Distributed Web Usage Mining Based Recommender System in Big Data Analytics Using Firefly Algorithm

AC. Priya Ranjani[1], Dr.M.Sridhar[2]

[1] Research Scholar, Dept. Of Computer Science, AcharyaNagarjunaUniversity, Guntur, Andhra Pradesh, India.
[2] Associate Professor, Dept. Of Compute Applications, RVR&JC CE, Guntur, Andhra Pradesh, India.
(E-mail:acpranjani@gmail.com)

Abstract--Big Data Analytics is used extensively for analysing data generated dynamically from various sources and employ it in the delivery of solutions that are useful to organizations. Personalizing of web pages in e-commerce sites that make use of the web links is now prevalent among recommender systems. In this work, a recommender system based on mining of web usage data with collaborative filtering based on individual needs has been put forward. The work further proposes a novel framework that makes use of Firefly Optimization to build a recommendation system effectively by using K-Means clustering on the buyers with similar interests and is followed by identifying the associations among pages and clusters. The experiments were carried out by employing web links and the contents of the simulated e-commerce websites with the aid of Hadoop framework. The proposed system showed better results.

Keywords--*Big Data Analytics, Recommender System, Firefly Optimization, K Means Clustering.*

I. INTRODUCTION

Big Data [1-3] has been gaining plenty of attention from the IT industry and academia. In the world of digital computing, information is normally generated and also collected which was at a range which had exceeded the range of the boundary. Now, more than 2 billion people all the world over have been connected by the Internet and more than 5 billion individuals have mobile phones. By the year 2020, over 50 billion devices will be connected to the Internet. When this happens, the data production predicted will be more than 44 times than it was in the year 2009. As the information keeps getting shared based on the wireless networks and the optic fibre, the data volume and market growth speed keep on increasing. But the rate of growth of the large data will generate various challenges like rapid data growth, security, diverse data and speed of transfer.

Big Data has been characterized by means of a large size of the sample and a high dimensionality. Both these features tend to raise three different unique challenges which are: noise accumulation, incidental homogeneity and spurious correlations. A high dimensionality which is combined with a sample size that is large will create issues like cost of computation along with algorithmic instability. There is a massive sample found in Big Data and this is typically aggregated from various sources at various points of times that make use of various technologies. This results in statistical biases, experimental variations and heterogeneity and this needs developing more robust and adaptive processes [4].

The growth of the World Wide Web (WWW), has made it crucial to identify information that is useful from the amount of data which may be voluminous. The Web consists of dynamic and rich collections of information on the hyperlink, access to webpage and information on usage providing valuable data mining source information. The web further poses a great number of challenges in knowledge discovery and applications of data mining to webpage and information on usage providing valuable data mining source information. The web further poses a great number of challenges in knowledge discovery and applications of data mining. Web Mining is a set of techniques of data mining that can automatically extract and discover certain useful knowledge found in different forms like video, audio, images and web documents [5]. Web mining is categorized into three types on the basis of knowledge extraction: The Web Content Mining, the Web Structure Mining, and Web Usage Mining.

Web Content Mining: this is referred to as mining, extraction and finally integrating useful data, knowledge and information from the web page contents or the web resources like the HTML - XML documents, query responses to the database and digital libraries. All of these resources have been related to the traditional techniques of information retrieval [6].

Web Structure Mining: also known as structured data mining, denotes the process of discovery of structural information derived from the link structure topology among hyperlinks or web documents. This structural mining may be performed either at a hyperlink or document level. The Graph theory is employed in mining for analysing node and structure of connection of websites [7].

Web Usage Mining: this denotes data mining technique applied to extract and discovers the interesting patterns of usage from the weblogs or web data for understanding and serving the needs of the web-based applications. The patterns of usage also include the understanding and the serving of the usage patterns and their need that include data of server such

as the IP address, the data of application servers such as the Web logic and the application level data such as events.

The primary goal behind a Web Usage Mining (WUM) is the understanding of website users and their behaviour by means of data mining process of web access information. The knowledge that is obtained from the web usage mining is used for enhancing website design and introduce a personalization service to facilitate better browsing [5]. Web Data Mining focuses on gaining business information and in order to make right predictions or decisions for the development of business, the process of production and workflow, the models of business from web data mining are as lots that are available on the World Wide Web. It also identifies some rules and characteristics belonging to users and their visiting behaviour for improving the user service quality. There is also a user profile which is used for classifying the user within a predefined user segment (either by taste or demographics) for capturing the user behaviour online and this may also include the preferences and interests of the user. Scientifically speaking, cloud computing may be the alternate word for a distributed computing which is over a wireless or wired network. this depends on a shared resource for accomplishing scale of economies and coherence to be the same for using the resource in a network. This means its capacity to either run a program or an application on the cluster of various computers at the same time.

The recommendation systems have been employed by e-commerce organizations for suggesting various products to customers. These products are recommended on the basis of the top sellers for a particular site, past preferences of customers and the customer demographics. These recommendation systems may generate them for the users by means of exploring their likes and needs. They further generate various recommendations that suit various users giving a web interface that is customized to users. This way, the web has been personalized for every user with a recommendation system. In the present situation, many e-commerce organizations rely on recommendation systems to offer better services to their customers.

The web recommendation systems [8] being an important tool is used for analysing user behaviour. It further supports the organizations to ensure intelligent decision making in relation to customers and their needs by means of automating their recommendations based on preferences. Thus this may work as a system of decision support to organizations.

Hadoop was a flexible infrastructure most suitable for large-scale computations along with data processing on commodity hardware network. This permits the applications to be able to work using many thousands of independent computers embedded with petabytes of data. The primary principles of Hadoop is to move computations based on data as opposed to moving the data for computation. It was used for breaking down large volumes of input data into smaller chunks and are processed separately on various machines. The basic building blocks of Hadoop architecture are Hadoop distributed File System(HDFS) and Map Reduce. HDFS [9] is a block-structured file system where each file is divided into blocks of specific size. It further adopts a Master/Slave architecture, in which there will be a single NameNode (master) and multiple DataNodes (slaves). HDFS is deployed based on a broad spectrum of the machines supporting Java. Usually the DataNodes will be spread across several machines to process large datasets.

In order to achieve parallel execution, Hadoop implements a programming model called Map Reduce [10] which actually comprises of two different tasks, the map task and the reduce task. Large data sets in the form of key-value pairs are fed as input to map task which are then processed and intermediate key-value pairs will be generated as output. The reduce phase accepts these non-persistent data , sorts and shuffles between different reduce phases and eventually the aggregated key-value pairs will eject as output. The input and output data will be stored on HDFS. Figure 1 describes the basic functioning of a Map Reduce component. The single Job Tracker running on Master will share the work among multiple Task Trackers running on slaves.

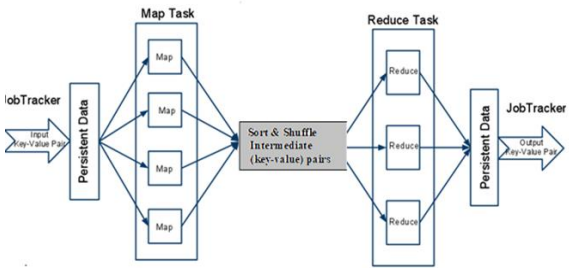

Fig. 1 Architecture of Map Reduce in Hadoop

Nature is an inspiration for introducing various meta heuristic algorithms and this has been able to find solutions to various problems by means of experience. A natural selection along with the survival of the fittest is the primary motivation to this. Animals tend to communicate with one another by means of various modes of communication. The fireflies use their property of flashing for purposes of communication. There are about 2000 species of fireflies that have distinct patterns of flashing. They normally tend to produce a flash that is short having a particular pattern. There is a light which is then produced by means of a biochemical process known as bioluminescence. A flashing communication has been used for attracting their mates and for warning predators. On the basis of the light pattern, there will be a suitable mate that communicates back by means of mimicking a similar pattern or by responding with a particular pattern. There is also a need to note the fact that the intensity of light can decrease through the distance. Thus, the flashing light which emanates from the firefly will get a response from the fireflies that are around it inside the same visual range of its flash.

For this work, there was a recommender system which was based on a distributed web-usage mining that had been proposed by using a firefly algorithm. The rest of the work has been organized as follows. Section 2 explains briefly all literature connected to this work. The details of all techniques used are explained in section 3. The results are elaborated in section 4 and section 5 has concluded the work.

II. LITERATURE SURVEY

An analysis of the regularities of websites and patterns used in navigation has been gaining plenty of attention from the community of business and researchers where web

browsing has become an everyday task for people all over the world. This is a data which exists in large scale known as Big Data and characterised by complexity, huge quantity, high costs of processing and requires computational intelligence, cloud computing, web-based computing and cognitive informatics. The actual size of the data collected from the web and the users of mobile devices is higher. In order to provide the capacity of making sense and to maximize the usage of these vast amounts of knowledge discovery, web data along with decision making is critical for scientific advancement. New tools were proposed for data mining of Big Data by Mohata and Dhande [11]. A tool which is effective for getting insight into Big Data is known as Visualization. The Apache Hadoop along with other technologies have been emerging for supporting concerns in the back-end and also in the front end to help businesses to explore data in an easier manner and understand them better.

Chen et al [12] had reviewed the state-of-the-art and the background of Big Data. Firstly, the background of Big Data was introduced and then reviewed along with the related technologies. These may be the Hadoop, the data centres, the Internet of Things and Cloud computing. The focus then shifts to four different phases in the Big Data value chain which are data generation, acquisition, data storage and analysis. Lastly, various representative applications in Big Data are analysed and this includes smart grid, enterprise management, medial applications, online social networks and the Internet of Things. All these aim at providing an overview that is comprehensive along with the big picture to the readers. The survey will be then concluded by discussing all open problems along with future directions.

Barot et al [13] had proposed another new approach to web page recommendation with the generation of the user profile. The approach further employs an evolutionary technique of bi-clustering for the purpose of web page recommendation. There are different datasets proposed to perform operations of data mining and one among them is the clickstream data whereas the other is a web access log file belonging to Kadi Sarva Vidhyalaya (KSV) University. The approach has an outcome of using the bi-clusters along with techniques of evolutionary bi-clustering based techniques that are analysed in the paper. The results are extremely useful in taking ideal decisions that are strategic in applications such as direct and target marketing.

In the last few decades, there have been several studies made in recommender systems focusing on the designing of the new recommender algorithms on the basis of computational intelligence based algorithms. These recommender systems are successful and aside from the algorithm's quality, which is dependent on various factors such as the sparseness, Cold start and the expandability used in the performance of various recommender systems affecting their quality. Khorsand and Mortazavi [14] had the primary motivation of providing an effective algorithm which was also a meta heuristic and was based on the firefly algorithms and imperialist competitive algorithms. The results of simulation of this algorithm proposed on datasets such as Film Trust and Move Lens have been showing a better accuracy in terms of item recommendation to the users aside from the other algorithms that are presented in the case of subject literature. Furthermore, this proposed algorithm will be able to choose some suitable items in the wide data range for giving to the output inside a time which is reasonable.

Sharma [15] had made a proposal for another novel architecture for integrating product information and the access log data of users to further generate another new set of suggestions for a particular user. His model used various data mining methods such as clustering and pattern matching. In its earlier work, the K-means clustering along with other matching algorithms had been employed. This could be the Boyer Moore Pattern Matching Algorithm. In order to further elevate this work, there had been another algorithm used in a soft computing algorithm which was based on Artificial Intelligence. This was able to bring about an improvement using a new mobile database created on the basis of some technical specifications. This had created some more new clusters that were based on the similarity concept and also on the concept of a random population with some more of these specifications. A product that had the minimum probability was thus kept above and was sorted in a manner which was top to bottom.

III. METHODOLOGY

The section below provides details on the mining usage data for web personalization, K-Means clustering, Firefly Algorithm (FA) with K-Means clustering and the Hadoop Map Reduce.

A. Mining usage data for web personalization

Normally, the web personalization systems that are based on usage involve three phases: data preparation and transformation, pattern identification, and recommendations.

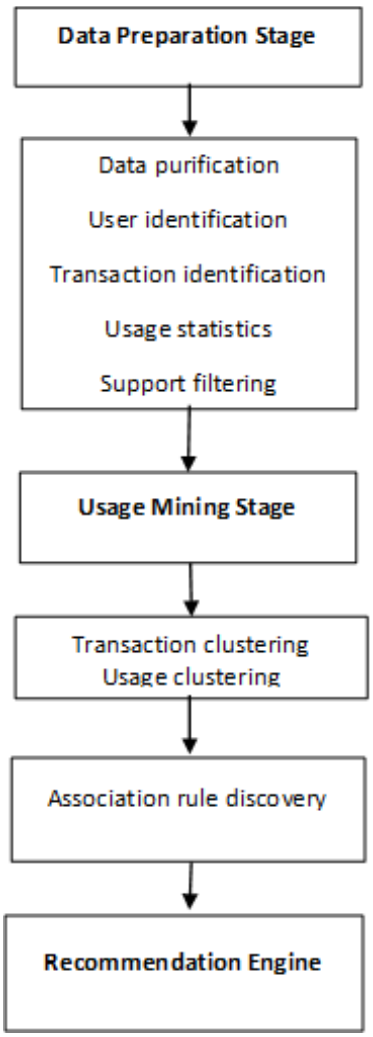

Fig. 2 General architecture for usage-based web personalization

Figure 2 shows the architecture for usage-based personalization of web. Among these, the last one is a real-time component, while the other two phases are conducted offline. The pattern discovery phase may include the detection of associations, patterns of sequential navigations , grouping of users or sessions, and grouping of page views or products. A recommendation engine takes into account the active user sessions and the uncovered patterns to furnish personalized content. This personalized content may be in the form of the products, links or advertisements targeted for the preferences of users which is determined by means of matching the patterns of usage [16]. Preprocessing of raw data is an essential stage for successful personalization of web pages [17-18] . The variant tasks involved in this stage are data cleaning, filling out the missing values, identification of page views, session, user and so on. The final step in preprocessing is transaction identification followed by pattern discovery for focusing on all relevant subsets of the page views in every session.

B. K-Means clustering algorithm

Clustering is a technique which groups the complete data set into clusters such that data objects in each cluster will have some similarity. K-means is one such well known and most used clustering method. Usually K-means is used to cluster very large set of data. In K-Means algorithm, it calculates the distance between each of data object and cluster centre in every iteration. As K-means is a partitioning and iterative method, the clusters formed will be independent and compact. There will be random selection of 'k' centres in the first iteration and obtain clusters by taking data objects which is nearest to the cluster centres. In the subsequent iterations cluster centres are updated based on the nearest data objects. Generally Euclidean distance is considered to find out distance between each data object and cluster centre. So in this way similar data objects are brought together to form clusters [19].

The K- means algorithm is summarized as follows:

Input: The session weight matrix which is obtained in, pre-processing stage, Number of clusters to be formed, cluster centre initialization and distance matrix.

Output: various clusters with closest centroid.

The K-means clustering algorithm for Web usage data comprises the following four steps.

(1) Choose 'k' initial cluster centres (representing k groups) randomly from the data set which includes around 2398 sessions formed in pre-processing steps.

(2) Assign all sessions to their closest cluster (measuring from the cluster centre). This is done by presenting each data point x and calculate the similarity (distance) d of this input entries y of each cluster centre j. the closest cluster centre to a data point x is the cluster centre with minimum distance to the data point x.

(3) Recalculate the centre of each cluster as the centroid of all the sessions in each cluster.

(4) If the new centres are different from the previous ones, repeat steps 2, 3 and 4. Otherwise terminate the algorithm.

So K-means starts with initial partition of K clusters and assign patterns to these clusters in order to reduce or decrease the squared error.

C. Firefly algorithm with k-means clustering algorithm

The Firefly algorithm is yet another metaheuristic algorithm that is swarm based and was presented by Yang [20]. It mimics the manner in which the fireflies interact by making use of their flashing lights. It assumes that fireflies are all unisex and this means they may be attracted to any firefly irrespective of their sex. The firefly's attractiveness is assumed to be directly proportional to the brightness and this further depends on its objective function [20]. Fireflies tend

to be attracted to the one which is brighter and the brightness continues to decrease and the distance is based on the inverse square law as per equation (1)

$$I \propto 1/\gamma^2 \quad (1)$$

In case the light passes through a new medium having a coefficient of light absorption γ, the intensity in a distance r from its source is as per equation (2)

$$I = I_0 e^{-\gamma r^2} \quad (2)$$

Wherein, I_0 is the intensity of light at the source and brightness β shown in equation (3):

$$\beta = \beta_0 e^{-\gamma r^2} \quad (3)$$

There is a generalized function for brightness for that of $\omega \geq 1$as in [21]. In reality, any function that is monotonically decreasing may be employed.

$$\beta = \beta_0 e^{-\gamma r^{\omega}} \quad (4)$$

In this algorithm, there is a feasible solution which is generated randomly known as fireflies that are assigned along with an intensity of light that is based on the objective function performance. This is used to compute the firefly's brightness and for problems of minimization, a solution having the smallest functional value is assigned along with its highest intensity of light. As soon as this intensity or the brightness belonging to solutions has been assigned, every firefly follows the other fireflies using better intensity of light. In the case of firefly which is the brightest, it performs a new local search by means of moving it within the neighbourhood randomly. Thus, in the case of two fireflies, in case the firefly j is found to be brighter than the firefly i, the latter will move towards the former. γ now denotes the algorithm parameter determining the actual degree where the process of updating is dependent on. All these updates of the Firefly location will continue with the same iteration until such time a termination criterion has been met. This termination criterion may either be the maximum number of iterations and a tolerance from its optimum value in case it is found that there is no improvement which is achieved in the consecutive iterations.

Advantages of the Firefly Algorithm

- A firefly will be able to handle problems of multi-modal optimization that are non-linear in an efficient and natural manner.
- The firefly will not make use of the velocities and there may not be any problem connected to the PSO velocity.
- The convergence speed of FA is high in terms of the probability of identifying its globally optimized answer.
- It also has a flexibility of integration with that of the other techniques of optimization that form the hybrid tools.
- It has no need for a proper initial solution for starting the process of iteration.

The pseudo code of K-means algorithm combined FA is shown below

(1) Consider a population of fireflies X_i(i = 1,2,….,N_f)

(2) Choose k cluster centers and initialize each c_k(k = 1,2,….,N_c)

(3) Initialize an object o_n (n=1,2, . . . N_o)

(4) Do while t < MaxGeneration

(5) Assign each object to the closest cluster center

(6) Compute J_i (i = 1, 2, ... N_f)

(7) Light intensity I_i is determined by J_i

(8) If $I_i > J_i$ then

(9) Move firefly i towards firefly j

(10) End if

(11) Find the current best

(12) End while

The clustering algorithm combined K-means algorithm and the FA has been proposed in 2011 by Mr Senthilnath et al. In this algorithm, the position vector of firefly i X_i is, $(c_1, c_2, c_3, \ldots\ldots c_k)$ that is, each firefly has the places of all cluster centers. The attractiveness of each firefly is defined by $\alpha(t) = \alpha(0)(10^{-4}/0.9)^{t/t_{max}}$ objective function, where α (t) is the randomization parameter and t is the number of iteration). Numerical experiments have indicated that this algorithm is more efficient algorithm than K-means algorithm and other algorithms using well known typical benchmark data sets [22].

IV. RESULTS AND DISCUSSION

The Hadoop [23] has been found to be an open source which was distributed as a computing framework used in the process of large distributed datasets. Hadoop is widely used in the platforms of the cloud recently and is further adopted by many research institutions and Internet companies. The Hadoop cluster contains two different paths: first being the Hadoop Distributed File System (HDFS) and next the Map Reduce. The Hadoop Map Reduce is a new framework that writes applications processing large amounts of data (the

multi-terabyte data-sets) that are processed in parallel on the large clusters with thousands of nodes found in the commodity hardware in a fault-tolerant and reliable manner. The job of a Map Reduce normally splits an input dataset into certain independent chunks that have been processed using map tasks and reduce tasks in parallel. Ideally, the input and also the output part of the job get sorted within a file-system. This framework further takes care of the tasks of scheduling and monitors them aside from re-executing the tasks which have failed. Map reduce framework has been adopted for execution of the data sets inorder to test the proposed algorithm.

Table 1 shows the parameters of Firefly Algorithm. Tables 2 and 3 shows the precision and coverage with Recommendation Threshold @ Support = 0.06 for window size 2, 3 and 4.

Table 1 Parameters of firefly algorithm

Parameter	**Value**
Maximum Iteration Time	100
Firefly Population	30
Dimension of population	25
Firefly size	20
Alpha	0.25
Beta	0.8
Gamma	1
Lower Boundary	0
Upper Boundary	8

Table 2 and figure 2 shows that the precision for window size 4 performs better by 6.45% and by 4.26% at recommendation threshold 0.1 than window size 2 and 3 respectively. The precision for window size 4 performs better by 7.4% and by 2.89% at recommendation threshold 0.5 than window size 2 and 3 respectively. The precision for window size 4 performs better by 10.13% and by 7.5% at recommendation threshold 1 than window size 2 and 3 respectively.

Table 2. Precision

Recommendation Threshold @ Support = 0.06	Window size=2	Window size=3	Window size=4
0.1	0.45	0.46	0.48
0.2	0.49	0.5	0.51
0.3	0.57	0.59	0.61
0.4	0.62	0.64	0.66
0.5	0.65	0.68	0.7
0.6	0.66	0.68	0.74
0.7	0.7	0.73	0.78
0.8	0.72	0.74	0.79
0.9	0.73	0.76	0.81
1	0.75	0.77	0.83

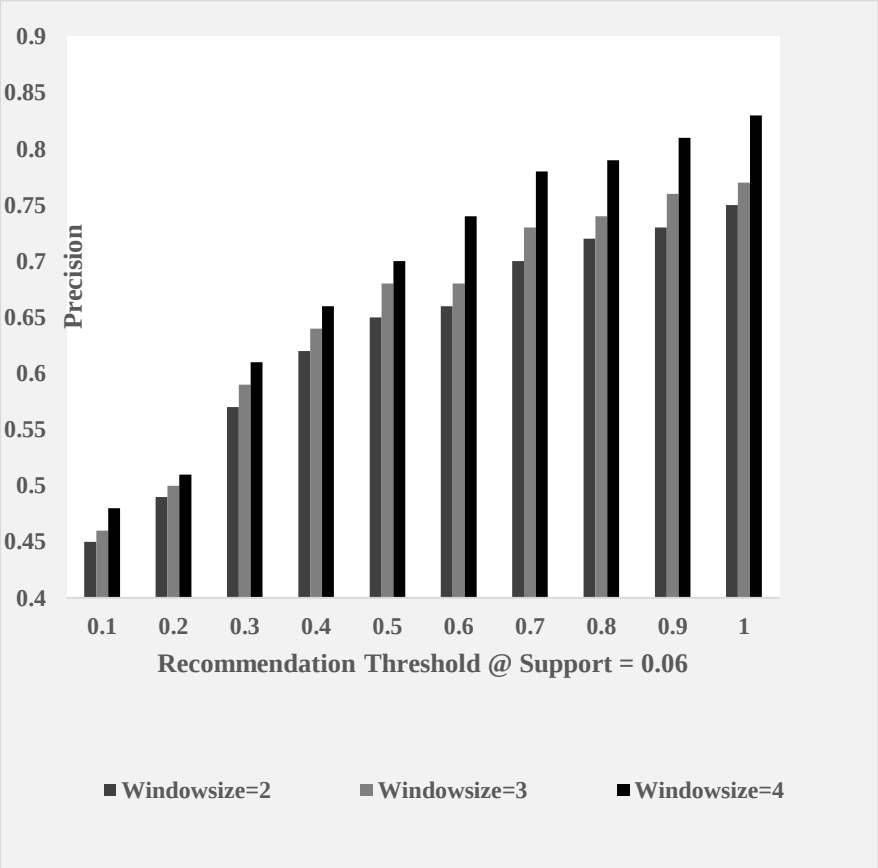

Fig. 2 Precision

Table 3 and figure 3 shows that the coverage for window size 2 performs better by 16.35% and by 42.25% at recommendation threshold 0.1 than window size 3 and 4 respectively. The coverage for window size 2 performs better by 28.57% and by 55.1% at recommendation threshold 0.5 than window size 3 and 3 respectively. The coverage for window size 2 performs better by 9.52% and by 31.58% at recommendation threshold 1 than window size 3 and 4 respectively.

Table 3 Coverage

Recommendation Threshold @ Support = 0.06	Window size=2	Window size=3	Window size=4
0.1	0.86	0.73	0.56
0.2	0.78	0.59	0.46
0.3	0.68	0.54	0.41
0.4	0.52	0.45	0.4
0.5	0.44	0.33	0.25
0.6	0.32	0.24	0.21
0.7	0.28	0.23	0.19
0.8	0.21	0.18	0.16
0.9	0.14	0.11	0.08

1	0.11	0.1	0.08

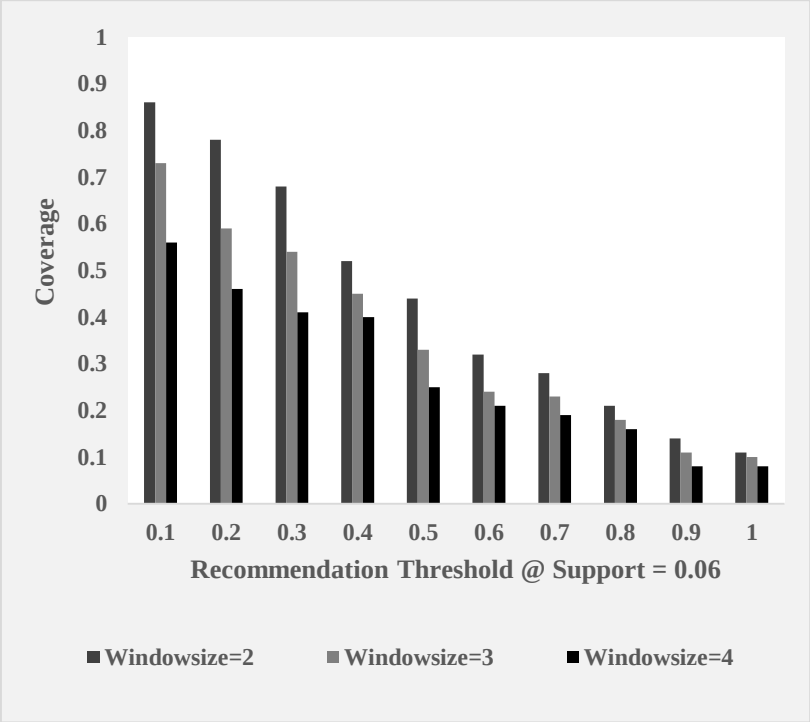

Fig. 3 Coverage

V. CONCLUSION

Big Data has been used for the identification of datasets that has a size which is beyond the capacity of the typical tools of the database software for storing, analysing and managing. Web mining is that application of the techniques of data mining for extracting some useful knowledge from the web data which includes all web documents, usage logs in websites and the hyperlinks that are between the documents. A web usage mining denotes applying the techniques of data mining for discovering the interesting usage patterns that are from the data of web usage for understanding and serving the needs of all web-based applications. A recommendation system has been a decision support system that provides desirable information to customers according to their requirements. The Firefly Algorithm is a swarm-based algorithm which is well-known and this has gained popularity in a very short time with various applications. The proposed recommender system was based on the firefly optimization. The results have proved

that a precision for a window size 4 which performs better by about 6.45% and further by about 4.26% at a recommendation threshold of 0.1 than a window size 2 and 3. The actual precision for a window size 4 will perform better by about 7.4% and further by about 2.89% at a recommendation threshold of 0.5 than the window size 2 and 3. The actual precision for a window size of 4 performs better by about 10.13% and further by about 7.5% at a recommendation threshold of 1 than a window size of 2 and 3 respectively.

References

[1] N.Khan, I.Yaqoob, I.A.T. Hashem, Z.Inayat, M.Ali, W.Kamaleldin and A.Gani,, " Big data: survey, technologies, opportunities, and challenges," The Scientific World Journal, vol.18, doi: 10.1155/2014/712826 , Jul 2014.

[2] A.Oussous, F.Z.Benjelloun, A.A.Lahcen and S. Belfkih, "Big data technologies: A survey," Journal of King Saud University-Computer and Information Sciences, vol. 30, no. 4, pp. 431–448, 2018.

[3] M.D.A.Praveena and B.Bharathi, "A survey paper on big data analytics," Proc. International Conference on Information Communication and Embedded Systems (ICICES), 2017, pp. 1-9.

[4] J.Fan, F.Han, and H.Liu, "Challenges of big data analysis," National Science Review, vol.1, no. 2, pp. 293-314, 2014.

[5] Anurag Kumar and Ravi Kumar Singh, "A study on web structure mining," International Research Journal of Engineering and Technology, vol. 4, no. 1 pp. 715-719 , Jan 2017.

[6] P.Sukumar, L.Robert and S.Yuvaraj, "Review on modern data pre-processing techniques in web usage mining (WUM)," Proc. International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS), 2016, pp. 64-69.

[7] Han, Jiawei, Jian Pei, and Micheline Kamber, Data mining: concepts and techniques. Elsevier, 2011.

[8] R.Mishra, P.Kumar and B.Bhasker, " A web recommendation system considering sequential information," Decision Support Systems, vol. 75, pp. 1-10, Jul 2015.

[9] M.R.Ghazia , D.Gangodkara , "Hadoop, MapReduce and HDFS: A Developers Perspective," Proc. International Conference on Intelligent Computing, Communication & Convergence (ICCC-2014) Procedia Computer Science, Vol. 48, 2015, pp, 45-50.

[10] Jeffrey Dean and Sanjay Ghemawat, "MapReduce: simplified data processing on large clusters", Communication of ACM, vol. 51, no. 1, Jan 2008

[11] P.B.Mohata and S.Dhande, "Web data mining techniques and implementation for handling big data," International Journal of Computer Science and Mobile Computing, vol.4, no. 4, pp. 330-334, 2015.

[12] M.Chen, S.Mao and Y.Liu, "Big data: A survey," Mobile Network Applications, vol. 19, no.2, pp. 171-209, 2014.

[13] M.Barot, K.H.Wandra and S.B.Patel, "Web usage data based web page recommender system," international journal of engineering development and research, vol. 5, no. 2, pp. 1769- 1775, 2017.

[14] Z.Khorsand, and R.Mortazavi, "An effective algorithm in a recommender system based on a combination of imperialist competitive and firefly algorithms".

[15] S.Sharma, "Enhanced swarm based optimized recommendation system," International Journal of Computer Applications, vol.152, no. 5, pp. 15-19, Oct 2016.

[16] J.Srivastava, R.Cooley, M.Deshpande and P.N.Tan, "Web usage mining: Discovery and applications of usage patterns from web data," SIGKDD Explorations, vol. 1, no. 2, pp. 12-23, 2000.

[17] B.Mobasher, H.Dai, T.Luo and M.Nakagawa," Effective personalization based on association rule discovery from web usage data," Proc. 3rd ACM Workshop on Web Information and Data Management (WIDM01), held in conjunction with the International Conference on Information and Knowledge Management (CIKM 2001). ACM, New York, Nov 2001, pp. 9-15.

[18] B.Mobasher, R.Cooley and J.Srivastava, "Automatic personalization based on web usage mining," Communications, ACM, vol. 43, no. 8 , pp. 142–151. Aug 2000.

[19] H.S.Guruprasad, " Comparative analysis of web usage data using som and k-means algorithms," International Journal of Advanced Research, vol. 3, issue 10, pp. 486-493, 2015.

[20] X-S,Yang, "Nature-inspired metaheuristic algorithms," Luniver Press, UK, 2010.

[21] W.A.Khan, N.N.Hamadneh, S.K.LTilahun and J.M.Ngnotchouye, "A review and comparative study of firefly algorithm and its modified versions," In Optimization Algorithms-Methods and Applications. IntechOpen, https://doi.org/10.5772/62472, 2016.

[22] M.Takeuchi, T. Ott, H. Matsushita, Y. Uwate and Y. Nishio, "K-means algorithm using improved firefly algorithm," RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing (NCSP'17), pp. 225-228.

[23] C.Wang, C.Tsai, C.Fan and Sh.Yuan, "A Hadoop based weblog analysis system," Proc. Seventh International

Conference on Ubi-Media Computing and Workshops(UMEDIA-2014), 2014, pp. 72-77.