IMPROVISING OF WEB RECOMMENDATION SYSTEM BY USING K-MEANS AND BITAP ALGORITHM

[1] **Dipali Wankhede,**[2] **S.G.Tuppad**
[1]ME Student,[2]Professor
[1]Computer Science,
[1]MSSCET, Jalna,India

Abstract*— The increase in the amount of information over the Internet in recent years has led to the risk of data flooding, which has resulted in problems accessing user-related information. In addition, the increase in the number of websites and web pages has made it difficult for webmasters to determine the content of their content. The online user's information needs can be obtained by evaluating the user's web usage. Web Based Data Mining (WUM) is used to extract knowledge from web users. Access logs are accessed by using data mining techniques. One of the WUM applications is a guideline, a privacy filtering technique used to determine whether a user approves a specified item or identifies a list of things that may be important to the user. This article is an architecture that combines product data with user access log data and generates a set of instructions for that user. The operation records interesting results in terms of precision, recall, and metrics.

***Index Terms*— Web Usage mining, Online Web Recommendation System, Clustering, Pattern Matching, Boyer Moore, K- means, Recommendation**

I. INTRODUCTION

In recent years, eCommerce, web services and web-based information systems have been used rapidly. The explosion of the web and the emergence of ecommerce have encouraged designers to develop. The ecommerce product introduction system has transformed the global business landscape. Online businesses are gaining popularity. Nowadays people usually transact via the internet. Web users demonstrate a variety of navigation styles by clicking on a set of pages. These variables can be understood by recording mine users using WUM. One of the most commonly used applications in web mining is online hints and forecasts. Web mining is a strategy for grouping web pages and web clients by viewing web site content and Previous Web mining client support Web client support in respect of pages to be seen in the future. Web mining includes Web Content Mining (WCM), Web Mining Mining (WSM), and Web Usage Mining (WUM). [16] Web mining is a system for extracting valuable information from search results. Extract from customer relationship while browsing the web. Recorded data collected in server access logs, logs, process records, client-side processing, client profiles, and metadata. The WUM process is a useful compilation process from the server log. In general, all guidance systems will follow the framework for creating effective recommendations. Different guidance systems use different methods based on the source of information used. Accessible data sources are user data. (Demographics) Product information [Types of keywords] and user ratings. [3] Current guidance systems have limitations such as intelligence, adaptability, precision, limitations. These disadvantages can be overcome by using a hybrid architecture. What's Combining product data with user access records and creating user sets using the Boyer-Moore Pattern Matching Algorithm and K-Means clustering algorithms.

II. RELATED WORK

Sneha Y.S, Mahadevan G., Madhura, [1] [2] offers an engineering structure that includes semantic information on exploiting web-based data mining. Use the longest subdirectory to create a suggested list. This system improves the performance of the existing Recommender system by overcoming new problems. The system consists of both online and offline phases. The RDF format is used for Semantic Data Integration. The system does not involve grouping user profiles, which leads to pattern searches by polls of all usage logs which will lead to the use. More time and lower overall system performance.

Brute force (BF) [1] or algorithm Naïve is the logical place to begin reviewing exact string matching algorithms place. That compared with a pattern with all the text substrings proposal in any case a complete match or a mismatch. It has no pre-processing phase and does not require additional space. The time complexity of the search phase brute force algorithm is O (mn). Knuth-Morris-Pratt (KMP) [2] algorithm was proposed in 1977 to accelerate the process of exact match patterns by improving the lengths of shifts. Characters from left to right pattern are compared. In case of coincidence or mismatch comparisons using prior knowledge to calculate the next position pattern with text. The complexity of preprocessing time is O (m) and the search phase is O (nm)

Boyer-Moore (BM) [3] algorithm published in 1977 and that time is considered as the search algorithm more efficient chain. It is performed in character comparisons reverse the order from right to left and did not require pattern around the pattern to look for in case of a mismatch. In case of a match or mismatch, this uses two changing the rules to change the correct pattern. Time and space complexity preprocessing is O (m + | Σ |) and the worst time to seek execution phase is O (nm + | Σ |). The best algorithm Boyer-Moore case is O (n / m). Boyer-Moore Horspool (BMH) [4] has not used the heuristic displacement as Boyer-Moore algorithm used. Only use heuristic occurrence to maximize the life of the characters changes corresponding to the right most character of text head. Is preprocessing time complexity is O (m + | Σ |) and search time complexity is O (mn).

Quick Search (QS) [5] algorithm comparisons from left to right, the criteria are changed one character to the right with the pattern and the misapplication of the rule changing character is examined. The worst case time complexity of QS is the same as the algorithm, but can take steps Horspool in practice.

Boyer-Moore Smith (MBS) [6] realized that, to calculate the change of BMH, sometimes moving maximize QS changes. Use the changing nature of the BMH evil ruler and QS bad character rule to change the pattern. Its time complexity is O preprocessing (m + | Σ |) and search time complexity is O (mn).

Hadi Khosravi et al [4] conducted a meaningful introduction process for electronic catalogs. The key view consists of displaying and arranging web articles, matching between the list and the set of activities that will guide you for personalization. This system uses key elements of your web presence and personalization, such as web page or web modeling and customer modeling, mapping between customers and the right product, and set up a set of recommendations. Ontology and OWL (Web Ontology Language) for product classification on the Web. This system avoids false positives. Suggestions such as products are recommended, even if they are not relevant to the customer.

Mehradad Jalali et al. [5] propose an online guidance system using the LCS algorithm. Involves two steps that work with each other. That is, the online and offline steps of pretreatment and navigation mining are carried out in offline form while forecasting takes place in an online process. Suleyman Salin and Pinar Senkul used an information-driven architecture as a model for web-based data mining. Generate access to normal Xin Sui, Suozhu Wang, Zhaowei Li [7] have conducted research and proposed a model that incorporates a Web-based recommendation system and a personal recommendation system (SWARPS). Ecommerce This involves the use of AI-Multi Agent techniques. (Agent, agent, web analytics, agent change, semantic agent, data mining agent, analytics analyst, and semantics analyst). Agents work together to make recommendations. The system has a level of intelligence, autonomy, and flexibility.

Himangni Verma and Hemant Verma [19] have brought their end-users to the customer web.The implementation of the customer framework will be evaluated. This assessment will help to develop a transactional model that will help in personalization and behavioral analysis. Frequently used access pattern algorithms are used to break the data into pieces of time. This data was prepared using the Hidden Markov Model (HMM) to find the data model that helps to generate accurate and effective recommendations.

III. PROPOSED SYSTEM

Current recommendation systems exhibits certain limitations such as intelligence, adaptability, flexibility, limited accuracy. These disadvantages can be overcome by implementing a hybrid architecture that integrates product information with user's access log data and then generates a set of recommendations for that particular user. This system handles most of the drawbacks and gives more efficient and more accurate result than previous systems.

Recommendation Systems can utilize data mining strategies for making suggestions utilizing learning gained from the activity and qualities of the clients. The architecture of an online web recommendation system based on web usage mining basically consists of three phases : Data Preprocessing, Pattern detection and generating recommendations. Data preprocessing and Pattern detection phases are performed offline and the recommendations are generated online. Data preprocessing involves transforming the web access logs and user profiles into format appropriate for the system. Pattern detection involves using data mining techniques like clustering, sequential pattern mining or association rule mining. Lastly the detected patterns are used to generate recommendations which provide customized links or data to the user.

Architecture Overview

Recommender framework helps clients to discover and evaluate their investments. The Recommender system can utilize data mining strategies to provide guidance based on knowledge gained from the activity and quality of the customer. The Miner's Guide to online guidance systems on the web. Internet use generally consists of three steps: data processing, pre-detection, pattern generation, and hinting. The process of data processing and pattern detection is performed offline and instructions are generated online. Previous data processing involved converting log files to web access and user profiles in a form suitable for the system. Includes pattern validation using data mining techniques such as custom mining groupings or mining rules. Finally, the detected pattern is used to generate instructions that provide links or custom information to the user. Each column has a subset of highly relevant attributes. Horizontal partitioning [8] [9] is achieved by grouping the teapot into the tank. Finally, within each group, the values of each column are random. (Or sort) to break the links between the columns. The basic idea of overlapping overlays is to divide the column of transverse relationships. But to maintain the relationship within each column, which will reduce the dimensionality. Of the data and maintain a better utility. Generalization and overlapping overlays maintain the utility because of the high relevancy of attributes and maintain the relationship between these attributes. The overlapping summit protects personal information because it distinguishes between unreliable and infrequent links. Keep in mind that when a data set has IQs and SAs cleared, they need to interrupt their relationship. Overlays, on the other hand, can segment certain QI features with SA by maintaining an attribute relationship with a delicate attribute. Intriguing stacked overlays help protect personal information that overlapping stacking ensures that for any track, there are several buckets.

We consider common parameters for data dissemination with horizontal partitioned data between multiple data providers, each of which provides sub-data of Ti records. As a special case, the data provider may be the data owner. Who owns their data? This is a common situation in social networking and hint systems. Our goal is to publish the anonymous information of embedded information so that the recipient of the information, including the data provider, will not affect the confidentiality of the personal information by any other person.

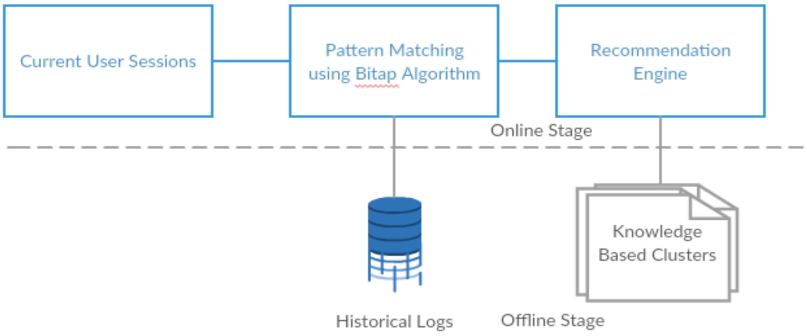


Figure 1 Architectural Flow Diagram

Figure 1 shows the architecture of existing recommendation system that uses the user information stored in the web log files. The improved system architecture above the instructions involves additional integrated user data. (Such as user profiles). This system includes more data mining algorithms such as clustering and pattern matching algorithms. As a result, users with common behavior are grouped first and then grouped into groups. This type of recommendation system will generate your own recommendations. New users are ranked first in a group, and then use the corresponding group format to set current user and other similar user guidance in the top cluster. It is divided into two main stages. Offline phase and online phase. Both of these steps clearly cooperate with each other.

Offline Phase of the Architecture

This process consists of two main modules: data processing and knowledge base of pre-processing products. In the process, I started with offline pre-processing basics. Web-Access-Log This includes splitting the client session and entering important data in the database.

1) Data Processing: At this stage, the source files are formatted to find the Web access range. Web server logs are generated throughout the user's web server access. There are different types of web logs based on different server parameters. These records include information such as URLs, IP addresses, clients, etc. Pre-processing features, such as session cleanup data, are performed prior to using Web mining algorithms on web server logs. Grouping is also carried out in the process. Here k refers to the clustering algorithm used. In the k-guide system, means can be used in the pre-processing process for identifying groups of users appears to be similar. Used to collect user profiles.

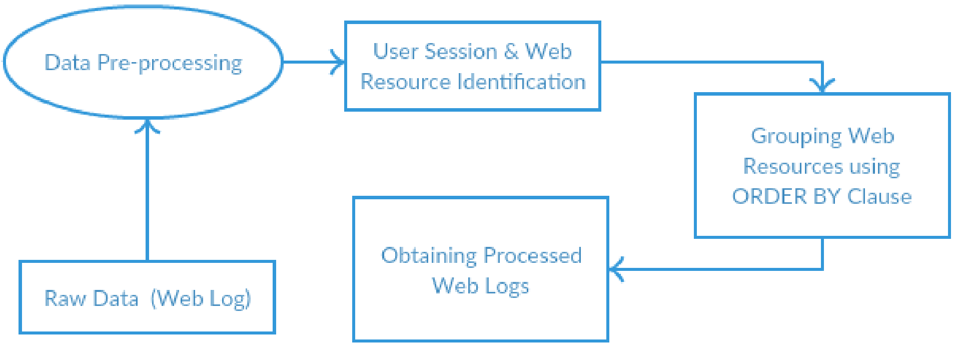

Data Pre-processing
User Session & Web Resource Identification
Grouping Web Resources using ORDER BY Clause
Obtaining Processed Web Logs
Raw Data (Web Log)

Figure 2 Data Pre Processing & Processed Web Logs

2) Knowledge base: After the data processing, product data information is merged with the user session data extracted from the record. These features include brand prices, user details, and transactions. It performs in a table in an advanced database system.

Online phase Architecture

In this session, when the user logs on to the server, the instructions are controlled with the knowledge base for the above transaction with the user. A list of recommended products is created based on the user's previous history and the type of group the user is a member of.

1) Creating Suggestions: An important system utility is to create recommendations using some refining parameters such as brand valuation and other customizable parameters to get a certain set of support values. Defined elements of the database. To get a summary of the pattern using the Boyer-Moore pattern matching algorithm, the pattern search can be used to find elements of customer interest that focus on current customer fitness to forecast and recommend. Appeal of future customers Moore's Boyer search algorithm is used in the guidance architecture.

Bitap Algorithm

The Bitap algorithm (also known as substitution or substitution or algorithm-and Baeza-Yates-Gonnet) is a vague match-layout. The algorithm says that if the specified string contains a substring The equation is determined by the approximation of the equation in terms of the distance Levenshtein - regardless of whether the substring and the form are at the distance k or not. The algorithm starts with a set of bit masks containing bits for each element of the pattern. He is then able to work most of the bit-intensive operations, which are very fast. The Bitap algorithm is well known as one of the algorithms that use the universal grep utility. Written by Udi Manber, Sun Wu, and Burra Gopal, the manuscripts of Manber and Wu provide extended algorithms to correct common expressions that are not the same.

Exact Searching

The bitap algorithm for exact string searching, in full generality, looks like this in pseudo code:

```
Algorithm bitap_search (text: string, pattern: string) returns string

m: = length (pattern)
if m = = 0
return text
/*Initialize the bit array R. */
R: = new array [m+1] of bit, initially all 0
R [0] = 1
for i=0; i<length (text); i+=1;
/*Update the bit array.*/
for k=m;k>=1;k-=1:
R[k] =R [k-1] & (text[i] = = pattern [k-1])
if R[m]:
return (text+i - m) +1
return nil
```

Fuzzy Searching

To perform fuzzy search string using the algorithm BITAP, it is necessary to expand the array of bits R in a second dimension. Instead of having a single matrix R which changes throughout the text, we now have k different matrices R1 ... k. Ri has a matrix representation of pattern prefixes that match any of the current string suffix i or fewer errors. In this context, an "error" can be an insertion, deletion or substitution; see Levenshtein distance for more information on these operations. The application then performs fuzzy matching (returning the first match with up to k errors) using the algorithm bitap diffuse. However, only pays attention to the substitutions, insertions or deletions for not - in other words, a Hamming distance of k. As before, the semantics of 0 and 1 are reversed from their intuitive meanings.

IV. K-MEANS CLUSTERING ALGORITHM

Clustering is an unsupervised or dividing pattern in groups or subgroups classification (i.e. clusters). Here the objects are grouped into classes of similar objects based on their location and connectivity within a space of dimension n. Mainly the principle of the grouping is to maximize similarity within a cluster, and to minimize the similarity between the groups. Although there are many clustering algorithms available, one of the most used it is the k means algorithm. Its aim is to minimize the distance of objects from the centroid of each group. One of the most clustering algorithms used is k-means clustering, which is a partitioning method. Information of a set of N elements is divided into k disjoint subsets Sj containing Nj questions which are so close to each other as could reasonably be expected to agree on a certain measured distance. Each cluster is characterized by over New Jersey, and its centroid λ j. The centroid is a point at which the sum of the distances of all objects of that group is minimized. Therefore, we can characterize the k-means clustering algorithm as an iterative methodology to minimize E = Σk 1σn∈sjd (xn, λ j), where xn is a vector of talking to the nth object, λ j is the centroid of the object Sj d is the measured distance. The k-means clustering moving objects between groups until you cannot decrease even more [15], [16].

V. EXPERIMENTAL RESULTS

In our experiments, we obtain data set from the Amazon SNAP web forum. This dataset includes reviews (ratings, text, helpfulness votes), product metadata (descriptions, category information, price, brand, and image features), and links (also viewed/also bought graphs).Our model is semi supervised and can do without any labeled data, but the bitap algorithm needs some regular expressions for the input to find the rest of the matching data . Therefore we introduce K means metric to extract a K items from the tuples as seed information. In our experiments we found that by using k-means , the top 10 highest values of the items could have perfect precision on the Dataset. Hence, we focus on selecting some aspects from the dataset as seed items by using an semi supervised metric.

In result analysis with binary classification, precision (also called positive predictive value) is the fraction of retrieved instances that are relevant, while recall (also called sensitivity) is the fraction of the relevant instances that are retrieved. Precision and recall are therefore based on understanding and measuring relevance.In simple terms, high accuracy means that an algorithm returns significantly more relevant than irrelevant results, while a high recall means that an algorithm has yielded the most relevant results.

The most important category measurements for binary categories are:

Precision	Recall	F Measure
$P = TP/(TP + FP)$	$R = TP/(TP + FN)$	$tp + tn/tp + tn + fp + fn$

Table 5.1 Precision, Recall F-measure Readings

Precision	98.69
Recall	91.01
F Measure	94.70

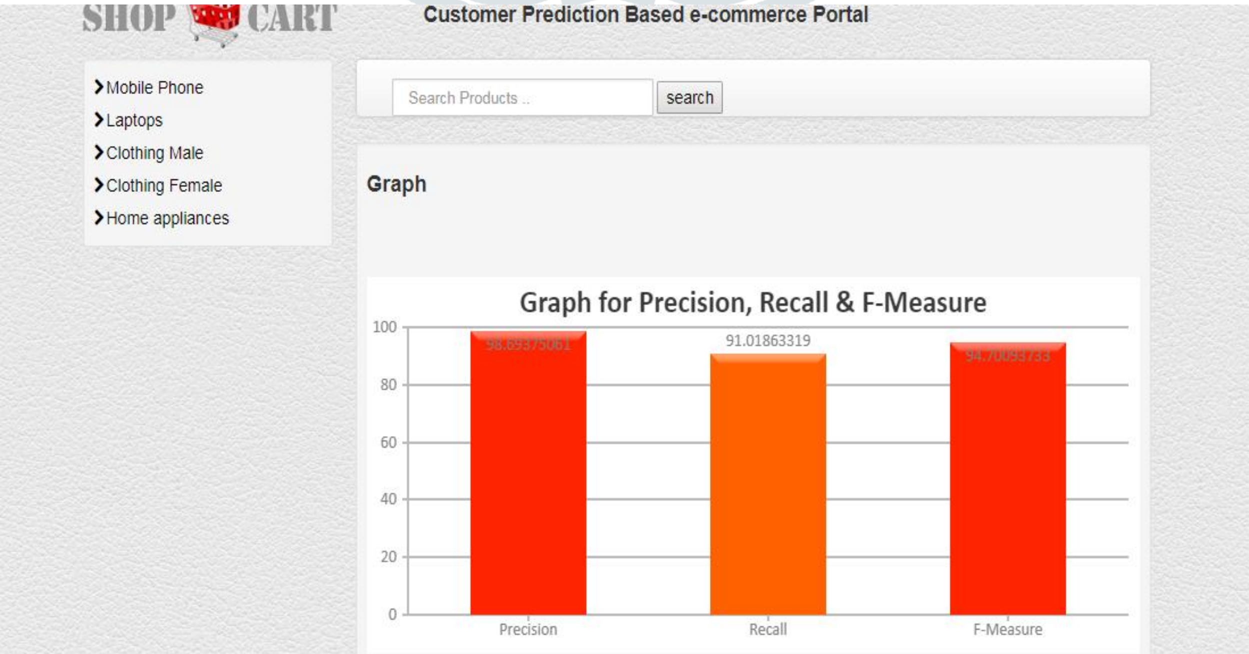
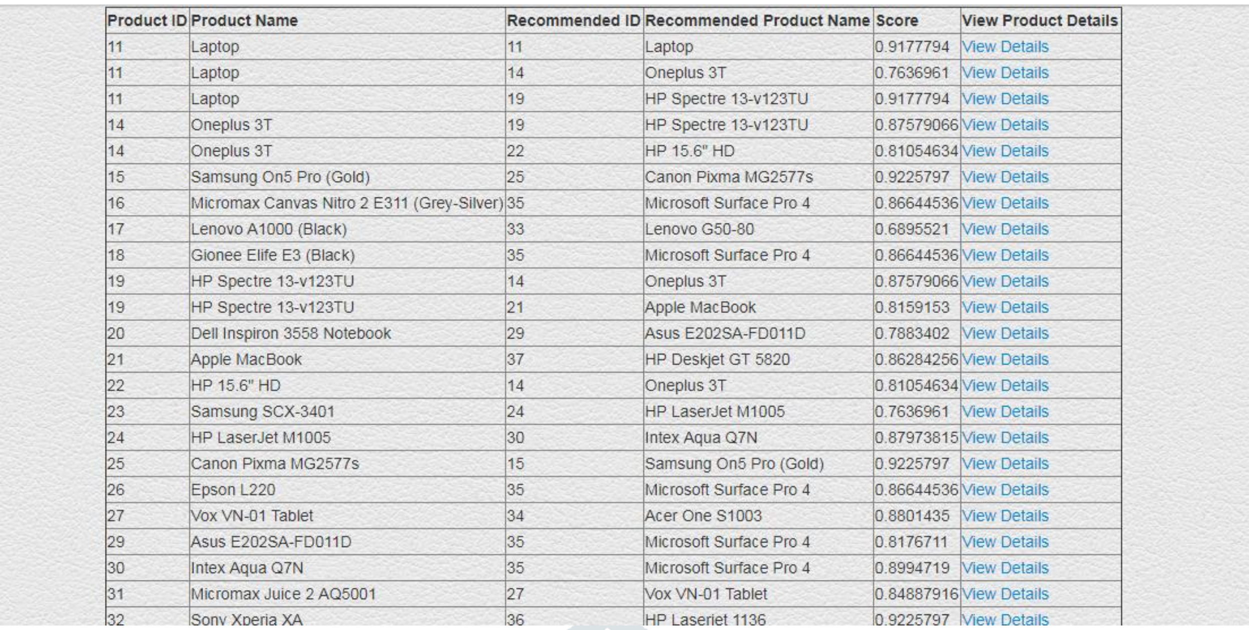
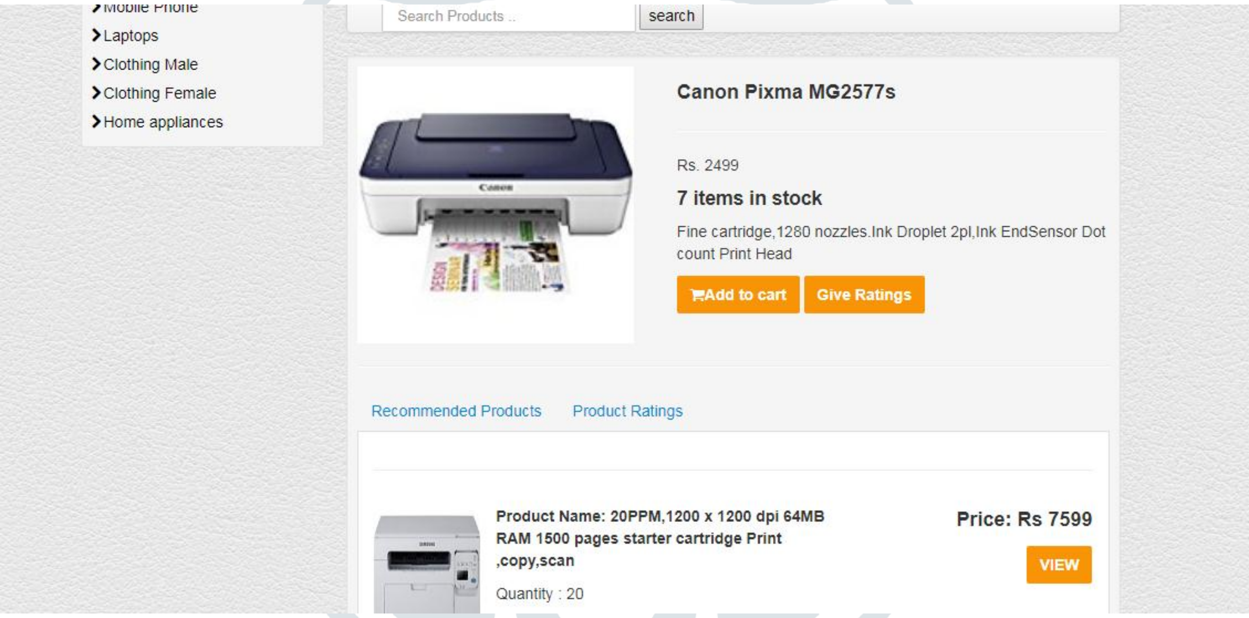


Figure 3 Graph for Precision, recall,F-measure



Product ID	Product Name	Recommended ID	Recommended Product Name	Score	View Product Details
11	Laptop	11	Laptop	0.9177794	View Details
11	Laptop	14	Oneplus 3T	0.7636961	View Details
11	Laptop	19	HP Spectre 13-v123TU	0.9177794	View Details
14	Oneplus 3T	19	HP Spectre 13-v123TU	0.87579066	View Details
14	Oneplus 3T	22	HP 15.6" HD	0.81054634	View Details
15	Samsung On5 Pro (Gold)	25	Canon Pixma MG2577s	0.9225797	View Details
16	Micromax Canvas Nitro 2 E311 (Grey-Silver)	35	Microsoft Surface Pro 4	0.86644536	View Details
17	Lenovo A1000 (Black)	33	Lenovo G50-80	0.6895521	View Details
18	Gionee Elife E3 (Black)	35	Microsoft Surface Pro 4	0.86644536	View Details
19	HP Spectre 13-v123TU	14	Oneplus 3T	0.87579066	View Details
19	HP Spectre 13-v123TU	21	Apple MacBook	0.8159153	View Details
20	Dell Inspiron 3558 Notebook	29	Asus E202SA-FD011D	0.7883402	View Details
21	Apple MacBook	37	HP Deskjet GT 5820	0.86284256	View Details
22	HP 15.6" HD	14	Oneplus 3T	0.81054634	View Details
23	Samsung SCX-3401	24	HP LaserJet M1005	0.7636961	View Details
24	HP LaserJet M1005	30	Intex Aqua Q7N	0.87973815	View Details
25	Canon Pixma MG2577s	15	Samsung On5 Pro (Gold)	0.9225797	View Details
26	Epson L220	35	Microsoft Surface Pro 4	0.86644536	View Details
27	Vox VN-01 Tablet	34	Acer One S1003	0.8801435	View Details
29	Asus E202SA-FD011D	35	Microsoft Surface Pro 4	0.8176711	View Details
30	Intex Aqua Q7N	35	Microsoft Surface Pro 4	0.8994719	View Details
31	Micromax Juice 2 AQ5001	27	Vox VN-01 Tablet	0.84887916	View Details
32	Sony Xperia XA	36	HP Laserjet 1136	0.9225797	View Details

Figure 4 Recommended Output



Canon Pixma MG2577s

Rs. 2499

7 items in stock

Fine cartridge,1280 nozzles.Ink Droplet 2pl,Ink EndSensor Dot count Print Head

Product Name: 20PPM,1200 x 1200 dpi 64MB RAM 1500 pages starter cartridge Print ,copy,scan

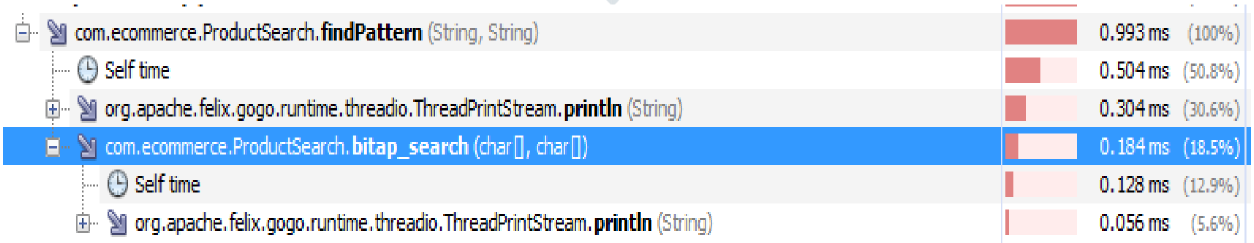
Quantity : 20

Price: Rs 7599

Figure 5 Recommended Product with Ratings

VI. COMPUTATIONAL EFFICIENCY

We compare the response time of bitap algorithm with generalization and the reduction of the consumption in terms of calculation efficiency.



```
com.ecommerce.ProductSearch.findPattern (String, String)                      0.993 ms  (100%)
    Self time                                                                 0.504 ms  (50.8%)
    org.apache.felix.gogo.runtime.threadio.ThreadPrintStream.println (String)  0.304 ms  (30.6%)
    com.ecommerce.ProductSearch.bitap_search (char[], char[])                  0.184 ms  (18.5%)
        Self time                                                             0.128 ms  (12.9%)
        org.apache.felix.gogo.runtime.threadio.ThreadPrintStream.println (String)  0.056 ms  (5.6%)
```

Figure 6 Java Profiling for find pattern and bitap searching method.

VII. FUTURE SCOPE AND CONCLUSION

This Online Web Recommendation System displays a list of recommended products based on the user’s recent history. One of the most popular clustering algorithms is k-means clustering algorithm, but in this method the quality of the final clusters rely heavily on the initial centroids, which are selected randomly moreover, the k-means algorithm is computationally very expensive also. As the same enhanced method also chooses the initial centroids based upon the random selection. So this method is very sensitive to the initial starting points and it does not promise to produce the unique clustering results. Finally this proposed method i.e. Bitap algorithm which focuses if the substring and pattern are within the distance k of each other, them the algorithm considers them equal.

This Advance System displays a list of recommended items. It involves clustering of pattern items which leads to searching of patterns in clusters rather than searching whole items from databases. It reduces execution time and thus increasing performance of the overall system. The system doesn’t suffer from searching items in whole database .it just find from pattern cluster which is to be recommended. Future scope includes using K-Means clustering the performance values obtained as a new parameter which is going to improving the performance of this advance recommendation system.

REFERENCES

[1] Dipali Wankhede, S. G. Tuppad “Improvising of web Recommendation System by using K-Means and Bitap Algorithm”, ISSN: 2249 – 8958, Volume-6 Issue-3, February 2017.

[2] Sneha Y.S, G. Mahadevan, Madhura,”An Online Recommendation System based on web usage mining and semantic web using LCS Algorithm”, IEEE 2011

[3] Sneha Y.S, G. Mahadevan, Madhura,,”A Personalized Product Based Recommendation System Using Web Usage Mining and Semantic Web”, International Journal of Computer Theory and Engineering (IJCTE) Vol. 4, No. 2, April 2012

[4] Mehrdad Jalali1, Norwati Mustapha, Md. Nasir B Sulaiman, Ali Mamat, “A Web Usage Mining Approach Based on LCS Algorithm in Online Predicting Recommendation Systems” 12th International Conference Information Visualisation IEEE 2008.

[5] SuleymanSalim et al “Using Semantic Information for web usage mining based recommendation” International Conference IEEE 2009

[6] Xin Sui, Suozhu Wang, Zhaowei Li “Research on the Model of Integration with Semantic Web and Agent Personalized Recommendation System “Proceedings of the 2009 13th International Conference on Computer Supported Cooperative Work in Design.

[7] Himangni Rathore, Hemant Verma, “Analysis on Recommended System for Web Information Retrieval Using HMM”, International Journal of Engineering Research and Applications ISSN : 2248-9622, Vol. 4, November 2014.

[8] Hadi Khosravi Farsani, and Mohammadali Nematbakhsh “A Semantic Recommendation Procedure for Electronic Product Catalog”, World Academy of Science, Engineering and Technology 22 2006.

[9] Hiral Y Modi, Meera Narvekar .”Enhancement of Online web Recommendation System Using A Hybrid Clustering And Pattern Matching Approach”: 2015 International Conference on Nascent Technologies in the Engineering Field (ICNTE-2015)

[10] RVVSV Prasad, V Valli Kumari “A Categorical Review of Recommender Systems”, International Journal of Distributed and Parallel Systems (IJDPS) Vol.3, No.5, September 2012.

[11] Jiawei Han and Micheline Kamber, “Data Mining: Concepts and Techniques”, Morgan kaufmann, 2nd Edition.

[12] Jaideep Srivastava, Prasanna Desikan, Vipin Kumar ”Web Mining - Concepts, Applications & Research Directions”.

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

[14] Bamshad Mobasher, Robert Cooley, Jaideep Srivastava “Automatic Personalization Based on Web Usage Mining”, Communications of the ACM Volume 43 Issue 8, Aug. 2000.

[15] Cormen, T.H., Leiserson, C.E., Rivest, R.L., Introduction to Algorithms, Chapter 34, MIT Press, 1990, pp 853-885. 2. Knuth, D., Morris, J. H., Pratt, V., "Fast pattern matching in strings," SIAM Journal on Computing, Vol. 6, No. 2, doi: 10.1137/0206024, 1977, pp.323–350.

[16] R.S. Boyer, J.S. Moore, "A fast string searching algorithm, “Communication of the ACM, Vol. 20, No. 10,1977, pp.762– 772.

[17] R. N. Horspool, "Practical fast searching in strings," Software—Practice and Experience, Vol. 10, No. 3, 1980, 501–506.

[18] Sunday, D.M., "A very fast substring search algorithm," Communications of the ACM, Vol. 33, No. 8, 1990, pp. 132- 142.

[19] Smith, P.D., "Experiments with a very fast substring search algorithm,"Software-Practice and Experience, Vol. 21, No. 10, pp.1065-1074.

[20] Crochemore, M., Czumaj, A., Gasieniec, L., Jarominek, S., Lecroq, T., Plandowski. W., Rytter, W., "Speeding up two string matching algorithms," Algorithmica, Vol. 12, No. 4/5, 1994, pp.247-267.

[21] RVVSV Prasad, V Valli Kumari “A Categorical Review of Recommender Systems” , International Journal of Distributed and Parallel Systems (IJDPS) Vol.3, No.5, September 2012

[22] Hadi KhosraviFarsani, and Mohammadali Nematbakhsh “A Semantic Recommendation Procedure for Electronic Product Catalog”, World Academy of Science, Engineering and Technology 22 2006.

[23] Neelam Duhan, A. K. Sharma and Komal Kumar Bhatia, “Page Ranking Algorithms: A survey”, in proceedings of the IEEE International Advanced Computing Conference (IACC), 2009.