Performance Enhancement of Query Based Information Extraction Using Firefly Optimization Clustering Technique

PriyankaDhiman[1], A.J. Singh[2]
[12]*Department of Computer Science, Himachal Pradesh University*

Abstract- Internet is a resource of huge information which is growing with a high speed as a large amount of data added to the world wide web. It is very difficult task to search the useful information from the web because of excessive information. It has also been found that XML tags plays effective role in information retrieval. The data on web can be organized by using clustering method. Clustering is the process in which data is divided into groups on the basis of their similarity. In this paper K-means clustering algorithm and Firefly algorithm is used where K-means algorithm makes effective clusters of the web dataBut it has some limitations such as random selection of initial centroid and its results in local optimal. Whereas Firefly algorithm is an optimization algorithm that utilizing lower boundary intensity attract towards outer boundary intensity.Through thisFirefly optimization algorithm, optimal solutionis opportunity to find the initial cluster centriod for K-Means clustering.The input is given in the proposed model is in the form of XML files and collected from the search engines like Google and Yahoo. The data is related to the universities and medical field.The result is generalized on the basis of precision, recall and f-measure. The outcomes of the metrics are effective and improved.

Keyword- clustering, optimization, query, K-means, firefly algorithm

I. INTRODUCTION

Nowadays, most of the people extract the information and gain knowledge by navigating different websites on World Wide Web. The WWW helps people to interact with each other for communication and business purpose. As the field is growing rapidly the difficulties are also increased in interaction with web content. Web data clustering is a processing in which web data is divided into groups according to their similarity and these groups are called clusters. The web data with dissimilarity has different group. The main goal is to organize the web data which provides the effective data availability and accessing. The clustering improves the data accessibility for the users, provides content on time, and improves the content delivery on web [1].In classical methods clustering is based on the web logs in which clusters are made on the basis of navigation history of the users. This clustering is based on the activities performed by the user when it enters and leaves the web. This type of clustering is mainly based on the heuristic approaches in which IP session in and out is considered [2].

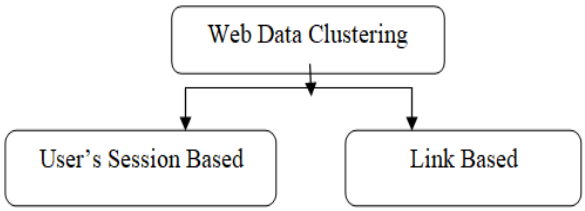

Fig.1: Web Data Clustering:[3]

Query extraction is a process in which user query is processed for the effective results from the clusters. The clusters are the collection of similar data stored on the network and used for effective data results. The data extraction is based on the keywords and query words. The words are used to analyze the clusters for pattern recognition, data analysis and image processing. Clustering of data improves the scalability and high dimensionality of data.Web data clustering process basically based on the user's session based and link based. In users session based clustering, clusters are made on the basis of user interest during the active session. In link based clustering, , user opens the link according to the interest and their requirementwhere data clusters are based on the link visit by the user this is mainly used in the e-commerce sector. According to the link based clustering other related links are also displayed for the user for better selection.

II. RELATED STUDY?

Alswaittiet. al. (2018) proposed the density based technique of clustering by using particle swarm optimization. This technique based on the kernel density and it also includes the gravitational learning factors. This work is mainly done to balance the exploration and exploitation [1]. Zhang, Qingchen, et al.(2017) presented the possibilistic c-mean algorithm for clustering of big data on cloud computing. This technique is used to optimize the objective function of tensor space. This approach also preserves the privacy and map reduces the huge amount of heterogeneous data. [2]. Miranda et al. (2017) proposed a clustering algorithm using the spatial contiguity to solve the problem regionalization problem. This clustering is a semi-supervised clustering which is done by using the k-mean algorithm and it called RegK-mean because it is applicable on regions [3]. Sheng, Weiguo, et al. presented the adaptive multi sub-population competition for effective search effectively over solution subspace in the different

number of cluster. In this network, individuals migrated between the subpopulation according to performance. The diverse search is promoted by using adaptive multisite crowding technique. It worked more effectively because value of parameters adjusted dynamically. [4]. Qian, Yuhua, et al. (2016) presented a data representation technique for categorical data in which it maps the data into categorical objects in Euclidean space. By using this framework Categorical clustering algorithm is designed which provides the effective clustering of the data as compared to existing approaches [5].

Suresh. K. et. al. (2016) proposed improved fuzzy c-mean clustering algorithm for clustering in the web usage mining. It is basically a process of web log data repositories. Repositories are used to find the patterns of user's access from web access log. [6].Pedrycz et al. (2011) also used the fuzzy c-mean algorithm for clustering but it enhanced the work by using proximity feature with algorithm. This algorithm helped in search of patterns and gives a rise in 2-phase Optimization Process. In this process FCM is implemented in the first phase and in second phase minimization of proximity values by using gradient [7]. Clustering of web results is also based on the meta-heuristic algorithm and proposed by Cobos et al.(2014) in which cuckoo search algorithm is used with k-mean and Bayesian algorithm. These algorithms are mainly based on the split and merge approach on clusters. This approach gives effective search results between the clusters because cuckoo algorithm has ability to search locally and globally in the clusters. The fitness of cluster is determined automatically by using Bayesian function [8]. Wang et al. (2014) proposed link based clustering of web results. This work based on the coupling and co-citation analysis of the links shared by the pages. The k-means clustering algorithm is used to improve the results by reducing the noise. This approach filtered the effective result by using the accessed and browsed pages by users [9]. Osińskiet. al. (2001)proposed the lingo algorithm for clustering search results in cluster based on their quality. In this algorithm algebraictransformation of the term document matrix is performed and suffix arrays are used for phrase extraction frequently. The concept is mainly based on the similarity between the documents [10].Liu, Wei et al. (2010) proposed a vision based approach for web data extraction and it is language independent. This approach is mainly used for data item and data record extraction. This approach gives highly accurate results in data and record extraction [11].Biancoet al. (2018) in the proposed approach, after applying clustering user sessions was identified without using priori of threshold values. The user session identification is based on the statistical properties of identified sessions. This approach is also helpful in other traffic properties also [12].

III. PROPOSED METHODOLOGY

This section of the paper describes the proposed methodology of the work by using the flow chart and the algorithms used for the optimization and clustering. The optimization is performed by using the firefly algorithm and clustering is performed by using the K-means clustering algorithm.

a. Firefly Algorithm: The Firefly algorithm was introduced by XIN SHE YANG in 2008.Firefly is a bio-inspired algorithm which is used for the optimization process. It is a meta-heuristic algorithm and supports the global optimization of the solutions and provides an optimal solution related to the problem. The working process is based on the flashing behavior of flies to attract the other flies. The brightness of the fly is directly proportional to the attractiveness becauseif the distance is more than brightness is low, and it attract fewer amounts of flies. If brightness is high then it attract a large amount of flies and gives effective solutions. The main feature of firefly algorithm is multimodality, and automatic subdivision.General steps firefly algorithm discussed below.

Step1 : Firstly initialize the objective function.

Step2 : Generate the initial firefly population.

Step3 : Determine the intensity of light.

Step4 : Calculate the attractiveness of flies.

Step5 : Update the intensities of light and rank the flies on the basis of it and find the current best.

b. ***K-means Algorithm:*** K-means clustering algorithm is used in this approach to make effective clusters of the web data.. It is basically unsupervised algorithm and it worked well on the unlabelled data. The main goal of this algorithm is to find the similar data and make the groups according to their features. The data with same feature is stored in single cluster. General steps of k-means algorithm discussed below.

{

Initialization: Initialize the K Centroids of cluster.

Do

Assign each data points to its closest point.

Re-compute the Centroids of cluster

}

Algorithm

1. Input database of XML files.
2. Extract the TF_IDF features

$$\boldsymbol{tf(t,d)} = \frac{\boldsymbol{f_d(t)}}{\min_{\omega \epsilon d} \boldsymbol{f_d}\,(\omega)}$$

$$\boldsymbol{idf(t,d)} = \boldsymbol{ln}\left(\frac{|\boldsymbol{D}|}{|\{\boldsymbol{d \epsilon D} : \boldsymbol{t \epsilon d}\}|}\right)$$

$$\boldsymbol{tfidf(t,d,D)} = \boldsymbol{tf(t,d)} \cdot \boldsymbol{idf}\,(\boldsymbol{t},\boldsymbol{D})$$

$$\boldsymbol{f_d(t)} := \boldsymbol{frequency of term t in document d}$$

$$\boldsymbol{D} := \boldsymbol{Corpus of documments}$$

3. Define K
4. Initialize initial population of flies.

Parameters	values
Max Generation	100
Test Number	100
Randomness	0.2
Randomness reduction	0.98
Population	50
A	[-10,10]
B	[0,3]
C	[-100, 100]
Replaceable Population	10
Absorption Coefficient	1

Check the fitness of fireflies using fitness function

$$x_t(t+1) = x_i(t) + \beta_0 e^{-\gamma r^2}(x_i - x_3) + \alpha \in_i$$

$\beta_0 e^{-\gamma r^2} \leftarrow centroid\ of\ k-mean\ distribution$
$x_t(t+1) \leftarrow att+1\ time\ centroid$
$x_i(t) \leftarrow at\ t\ time\ centroid$
$\alpha \leftarrow Starting\ T\square res\square old$
$\in_i \leftarrow Centroid\ value\ at\ star$

Update according to the intensity of light
Rank the fireflies and update the position.
Check the result is optimal or not

5. Apply Clustering using K-means clustering algorithm.
6. Analyze the precision, recall, and accuracy.

Flow Chart of Proposed Methodology-

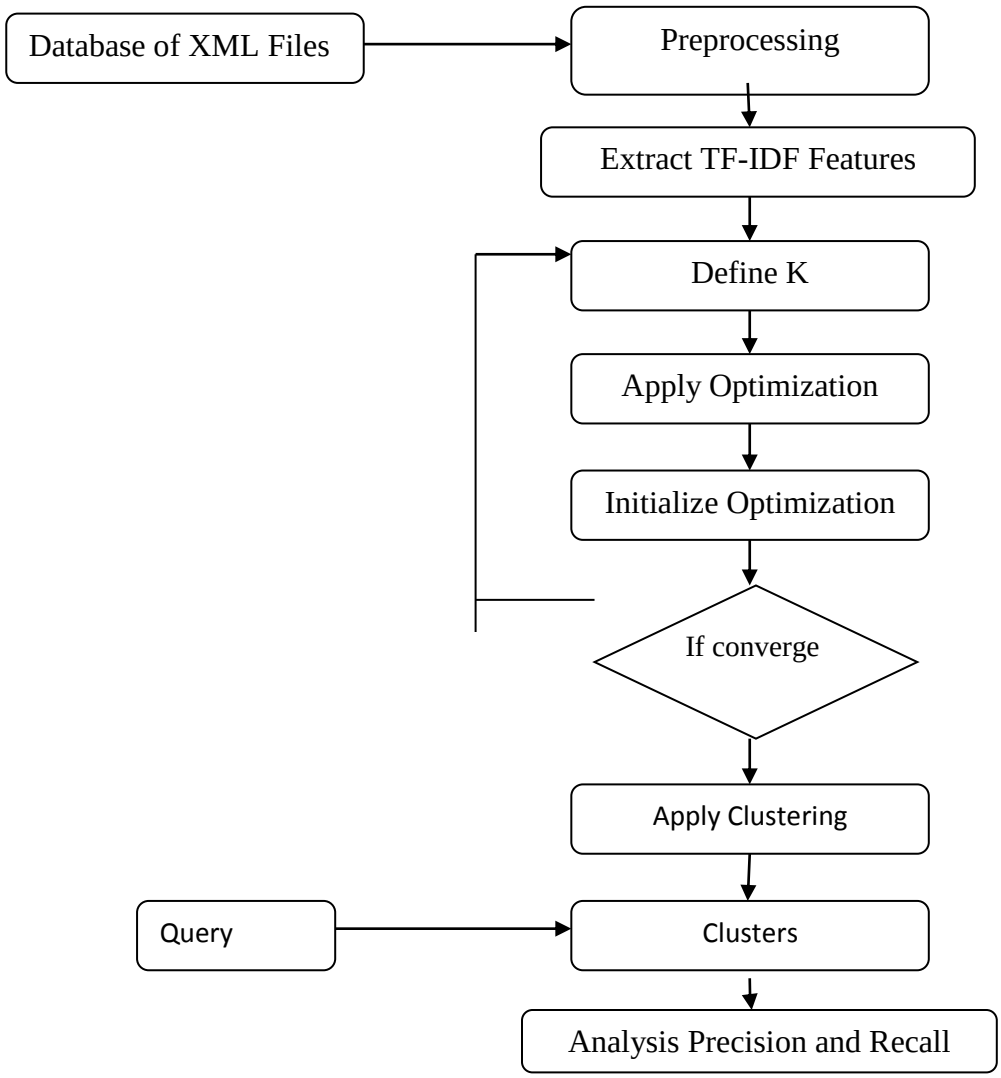

Fig.2:

General steps of Flow Chart-

Step1: Input the data base XML file.

Step 2: Preprocessing of XML file in which removal of duplicate data by using tokenization, stemming and stop word removal.

Step 3: Extract the TF-IDF features from the file and define the value of K.

Step 4: Apply the firefly optimization algorithm.

{

Firstly initialize the objective function.

Generate the initial firefly population.

Determine the intensity of light.

Calculate the attractiveness of flies.

Update the intensities of light and rank the flies on the basis of it and find the current best.

}

Step 5: Check the convergence of the output given by firefly algorithm.

Step 6: Apply the clustering and then query on the clusters.

Step 7: Analyze the precision, recall and accuracy.

IV. EXPERIMENTAL RESULTS

In theexperiment Performance Enhancement of Query based Information Extraction done on the basis of tool named Java Eclipse. Java is the high level language and interactive environment.Machine configured with core i5 processor, 4 GB RAM and Window-7 OS. Java provides support for web applications and also finds application in development of e-commerce web applications using open source platforms and also develops android applications. For Information extraction the content is taken from the search engines like Google and Yahoo in the form of XML files. These files are related to the university, and medical field. Some performance metrics are used for the query optimization are precision, recall and f-measure which are described as follows:

Precision: In the field of information retrieval, precision is the fraction of retrieved documents that are relevant to the query:In the Experiment Precision results shown in figure 1.3

$$Precision = \frac{|\{Relevent\ Document\} \cap \{Retrieved\ Document\}|}{Retrieved\ Document}$$

Recall: In information retrieval, recall is the fraction of the relevant documents that are successfully retrieved. In the Experiment Recall results shown in figure 1.4

$$Recall = \frac{|\{Relevant\ Document\} \cap \{Retrieved\ Document\}|}{Relevent\ Docment}$$

F-Measure:A measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure or balanced F-score: In the Experiment F-Measure results shown in figure 1.4

$$\mathrm{F-Measure} = 2 * \frac{Precision.Recall}{Precision+Recall}$$

Table I shows the computational results of K-Means and Firefly algorithm:

Query	Precision(K-mean)	Recall(K-mean)	F-measure(k-mean)	Precision(K-firefly)	F-measure(K-firefly)	Recall(K-firefly)
query1	45.6	48.34333333	46.95163469	78.34	72.29150296	66.71
query2	56.2	47.18333333	51.49469228	65.34	65.02423702	64.71
query3	43.23	42.86	43.04460245	56.45	61.3659474	66.71
query4	42.12	46.52666667	44.2685351	72.34	71.83826	71.34
query5	43.23	46.23	44.70484202	71.34	70.67018891	70.00666667
query6	54.23	46.45	50.18947599	70.34	69.27527505	68.22666667
query7	41.23	42.55	41.88480035	68.34	68.22657254	68.11333333
query8	43.89	46.25333333	45.05617383	66	65.72275101	65.44666667
query9	42.53	49.10666667	45.70018089	70	69.77965319	69.56
query10	52.34	50.13	51.22308269	60.34	63.58602257	67.00666667
query11	52.45	49.025	50.70859148	78.34	74.23230833	70.34
query12	45.6	45.6	45.6	62.34	62.34	62.34

TABLE1:COMPUTATIONAL RESULTS

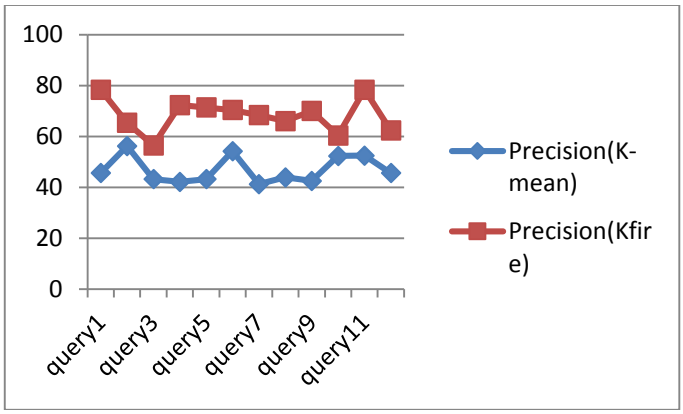

Fig.3: Precision on proposed and existing approach.

The above given figure 1.3 depicts the precision of the k-mean and k-firefly algorithm on the different queries. The red curve represents the precision of k-firefly and blue line depicts the precision with k-means. The precision of proposed approach k-means with firefly is better than existing approach.

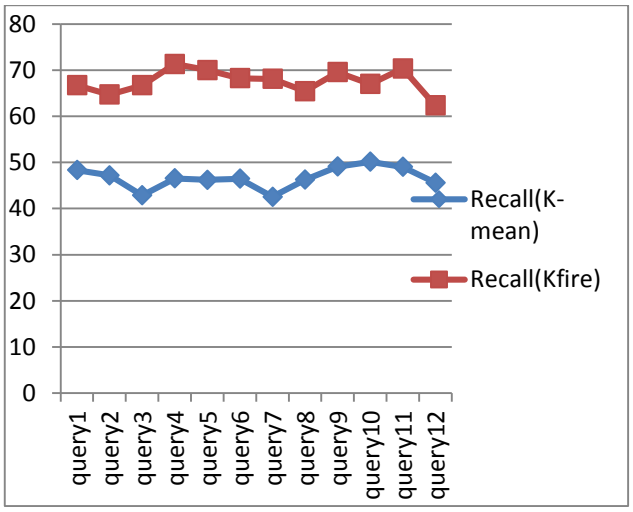

Fig.4: Recall of proposed and existing approach.

The above given figure 1.4 depicts the recall of the k-means and k-firefly algorithm on the different queries. The red curve represents the recall of k-firefly and blue line depicts the recall with k-means. The recall of proposed approach k-means with firefly is better than existing approach.

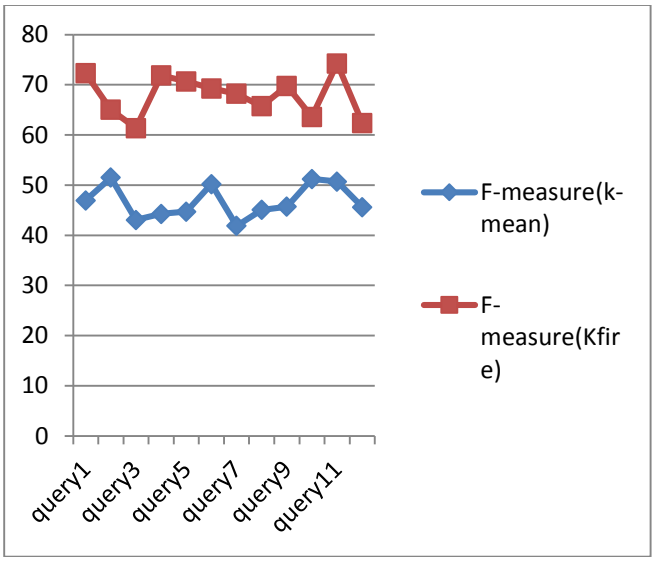

Fig.5: F-measure of proposed and existing approach.

The above given figure 1.5 depicts the F-measure of the k-means and k-firefly algorithm on the different queries. The red curve represents the F-measure of k-firefly and blue line depicts the F-measure with k-means. The F-measure of proposed approach k-means with firefly is better than existing approach.

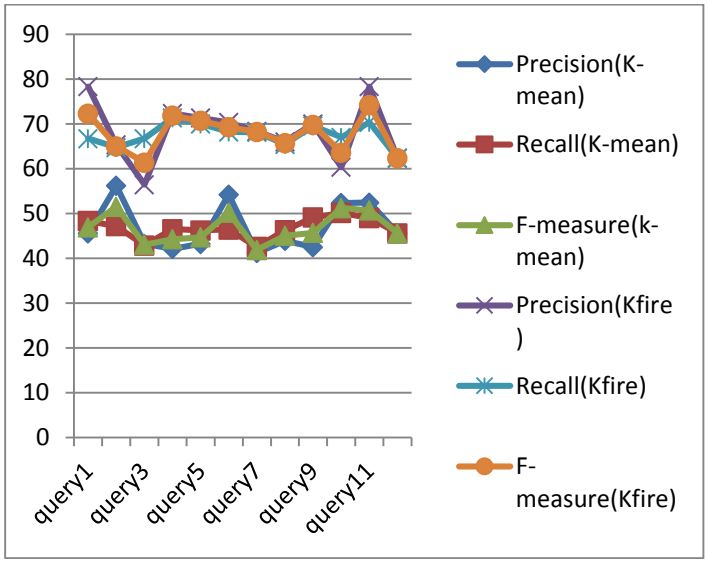

Fig.6: Comparison of Precision, Recall, and F-measure of proposed and existing approach.

The above given figures 1.3-1.6 represented the precision, recall, and f-measure on 12queries. The graph curves show that the results by using firefly algorithm with k-means are more effective than k-means alone. The firefly algorithm gives

effective results because it is capable in global optimization and provides more optimal results.

V. CONCLUSION

Clustering of web data is an effective approach of information retrieval. Documents in the clusters are similar or Dis-similar to the other cluster. This work is based on the query retrieval from the clusters and feedback on the database samples.In the proposed work clustering is done by weighted featuresusing k-means clustering algorithm andoptimization by firefly algorithm. In this work features are used in two differentways that are weighted features and without weighted features.The result of the firefly algorithm is compared with the existing k-means algorithm and the novel algorithm performs better with high precision, recall, and f-measure. The combination of two technologies machine learning and data management makes the results more effective and efficient by providing good prediction with high accuracy.

VI. REFERENCES

[1]. Alswaitti, Mohammed, MohanadAlbughdadi, and NorAshidi Mat Isa. "Density-based particle swarm optimization algorithm for data clustering." *Expert Systems with Applications* 91 (2018): 170-186.

[2]. Zhang, Qingchen, et al. "PPHOPCM: Privacy-preserving high-order possibilistic c-means algorithm for big data clustering with cloud computing." *IEEE Transactions on Big Data* (2017).

[3]. Miranda, Leandro, José ViterboFilho, and Flavia Cristina Bernardini. "RegK-Means: A clustering algorithm using spatial contiguity constraints for regionalization problems." *Intelligent Systems (BRACIS), 2017 Brazilian Conference on*. IEEE, 2017.

[4]. Sheng, Weiguo, et al. "Adaptive multisubpopulation competition and multiniche crowding-based memetic algorithm for automatic data clustering." *IEEE transactions on evolutionary computation* 20.6 (2016): 838-858.

[5]. Qian, Yuhua, et al. "Space structure and clustering of categorical data." *IEEE transactions on neural networks and learning systems* 27.10 (2016): 2047-2059.

[6]. Suresh, K., et al. "Improved FCM algorithm for clustering on web usage mining." *Computer and Management (CAMAN), 2011 International Conference on*. IEEE, 2011.

[7]. Pedrycz, Witold, Vincenzo Loia, and Sabrina Senatore. "P-FCM: a proximity—based fuzzy clustering." *Fuzzy Sets and Systems* 148.1 (2004): 21-41.

[8]. Cobos, Carlos, et al. "Clustering of web search results based on the cuckoo search algorithm and balanced Bayesian information criterion." *Information Sciences* 281 (2014): 248-264.

[9]. Wang, Yitong, and Masaru Kitsuregawa. "Link based clustering of web search results." *International Conference on Web-Age Information Management*. Springer, Berlin, Heidelberg, 2001.

[10]. Osiński, Stanisław, Jerzy Stefanowski, and Dawid Weiss. "Lingo: Search results clustering algorithm based on singular value decomposition." *Intelligent information processing and web mining*. Springer, Berlin, Heidelberg, 2004. 359-368.

[11]. Liu, Wei, XiaofengMeng, and WeiyiMeng. "Vide: A vision-based approach for deep web data extraction." *IEEE Transactions on Knowledge and Data Engineering* 22.3 (2010): 447-460.

[12]. Bianco, A., et al. "Web user session characterization via clustering techniques." *Global Telecommunications Conference, 2005. GLOBECOM'05. IEEE*. Vol. 2. IEEE, 2018.