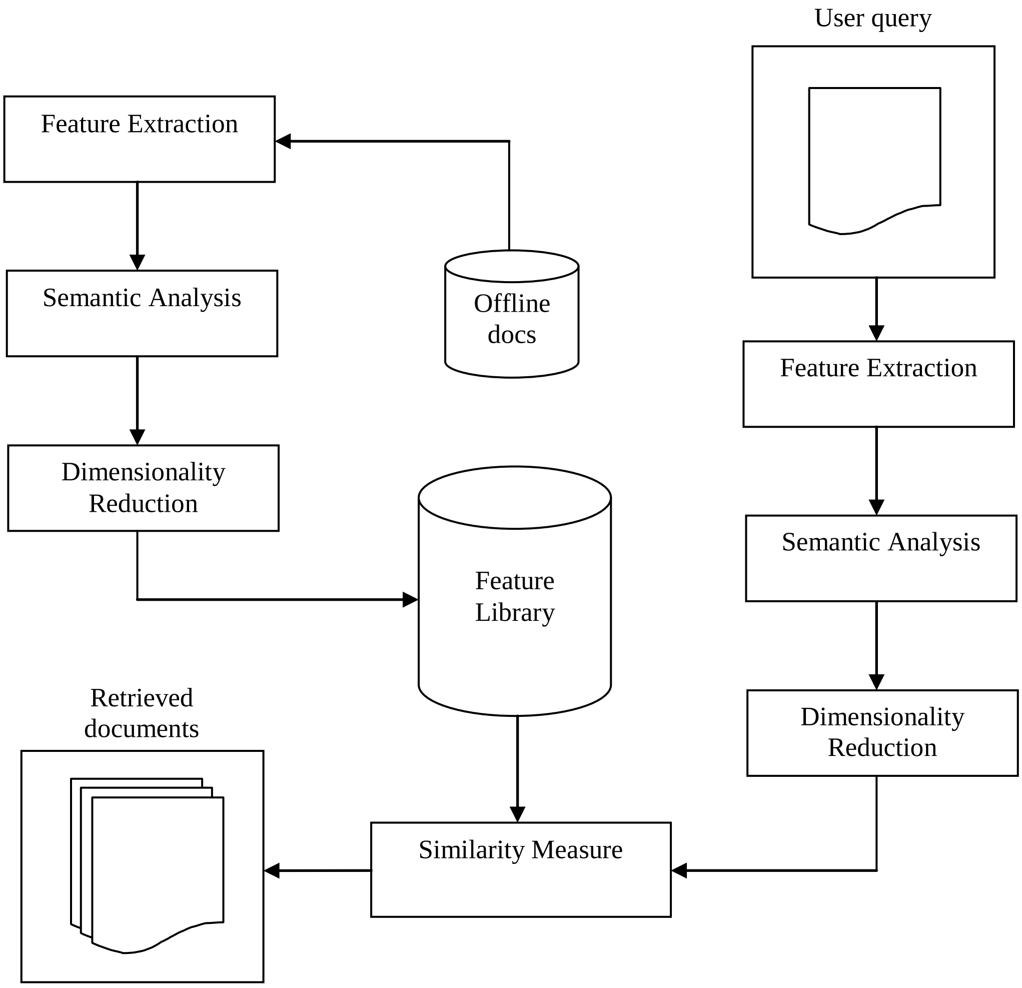

Fig. 1: Illustration of Information retrieval system

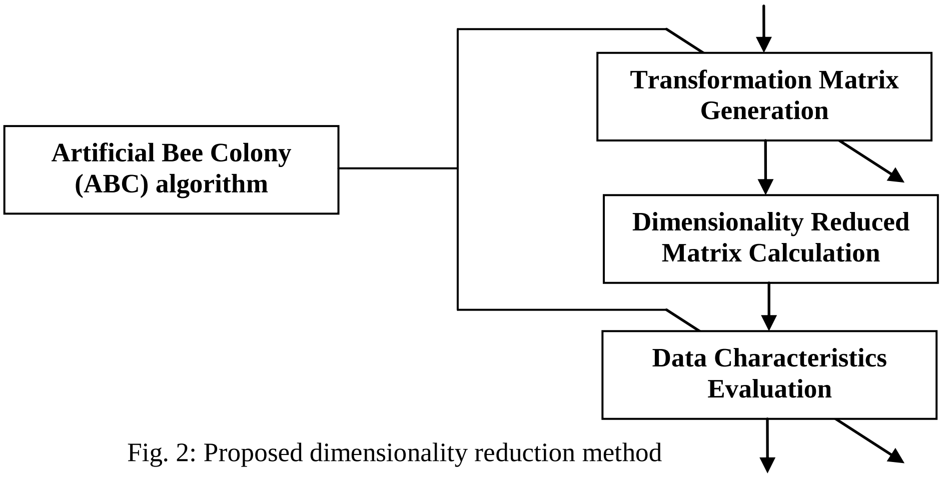
Artificial Bee Colony (ABC) algorithm
Transformation Matrix Generation
Dimensionality Reduced Matrix Calculation
Data Characteristics Evaluation

Fig. 2: Proposed dimensionality reduction method

The block associated with the dimensionality reduction technique in Fig 1 is exchanged with the dimensionality reduction scheme portrayed in Fig 2. Initially, a discussion on the formulation of the problem is made. Then, the way the problem can be mapped to optimization is analyzed and ABC is employed at the later stages to provide solution to the issue.

4.1. Problem Formulation

Assume the document set to be,

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

where, D specify a rectangular matrix enclosing the documents along with the terms. The application of dimensionality reduction technique is highly necessary for decreasing the value of m. The reason is that m indicates the volume of D and it is found to have a larger value. The main aim of dimensionality reduction is to establish a matrix of size $p\times n$. Here, $p \ll n$ and D' denotes the document matrix with decreased dimension, which can be expressed as in Eq. (2).

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

In this equation, V point to the transformation matrix that is required to transform the high dimensional matrix to a lower dimensional space and its size is $m\times p$.

4.2. Optimization Model

The problem related to the finding of D' can be seen from an optimization perspective. The goal of this particular optimization problem is to decide on the optimal factors, which satisfy the attributes of D'. Thus, the objective function of the optimization problem can be stated as in Eq. (3).

$$V^* = \underset{V}{\arg\max}\, f\left(V,D'\right) \qquad (3)$$

where, V^* is the optimal transformation matrix used for reducing the dimensionality and $f\left(V,D'\right)$ specify the function that has to undergo maximization.

Then, the maximization function can be represented as in Eq. (4).

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

where, $d'_{ij}\in D'$ was computed using Eq. (2) and $\bar{d}'_j$ indicates the mean vector that can be calculated as follows:

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

4.3. Dimensionaltiy Reduction using ABC

ABC has been proposed by Karaboga in 2005 [19]. ABC is an optimization model inspired from the swarming behavior of bee colony. The model consists of colony of artificial bees that contains three groups of bees: employed bees, onlookers and scouts. A bee waiting on the dance area for making decision to choose a food source is called an onlooker and a bee going to the food source visited by itself previously is named an employed bee. A bee carrying out random search is called a scout. In the ABC algorithm, first half of the colony consists of employed artificial bees and the second half constitutes the onlookers. For every food source, there is only one employed bee. In other words, the number of employed bees is equal to the number of food sources around the hive. The employed bee whose food source is exhausted by the employed and onlooker bees becomes a scout [19].

The basic steps of ABC can be illustrated as

- Initialize.
- REPEAT.

  - Place the employed bees on the food sources in the memory;
  - Place the onlooker bees on the food sources in the memory;
  - Send the scouts to the search area for discovering new food sources.
- UNTIL (requirements are met).

The pseudo code of ABC to perform dimensionality reduction is illustrated in Figure 3.

At the first step, the ABC generates a randomly distributed initial population of various transformation matrices, termed as V . The term, 'cycle', refers to the number of optimizing steps. It is initialized as one followed by calculating the fitness of the initial population. Each employed bee produces a new solution using the following equation.

$$V_{ij}^{*} = V_{ij} + \phi_{ij}\left(V_{ij} - V_{kj}\right):\ 0 \le i,k \le N_p \text{ and } 0 \le j \le |V| \qquad (6)$$

where, N_p is the population size and $|V|$ is the dimension of the transformation matrix. For every such new solution, the fitness is calculated. After the solution sets are updated based on the greed selection process, the probability value for each solution is calculated using the following equation.

$$p_i = \frac{f_i}{\Sigma f_i} \qquad (7)$$

```
Initialize population $V$
Initialize current cycle as one
Calculate fitness $f$ of the population
REPEAT
        For each employed bee {
                Produce new solution $V^*$ using Eq. (6)
                Calculate the fitness $f^*$
                Apply greedy selection process}
        Calculate the probability values $\rho$ for the solutions $V$ using Eq. (7)
        For each onlooker bee {
                Select solutions $V$ based on $\rho$
                Produce new solution $V^*$
                Calculate the fitness $f^*$
                Apply greedy selection process}
        If there is an abandoned solution for the scout, then replace it with a new solution which will be randomly produced
        by Eq. (8)
        Memorize the best solution so far
        Increment cycle by one
UNTIL cycle = maximum number of cycles
```

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

Every onlooker bee is defined based on the probability calculation. In other words, the onlooker bees are the bees that have good probability rate, as per equation (7). For every onlooker bee, the similar new solution upgrading process is performed followed by fitness calculation and the greedy selection process. The process is repeated, while each onlooker bee is monitored for improvement. If any onlooker bee does not show any improvement, its value is memorized and it is replaced as scout bee. The scout bee generation is performed based on the following equation.

$$SC_i = SC_i^{\min} + r_1 * \left(SC_i^{\max} - SC_i^{\min}\right) \qquad (8)$$

where, SC_i, $SC_i^{\min}$ and $SC_i^{\max}$ refer to the generated scout bee, minimum and maximum limits for the scout bees, respectively.

Once this process reaches the maximum number of cycles, we can obtain $V*$, which is the optimal transformation matrix to apply with D so that the dimensionality reduced matrix D' can be obtained.

5. 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 ABC 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 ABC, which is a recent promising optimization problem. The theoretical description of using ABC to solve the problem is described further. In the future works, we attempt to study the performance of ABC on dimensionality reduction over other dimensionality reduction methods such as PCA, ICA, etc.

6. REFERNCES

[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

[22] D. Karaboga, An Idea Based On Honey Bee Swarm for Numerical Optimization, Technical Report-TR06,Erciyes University, Engineering Faculty, Computer Engineering Department 2005.