International Journal of Computing and Corporate Research
ISSN (Online) : 2249-054X
Volume 6 Issue 3 May 2016
International Manuscript ID : 2249054XV6I3052016-409

IMPROVED TWO LEVEL RULE MINING APPROACH WITH GENETIC ALGORITHM FOR GLOBAL OPTIMIZATION

Dr. Vijayaraghavan.A
Professor & Head
Dept. of CSE
HMS Institute of Technology
Tumkur, Karnataka

ABSTRACT

Recommender systems or recommendation systems (RS) refers to a subclass of information filtering system that seek to predict the 'rating' or 'preference' that a user would give to an item. Using Recommender Systems, the effective suggestions of simply recommendations can be given to the user in interest. For example, in case online shopping the user can be give the recommendations for purchase depending upon the historical shopping behavior. A number of algorithms were devised so far but still there is lots of scope of research. In classical way, the statistical methods are used for providing the recommendations. In this research work, the integration of advance data mining tools shall be implemented to optimize the results on cost optimization. Upto now, the cost optimization is not addressed by any research work in recommender systems. This work focus on the implementation of two rule mining approaches and then refining the results using GA.

Keywords – Recommender System, Nature Inspired Algorithms, Cost Optimization

RECOMMENDER SYSTEMS

Recommendation systems changed the way inanimate websites communicate with their users. Rather than providing a static experience in which users search for and potentially buy products, recommender systems increase interaction to provide a richer experience.

Recommender systems identify recommendations autonomously for individual users based on past purchases and searches, and on other users' behavior.

Most recommender systems take either of two basic approaches: collaborative filtering or content-based filtering. Other approaches (such as hybrid approaches) also exist.

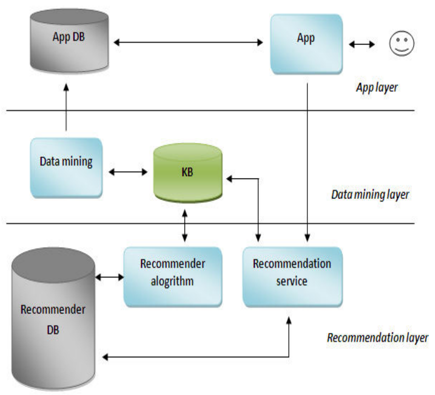

Figure 1 - Architectural Approach

Collaborative filtering

The new developer Works Premium membership program provides an all-access pass to powerful development tools and resources, including 500 top technical titles (dozens specifically for open source developers) through Safari Books Online, deep discounts on premier developer events, video replays of recent O'Reilly conferences, and more.

Collaborative filtering arrives at a recommendation that's based on a model of prior user behavior. The model can be constructed solely from a single user's behavior or — more effectively — also from the behavior of other users who have similar traits. When it takes other users' behavior into account, collaborative filtering uses group knowledge to form a recommendation based on like users. In essence, recommendations are based on an automatic collaboration of multiple users and filtered on those who exhibit similar preferences or behaviors.

For example, suppose you're building a website to recommend blogs. By using the information from many users who subscribe to and read blogs, you can group those users based on their preferences. For example, you can group together users who read several of the same blogs. From this information, you identify the most popular blogs that are read by that group. Then — for a particular user in the group — you recommend the most popular blog that he or she neither reads nor subscribes to.

Content-based filtering

Content-based filtering constructs a recommendation on the basis of a user's behavior. For example, this approach might use historical browsing information, such as which blogs the user reads and the characteristics of those blogs. If a user commonly reads articles about Linux or is likely to leave comments on blogs about software engineering, content-based filtering can use this history to identify and recommend similar content (articles on Linux or other blogs about software engineering). This content can be manually defined or automatically extracted based on other similarity methods.

Hybrids

Hybrid approaches that combine collaborative and content-based filtering are also increasing the efficiency (and complexity) of recommender systems. A simple example of a hybrid system could use the approaches shown in Figure 1 and Figure 3. Incorporating the results of collaborative and content-based filtering creates the potential for a more accurate recommendation. The hybrid approach could also be used to address collaborative filtering that starts with sparse data — known as *cold start*— by enabling the results to be weighted initially toward content-based filtering, then shifting the weight toward collaborative filtering as the available user data set matures.

Algorithms that recommender systems use

As demonstrated by the winning approach for the Netflix prize, many algorithmic approaches are available for recommendation engines. Results can differ based on the problem the algorithm is designed to solve or the relationships that are present in the data. Many of the algorithms come from the field of machine learning, a subfield of artificial intelligence that produces algorithms for learning, prediction, and decision-making.

Pearson correlation

Similarity between two users (and their attributes, such as articles read from a collection of blogs) can be accurately calculated with the *Pearson correlation*. This algorithm measures the linear dependence between two variables (or users) as a function of their attributes. But it doesn't calculate this measure over the entire population of users. Instead, the population must be filtered down to *neighborhoods* based on a higher-level similarity metric, such as reading similar blogs.

The Pearson correlation, which is widely used in research, is a popular algorithm for collaborative filtering.

Clustering algorithms

Clustering algorithms are a form of unsupervised learning that can find structure in a set of seemingly random (or unlabeled) data. In general, they work by identifying similarities among items, such as blog readers, by calculating their distance from other items in a *feature space*. (Features in a feature space could represent the number of articles read in a set of blogs.) The number of independent features defines the dimensionality of the space. If items are "close" together, they can be joined in a cluster.

Many clustering algorithms exist. The simplest one is *k*-means, which partitions items into *k* clusters. Initially, the items are randomly placed into clusters. Then, a *centroid* (or *center*) is calculated for each cluster as a function of its members. Each item's distance from the centroids is then checked. If an item is found to be closer to another cluster, it's moved to that cluster. Centroids are recalculated each time all item distances are checked. When stability is reached (that is, when no items move during an iteration), the set is properly clustered, and the algorithm ends.

Calculating the distance between two objects can be difficult to visualize. One common method is to treat each item as a multidimensional vector and calculate the distance by using the Euclidean algorithm.

Other clustering variants include the Adaptive Resonance Theory (ART) family, Fuzzy C-means, and Expectation-Maximization (probabilistic clustering), to name a few.

Other algorithms

Many algorithms — and an even larger set of variations of those algorithms — exist for recommendation engines. Some that have been used successfully include:

- **Bayesian Belief Nets**, which can be visualized as a directed acyclic graph, with arcs representing the associated probabilities among the variables.
- **Markov chains**, which take a similar approach to Bayesian Belief Nets but treat the recommendation problem as sequential optimization instead of simply prediction.
- **Rocchio classification** (developed with the Vector Space Model), which exploits feedback of the item relevance to improve recommendation accuracy.

Challenges with recommender systems

Taking advantage of the "wisdom of crowds" (with collaborative filtering) has been made simpler with the data-collection opportunities the web affords. But the massive amounts of available data also complicate this opportunity. For example, although some users' behavior can be modeled, other users do not exhibit typical behavior. These users can skew the results of a recommender system and decrease its efficiency. Further, users can exploit a recommender system to favor one product over another — based on positive feedback on a product and negative feedback on competitive products, for example. A good recommender system must manage these issues.

One problem that's endemic to large-scale recommendation systems is scalability. Traditional algorithms work well with smaller amounts of data, but when the data sets grow, the traditional algorithms can have difficulty keeping up. Although this might not be a problem for offline processing, more-specialized approaches are needed for real-time scenarios.

Finally, privacy-protection considerations are also a challenge. Recommender algorithms can identify patterns individuals might not even know exist. A recent example is the case of a large company that could calculate a pregnancy-prediction score based on purchasing habits. Through the use of targeted ads, a father was surprised to learn that his teenage daughter was pregnant. The company's predictor was so accurate that it could predict a prospective mother's due date based on products she purchased.

DATA MINING AND MACHINE LEARNING

Data mining refers to the analysis of the large quantities of data that are stored in computers. Data mining is known as exploratory data analysis. Masses of data generated from cash registers, from scanning, from topic specific databases throughout the company, are explored, analyzed, reduced, and reused. Searches are performed across different models proposed for predicting sales, marketing response, and profit. Classical statistical approaches are fundamental to data mining. Automated Artificial Intelligence (AI) methods are also used.

Data mining requires identification of a problem, along with collection of data that can lead to better understanding and computer models to provide statistical or other means of analysis. It is integrated and placed in some common data store.

Part of it is then taken and pre-processed into a standard format. This ‘prepared data’ is then passed to a data mining algorithm which produces an output in the form of rules or some other kind of ‘patterns’.

Knowledge discovery in databases (often called data mining) aims at the discovery of useful information from large collections of data.

Knowledge discovery in databases is interactive and iterative process with several steps and data mining is a part of this process.

CLASSICAL APPROACHES

There are many Data Mining algorithms to mine frequent patterns for finding association rules. The two widely used algorithms are FP-Growth and Apriori.

- **FP Growth**: Frequent pattern growth a very popular association rule mining algorithm for discovering itemsets in a database. The algorithm follows two step approaches for finding interesting rules. The Step1 of the algorithm builds a tree known as FP tree and in step2 frequent items are extracted from this FP tree. FP Growth algorithm is a 2-pass algorithm over database. Where one side FP Growth does not generates any candidate sets and thereby it is considered to be fastest than Apriori, the other side the drawbacks with this algorithm comes up in the form of expensive tree building and the uncertainty of fitting FP tree in memory.

- **Apriori:** Apriori Algorithm is a decisive algorithm for mining frequent itemsets for Boolean association rules. It uses prior knowledge of frequent itemset properties. Apriori employs an iterative approach known as a level-wise search, where k-itemsets are used to explore (k+1) itemsets. First the set of frequent 1 itemsets is found by scanning the database to accumulate the count for each item, and collecting those items that satisfy minimum support. The resulting set is denoted L1. Next L1, is used to find L2, the set of frequent 2- itemsets, which is used to find L3, and so on, until no more frequent k-itemsets can be found.

APRIORI ALGORITHM

Apriori algorithm is a classical algorithm used in data mining for learning association rules. It is the procedure for finding useful and potential knowledge in database. Association rules are associated with the prominent knowledge of data mining and results that can be defined as the relations and dependency between the data items with the usage of support and confidence. The core idea of the Apriori is scanning the database repeatedly. With the paradigm that the subset of the frequent data items are frequent patterns that can be gained with the length of frequent (k+1)-itemsets L_{k+1} from the frequent k-itemsets L_k. At the k time it scans the database only the candidate items C_{k+1} that generates from the L_k was concerned. Further, the appearance time of the C_{k+1} can be verified by another scanning database. The main idea of these algorithms is according the theory that the subset of frequent items is a frequent set and the superset of a infrequent set is an infrequent itemset.

APPLICATIONS OF ASSOCIATION RULES MINING

Association rule mining find applications in number of Business and Individual Intelligence areas. Following are the most widely used applications:

- Super-markets: Shopping centers use association rules to place the items next to each other so that users buy more items.
- Online shopping: Amazon use association mining to recommend the items based on the current item being bought.
- Web Search: The Google search engine has a functionality of auto-complete, where after typing a word it searches frequently associated words that user types after that particular word using association mining.

LITERATURE REVIEW

Base (2015) - Association Rule Mining technique that attempt to unearthing interesting pattern or relationship between data in large Database. Genetic Algorithm is a search heuristic which is used to generate useful solution for optimization and search problems. Genetic Algorithm based evaluation in Mining Technique is backbone for mining interesting

Rule based on GA parameters like fitness function, Crossover Rate, Mutation Rate. The key focus of this synthesize approach is to optimize the rule that generated by mining methodology and to provide more accurate results. The Proposed Approach is to generate rules based on Quantitative dataset, using the concept of threshold - frequent item sets are define as initial population which the first step of Genetic algorithm. Crossover & mutation is applied to generate more combination of rule & can identify Co-occurrence of item sets.

Kumar et al. [1] implements three phases of Web usage mining namely preprocessing, pattern discovery, and pattern analysis. Apriori algorithm is used to generate an association rule that associates the usage pattern of the clients for a particular website. The output of the system was in terms of memory usage and speed of producing association rules. A clustering algorithm to find out data clusters for both numerical and nominal data is proposed by Sharma et al. [2] by calculating the average and log values of data set. This algorithm improves the techniques of Web Usage Mining by first discover the log files of individual users at one place.

Martinez-Romo et al. [3] have analyzed different information retrieval methods for both, the selection of terms used to construct the queries submitted to the search engine, and the ranking of the candidate pages that it provides, in order to help the user to find the best replacement for a broken link. To test the sources, they have also defined an evaluation methodology which does not require the user judgments, what increases the objectivity of the results.

A new reactive session reconstruction method is given by Dohare et al. [4]. This algorithm is better than previously developed both time and navigation oriented heuristics as it does not allow page sequences with any unrelated consecutive requests to be in the same session. They have also implemented agent simulator for generating real user sessions. Das et al. [5] analyzed the web server user access logs of Firat University to help system administrator and Web designer to improve their system by determining occurred system errors, corrupted and broken links by using web using mining.

Fayyad et al. [6] have focused on web log file format, its type and location. Log files usually contain noisy and ambiguous data. Preprocessing involves removal of unnecessary data from log file. Data preprocessing is an important step to filter and organize appropriate information before using to web mining algorithm. They have also proposed two algorithms for field extraction and data cleaning. Preprocessing web log file is used in data mining techniques, also used in intrusion detection system as input to detect intrusion. Apriori - the first scalable algorithm designed for association-rule mining algorithm. Apriori is an improvement over the AIS and SETM algorithms stated, Das et al. [5]. The algorithm is based on the large itemset property which states: Any subset of a large itemset is large and if an itemset is not large and then none of its supersets are large.

The Apriori algorithm searches for large itemsets during its initial database pass and uses its result as the seed for discovering other large datasets during subsequent passes Brachman et al. [7] specifies that the rules having a support level above the minimum are called large or frequent itemsets and those below are called small itemsets. Agrawal et al. [8] derived Association Rules from data called the "market-basket problem". Given a set of items and a large collection of transactions which are sets (baskets) of items from a database, the task is to find relationships between the containments of various items within those baskets. The task in Association Rule mining involves finding all rules that satisfy user defined constraints on minimum support and confidence with respect to a given dataset. Most commonly used Association Rule discovery algorithm that utilizes the frequent itemset strategy is exemplified by the Apriori algorithm.

Data mining (DM) a step from Knowledge Discovery in Database (KDD) process, Andrassyova et al. [9] defines it as a "nontrivial process of identifying valid, novel, potentially useful and ultimately understandable pattern in data". The term pattern here refers some abstract representation of a subset data of the data, that is, an expression in some language describing a data subset or a data subset or a model applicable to that subset.

Kleinberg [10] categorized web mining into three areas of interest based on which part of the Web to mine; Web Content mining, Web Structure mining, and Web Usage Mining. In Web mining, data collected at the server-side, client-side, proxy servers or a consolidated Web/business database. Hedberg [11] provided data sources that can be used to construct several data abstractions, namely users, page-views, click-streams and server sessions.

Tang et al. [12] have used re-ranking method and generalized Association Rules to extract access patterns of the Web sites pattern usage.

Genetic Algorithm

Genetic Algorithm (GAs) are request frameworks considering models of typical determination and inherited qualities (Fraser, 1957; Bremermann, 1958; Holland, 1975). We start with a brief preface to essential inherited counts and related wording.

GAs encode the decision variables of a request issue into restricted length arrangement of letters all together of certain cardinality. The strings which are contender responses for the chase issue are suggested as chromosomes, the letters all together are insinuated as qualities and the estimations of characteristics are called alleles. For example, in an issue, for instance, the voyaging deals agent issue, a chromosome identifies with a course, and a quality may identify with a city. Rather than standard improvement systems, GAs work with coding of parameters, rather than the parameters themselves.

To create extraordinary game plans and to complete trademark decision, we require a measure for perceiving incredible courses of action from dreadful game plans. The measure could be an objective limit that is a numerical model or a PC entertainment, or it can be a subjective limit where individuals pick better game plans over all the more horrendous ones. In a general sense, the wellbeing measure must choose a candidate course of action's relative health, which will thusly be used by the GA to control the improvement of good plans.

Another basic thought of GAs is the considered masses. Not in any manner like standard request procedures, genetic figurings rely on upon a masses of cheerful courses of action. The masses size, which is normally a customer showed parameter, is one of the basic variables impacting the flexibility and execution of genetic figurings. Case in point, little people sizes may incite unfavorable blending and yield substandard plans. On the other hand, tremendous masses sizes lead to unnecessary utilization of noteworthy computational time. When the issue is encoded chromosomally and a wellbeing measure for isolating incredible courses of action from horrendous ones has been picked, we can start to create answers for the interest issue using the going with steps:

Instatement. The beginning masses of contender courses of action is for the most part created erratically over the request space. Regardless, space specific learning or other information can be easily combined.

Evaluation. Once the people is instated or a children masses is made, the health estimations of the contender game plans are surveyed.

Determination. Determination administers more copies of those courses of action with higher wellbeing qualities and thusly drives the survival-of-the-fittest segment on the candidate plans. The essential considered determination is to incline toward better responses for more awful ones, and various decision routines have been proposed to accomplish this idea, including roulette-wheel decision, stochastic exhaustive decision, situating decision and rivalry decision, some of which are portrayed in the accompanying territory.

Recombination. Recombination solidifies parts of two or more parental responses for make new, possibly better game plans (i.e. family). There are various strategies for completing this (some of which are discussed in the accompanying section), and gifted execution depends on upon a suitably formed recombination framework. The family under recombination won't be vague to a particular parent and will rather unite parental qualities novelly (Goldberg, 2002).

Change. While recombination deals with two or more parental chromosomes, change locally however heedlessly changes an answer. Yet again, there are various assortments of change, yet it generally incorporates one or more changes being made to a solitary's trademark or qualities. By the day's end, change performs a subjective walk around the locale of a confident course of action.

Substitution. The family people made by determination, recombination, and change replaces the first parental masses. Various substitution frameworks, for instance, elitist substitution, period clever substitution and continuing state substitution schedules are used as a piece of GAs.

Goldberg (1983, 1999a, 2002) has contrasted GAs with foolish types of particular strategies for human improvement and has shown that these directors when examined only are lacking, yet when joined together they can work outstandingly. This edge has been elucidated with the thoughts of the key nature and advancement intuition. The same study contemplates a blend of determination and change to steady change (a sort of incline climbing), and the mix of decision and recombination to progression (crossfertilizing).

PROBLEM STATEMENT

- The classical approach of the Recommender Systems makes use of the Association Rule Mining and Correlation Approach
- In the proposed approach, there is need to improve and enrich the classical algorithms with higher level of accuracy and integrity
- There is need to propose and implement the work based on cost factor optimization.
- The integration of Statistical Methods are implemented in classical way making use of correlation in the attributes.
- There is need to devise a new algorithm that will optimize the results and recommendations based on the cost factor of recommended products or objects.
- There is need to address and avoid the issue of Cold Start Problem of Recommender Systems

RESEARCH OBJECTIVES

- Implementation of two rule mining algorithms for fetching the recommendations
- Reducing the results using genetic algorithm based on the fitness functions.
- To evaluate the models of recommender systems and present the detailed comparative analysis

- Till now in the existing approach, the frequent associated datasets are mapped.
- The existing / classical papers show the recommendations based on the interconnection of various objects rather than their associated cost
- In the classical / existing work, the cost factor is not considered
- It our proposed work, the cost factor associated with each recommender shall be executed and implemented
- In our new approach, the recommendations shall be shown to the user as well as organization in cost effective aspects
- The parameters in our research are
    - Execution Time
    - Complexity
    - Cost
    - Overall Performance of the Algorithm

PROPOSED OUTCOME

- Improved Recommender System in association with multiple Apriori Algorithm
- The results from multiple rule mining approaches shall be shortened and refined using genetic algorithm
- Implementation of the Proposed Approach in E-Commerce Application
- Integration of the implementation with real data set from Open Data Portals for research and development
- Development and Implementation of a Unique Algorithm for Improved Recommendations based on the Cost Factor

RESEARCH METHODOLOGY

DATA COLLECTION

- Generation of Own Data Set for Research

EXISTING APPROACH

- Implementation of the Existing Algorithm of Recommendations based on Association Rule Mining and Statistical Methods
- Fetching of Results and Graphs on Existing Approach

PROPOSED APPROACH

- Marking the Feature Point of Cost in the DataSet
- Implementation of Genetic Algorithm for refining the results of rule mining
- Integration of GA based approach for multi-level refinement from the results of two rule mining approaches.
- Implementation of Cost Based Association Rule Mining in the Proposed Approach
- Fetching of Results and Graphs on Existing Approach

COMPARISON OF RESULTS IN EXISTING AND PROPOSED APPROACH

- Comparative Analysis between Classical and Proposed Approach based on multiple parameters

- Cost Factor
- Execution Time
- Complexity

CONCLUSION

This research work is having focus on the cost optimization in recommender system which is not adopted so far. The proposed approach and architecture is effective in terms of cost and complexity which is better than the classical work.

REFERENCES

[1] Kumar, B.S. and Rukmani, K.V., 2010, "Implementation of Web Usage Mining Using APRIORI and FP Growth Algorithms" *International Journal of Advanced Networking and Applications*, Vol. 1, Issue 6, pp. 400-404.

[2] Sharma, P. and Bhartiya, R., 2011, "An efficient Algorithm for Improved Web Usage Mining" *International Journal of Computer Technology & Applications*, Vol. 3, No.2, pp. 766-769.

[3] Martinez-Romo, J. and Araujo, L., 2010, "Analyzing Information Retrieval Methods to Recover Broken Web Links", I*n Proceedings of the 32nd European Conference on Information Retrieval, ECIR 2010,* Milton Keynes, UK, pp. 26-37.

[4] Dohare, M.P.S., Arya, P. and Bajpai A., 2012, "Novel Web Usage Mining for Web Mining Techniques" *International Journal of Emerging Technology and Advanced Engineering*, vol. 2, Issue 1, pp. 253-262.

[5] Das, R., Turkoglu, I. and Poyraz, M., 2007, "Analyzing of System Errors for increasing a web server performance by using web usage mining", *Journal of electrical & electronics engineering*, Vol. 7, No. 2, pp. 379 – 386.

[6] Fayyad, U.M., Piatetsky-Shapiro, G. and Smyth, P., 1996, "The KDD Process for Extracting Useful Knowledge from Volumes of Data", *Communications of the ACM*, Vol. 39, No. 11, pp. 27-34.

[7] Brachman, R.J., Anand, T., 1996, "The Process of Knowledge Discovery in Databases", *Advances in Knowledge Discovery & Data Mining,* Fayyad, U.M. - Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R., Eds. AAAI/MIT Press, Cambridge, Massachusetts, pp. 37-57.

[8] Agrawal, R., Imielinski, T. and Swami, A., 1993, "Mining association rules between sets of items in large databases", *In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, Washington, D.C., pp 207-216.

[9] Andrassyova, E. and Paralic, J., 2000, "Knowledge Discovery in Databases: A Comparison of Different Views", *In Journal of information and organizational sciences,* Varazdin, Croatia, Vol. 23, No. 2, pp. 95 - 102.

[10] Kleinberg, J.M., 1999, "Authoritative sources in a hyperlinked environment", *Journal of the ACM*, Vol. 46*,* No. 5, pp. 604-632.

[11] Hedberg, S.R., 1996, “Searching for the mother lode: tales of the first data miners”, *IEEE EXPERT*, Vol. 11, No. 5, pp. 4-7.

[12] Tang, C., Lau R.W.H., Li, Q., Yi, H., Li, T. and Kilis, D., 2000, “Personalized Courseware Construction Based on Web Data Mining”, *In Proceeding of the First International Conference on Web Information Systems Engineering (WISE 2000),* vol. 2, pp. 204-211.

[13] A Context-Aware Framework for an Intelligent Mall based on Recommender System, Yuxiang Ye, Run Zhao and Dong Wang