Survey and Analysis of Searching Algorithms

Tahira Mahboob[1]
Assistant Professor

Fatima Akhtar[2]

Moquaddus Asif [3]

Nitasha Siddique[4]

Bushra Sikandar[5]

[1, 2 ,3, 4] Department of Software Engineering
[5] Department of Computer Science
[1, 2 ,3, 4, 5] Fatima Jinnah Women University, Pakistan

ABSTRACT: ***The data mining is vast field with application found in many areas such that science, industrial problems and business. The data mining system architecture has the several main components database, data warehouse, or other information repositories, a server that extract the related data from repositories based on user request. The classification tasks in massive data set have been done by Varity of proposed algorithm. Majority of algorithm has proved effective but not all of them are easily extensible and flexible. This research introduce various approaches such to find useful result. This can be overcome by grouping of similar so clustering is a distinct problem. Then there is full-text search in large text collections which is robust against errors on both query and documents side. The paper discusses content-based recommendation systems that recommend an item to a user based upon a description of the item and a profile of the user's interests. The paper also uses improved Boyer Moore Horspool algorithm for evaluation of Enhanced pattern matching performance and lastly Association rules mining that finds interesting relations or associations relations between the item sets among massive amount of data.***

KEYWORDS-Apriori,Association rule, FORBD Dynamic, Boyer Moore Horspool algorithm, Extensional definition of dynamic association rules (EDAR), DARPA (Defense Advanced Research Projects Agency),Partial Decision Tree (PART,Evolutionary Algorithms (EAs),data mining; evolutionary algorithms; rule-based classifiers; decision trees

I- INTRODUCTION

Data mining is a process of discovering interesting and hidden patterns from huge amount of data where data is collected in data warehouse such as on line analytical process, databases and other information repositories .Data mining is a knowledge discovery in databases. The data mining consist of an integration of techniques from different disciplines like database technology, statistics, neural networks, information retrieval and machine learning.

A search engine return results that has a problem of finding useful result this can be overcome by grouping of similar documents in a search results list so clustering is a distinct problem. Lingo algorithm that is presented is a novel algorithm for clustering search results, which focus on quality of cluster description. There is also problem of full-text search in large text collections which is robust against errors on both query and documents side. This paper discusses content-based recommendation systems that recommend an item to a user based upon a description of the item and a profile of the user's interest s. may variety of domains used Content-based recommendation systems ranging for recommending web pages, research papers, articles, and items for sale. Various systems are different in details but share a common means to describe the items that may recommend an article by creating a profile of the user that describes the types of items the user Likes or dislikes. The paper also uses improved Boyer Moore Horspool algorithm for evaluation of Enhanced pattern matching performance. It combines the deterministic finite state to match information to skip several characters. It best fit in cases that contain lots of characters sets. Another algorithm is Association rules mining that finds interesting relations or associations relations between the item sets among massive amount of data.. The association rule produces candidate frequent item sets based on equivalence class and equivalence that reduces the system expense.

The core problem of mining association rules is how to form association rule whose value of confidence and support is no less the user specified minimum confidence and minimum support respectively. The

most challenging factor in association rules mining is frequently mining pattern. Different searching methods and various types of techniques have been used to increase the performance of searching such as clustering, association, Rule based algorithms and tree based classifiers. The degree to find suitable solution for each time does not guarantee by insatiable searching method

ALGORITHMS FOR SEARCHING INFORMATION FROM DATABASE

I. An Improved Apriori Algorithm for Association Rules

Association rule has several mining algorithms. The Apriori algorithm is most important one. The Apriori algorithm is used to extract frequent itemset from large dataset. The association rule is also defined for these item sets for discovering the knowledge. Based on this algorithm, this paper presents the limitation and improvement of Apriori algorithm. Apriori algorithm wasting a lot of time for scanning the whole database searching on frequent item sets. The improved Apriori algorithm by reducing the number of transaction to be scanned reduces the time consumed in transaction scanning for candidate itemsets. From the view of time consumed whenever m of m-item set increases, performance gap between original Apriori and the improved Apriori increases and whenever the minimum support value increases, the gap between original Apriori and the improved Apriori decreasesThe paper shows by experimental results with various groups of transactions, and with various values of minimum support that applied on the original Apriori and improved Apriori, that improved Apriori reduces the time consumption by 67.38% in comparison with original Apriori. The improvement makes the Apriori algorithm more efficient and less time consuming.

II. Improved Algorithms Research for Association Rule Based on Matrix

Mining the association rules are of great importance, it is applied widely in the field of data mining. The working efficiency of mining algorithm of association rules becomes very important because of the huge scale of event database mined. Although Apriori algorithm in association rule uses cut-technology when it produces the candidate item sets , while scanning the transaction database each time it has to scan the whole database. For massive data the scanning speed is very slow. The Apriori algorithm basic approach is transforming the event database into database of matrix so as to get the matrix item set of maximum item set. The matrix based algorithm only scan the database once and then convert the whole event database into matrix database.When finding the frequent m-item set from the frequent m-item set, only its matrix set is found. So to get frequent k item set only corresponding data are calculated. That's why the computing time of improved Apriori algorithm is very fast. The speeds of the improved Apriori algorithm based on matrix and the original Apriori algorithm based on association is compared by simulation data. The efficiency of improved Apriori algorithm is proved by experiments.

Support(Ij ,Ik)=($\sum_{i=1}^{m} Mat(Ij, Ik)$ *) / m*

III. An Approach for Finding Optimized Rules Based on Dynamic-Characteristics

In data mining important research field is mining association rules. The traditional algorithms of association rules consider the effectiveness of rules in the database, and pay less attention to relevant dynamic information between rules and time, and the changing trend of the rules over time. The traditional association rules algorithms are lack of timeliness and predictability. Dynamic association rules are studied according to the time characteristics of rules in the database. The definition of FORBD [1] is extended by defining the confidence information gain and support information gain. Overall change trends of rules with time are reflected by support information gain and confidence information gain .The optimized rules can be finding by using an approach based on dynamic characteristic.

$$\mathrm{conf}(X \Rightarrow Y) = \mathrm{supp}(X \cup Y)/\mathrm{supp}(X)$$

IV. A Comparison between Rule Based and Association Rule Mining Algorithms

Recently the data mining problem has been solved by using association rule mining algorithm in a very efficient manner. Rule based mining can be implemented both by supervised learning or unsupervised learning techniques. can be compared by comparing Apriori of association rule mining and PART (Partial Decision Among the vast range of available approaches, it is always challenge to select the suitable algorithm for rule based mining task. Rules in PART are based on class attribute. PART has selected up more classes than Apriori .The core idea of this research is to do comparison between the

performance association rule mining algorithm and the rule based classification. Their comparison is based on their rule computational complexity and based classification performance. The performance of association rule mining algorithm and the rule based classification Tree)[2] of classification algorithmThe performance of these two algorithms is used to measure by the DARPA (Defense Advanced Research Projects Agency) [2] data, is a well known intrusion detection problem. The training rules are compared with already defined test sets. Apriori is a better choice in terms of accuracy and computational complexity. The Apriori algorithm requires less computational time. But the rules relating to class attribute does not produce by Apriori each time. If such features are included with Apriori , its performance increases for rule based classification.

V. Evolutionary Data Mining Approaches for Rule-based and Tree-based Classifiers

The algorithms used in this paper focus on synthesizing classifiers with Evolutionary Algorithms (EAs)[3] in supervised data mining .The first one method that is based on encoding rule sets with bit string genomes and other one use Genetic Programming to build the decision trees with arbitrary expressions connected to the nodes. This technique has been compared to some standard. The comparison results show that the performance of the proposed classifiers can be very competitive. In different configuration of the EAs both approaches work well. This algorithm outperformed the other algorithm in at least one area by obtaining highest precision.

VI. Lingo: Search Results Clustering Algorithm: Based on Singular Value Decomposition

In Vector Space Model (VSM) which is a procedure of information retrieval , linear algebra operations are used to calculate similarities among the unique documents that transforms the difficulty of comparing textual data into a crisis of comparing algebraic vectors in a multidimensional space. Once the modification is done, every unique term (word) from the compilation of analyzed documents forms a separate aspect in the VSM and each document is represented by a vector spanning all the factors.
Lingo reverses the process to avoid the problem of recurring ordered sequences of expressions , it first ensure that humans can create a human-perceivable cluster label and only then allocate documents to it. It extracts recurring phrases from the input documents, in suspense that they are the most informative source of human-readable topic descriptions. Next, by performing decrease of the unique term-document matrix using SVD, lingo discovers any existing latent structure of varied topics in the search result. Finally, in research paper we harmonize group descriptions with the extracted topics and assign related documents to them. The clusters are sorted for show, designed using the following simple formula based on their score: Cscore = label score× kCk, where kCkis the number of documents given to clusterC.

VII. Occurrences Algorithm for String Searching Based on Brute-force Algorithm

Brute-force Algorithm compares pattern and text character by character; until a match is found or the end of the text is reached the pattern is shifted one location to the right and comparison is repeated . The algorithm computes with two pointers; a “text pointer” i and a “pattern pointer” j. pattern and text are compared for all (n-m) suitable shifts, the pattern pointer is incremented while text and pattern characters are equal. If a difference occurs, i is incremented, j is reset to zero and the comparing process is restarted. The algorithm gives the position of the pattern if match is found, if not, it returns not found message. It consists of three steps:
Preprocessing the pattern algorithm calculates the number of occurrences and number of repetitions, for each character in the example. The algorithm finds the character that is of the highest occurrence in the pattern.
Preprocessing the text stores the segment index in an array by finding the character that is of the highest occurrence in the pattern by character found in the previous process of highest number of occurrences, then calculate the number of occurrences of that character in the segment. Searching algorithm compare pattern and the segments that their indexes are stored in the array. The algorithm depends on the first character in the pattern to search, when the pattern does not include a Repetitive character. Searching has same method as searching a pattern having characters repeated.

VIII. Content-based Recommendation Systems

Based upon a kind of the item and a profile of the user’s interests Content-based recommendation

systems recommend an item to a user. User profile may be entered by the user, but is usually learned from feedback the user gives on items. A diversity of learning algorithms have been modified to learning user profiles, and the preference of learning algorithm depend upon the content. Reviewing a number of classification l earning algorithms are the key component of content- based recommendation systems, as they study a function that models each user's interests. The function predicts the user's interest in the item by giving a new item in the user model. Probability may be used to sort a list of recommendations by creating a function that will present an guess of the probability that a user will like an unseen item or not. This algorithm create a function that directly predicts a numeric value such as the degree of interest

IX. Enhanced Pattern Matching Performance Using Improved Boyer Moore Horspool Algorithm

This paper use information matching to skip several characters. It proposes Improved Boyer-Moore-Horspool Algorithm a new multiple patterns identical algorithm, That combines deterministic finite state automata (DFSA) with MBMH algorithm. Not only realizes multiple patterns accuracy matching, but also immediately shift by using bad character heuristic. It increases matching speed, when form string character sets are less than text string character sets. The strings aligns substrings of text string ti ti+1…. ti+m-1 according to t i+m and ti+m-1 ascertain shift distance when mismatching occur in string and pattern. It has Need to build two shift tables, set up shift and shift0. Structure processing of shift table is the same as modified table of BMH algorithm. If ti+m-1 appear the first position from right to left in the substrings of sample strings (p0 p1p2 p m-2) ascertain shift value it then shift pattern strings to right, make ti+m align the first same character of sample strings. If disappear, then shift pattern strings to right m+1 characters distances.

X. Efficient Fuzzy Search in Large Text Collections

A matching query q , a threshold, and a dictionary of words W are given in Fuzzy word / auto completion algorithm. Its formula is:$LD(q;w) \leq \rho$, where LD is the word prefix Levenshtein distance that effectively find all words. It is a new methods that permit query suggestions based on the contents of the document collection instead of on pre-compiled lists .Fuzzy word matching problem has two practical algorithms with different trade-offs. The first algorithm is based on a method called truncated deletion neighborhoods that allows an algorithm with index that retains most of the effectiveness of deletion neighborhood-based algorithms, it is particularly efficient on short words. The second algorithm is particularly efficient on large words by making use of a signature based on the largest common substring between two words. Instead of q-gram indexes, our algorithm is based on permuted glossary, providing entrance to the word list via cyclic substrings of random lengths that can be computed in steady time.

Table-1 indicated the parameter selected for comparison and evaluation. With typical values of Y:Yes, N:No and ND: Not defined. Table-2 covers the analysis of these parameter selected in Table-1.

TABLE-1: QUALITY ASSURANCE PARAMETER, MEANING AND POSSIBLE VALUES

Serial NO	Evaluation Parameter	Meaning	Possible Values
	Quality Parameters		
1	Performance	The task accomplishment against accuracy, cost, speed and completeness that are known standards for evaluation.	High, Low
2	CPU utilization	Computer's consumption for processing resources is CPU utilization	Y,N,ND
3	Cost Effectiveness	Increase production reference to cost.	Y,N,ND
4	Memory space	When services are assigned with usage of physical memory its Memory is allocated.	Y,N,ND
5	Automatable	By increasing workload, system manage to give maximum performance	Y,N,ND
6	Efficiency	Resources consumed by the algorithm	Y,N,ND
7	Effectiveness	Ability to produce a preferred result.	Y,N,ND
8	Query Processing time	The effective process to run query by considering all query plans	High, Low
10	Robust against mistakes	System performance that is not cover by specification	Y,N,ND
11	Precision	The ability of a system to avoid change without considering initial stable state.	Y,N,ND
12	Time complexity	The amount of time taken by algorithm to run the input string function.	Y,N,ND
13	Computational complexity	The crisis situations that are solved by mechanical algorithms	Y,N,ND
	Machine learning Parameters		
14	Receiver Operating Characteristic (ROC) graph	It is technique to organize, visualize, and select classifiers that depend on their performance in 2D space.	Y,N,ND
15	Accuracy	It is the ratio of the number of correct predictions and the total number of predictions.	Y,N,ND
16	Error rate	It measures the total number of incorrect predictions against the total number of predictions	Y,N,ND
17	Precision	It is defined where datasets are much unbalanced.	Y,N,ND
18	Recall	It is the proportion of the number of data items that system selected as the positive.	Y,N,ND
19	F1-Score	For optimization F1 score combines both recall and precision with equal importance into a one parameter.	Y,N,ND

Note: Y: Yes, N: No, ND: Not defined

TABLE-2: ANALYSIS TABLE OF PARAMETERS AGAINST AUTHORS

Serial #	*Authors*	*ROC Graph*	*Accuracy*	*Error rate*	*Precision*	*Recall*	*F1-score*	*Performance*	*Reduce time*	*Cost*	*Computational*	*Automatable*	*Robustness*	*Cost effective*	*Performance*	*CPU utilization*	*Memory space*	*Efficiency*	*Effectiveness*	*Query pros. t*
1	D. Hong	Y	Y	ND	N	N	N	H	Y	Y	ND	N	L	ND	Y	Y	N	N	N	L
2	A. Mohammad , O. Saleh,R. A. Abdeen	N	ND	Y	N	N	ND	H	Y	Y	Y	N	Y	N	Y	Y	N	N	N	H
3	M. J. Pazzani , D. Billsus	Y		ND	N	N	N	ND	Y	Y	Y	ND	N	Y	Y	N	N	Y	Y	H
4	H. Bast,M. Celikik	Y	Y	N	N	N	N	H	Y	Y	Y	ND	N	Y	Y	Y	N	Y	N	L
5	I. Osi´nski, J. Stefanowski, D. Weiss	N	N	Y	ND	N	N	ND	Y	ND	Y	N	N	N	ND	Y	ND	Y	ND	L
6	H.M.Najadat, M.Al-Malogei B.Arkok	Y	ND	Y	N	N	N	ND	N	Y	Y	ND	Y	Y	Y	N	N	N	H	H
7	L.X.Wen and W.WeiQing	Y	Y	Y	N	N	N	L	Y	ND	Y	N	N	Y	ND	Y	ND	Y	Y	Y
8	J H.Mei and J.Zhou	Y	Y	Y	N	N	N	H	Y	Y	Y	ND	ND	Y	Y	Y	Y	ND	N	L
9	M.M.Mazid and A.S.Ali	Y	Y	Y	N	N	N	H	Y	Y	Y	N	N	N	Y	Y	Y	ND	N	L
10	T.Weise and R.Chiong	Y	Y	Y	N	N	N	H	Y	Y	Y	ND	N	ND	Y	N	N	Y	Y	L

CONCLUSION

Search engines are designed to be a scalable search engine. Their primary goal is to provide high quality search results over a rapidly growing World Wide Web. The searching techniques used nowadays gives major results over the last decade when numerous content-based, collaborative, rules based, Associations and clustering were planned and several "Search Engines" have been developed. But still, the recent generation of search systems surveyed in this paper still requires further enhancements to make Searching methods more useful in a wider range of applications.

In the Future the problems presented in this paper would advance and continue the discussion in community about the next generation of Search engines. In this paper, we evaluate these systems on the basis of evaluation criteria of the current Search algorithms and review possible extensions that can provide improved capabilities in future research.

ACKNOWLEDGMENT

We would like to thank our Supervisor and friends for their encouragement and moral support given to us. We would also like to extend our thanks to our family members for all the help, direct or indirect.

REFERENCES

[1]. H.M.Najadat, M.Al-Malogei and B.Arkok. "An Improved Apriori Algorithm for Association Rules." *International Research Journal of Computer Science and Application,* Vol. 1, No. 1, June 2013.Available: arxiv.org/pdf/1403.3948

[2]. L.X.Wen and W.WeiQing. "Improved Algorithms Research for Association Rule Based on Matrix." *International Conference on Intelligent Computing and Cognitive Informatics,* 2010. Available: http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=5565944&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D5565944

[3]. H.Mei and J.Zhou. "An Approach for Finding Optimized Rules Based on Dynamic Characteristics." *International Journal of Advancements in Computing Technology(IJACT),Volume4, Number14, August 2012.* Available: www.aicit.org/IJACT/ppl/IJACT1174PPL.pdf

[4]. M.M.Mazid and A.S.Ali. "A Comparison Between Rule Based and Association Rule Mining Algorithms." *Third International Conference on Network and System Security, 2009.* Available: http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=5319344&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D5319344

[5]. T.Weise and R.Chiong. "Evolutionary Data Mining Approaches for Rule-based and Tree-based Classifiers." *Proc. 9th IEEE Int. Conf. on Cognitive Informatics (ICCI'10),*2010IEEE. Available: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=5599821

[6]. D. Hong." Enhanced Pattern Matching Performance Using Improved Boyer Moore Horspool Algorithm." *Journal of Convergence Information Technology,* vol. 7, Num. 4, (issue4.9), Mar 2012 .Available: www.aicit.org/JCIT/ppl/JCIT%20Vol7%20No4_part9.pd

[7]. Mohammad , O. Saleh and R. A. Abdeen. "Occurrences Algorithm for String Searching Based on Brute-force Algorithm." *Journal of Computer Science* , pp. 82-85, 2006. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.165.7421&rep=rep1&type=pdf

[8]. M. J. Pazzani and D. Billsus "Content-based Recommendation Systems.".Internet: http://citeseerx.ist.psu.edu/viewdoc/download doi=10.1.1.130.8327&rep=rep1&type=pdf

[9]. H. Bast and M. Celikik." Efficient Fuzzy Search in Large Text Collections*." ACM Transactions on Information Systems*, Vol. 9, No. 4, Article 39,Mar. 2010 .Available: http://ad-publications.informatik.uni-freiburg.de/TOIS_fuzzy_BC_2013.pdf

[10]. Osinski, J. Stefanowski, and D. Weiss. "Lingo: Search Results Clustering Algorithm Based on Singular Value Decomposition.". Internet: http://www.researchgate.net/publication/221501367_Lingo_Search_Results_Clustering_Algorithm_Based_on_Singular_Value_Decomposition