A comprehensive and heuristic approach for Personalized Web Search using Greedy Algorithm

K.Saranya

M. Phil Scholar, School of Computer Science Engineering and Applications, Bharathidasan University, Tiruchirappalli, Tamil Nadu, India, ksaranyakarunanithi.sk@gmail.com

Abstract - *Personalized web search (PWS) used for developing the quality of various search services on the Internet. Users might experience failure when search engines return unrelated results that do not meet their real intentions. Such irrelevance is largely due to the huge variety of users' contexts and backgrounds, as well as the ambiguity of texts. However, evidences show that user's private information during search has become known to publicly due to proliferation of PWS. We proposed a PWS framework so-called UPS that can adaptively generalize profiles by queries while respecting user specified privacy requirements. This project presents two greedy algorithms, namely Greedy DP and Greedy IL, for runtime generalization. It also provides an online prediction mechanism for determining whether personalizing a query is beneficial. Rough set theory, which has been used victoriously in solving problems in pattern recognition, machine learning, and data mining, centers around the idea that a set of individual objects may be approximated via a lower and upper bound. In order to obtain the profits that rough sets can provide for data mining and related tasks, efficient computation of these approximations is vital. Compared with the classic Set theory, Rough Set is a mathematical approach to describe imprecision, vagueness, and ambiguity in data analysis, and it was earliest invented.*

Key Words: **Classifier Ensemble Selection, Rough Sets, Feature Selection, Harmony Search, Fuzzy-rough Sets.**

1. INTRODUCTION

Recommender systems can use data mining techniques in order to make recommendations using knowledge learnt from the action and attributes of users. The main aim of data mining is to discover new, interesting and useful knowledge or information using a variety of techniques such as prediction, classification, clustering, association rule mining and sequential pattern discovery. Currently, there is a rising interest in data mining and educational systems, making educational data mining a new and increasing research community. The data mining approach to personalization uses all the information about users/students which is available on the web site (in the web course) in order to learn user models and to make use of these models for personalization. These systems can use different recommendation techniques in order to recommend online learning actions or optimal browsing pathways to students, based on their preferences, knowledge and the browsing history of other students with identical characteristics.

Large amounts of data are generated every day and the ability to analyses them is normally a challenge. Experts need efficient data mining methods to extract useful information and to perform the analysis of the data. This is the case of the Rough Sets Theory (RST); Pawlak introduced mathematical rough set theory in the bit previous 1980"s. The theory was based on the distinguishability of objects. Rough set theory affords systems designers with the ability to handle ambiguity. If a concept is „not definable" in a given information base, rough sets can „approximate" with honor to that knowledge. From a medical point of view, the attribute-value boundaries are generally vague.

The rough set philosophy is established on the assumption that with every item of the universe of discourse we associate certain information (data, knowledge). For example, if objects are patients suffering from some disease, symptoms of the disease form information about patients. Objects characterized by the same information are indiscernible (similar) in sight of the available information about them. The in discernibility relation generated in this approach is the mathematical basis of rough set theory. This understanding of indiscernible is related to the concept of Gottfried Wilhelm Leibniz that objects are indiscernible if and only if all available functional take on them identical values (Leibniz"s Law of Indiscernible The Identity of Indiscernible). However, in the rough set approach indiscernible is defined relative to a given set of functional (attributes).

A weak aspect of RST is the unavailability of free RST software, except for limited implementations. On the other hand, there is RST proprietary software. RST is an extension of the set theory and has the implicit feature of compressing the dataset. Such compression is due to the definition of sameness classes based on indiscernibility relations and to the elimination of redundant or worthless attributes. A central notion in RST is attribute reduction, which generates reduces.

To be as adaptive as possible, a RS should consider previous interactions with different users before providing recommendations to the current user. The problem is then to set up a policy allowing to learn from item relevance while continuing to recommend related items to users. When the entire dataset is available at once, we can estimate all items' relevance : this is a supervised learning atmosphere (Hastie, Tibshirani, and Friedman 2009). This is usually not the case in real RS: new users and new items occur continuously; Copyright © 2015, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights earmarked. moreover, the choice of the items to be recommended at each interaction must be decided with the information from past interactions only. Such an environment is called ”reinforcement learning” (Sutton and Barto 1999). It requires implementing a strategy to gain information on the relevance of each item (exploration) while ensuring that the RS will continue to recommend relevant items (exploitation). This problem is known as the exploration/exploitation dilemma. Bandit algorithms are known to offer solutions to this dilemma (Bubeck and Cesa-Bianchi 2012).

2. LITERATURE SURVEY

MarzenaKryszkiewicz[1] presented Rough Set approach to incomplete information systems, i.e. to systems in which attribute values for objects may be unidentified (missing, null). His main concern was devoted to find rules from suchsystems. He suggested reduction of knowledge that eliminates only that information, which is not essential from the point of view of classification or result making.

PuntipPattaraintakorna, Nick Cerconeb [2], described Medical science is not an exact science in which processes can be simply analyzed and modeled. Rough set theory has proven well suited for accommodating such roughness of the medical profession. As rough set theory matures and its theoretical perspective is extended, the theory has been also followed by improvement of innovative rough sets systems as a result of this maturation. Unique concerns in medical sciences as well as the requirement of integrated rough sets systems are discussed. We present a short survey of current research and a case study on integrating rough set theory and medical application. Issues in the current state of rough sets in progressing medical technology and some of its challenges are also highlighted.

Alex SandroAguiar Pessoa, Stephan Stephany[3], offered the innovative use of two known met heuristics for the calculation, the Variable Neighborhood Search, the Variable Neighborhood Descent, besides a third heuristic called Decrescent Cardinality Search. The last one is a new heuristic specially proposed for reduct calculation. Considering some database commonly found in the literature of the field, the reducts that have been obtained present lower cardinality, i.e., a lower number of elements.

Y.H.Qian, J.Y.Liang, W.Pedrycz, and C.Y.Dang[4], Feature selection (attribute reduction) from large-scale unfinished data is a challenging problem in areas such as pattern recognition, machine learning and data mining. In rough set theory, feature selection from unfinished data aims to retain the discriminatory power of original features. To address this problem, many feature selection algorithms have been proposed, however, these algorithms are often computationally time-consuming. To overawed this shortcoming, they introduced a theoretic framework based on rough set theory, which one is called positive approximation and can be used to accelerate a heuristic process for feature selection from unfinished data. As an application of the proposed accelerator, a general feature selection algorithm was designed. By join in the accelerator into a heuristic algorithm, they obtain several modified representative heuristic feature selection algorithms in the rough set theory.

J.B.Zhang,T.R.Li,D.Ruan,Z.Z.GaoandC.B.Zhao [5], The effective computation of approximations is vital for developing the performance of data mining or other related tasks. The recently introduced Map Reduce technique has multiplied a lot of attention from the scientific community for its applicability in massive data analysis. The authors wished-for a parallel method for computing rough set approximations. Consequently, algorithms corresponding to the parallel method built on the MapReduce technique are put forward to deal with the massive data. An extensive experimental evaluation on different large data sets appearances that the proposed parallel method is effective for data mining.

G.L.Liu[6] planned a new matrix view of the theory of rough sets, they started with a binary relation and redefine a pair of lower and upper approximation operators consuming the matrix representation. Different classes of rough set algebras are obtained from various types of binary relations. Various classes of rough set algebras are characterized by different sets of axioms. Axioms of upper approximation operations guarantee the existence of certain types of binary relations (or matrices) producing the similar operators. The upper approximation of the Pawlak rough sets, rough fuzzy sets and rough sets of vectors over an arbitrary fuzzy lattice are categorized by the same independent axiomatic system.

3. PROPOSED SYSTEM

It can adaptively generalize profiles by questions while respecting user specified privacy requirements. It presents two greedy algorithms, that is Greedy DP and Greedy IL, for runtime generalization. It aims at protecting the privacy in individual user profiles while retaining their helpfulness for PWS. (User customizable Privacy-preserving Search).

Every query from the client user were provided by the specific requests to the server, this hides the frequent click through logs or content based mechanism, from this client can protect the data from the server. In the same case our mechanism maintains the online profiler about the client hence it hides the click logs and provides a safeguard to the user data. After that, online profiler question were processed in the manner of generalization process, it is used to meet the separate prerequisites to handle the user profile and it is based on the preprocessing the user profiles.

Our architecture, not only the user's search presentation but also their background activities (e.g., viewed before) and private information (e.g., emails, browser bookmarks) could be added into the user profile, permitting for the structure of a much richer client model for personalization. The sensitive contextual information is usually not a main aspect since it is strictly stored and used on the user side. A user's private information including user queries and click logs history resides on the user's private computer, and is exploited to better suppose the user' information require and provide a relevant examine results.

Our proposed algorithm uses the greedy method based on the discriminating power and information loss protection to receive the relations. Here it uses the inherited method to generalize the query. It allows performing the customization progression to protect the data and use the User customizable Privacy-preserving Search framework addressed the privacy issues. This aims at protecting the privacy in individual user profiles.

3.1. Greedy Algorithm

A greedy algorithm is follows the problem solving heuristic of making the locally optimal choice at each step with the hope of finding a global optimum. In many issues, a greedy strategy does not in general produce an optimal solution, but nonetheless a greedy heuristic may yield locally optimal solutions that approximate a global optimal solution in a reasonable time.

In general, greedy algorithms have five components:

- ❖ A candidate set, from which a solution is generated
- ❖ A selection function, which chooses the best candidate to be added to the solution
- ❖ A feasibility function, that is used to determine if a contestant can be used to contribute to a solution
- ❖ An objective function, which assigns a value to a solution, or a incomplete solution, and
- ❖ A solution function, which will indicate when we have discovered a complete(final) solution

Greedy algorithms produce good solutions on some mathematical problems, but not on others. Most issues for which they work will have two properties:

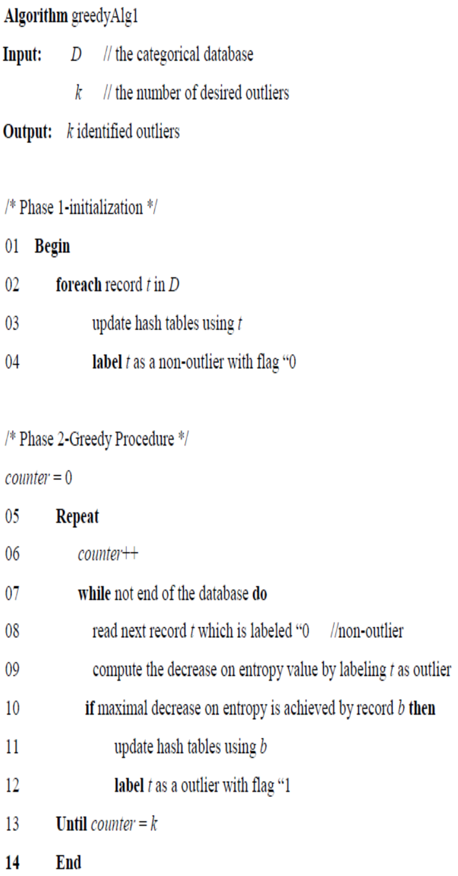
Greedy choice property

We can make whatever choice seems greatest at the instant and then solve the subproblems that arise later. The choice made by a greedy algorithm may depend on selections made so far, but not on future choices or all the solutions to the subproblem. It iteratively creates one greedy choice after another, reducing each given problem into a smaller one. In other words, a greedy algorithm never reevaluates its choices. This is the main difference from dynamic programming, which is exhaustive and is guaranteed to find the result.

After every stage, dynamic programming makes decisions based on all the decisions made in the preceding stage, and may reconsider the previous stage's algorithmic path to solution.

Greedy algorithms can be categorized as being 'short sighted', and also as 'non-recoverable'. They are ideal only for problems which have 'optimal substructure'. Despite this, for numerous simple problems (e.g. giving change), the best suited algorithms are greedy algorithms. It is main, however, to note that the greedy algorithm can be used as a selection algorithm to prioritize options inside a search, or branch-and-bound algorithm.

Algorithm



```
Algorithm greedyAlg1
Input:   D   // the categorical database
         k   // the number of desired outliers
Output:  k identified outliers

/* Phase 1-initialization */
Begin
    foreach record t in D
        update hash tables using t
        label t as a non-outlier with flag “0

/* Phase 2-Greedy Procedure */
counter = 0
    Repeat
      counter++
      while not end of the database do
        read next record t which is labeled “0    //non-outlier
        compute the decrease on entropy value by labeling t as outlier
       if maximal decrease on entropy is achieved by record b then
          update hash tables using b
          label t as a outlier with flag “1
    Until counter = k
    End
```

4. EXPERIMENTAL RESULTS

The main type of experiment aims to compare the effect on sales between the M-list and the B-list, together with PARs. This can be done with live evaluations such as A/B tests, where the users would be split in two groups and each group is shown a different alternative. However, live testing requires control over what is shown on a website and can be costly if one strategy performs significantly worse than the other strategy. Instead, we evaluate the two strategies using logged sales data. This problem of evaluating a new behavior, or policy, using only observations collected during the execution of another policy is known as policy evaluation [15]. To simplify our experiments, we compare the M-list and B-list with PARs as they would appear on the sites' top pages but we are confident that our method works equally well under the general situation that the customer navigates through the site, dynamically changing the part of the product catalogue currently considered by the recommendation system.

When compared against the use of supervised system, CES equipped with unsupervised system finds more compact ensembles with an average size of 8. This is feasibly due to the fact that the class label is no longer considered in the dependency calculation, and therefore less consistency constraints are located upon the construction of ensembles (artificial feature subsets).

The reduced ensembles still maintain acceptable classification accuracies, in comparison to the use of supervised system, and the base pool. Both approaches generally deliver healthier results than randomly picked ensembles, except for ecoli and glass datasets, the unverified method seems to have an equal performance as random picking. One possible explanation is that the amount of classes for a given training dataset has direct impact upon unsupervised performance. The more classes there are, the more several the classifier predictions become, thereby providing more available values for the artificial features.

Performance Analysis

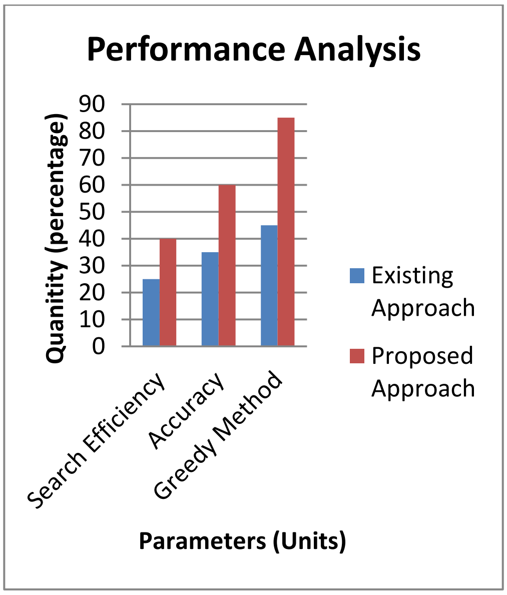


Fig 4.1

Fig 4.1 by applying fuzzy-rough feature selection technique to minimize redundancy in an artificial dataset generated via transforming a given classifier pool's decision matrix. The aim is to further reduce the size of a classifier ensemble, while maintaining and improving classification accuracy, making the ensemble more efficient. Experimental comparative studies show that Greedy Method approaches can entail good solutions. It takes parameter as search efficiency; accuracy and method based on greedy. Reduced ensembles are found with comparable classification accuracies as the base pools, and in most cases provide good improvements over the performance achievable by the base algorithms. In particular, the use of greedy method can help create smaller ensembles, especially when complex mixed classifiers are used.

5. FUTURE WORK

In the near future, it will be installed in Apache Server and so it will be available in internet.

Datasets will be updated continuously and it will make online actual rating predictions to the users whose habits are changing day by day. As a result, it can be sensitively satisfying current user tastes.

Web services in particular suffer from producing recommendations of millions of items to millions of users. The time and computational power can even limit the performance of the best hybrid systems. For larger dataset, we can work on scalability problems of recommendation systems.

The Prediction approach can also be tried in different datasets to test harmony performance of system scalability problems of recommendation systems.

6. CONCLUSIONS

Web users were increases because of accessible of information's from the web browser based on the search engine. With the increasing number of user service engine must supply the related search result based on their behavior or based on the user performance. Providing related result to the user is based on their click logs, query histories, bookmarks, by this secrecy of the user might be loss. For providing relevant search by using these approaches the privacy of the user may loss. Most existing system provides supply a major barrier to the private information during user search. That method does not protect privacy issues and rising information loss for the user data. For this issue this paper proposes user based architecture based on the greedy algorithm to prevent the user data and provide the relevant search result to the user in future it can contain this work in mobile application. The experiment results show that the proposed method is effective and practical, and has potential for further development. In addition, combining the often used popular evaluation methods and the specific situation in which the successive approximation approach based on Rough Set theory have been built up, an innovative evaluation metric named CEI has been proposed and proven very effective for the assessment of the similar research.

REFERENCES

[1]. M. Kryszkiewicz, "Rough set approach to incomplete information systems," Inf. Sci., vol. 112, no. 1–4, pp. 39–49, Dec. 1998.

[2]. PuntipPattaraintakorna, Nick Cerconeb "Integrating rough set theory and medical applications", Volume 21, Issue 4, Pages 400–403, April 2008

[3]. Pessoa, A.S.A. and Stephany, S. "An Innovative Approach for Attribute Reduction in Rough Set Theory". Intelligent Information Management, 6, 223-239. (2014)

[4]. Y. H. Qian, J. Y. Liang, W. Pedrycz, and C. Y. Dang, "An efficient accelerator for attribute reduction from incomplete data in rough set framework," Pattern Recognit., vol. 44, no. 8, pp. 1658–1670, Aug. 2011.

[5]. J. B. Zhang, T. R. Li, D. Ruan, Z. Z. Gao, and C. B. Zhao, "A parallel method for computing rough set approximations," Inf. Sci., vol. 194, pp. 209–223, Jul. 2012.

[6]. G. L. Liu, "The axiomatization of the rough set upper approximation operations," FundamentaInformat., vol. 69, no. 3, pp. 331–342, 2006.

[7]. G. L. Liu, "Axiomatic systems for rough sets and fuzzy rough sets," Int. J. Approximate Reasoning, vol. 48, no. 3, pp. 857–867, Aug. 2008.

[8]. J. W. Grzymala-Busse, and W. Ziarko, "Data mining and rough set theory," Commun. ACM, vol. 43, no. 4, pp. 108–109, Apr. 2000.

[9]. Y. H. Qian, C. Y. Dang, J. Y. Liang, and D. Tang, "Set-valued ordered information systems," Inf. Sci., vol. 179, no. 16, pp. 2809– 2832, Jul. 2009.

[10]. T. R. Li, D. Ruan, W. Geert, J. Song, and Y. Xu, "A rough sets based characteristic relation approach for dynamic attribute generalization in data mining," Knowl.-Based Syst., vol. 20, no. 5, pp. 485–494, Jun. 2007.

[11]. Shen, Xuehua, Bin Tan, and Cheng Xiang Zhai. "Implicit user modeling for personalized search." Proceedings of the 14th ACM international conference on Information and knowledge management. ACM, 2005.

[12]. T. Joachims, L. Granka, B. Pang, H. Hembrooke, and G. Gay, "Accurately Interpreting Clickthrough Data as Implicit Feedback," Proc. 28th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '05), pp. 154-161, 2005.

[13]. Shen, Xuehua, Bin Tan, and Cheng Xiang Zhai. "Context-sensitive information retrieval using implicit feedback." Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2005.

[14]. Xu, Yabo, et al. "Online anonymity for personalized web services." Proceedings of the 18th ACM conference on Information and knowledge management. ACM, 2009.

[15]. A. Viejo and J. Castell_a-Roca, "Using Social Networks to Distort Users' Profiles Generated by Web Search Engines," Computer Networks, vol. 54, no. 9, pp. 1343-1357, 2010