ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)
Vol. 5, Issue 1 (Jan. - Mar. 2017)
International Journal of Advanced Research in Computer Science & Technology (IJARCST 2017)

# Ant Optimized Social Behavior Model To Evaluate Web Document Key Word Matching

[I]Dr R.S. Vetrivel, [II]Sivaranjini.C, [III]P Jeyanthi Rani
[I]Professor, [II]Research Scholar, [III]Assistant Professor
[I,II,III]Computer Science, Subramanya College of Arts & Science, Palani

**Abstract**
*Web mining has become a more demaning issue in resolving the effects of web document key word matching. Information domain with keywords matching technique has an essential growth to support highest similarity for particular input data. An ant optimized model is applied to choose the appropriate keyword in Web documents. Collection of keywords is, however, not considered in the evaluation process, thus raising the probability of inaccurate keyword matching. The existing work introduced a fuzzy clustering algorithm that performed on related input data. It designs an Expectation-Maximization framework graph as centralized of an object in the graph which is described as likelihood. In sentence clustering operation, fuzzy cluster algorithm was employed to determine the related clusters of similar correlated sentences with the help of text mining tasks. However, existing work fails to operate the hierarchical structured documents and works only on flat clustered patterns of the document. The fuzzy clustering algorithm does not evaluate keyword matching process. In order to overcome these issues, Ant Optimized Social Behavior Model to evaluate web document key word matching (AOSBM) technique is presented to provide accurate keyword matching for data mining technique. At first, Web document collection is performed to collect the documents from the Web. Then, the keyword search model is used to discover the single word include suitable meaning. Next, Ant Optimized model is applied to evaluate the Web document keyword matching process effectively this in turn trust level is improved in proposed AOSBM technique. Lastly, Correlation based keyword matching process is employed to calculate the high correlation among the keyword from user input queries.*

**Key Terms**
*Web Documents, Keyword Matching, Ant Optimized Social Behavior Model*

## I. Introduction

### A. Web Mining

Web mining is an important operation in data mining method that used to identify the patterns from the Web. Web mining is widely used for extracting the user behaviour information or other changing information from multiple data sets. The basic structure of Web mining is divided into Web content mining, Web structure mining and Web usage mining. Web content mining is the most importance one for extracting relevant information from Web. Web content mining introduces searching effect of relevant information for retrieving exact users quires and terms. Web content mining includes the text, images, audio, video data or structured records and so on. It is also called Web document mining. Web document mining technologies are the most important one for extracting knowledge on the Web.

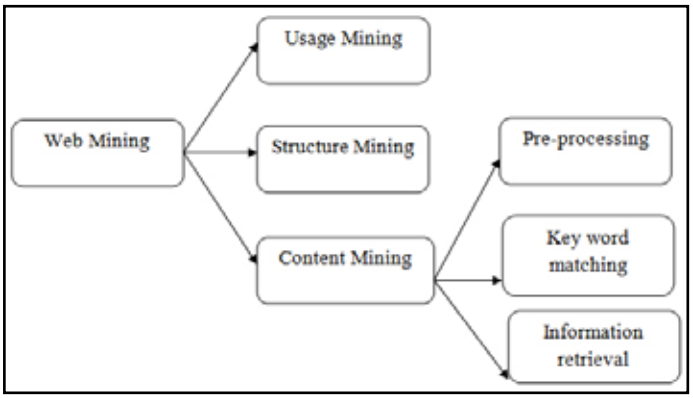


Fig. 1.1 Taxonomy of Web mining

Figure 1.1 shows the classification of web mining for retrieving information from Web documents. Application of text mining to Web content has severely investigated. Text mining concerned about topic discovery, extracting correlation patterns, clustering of web documents and classification of Web Pages. Web Structure Mining involves the process of identifying structure information from the Web. The structure of a common Web graph includes Web pages as nodes and hyperlinks as edges that associate related pages. Web structure mining further separated into two kinds based on the kind of structure information used.

Web usage mining is the most important operation of data mining method used for determining interesting usage patterns from Web data. Usage data confines the uniqueness or source of Web users along with their browsing behavior at a Web site. Web usage mining is further divided into three types based on the kind of usage data. The type of usage data is Web server data, application server data and application level data. In Web server data, the user logs are gathered by Web server. It contains IP address, access time and page reference. Application server data has the potentiality to record different kind of business events. A new kind of events is described in an application and provides the records of these specially described events.

### B. Ant Optimization

Ant optimization is a recently planned metaheuristic technique to deals with hard combinatorial expansion problems. The stimulating source of ant optimization is called trail tracking operation and follows the behaviour of real ants which employ a communication medium. In ant optimization technique, artificial ants produce uncertain construction trials that probabilistically generate a solution by continuously implementing solution components to partial solutions. An uncertain component in ant optimization permits the ants to construct huge solutions and hence investigate a more number of solutions.

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

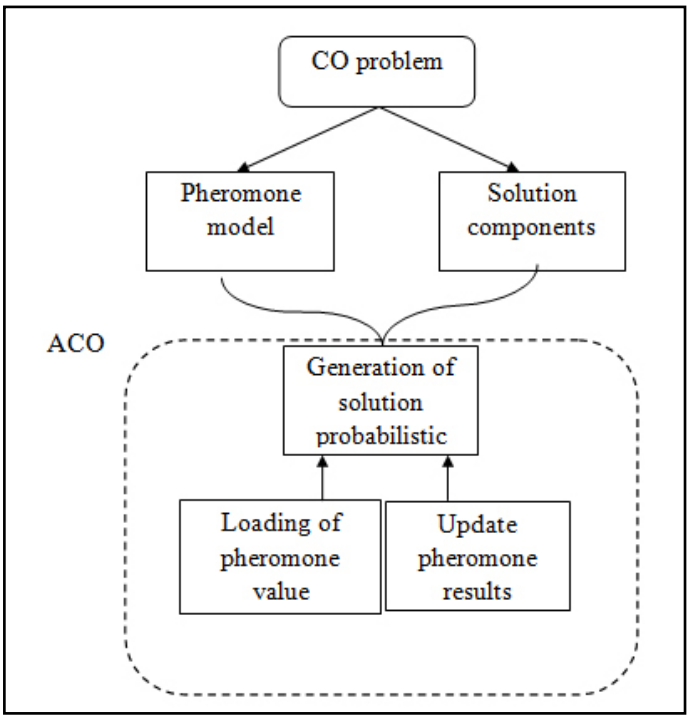

Fig. 1.2 : Process of ant colony optimization

The basic working of an ant colony optimization algorithm is shown in figure 1.2. Initially finite set of solution components are used to solve the colony optimization (CO) problem by collecting the solutions to the CO problem. Next, a set of pheromone values are described and represented as T. This set of values is generally called the pheromone model which is observed from a practical view parameterized probabilistic model. The pheromone model is one of the important components of the ant colony optimization metaheuristic. Typically, pheromone values are interrelated to solution components. In general, ant colony optimization scheme tries to determine an optimization problem with the help of processing following steps. The candidate results are generated with the help of pheromone model i.e., a parameterized probability allocation done with the solution space. The candidate results are applied for altering the pheromone values in such a way that is considered to bias potential sampling.

### C. Keyword matching

Web mining is the process of extracting useful information from the Web documents. Due to the large amount of available data, computation and digitalization associated to the domain like business, industry, science, finance, banking, and healthcare are increased. Therefore ant optimization is required to be constructed which matches the query to the appropriate synonyms and thus reduce the computation related to the above said domain. Hence ant optimization technique develops appropriate results in Web document keyword matching. Most applicable keyword produce effective results since, it search for more documents compressing the relevant data. The document keyword matching is explained as follows. It contains the phases of query input, association of relevant keyword, matching keyword and displays the relevant results phases.

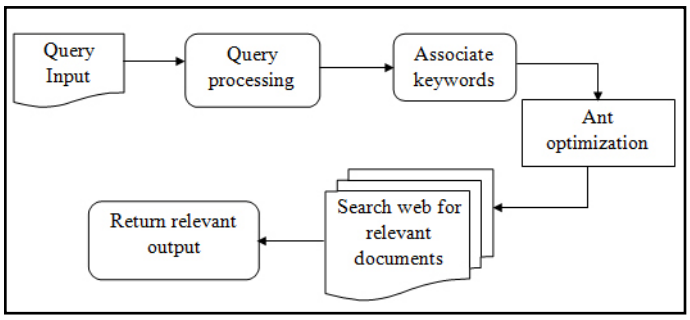

Fig. 1.3 : Web document keyword matching process

Figure 1.3 shows the Web document keyword matching process. Initially, the query is entered by user for performing the search operation. The entered query is highly mention relevant keywords to his query. After entering the query, it used for further processing. In association of relevant keywords phase, ant optimization is applied for entering keyword and thus obtains relevant words with the help of social behaviour model. When search is completed, the list of synonyms relevant to keyword is retrieved. Web is searched for retrieving relevant documents depending on the keyword and synonym. Next, phase of matching keywords on web is look for relevant documents depended on the keyword. For example, the search for data mining contains the search for data mining with knowledge discovery, classification, etc. The final phase of displaying relevant results are performed to determine the relevant to query and are returned and displayed on search page.

## II. Literature Survey

### 1. Keyword Extraction and Clustering for Document Recommendation in Conversations

In this paper [1], an algorithm designed extracts keywords from output of ASR system. The system employs the topic modeling methods. Submodular reward function helps diversity in keyword to match potential diversity and minimizes the ASR noise.
A technique is designed to derive many separated queries from keyword set for increasing the chances of making relevant recommendation while searching for queries in English Wikipedia.

### 2. Hybrid Ontology for Semantic Information Retrieval Model using Keyword Matching Indexing System

This paper [2], presented a query mechanism for information retrieval that combines ontology queries with keyword search. Ontology is the process of clarification of ideas of information domain common for many users. An ontology is established into information retrieval is used to introduce searching effects of relevant information users.
Keyword matching process with historical domain is essential one for identifying the best match for input queries. The ontology-based query is arranged to predicate the logic uncertainty for routing the query to servers. Matching algorithms classifies warm area in computer science and artificial intelligence. The semantic matching results are studied based on the input queries and information in ontology field.

### 3. Machine Learning-Based Web Documents Categorization by Semantic Graphs

In this paper [3], Web pages categorization was carried out with help of semantic graphs and machine learning techniques. Semantic

graph presents compact and structured concepts representation in document based on semantic information.
The semantic graph identifies the map of semantic areas in document. The semantic measure between terms is computed with lexical database (WordNet). The document categorization uses machine learning technique.

### 4. Semantic analysis of web documents for the generation of optimal content

This paper [4], presented the LDArank is a mechanism with Latent Dirichlet Allocation (LDA) for semantic analysis of web content and generation of optimal content for particular queries. Search Engine Optimization (SEO) manages the variations of search engines and web.
SEO's is used for design of efficient plan for optimal ranking of websites and web pages in search engines. *LSHrank* is used for creating the web content against many SEO factors.

### 5. Context-Based Diversification for Keyword Queries over XML Data

This paper [5], presented the approach called diversify XML keyword search with its dissimilar contexts in XML data. A short, vague keyword query and XML data are found and derive the keyword search candidates of the query by feature selection model.
An effective XML keyword search diversification model calculates the quality of all candidates. Three efficient algorithms calculate generated query candidates with diversified search intentions to identify and to return top-k qualified query candidates same as keyword query with large distinct results.

### 6. Finding similar documents using different clustering techniques

This paper [6], presented a document clustering technique for data mining. Many models are designed to groups the capstone project documents by three clustering methods like k-means, k-means fast, and k-medoids.
Three similarity measures are verified like cosine similarity, Jaccard similarity and Correlation Coefficient. The quality of the obtained models is evaluated and compared. The results signify best performance attained using k- means and k-medoids combined with cosine similarity. The variation in quality of clustering depending on evaluation measure is employed.

### 7. K-Means Document Clustering using Vector Space Model

In this paper [7], a new K-Means Clustering technique is designed where Cosine Similarity of Vector Space Model is employed as centroid for clustering.
The documents are clustered when dimension is high as it uses vector space representation. Document Clustering is grouping of same documents into classes. The documents in same clusters are identical while documents in different clusters are not identical. It is famous technique in data analysis with statistics and image analysis. The traditional clustering approaches do not use the algorithmic approach while managing the high dimensional data.

### 8. Enhanced Associative Classification of XML Documents Supported by Semantic Concepts

In this paper [8], a new approach called supervised classification was designed to classify XML documents with help of rule based classifier through enriched structure and content features.
This methodology addresses the existing issues through accomplishing the classification with structure and content features. It uses ontological information into structural and content based features from XML documents and changes into transaction formats where FP-growth algorithm creates the association rules. An associative classifier eliminates the irrelevant rules from generated association rule.

### 9. Approximate XML structure validation based on document–grammar tree similarity

In this paper [9], document–grammar tree similarity method is designed for computing the structural similarity between XML document and XML grammar (DTD or XSD) that choose limitation on presence, repeatability and XML elements/attributes.
The designed approach uses idea of tree edit distance with new edit distance recurrence and dedicated algorithms for evaluating XML documents and grammar structures as ordered labeled trees. The designed method executes an exact validation with maximum similarity threshold on results.

### 10. Web Document Clustering System Using K – Means Algorithm

This paper [10], presented clustering technique which groups unordered text documents into meaningful and coherent clusters. Document clustering is an essential technique for data analysis. The documents similarity is computed through the Winnowing algorithm and Cosine algorithm. Topic modeling is a statistical model used for determining the latent topics in documents through machine learning.

### 11. On retrieving intelligently plagiarized documents using semantic similarity

In this paper [11], semantic similarity measure is designed for effectively retrieving plagiarized documents from Web.
Semantic similarity measure using Nearest Neighbor search and a kernel in multiclass support vector machine (SVM) to detect plagiarism in text documents. In this method, performance gets increased for the respective increase in size of database. Therefore, semantic similarity measure is employed to retrieve the documents with structured information.

### 12. A semantic similarity measure for linked data: An information content-based approach

This paper [12] presented a systematic measurement method based on semantic similarity among resources in Linked Data.
The degree of similarity among items was successfully achieved in this method. Hence, systematic measurement was utilized for classifying unstructured text documents with high clustering efficiency.

### 13. A Semantic Similarity Measure for Linked Data: Information Content-Based Approach

In this paper [13], a measurement model of semantic similarity is designed between resources in Linked Data.
With feature-based definition of Linked Data, a generalized information content-based approach increases the efficiency of existing methods limited specific knowledge representation models. A document representation method called WordNet-based lexical semantic VSM addresses the issues of similarity

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

measures.

### 14. A three-phase approach to document clustering based on topic significance degree

This paper [14], developed the three-phase approach for topic based document clustering. In first phase, best topic model is described and designed formal concept about significance degree of topics for identifying best number of suitable topics from original topic model by LDA.

An initial clustering centers are used by k-means++ algorithm. An initial clustering center is obtained and utilized k-means algorithm for document clustering.

### 15. A fuzzy document clustering approach based on domain-specified ontology

This paper [15], designed the domain-specific controlled vocabulary that explaining about the hazards with dairy products. Synonyms of controlled vocabulary in document set are essential for feature selection. With vector space model (VSM), singular value decomposition (SVD) converts all term-document vectors into concept space.

The mutual information between documents is attained through measuring the similarity of two document vectors in orthogonal matrix of right singular vectors. As information matrix is fuzzy compatible relation, fuzzy equivalence is derived through computing max–min transitive closure. Depending on fuzzy equivalence relation, all data sequences are assigned into clusters under cluster validation index.

## III. Ant Optimized Social Behavior Model to Evaluate Web Document Key Word Matching

Web mining has become a more demanding issue in resolving the effects of web document key word matching. Most of the Web mining technique adopting, document keyword matching changes the originality of the database and were designed to partially evaluate the document keyword matching. Collection of keywords is, however, not considered in the evaluation process, thus raising the probability of inaccurate keyword matching.

In this work, we plan to develop an Ant Optimized Social Behavior Model to evaluate web document key word matching (AOSBM) technique and expand the scope of accurate keyword matching for data mining technique. An ant optimized model is divided to perform following process. Initially, Web document collection is performed to collect the documents from the Web. It is used to further process the keyword search for obtaining relevant documents. Next, the keyword search model is employed to determine the single word include suitable meaning. The keyword search model is usually takes the input as queries.

Then, the Ant Optimized Social Behavior Model is designed to develop the Web document keyword matching from large database. Ant Optimized model also used to evaluate the Web document keyword matching process effectively this in turn trust level is improved in proposed AOSBM technique. Finally, Correlation based keyword matching process is applied calculate the high correlation among the keyword from user input queries this in turns ensures the time for keyword matching in data mining technique. In addition the collected Web documents are used to minimize the search time for keyword in databases.

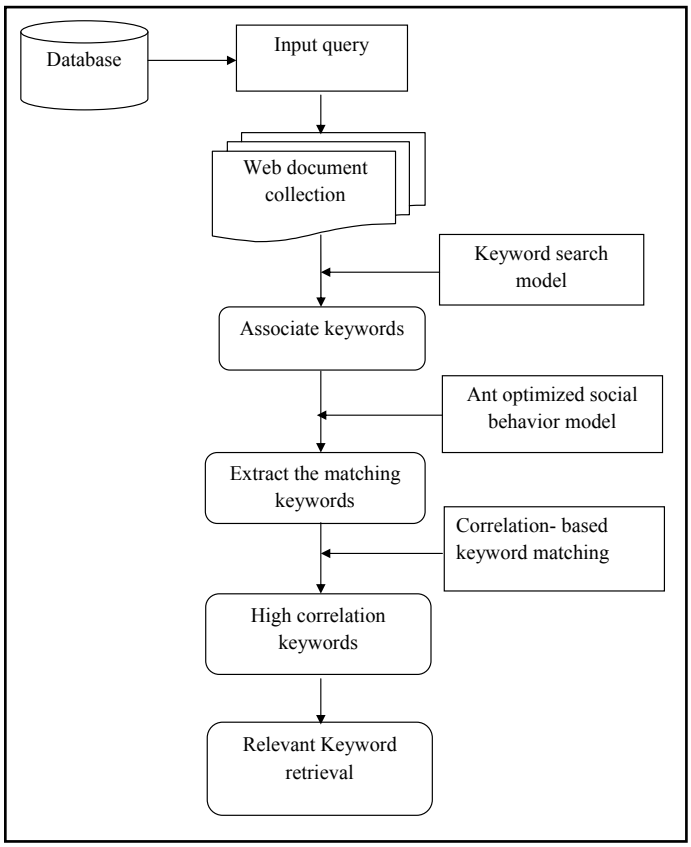


Fig. 3.1 : Architecture of Web document keyword matching process based on the Ant optimized social behavior model

Figure 3.1 shows the Web document keyword matching process. It consists of four steps for providing accurate keyword matching on the Web documents. At first, Web documents are collected with the help of Web document collection process. It is further used for keyword search model. Based on the keywords the related documents are searched by keyword search model. Next, ant optimized social behavior model is employed to determine the matching keyword in the Web documents. Finally, correlation-based keyword matching is applied to calculate the correlation among the matching keyword. This in turns (high correlation) relevant document is extracted from Web in data mining technique.

The Web document key word matching process is divided into:

a) Web document collection
b) Keyword search model
c) Ant optimized social behavior model
d) Correlation-based keyword matching

### a) Web document collection

Web document collection has the significant role for finding appropriate information from large database. Web document collections have grown through an interactive and time-dependent process. The Web is unique in the sense that it is self organized. Document collections are important data sets in many applications. It has the set of queries or databases or documents from Web. Most of the Web document collections require meaningful source and hyperlink structure for providing related information. These collected documents are matched with relevant queries. Information retrieval is used to search the keyword within a document collection for obtaining most relevant information to a user's query. The type of document collection is severely

affects the schemes and algorithms used to process queries. Web document collection is used to retrieve the search information in largest linked database. These collections are taken into the account of further process the document keyword matching operation in data mining technique.

### b) Keyword search model

After, the Web document collection, keyword search is performed in document collection. In keyword search, queries are generally used to search the keywords from Web to extracting relevant documents. Querying of databases is depended upon the difficult query languages that are unsuitable for casual end-users, since they are difficult and inflexible to learn. The development of keyword search and the wide variety use of databases are the central part of data published on the Web. The document or web keyword search process is employed to identify the single word with appropriate meaning in different relations. Keyword search model is easily extended to reduce the weights documents for finding accurate match or keywords. Keywords are especially appropriate in the context of matching metadata. Systems holding keyword search on comparative data minimize the difficulties while publishing relevant data on the Web and producing it searchable. Keyword search has provides number of features which allow users without understanding the database or scheme to search the relevant document with ease.

### c) Ant optimized social behavior model

An Ant optimized social behavior model is developed to expand the scope of improving the Web document keyword matching and reduces the searching time in data mining technique with large database. Initially, keyword matching web document from collected databases are evaluated using ant optimized based relative keyword search. Based on the evaluated Web documents, optimal matching of relevant document is arrived with ant optimized social behavior model even for larger databases, ensuring time for keyword matching. The keyword matching on the collected Web documents are checked for several users' requested database or keywords.

The basic concept of ant principle is that the random wandering nature and upon successful identification of food return to their colony while laying down pheromone trails. On the other hand, if other ants discover those paths, the ants again does not travel at random manner, but it blindly follows the trail provided by the earlier ants. In a similar manner, if the keywords in the documents is said to occur repeatedly, then it is said to be appropriate keyword. By changing the keyword in a random manner with the aid of probability functions, frequent appropriate keywords are matched in an efficient manner.

### d) Correlation-based keyword matching

After performing ant optimized social behavior model, the correlation-based approach is designed to efficiently match the keyword from Web document collection. A matching of appropriate keywords investigates different user requested query on the database that enhances the user trust level by using Web document keyword matching in data mining. The correlation between the matched keywords and the databases (i.e. queries) which aims to reduce the complexity of Web document keyword matching using ant optimized social behavior model method.

| Input: Database 'D' containing Text documents ,Input query with number of terms |
|---|
| Output: Relevant keyword matching |
| Step 1: Begin<br>Step 2:Input is obtained from database 'D'<br>Step 3: Documents are collected from input query<br>Step 4: Related keywords are identified using keyword search model<br>Step 5: Extract the matching keywords using ant optimized social behavior model<br>Step 6: High correlation keyword extraction Based on the correlation keyword matching<br>Step 7: Relevant keyword matching document is retrieved<br>Step 8: End |

Fig. 3.2 : Algorithm of proposed AOSBM method

## IV. Results and Discussion

### A. Performance Metrics

In this section we evaluate performance Ant Optimized Social Behavior Model to evaluate web document key word matching through java environment. To confirm the analytical results, we implemented Ant Optimized Social Behavior Model in java environment and evaluated the Ant Optimized Social Behavior performance of technique. The performance of is evaluated by the following metrics.

- Correlation Coefficient
- Time for keyword matching
- Document collection efficiency

### 1. Correlation coefficient

Pearson product-moment correlation coefficient is used to measure the strength and direction of the linear relationship between two variables that is defined as the (keywords) covariance of the variables divided by the product of their (keyword) standard deviations.

Table 4.1.Correlation Coefficient

| Web document information | Correlation Coefficient | |
|---|---|---|
| | fuzzy clustering algorithm (FCM) | Ant Optimized Social Behavior Model (AOSBM) |
| 5 | 21.45 | 27.54 |
| 10 | 24.56 | 34.72 |
| 15 | 29.91 | 41.82 |
| 20 | 31.89 | 46.90 |
| 25 | 34.71 | 49.9 |

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

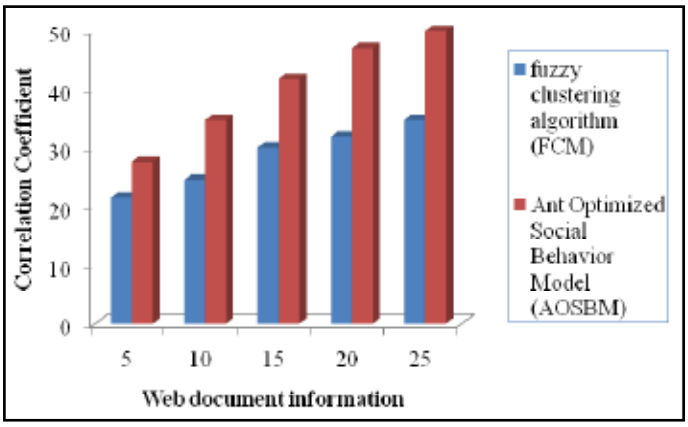


Fig. 4.1 : Measure of Correlation Coefficient

Figure 4.1 demonstrates Correlation Coefficient. X axis represents the Web document information whereas Y axis denotes Correlation Coefficient using both the existing fuzzy clustering algorithm (FCM) and Ant Optimized Social Behavior Model (AOSBM). When the Web document information increased, Correlation Coefficient gets increases accordingly. Figure 4.1 shows better performance of Ant Optimized Social Behavior Model (AOSBM) in terms of Web document information than existing fuzzy clustering algorithm (FCM). Ant Optimized Social Behavior Model (AOSBM) achieves 40% higher Correlation Coefficient variation when compared to existing system.

### 2. Time for keyword matching

The time for keyword matching is measured based on the total number of keywords and the time taken to require keyword matching using proposed AOSBM scheme. The time for keyword matching is measured in terms of milliseconds (ms).

Table 4.2 : Tabulation for Time for keyword matching

| Number of search | Time for keyword matching (ms) | |
|---|---|---|
| | fuzzy clustering algorithm (FCM) | Ant Optimized Social Behavior Model (AOSBM) |
| 5 | 7.65 | 10.55 |
| 10 | 9.15 | 12.3 |
| 15 | 10.85 | 13.9 |
| 20 | 12.41 | 15.35 |
| 25 | 12.3 | 15.7 |

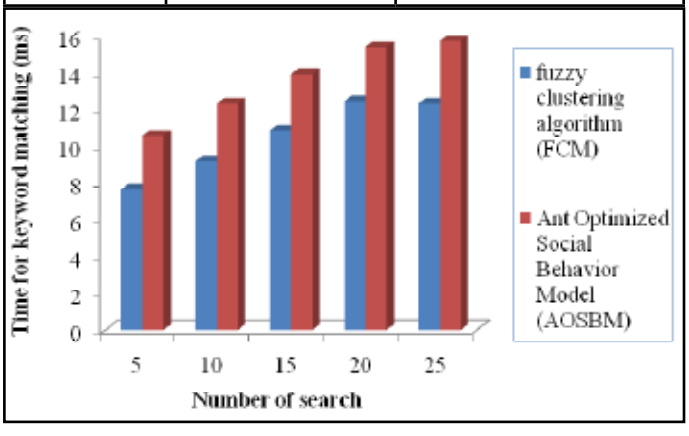


Fig. 4.2 : Measure of Time for Keyword Matching

Figure 4.2 shows Time for Keyword Matching. X axis represents the number of searches whereas Y axis denotes time for keyword matching using both the existing fuzzy clustering algorithm (FCM) and Ant Optimized Social Behavior Model (AOSBM). When the number of searches increased, time for keyword matching gets increases accordingly. Figure 4.2 shows better performance of Ant Optimized Social Behavior Model (AOSBM) in terms of time for keyword matching than existing fuzzy clustering algorithm (FCM). Ant Optimized Social Behavior Model (AOSBM) reduces the time for keyword matching by 30% when compared to existing system.

### 3. Document collection efficiency

Document collection efficiency is measured based on the number of search used for collecting the documents.

Table 4.3 : Tabulation for Document efficiency

| Number of search | Document collection efficiency (%) | |
|---|---|---|
| | fuzzy clustering algorithm (FCM) | Ant Optimized Social Behavior Model (AOSBM) |
| 5 | 60.48 | 70.35 |
| 10 | 66.45 | 74.41 |
| 15 | 68.52 | 75.39 |
| 20 | 66.27 | 73.28 |
| 25 | 69.92 | 76.97 |

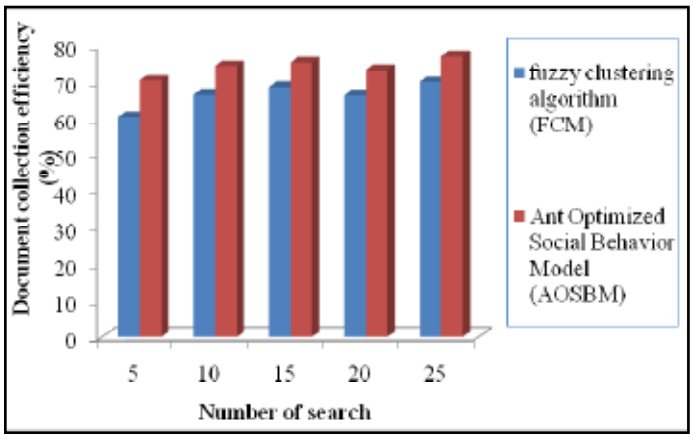


Fig. 4.3 : Measure of Document Collection Efficiency

Figure 4.3 shows Document collection efficiency. X axis represents the number of searches whereas Y axis denotes document collection efficiency using both the existing fuzzy clustering algorithm (FCM) and Ant Optimized Social Behavior Model (AOSBM). When the number of searches increased, document collection efficiency gets increases accordingly. Figure 4.3 shows better performance of Ant Optimized Social Behavior Model (AOSBM) in terms of document collection efficiency than existing fuzzy clustering algorithm (FCM). Ant Optimized Social Behavior Model (AOSBM) increases the document collection efficiency by 12% when compared to existing system.

### V. Conculsion and Future Work

The proposed system is Web document keyword matching method based on Ant optimized social behavior model for preserving high correlation on keyword matching. It is simultaneously maximizes the

correlation between the documents in the local patches and minimizes the correlation between the documents outside these patches. The calculation of exact correlation between documents is based on the accuracy of sentences. The proposed Ant optimized social behavior model is employed to enhances the Web document keyword matching and thus reduces the searching time in data mining technique. In addition, the evaluated Web documents are used to extract the optimal matching document in an efficient manner. This in turns Ant optimized social behavior model reduces the time for keyword matching in Web documents.

## Reference

[1] *Maryam Habibi and Andrei Popescu-Belis, "Keyword Extraction and Clustering for Document Recommendation in Conversations", IEEE/ACM Transactions on Audio, Speech, and Language Processing, Volume 23, Issue 4, April 2015, Pages 746-759.*

[2] *K. R. Uthayan and G. S. Anandha Mala, "Hybrid Ontology for Semantic Information Retrieval Model using Keyword Matching Indexing System", Hindawi Publishing Corporation, The Scientific World Journal, Volume 2015, October 2014, Pages 1-9.*

[3] *Francesco Camastra, Angelo Ciaramella, Alessio Placitelli, and Antonino Staiano, "Machine Learning-Based Web Documents Categorization by Semantic Graphs", Springer, Volume 37, 2015, Pages 75-82.*

[4] *Themistoklis Mavridis, Andreas L. Symeonidis, "Semantic analysis of web documents for the generation of optimal content", Elsevier, Engineering Applications of Artificial Intelligence, Volume 35, October 2014, Pages 114–130.*

[5] *Jianxin Li, Chengfei Liu and Jeffrey Xu Yu, "Context-Based Diversification for Keyword Queries Over XML Data", IEEE Transactions on Knowledge and Data Engineering, Volume 27, Issue 3, March 2015, Pages 1-14*

[6] *Sumayia Al-Anazi, Hind AlMahmoud, Isra Al-Turaiki, "Finding similar documents using different clustering techniques", Procedia Computer Science, Volume 82, 2016, Pages 28 – 34.*

[7] *R. Malathi Ravindran and Antony Selvadoss Thanamani, "K-Means Document Clustering using Vector Space Model", Bonfring International Journal of Data Mining, Volume 5, Issue 2, July 2015, Pages 10-14.*

[8] *Thasleena N.T, Varghese S.C, "Enhanced Associative Classification of XML Documents Supported by Semantic Concepts", Elsevier, Procedia Computer Science, Volume 46, 2015, Pages 194 – 201.*

[9] *Joe Tekli, Richard Chbeir, Agma J.M. Traina, Caetano Traina Jr., Renato Fileto, "Approximate XML structure validation based on document–grammar tree similarity", Information Sciences, Elsevier, Volume 295, 2015, Pages 258–302.*

[10] *Irwan Bastian, Rozaliyana and Metty Mustikasari, "Web Document Clustering System Using K – Means Algorithm", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 6, Issue 8, August 2016, Pages 181-186.*

[11] *Syed Fawad Hussain, Asif Suryani, "On retrieving intelligently plagiarized documents using semantic similarity", Elsevier, Engineering Applications of Artificial Intelligence, Volume 45, October 2015, Pages 246–258.*

[12] *Rouzbeh Meymandpour, Joseph G. Davis, "A semantic similarity measure for linked data: An information content-based approach", Elsevier, Knowledge-Based Systems, Volume 109, October 2016, Pages 276–293.*

[13] *Rouzbeh Meymandpour and Joseph G. Davis "A Semantic Similarity Measure for Linked Data: Information Content-Based Approach", Knowledge-Based Systems, Elsevier, Volume 109, 1 October 2016, Pages 276–293.*

[14] *Yinglong Ma, Yao Wang, Beihong Jin, "A three-phase approach to document clustering based on topic significance degree", Expert Systems with Applications, Elsevier, Volume 41, 2014, Pages 8203–8210.*

[15] *Lin Yue, Wanli Zuo, Tao Peng, Ying Wang, Xuming Hand, "A fuzzy document clustering approach based on domain-specified ontology", Data & Knowledge Engineering, Elsevier, Volume 100, November 2015, Pages 148–166.*