International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-6 Issue-3, February 2017

Enhancement of Online Web Recommendation System using a Hybrid Clustering and Pattern Matching Approach

Dipali Wankhede, S. G. Tuppad

Abstract: ***Increasing the amount of information over the Internet in recent years has led to the increased risk of flooding of information which in turn has created the problem of access to relevant data users. Also with the rise in the number of websites and web pages, webmasters find it difficult to make the content according to user need. Demand for information Users can imagine evaluating web user browsing behavior. Web Usage Mining (WUM) is used to extract knowledge from access logs Web user by using Data mining techniques. One of the applications is WUM recommendation system that is customized information filtering technique used to determine whether any of a user approved a particular article or to identify a list of items that it can be of great importance to the user. In this document architecture that integrates product information with the user access to log data and then generates a set of recommendations for it is presented that particular user. The application has registered encouraging in terms of precision, recall and F1 results metrics.***

Index Terms: ***Web Usage mining, Online Web Recommendation System, Clustering, Pattern Matching, Boyer Moore, K-Means, Recommendation.***

I. INTRODUCTION

In recent years, e-commerce, web services and web information system have been used explosively. Due to Massive explosion of the worldwide web and the emergence of electronic commerce, Designers are encouraged to develop recommender systems. E-commerce has changed all commercial business scenarios in the world. Subject on Internet is increasingly gaining popularity today. People are more likely to carry out transactions through Internet.

Internet users demonstrate a variety of navigation patterns by clicking series of web pages. These patterns can be understood by mining user starts using WUM. One of the Mining applications which have web usage is Online Recommendation and prediction. Web mining is the strategy to group pages and web clients to look at the bottom of the page and Web client drive previously. Web mining is compatible with customer regarding web pages to be viewed in the future. Web Mining is mining web content (WCM), Web Mining Structure (WSM), and the use of Web Mining (WUM) [9]. Web usage mining is the system to extract valuable the auxiliary information data resulting from the customer associations while surfing on the Web. That separates the information contained in the server access logs, reference records, operators, is the side of the customer, customer profile and Information.WUM goal is the process of harvesting useful Information server logs.

In general, all follow a recommendation systems efficient framework for generating recommendations. Various recommender systems use different approaches the sources of information they use. Accessible sources they are user information (demographics), product information (Keywords, genres) and classifications of user-elements [8].

Current recommendation system exhibits some limitations such as intelligence, adaptability, flexibility, limited accuracy. These disadvantages can be overcome by implementing a hybrid architecture that integrates product information with data log user access and then generates a set of recommendations for that particular user using Bitap algorithm and K-means Algorithm. The rest of the article is organized as follows: In Section 2, recent research developments in the field of recommender systems is reviewed. Section 3 explains the Recommendation system architecture and Section 4 It illuminates several algorithms used. Section 5 shows the performance of the recommendation system and the set of Recommended for a particular user product. Finally, section 6 personifies the role.

II. LITERATURE SURVEY

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.

Manuscript published on 28 February 2017.
***** Correspondence Author (s)
Dipali Wankhede, Department of Computer Science & Engineering, BAMU Matsyodari Shikshan Sansthas College of Engineering and Technology ,Jalna, Aurangabad (Maharashtra)-431203. India.
Prof.S.G.Tuppad, Assistant Professor, Matsyodari Shikshan Sanstha's College of Engineering and Technology, Jalna, Aurangabad (Maharashtra)-431203. India

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC-BY-NC-ND license http://creativecommons.org/licenses/by-nc-nd/4.0/

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).

Turbo Boyer Moore (TBM) [7] is the variation of Boyer Moore algorithm, reminiscent of the substring text string matches the pattern in recent comparisons suffix. You cannot compare the matched substring again; only compared to the other characters in the pattern with the text string.

III. PROPOSED SYSTEM

Current recommendations systems have certain limitations, such as intelligence, adaptability, flexibility, limited accuracy. These disadvantages can be overcome by implementing a hybrid architecture that integrates product information with data users access log and then generates a series of recommendations for that particular user. This system is responsible for most of the disadvantages and as more efficient and gives more accurate results than previous systems.

3.1. Architecture Overview

Recommendation frames help customers discover and evaluate things on investment. Recommender systems can use data mining strategies to make suggestions using knowledge gained from the activity and the qualities customers. The architecture of a miner in Web-based online recommendation system, Internet use basically consists of three phases: pre-processing of data, detecting patterns and generating recommendations. The phases of data pre-processing and pattern detection are performed offline and the recommendations are generated online. Data preprocessing involves the transformation of web access logs and user profiles in the appropriate format for the system. It includes pattern detection using data mining techniques such as clustering, pattern mining customized or mining association rules. Finally, the detected patterns are used to generate recommendations that provide customized links or data to the user.

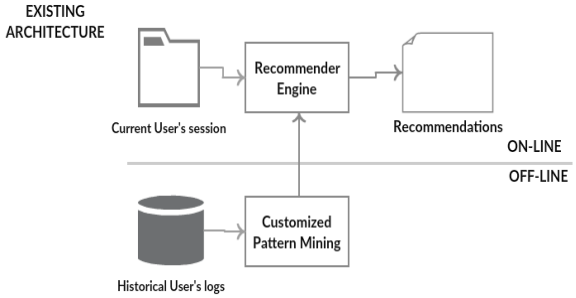

Figure 1. Existing Architecture for Online Web-Based Recommendation System

Figure 1 shows the architecture of existing recommendation system that uses the user information stored in the web log files. These systems are simple because they use a single data mining algorithm, usually a pattern algorithm customized mining. This algorithm applies to all user records to find patterns common navigation user so that the system can find out and predict the next page request. The disadvantage of these systems is that the new user gets only recommendations based on your current navigation. An alternative to these systems is a more advanced system that uses algorithms of data mining algorithms such as clustering and pattern matching. Figure 2 depicts one such advanced system.

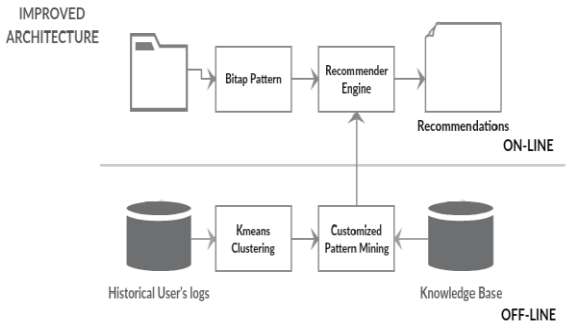

Figure 2. Improved Architecture for Online web-based Recommendation System

The improved system architecture above recommendation involves further integration user information (such as user profiles). This system incorporates more data mining algorithms such as clustering and pattern matching algorithms. Thus users who are of common behavior are grouped first and then each group patterns are discovered. This type of recommendation system generates personalized recommendations. New users are ranked first in one of the clusters and then patterns corresponding groups are used to customize the recommendations based on the current user navigation and other similar users in the cluster. The above architecture is divided into two main phases; offline phase and phase online. These two stages cooperate unequivocally relative to each other.

3.1.1. Offline Phase of the Architecture

This stage consists of two major modules: pre-processing of data and knowledge base of products. In stage I started with the basics offline Web-Access-Log preprocessing that includes extracting client session and enter important data in the database.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

1) Data preprocessing: In this phase, the records of the original fabric are reformatted to discover web access sessions. The web server logs generated throughout the store server web access activities Internet user. There are different types of web logs based on different server parameters. These records include information such as client IP addresses URL, etc.

The different features of pre-processing such as cleaning data, session identification is carried out before using web mining algorithms on the web server logs.

Clustering is also carried out in the preprocessing stage. Here k means clustering algorithm is used. In a recommendation system, k means can be used in the preprocessing stage for identifying user groups that seem to have similar preferences. It is used to aggregate user profiles.

2) Knowledge Base: After the pre-processing of data, various features of the products that are combined with user session data extracted from the records. These features include price, brand, etc. User details and the transaction is carried in tables in the database.

3.1.2. Online phase Architecture

During this phase when the user logs into the server, the recommendation engine controls with the knowledge base for the above transactions the user. The list of recommended products is generated based on the previous history of the user and the pattern of the group to which the user belongs.

1) Generation of Recommendation: The essential utility of recommendation system is generating recommendations using some refining parameters as value, brand rating and so on. Refining parameters to produce support exact set of prescribed elements of the information base. To produce summary of the Boyer-Moore proposed use pattern matching algorithm.

Pattern search algorithm could be used to discover the elements of customer interest focused around exercises current customers to anticipate and suggest future customer appeal. Pattern search algorithm, Boyer-Moore used within the recommendation piece of architecture.

3.2. Bitap Algorithm

The Bitap algorithm (also known as displacement or instead or algorithm-and Baeza-Yates-Gonnet) is a fuzzy matching algorithm chain. The algorithm says if a given text contains a substring that is "approximately equal" to a given pattern, where the approximate equality is defined in terms of Levenshtein distance - whether the substring and pattern are at a distance k given another, then the algorithm considers equal. The algorithm begins by precomputing a set of bit masks containing one bit for each element of the pattern. Then he is able to do most of the work with bitwise operations, which are extremely fast.

The Bitap algorithm is perhaps best known as one of the underlying algorithms UNIX grep utility, written by Udi Manber, Sun Wu, and Burra Gopal. Manber original and Wu document provides extensions algorithm to address general fuzzy matching regular expressions.

Because the data structures required by the algorithm performed best patterns within a constant length (typically the word length of the machine in question), and inputs also preferred over a small alphabet. Once you applied to a given alphabet and word length m, however, its runtime is completely predictable - it runs in O (mn) operations, regardless of the structure of the text or pattern.

The bitap algorithm for exact string searching was invented by Balint Domolki in 1964 [10] and extended by RK Shyamasundar in 1977 [11], before being reinvented in the context of fuzzy string search Manber and Wu in 1991 [12] based on the work done by Ricardo Baeza-Yates and Gaston Gonnet. The algorithm was improved by Baeza-Yates and Navarro in 1996 [13] and later by Gene Myers for long patterns in 1998 [14].

3.2.1. 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
```

3.2.2 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. ALGORITHMS USED

4.1. 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.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

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 $E = \Sigma k\ 1\sigma n \in sjd\ (xn, \lambda j)$, 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. CONCLUSION

This Online Wav 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.

REFERENCES

1. 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.
3. R.S. Boyer, J.S. Moore, "A fast string searching algorithm, "Communication of the ACM, Vol. 20, No. 10,1977, pp.762– 772.
4. R. N. Horspool, "Practical fast searching in strings," Software—Practice and Experience, Vol. 10, No. 3, 1980, 501–506.
5. Sunday, D.M., "A very fast substring search algorithm," Communications of the ACM, Vol. 33, No. 8, 1990, pp. 132- 142.
6. Smith, P.D., "Experiments with a very fast substring search algorithm," Software-Practice and Experience, Vol. 21, No. 10, pp.1065-1074.
7. 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.
8. 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 Hadi Khosravi Farsani, and Mohammadali Nematbakhsh "A Semantic Recommendation Procedure for Electronic Product Catalog", World Academy of Science, Engineering and Technology 22 2006.
9. 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.
10. Balint Domolki, An algorithm for syntactical analysis, Computational Linguistics 3, Hungarian Academy of Science pp. 29–46, 1964.
11. R. K. Shyamasundar, Precedence parsing using Domolki's algorithm, International Journal of Computer Mathematics, 6(2)pp 105–114, 1977.
12. Udi Manber, Sun Wu. "Fast text searching with errors." Technical Report TR-91-11. Department of Computer Science, University of Arizona, Tucson, June 1991.
13. R. Baeza-Yates and G. Navarro. A faster algorithm for approximate string matching. In Dan Hirchsberg and Gene Myers, editors, Combinatorial Pattern Matching (CPM'96), LNCS 1075, pages 1–23, Irvine, CA, June 1996.
14. G. Myers. "A fast bit-vector algorithm for approximate string matching based on dynamic programming." Journal of the ACM 46 (3), May 1998, 395–415.
15. Xavier Amatriain, Alejandro Jaimes, Nuria Oliver, and Josep M. Pujol, Data Mining Methods for Recommender Systems.
16. R.Suguna, D, Sharmila, "Clustering Web log Files - A
17. Review", International Journal of Engineering Research &
18. Technology (IJERT) Vol. 2 Issue 4, April – 2013 ISSN: 2278-0181]

Dipali Wankhede is pursuing her Masters in Engineering from MSSCET Jalna. Her hobbies include reading books and listening music. Special Thanks to Principal Dr. S.K.Biradar & HOD Prof. G.P. Chakote

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)
© Copyright: All rights reserved.