IJDCST @May-June, Issue- V-2, I-4, SW-02
ISSN-2320-7884 (Online)
ISSN-2321-0257 (Print)

Query Pattern Evaluation in User Search Histories

[1]Dr.J.Srinivasa Rao, [2]Mr. Hari Krishna.Deevi, [3]Mr. V.Shanmukha Rao.
[1]Director & Professor, Nova College of Engineering and Technology, Jupudi, Vijayawada, AP.
[2]Associate Professor, Department of CSE, Nova College of Engineering and Technology, Jupudi, Vijayawada, AP.
[3]M.Tech, CSE Department, Nova College of Engineering and Technology, Jupudi, Vijayawada, AP.

Abstract: Based on the client-server model, we present a detailed architecture and design for implementation of PMSE. In our design, the client collects and stores locally the click through data to protect privacy, whereas heavy tasks such as concept extraction, training, and re ranking are performed at the PMSE server. PMSE significantly improves the precision comparing to the baseline. If any technique present for improving the efficiency of the relative process in query patterns and travel patterns accessing. In this paper, we propose CPHC (Classification by Pattern based Hierarchical Clustering), a semi-supervised classification algorithm that uses a pattern-based cluster hierarchy as a direct means for classification. All training and test instances are first clustered together using an instance-driven pattern-based hierarchical clustering algorithm that allows each instance to "vote" for its representative size-2 patterns in a way that balances local pattern significance and global pattern interestingness. These patterns form initial clusters and the rest of the cluster hierarchy is obtained by following a unique iterative cluster refinement process that exploits local information. The resulting cluster hierarchy is then used directly to classify test instances, eliminating the need to train a classifier on an enhanced training set. Our experimental results show efficient processing of each query optimization in training data set.

Key Words: PMSE, CPHC, Cluster hierarchy, Cluster refinement, semi-supervised classification

I. INTRODUCTION

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics [1].

Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings (including values such as the distance function to use, a density threshold or the number of expected clusters) depend on the individual data set and intended use of the results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure. It will often be necessary to modify data preprocessing and model parameters until the result achieves the desired properties[1].

A "clustering" is essentially a set of such clusters, usually containing all objects in the data set. Additionally, it may specify the relationship of the clusters to each other, for example a hierarchy of clusters embedded in each other. Clusterings can be roughly distinguished as:

- hard clustering: each object belongs to a cluster or not
- soft clustering (also: fuzzy clustering): each object belongs to each cluster to a certain degree (e.g. a likelihood of belonging to the cluster).

Clustering algorithms can be categorized based on their cluster model, as listed above. The following overview will only list the most prominent examples of clustering algorithms, as there are possibly over 100 published clustering algorithms. Not all provide models for their clusters and can thus not easily be categorized. An overview of algorithms explained in Wikipedia can be found in the list of statistics algorithms.

There is no objectively "correct" clustering algorithm, but as it was noted, "clustering is in the eye of the beholder."[2] The most appropriate clustering algorithm for a particular problem often needs to be chosen experimentally, unless there is a mathematical reason to prefer one cluster model over another. It should be noted that an algorithm that is designed for one kind of model has no chance on a data set that contains a radically different kind of model.[2] For example, k-means cannot find non-convex clusters.

Connectivity based clustering (hierarchical clustering)

Connectivity based clustering, also known as hierarchical clustering, is based on the core idea of objects being more related to nearby objects than to objects farther away. These algorithms connect "objects" to form "clusters" based on their distance. A cluster can be described largely by the maximum distance needed to connect parts of the cluster. At different distances, different clusters will form, which can be represented using a dendrogram, which explains where the common name "hierarchical clustering" comes from: these algorithms do not provide a single partitioning of the data set, but instead provide an extensive hierarchy of clusters that merge with each other at certain distances. In a dendrogram, the y-axis marks the distance at which the clusters merge, while the objects are placed along the x-axis such that the clusters don't mix.

Connectivity based clustering is a whole family of methods that differ by the way distances are computed. Apart from the usual choice of distance functions, the user also needs to decide on the linkage criterion (since a cluster consists of multiple objects, there are multiple candidates to compute the distance to) to use. Popular choices are known as single-linkage clustering (the minimum of object distances), complete linkage clustering (the maximum of object distances) or UPGMA ("Unweight Pair Group

ISSN-2320-7884 (Online)
ISSN-2321-0257 (Print)

Method with Arithmetic Mean", also known as average linkage clustering). Furthermore, hierarchical clustering can be agglomerative (starting with single elements and aggregating them into clusters) or divisive (starting with the complete data set and dividing it into partitions).

II. RELATED WORK

Hassan H. Malik, and John R. Kender stated that The global pattern mining step in existing pattern-based hierarchicalclustering algorithms may result in an unpredictable number of patterns. In thispaper, we propose IDHC, a pattern-based hierarchical clustering algorithm thatbuilds a cluster hierarchy without mining for globally significant patterns.IDHC allows each instance to "vote" for its representative size-2 patterns in away that ensures an effective balance between local and global patternsignificance. The number of patterns selected for each instance is dynamicallydetermined using a local standard deviation based scheme, and the rest of thecluster hierarchy is obtained by following a unique iterative cluster refinementprocess. By effectively utilizing instance-to-cluster relationships, this processdirectly identifies clusters for each level in the hierarchy, and efficiently prunesduplicate clusters. Furthermore, IDHC produces cluster labels that are moredescriptive (patterns are not artificially restricted), and adapts a soft clusteringscheme that allows instances to exist in suitable nodes at various levels in thecluster hierarchy. We present results of experiments performed on 16 standardtext datasets, and show that IDHC almost always outperforms state-of-the-arthierarchical clustering algorithms in terms of entropy, and achieves better FScores in most cases, without requiring tuning of parameter values

Jianyong Wang and George Karypis stated that Many studies have shown that rule-based classifiersperform well in classifying categorical and sparse high dimensional databases. However, a fundamental limitation with many rule-based classifiers is that they find the rules by employingvarious heuristic methods to prune the search space, and select the rules based on the sequential database coveringparadigm. As a result, the final set of rules that they use may not be the globally best rules for some instances in the trainingdatabase. To make matters worse, these algorithms fail to fullyexploit some more effective search space pruning methods inorder to scale to large databases.In this paper we present a new classifier, HARMONY, whichdirectly mines the final set of classification rules. HARMONYuses an instance-centric rule-generation approach and it canassure for each training instance, one of the highest-confidencerules covering this instance is included in the final rule set,which helps in improving the overall accuracy of the classifier. Byintroducing several novel search strategies and pruning methodsinto the rule discovery process, HARMONY also has highefficiency and good scalability. Our thorough performance studywith some large text and categorical databases has shown thatHARMONY outperforms many well-known classifiers in termsof both accuracy and computational efficiency, and scales wellw.r.t. the database size.

Wenmin Li Jiawei Han Jian Pei stated that previous studies propose that associative classificationhas high classification accuracy and strong flexibility athandling unstructured data. However, it still suffers fromthe huge set of mined rules and sometimes biased classificationor overfitting since the classification is based ononly single high-confidence rule.In this study, we propose a new associative classificationmethod, CMAR, i.e., Classification based onMultiple Association Rules. The method extends an efficientfrequent pattern mining method, FP-growth, constructsa class distribution-associated FP-tree, and mineslarge database efficiently. Moreover, it applies a CR-treestructure to store and retrieve mined association rules efficiently,and prunes rules effectively based on confidence,correlation and database coverage. The classification isperformed based on a weightedanalysis using multiplestrong association rules. Our extensive experiments ondatabases from UCI machine learning database repository show that CMAR is consistent, highly effective at classificationof various kinds of databases and has better averageclassification accuracy in comparison with CBA andC4.5. Moreover, our performance study shows that themethod is highly efficient and scalable in comparison withother reported associative classification methods

Martin Ester stated that Text clustering methods can be used to structure large sets of textor hypertext documents. The well-known methods of textclustering, however, do not really address the special problems oftext clustering: very high dimensionality of the data, very largesize of the databases and understandability of the cluster

description. In this paper, we introduce a novel approach whichuses frequent item (term) sets for text clustering. Such frequentsets can be efficiently discovered using algorithms for associationrule mining. To cluster based on frequent term sets, we measurethe mutual overlap of frequent sets with respect to the sets ofsupporting documents. We present two algorithms for frequentterm-based text clustering, FTC which creates flat clustering's and HFTC for hierarchical clustering. An experimental evaluation onclassical text documents as well as on web documentsdemonstrates that the proposed algorithms obtain clustering's of comparable quality significantly more efficiently than state-of-the art text clustering algorithms. Furthermore, our methods providean understandable description of the discovered clusters by theirfrequent term sets.

Bing Liu Wynne Hsu Yiming Ma stated that Classification rule mining aims to discover a small set ofrules in the database that forms an accurate classifier.Association rule mining finds all the rules existing in thedatabase that satisfy some minimum support and minimumconfidence constraints. For association rule mining, thetarget of discovery is not pre-determined, while forclassification rule mining there is one and only one predeterminedtarget. In this paper, we propose to integratethese two mining techniques. The integration is done byfocusing on mining a special subset of association rules,called class association rules (CARs). An efficientalgorithm is also given for building a classifier based on theset of discovered CARs. Experimental results show that theclassifier built this way is, in general, more accurate thanthat produced by the state-of-the-art classification systemC4.5. In addition, this integration helps to solve number of problems that exist in the current classification systems.

III. EXISTING SYSTEM

Design for PMSE by adopting the meta search approach which relies on one of the commercial search engines, such as Google, Yahoo, or Bing, to perform an actual search..

ISSN-2320-7884 (Online)
ISSN-2321-0257 (Print)

A personalization framework that utilizes a user's content preferences and location preferences as well as the GPS locations in personalizing search results. The user profiles for specific users are stored on the PMSE clients, thus preserving privacy to the users. PMSE has been prototyped with PMSE clients on the. The user profiles for specific users are stored on the PMSE clients, thus preserving privacy to the users. PMSE has been prototyped with PMSE clients on the GOOGLE Server.PMSE incorporates a user's physical locations in the personalization process. We conduct experiments to study the influence of a user's GPS locations in personalization.The results show that GPS locations help improve retrieval effectiveness for location queries (i.e., queries that retrieve lots of location information).

PMSE profiles both of the user's content and location preferences in the ontology based userprofiles, which are automatically learned from the click through and GPS data without requiringextra efforts from the user.PMSE addresses this issue by controlling the amount of information in the client's user profilebeing exposed to the PMSE server using two privacy parameters, which can control privacysmoothly, while maintaining good ranking quality.

PMSE incorporates a user's physical locations in the personalization process. We conductexperiments to study the influence of a user's GPS locations in personalization.

IV. PROPOSED SYSTEM

The quality of clustering achieved by traditional flat clustering algorithms (i.e., k-means clustering) relies heavily on the desired number of clusters value of k), which must be known in advance.We propose CPHC (i.e., Classification by Pattern based Hierarchical Clustering), a novel semi-supervised classification algorithm that uses a pattern-based cluster hierarchy as a direct means for classification.In addition, this approach uses a novel feature selection method that ensures that all training and test instances are covered by the selected features, uses parameters that are robust across datasets with varying characteristics, and also has the positive side effect of improving the chances of classifying isolated test instances on sparse training data by inducing a form of feature transitivity.

CPHC (i.e., Classification by Pattern based Hierarchical Clustering), a novel semi-supervised classification algorithm that uses a pattern-based cluster hierarchy as a direct means for classification. Unlike existing semi-supervised classification algorithms, CPHC directly uses the resulting cluster hierarchy to classify test instances and hence eliminates the extra training step.The remainder of this section briefly introduces the notations used in this paper, discusses the motivation for instance-driven pattern-based hierarchical clustering, discusses the significance of pattern lengths in these hierarchies and also provides abrief overview of the CPHC algorithm.

CPHC (i.e., Classification by Pattern-based Hierarchical Clustering), a novel semi-supervisedclassification algorithm that uses pattern-lengths as a way of establishing cluster (i.e., node) weights. CPHC first applies an unsupervised instance-driven pattern-based hierarchical clustering algorithm (i.e., IDHC, Section 1.2) to the whole dataset toproduce a cluster hierarchy. Unlike existing semi-supervised classification algorithms [8,9,10]. CPHC directly uses the resulting cluster hierarchy to classify testinstances and hence eliminates the extra training step. To classify a test instance,CPHC first uses the hierarchical structure to identify nodes that contain the testinstance, and then uses the labels of co-existing training instances, weighing them bynode pattern-lengths (i.e., by multiplying the node pattern-interestingness value withthe pattern-length) to obtain class label(s) for the test instance. This allows CPHC toclassify unlabeled test instances without making any assumptions about their distribution in the dataset.

V. EXPERIMENTAL RESULTS

We conclude that a broad experimental result gives us it is a pattern-based cluster hierarchy for classification. CPHC first uses the hierarchical structure to identify nodes that contain the test instance, and then uses the labels of co-existing training instances, weighing them by node pattern-lengths (i.e., by multiplying the node pattern-interestingness value with the pattern-length) to obtain class label(s) for the test instance. By Using CPHC we can classify test instances and we can eliminate the enhanced training set. By that results can show efficient processing of each query optimization in training data set.

VI. CONCLUSION

The semi-supervised approach first clusters both the training and test sets togetherinto a single cluster hierarchy, and then uses this hierarchy as a direct means forclassification; this eliminates the need to train a classifier on an enhanced training set.

In addition, this approach uses a novel feature selection method that ensures that alltraining and test instances are covered by the selected features, uses parameters thatare robust across datasets with varying characteristics, and also has the positive sideeffect of improving the chances of classifying isolated test instances on sparsetraining data by inducing a form of feature transitivity. Lastly, this approach is very robust on very sparse training data.

VI. REFERENCES

[1]. Clustering algorithms from Wikipedia

[2]. Estivill-Castro, Vladimir (20 June 200202). "Why so many clustering algorithms — A Position Paper". ACM SIGKDD Explorations Newsletter 4 (1): 65–75. doi:10.1145/568574.568575

[3].Malik, H. H., Kender, J. R.: Instance Driven Hierarchical Clustering of DocumentCollections. In: From Local Patterns to Global Models Workshop, European Conference onMachine Learning and Practice of Knowledge Discovery in Databases (2008).

[4].Wang, J., Karypis, G.: On Mining Instance-Centric Classification Rules. IEEE Transactionson Knowledge and Data Engineering, Volume 18, No. 11 (2006).

[5].Li, W., Han, J., Pei, J.: CMAR: Accurate and Efficient Classification based on multipleclass-association rules. In: First IEEE International Conference on Data Mining.

ISSN-2320-7884 (Online)
ISSN-2321-0257 (Print)

[6].Beil, F., Ester, M., Xu, X.: Frequent term-based text clustering. In: International Conferenceon Knowledge Discovery and Data Mining, pp. 436-442.

[7].Liu, B., Hsu, W., Ma, Y.: Integrating Classification and Association Rule Mining. In: FourthACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

[8]. Kyriakopoulou, A., Kalamboukis, T.: Using clustering to enhance text classification. In:30th annual international ACM SIGIR conference on Research and development ininformation retrieval (2007).

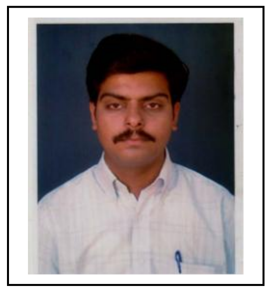
[9]. Raskutti, B., Ferr, H., Kowalczyk, A.: Using unlabeled data for text classification throughaddition of cluster parameters. In: 9th International Conference on Machine Learning.

[10]. Zeng, H. J., Wang, X.H., Chen, Z., Lu, H., Ma, W. Y.: CBC: Clustering based textclassification requiring minimal labeled data. In: Third IEEE International Conference on Data Mining.

BIOGRAPHY

Dr. J.SRINIVAS RAO, M. Tech, PhD. Received his M. Tech in computer science & engineering from KL University in 2008, Ph D from CMJ University Meghalaya, INDIA. He is an Outstanding Administrator & Coordinator. He is having 16 years of experience and handled both UG and PG classes. Currently he is working as a Director & Professor in NOVA College of Engineering Technology, Vijayawada, A.P, INDIA .He has Published 30 research Papers in various international Journals and workshops with his incredible work to gain the knowledge for feature errands.

Mr. Hari Krishna.Deevi is a qualified person holding M.Sc (CSE) & M.Tech Degree in CSE from Acharya Nagarjuna university, He is an Outstanding Administrator & Coordinator. He is working as an Associate Professor in NOVA College of Engineering Technology. He guided students in doing IBM projects at NOVA ENGINEERING College. Who has Published 10 research Papers in various international Journals and workshops with his incredible work to gain the knowledge for feature errands.



Mr. V.Shanmukha Rao, was born in 1977 in AP, India. He completed his Master of Computer Applications in AMA College of Engineering & Technology, affiliated to Madras University in 2001, MPhil (CS) degree from Madurai Kamaraj University, M.Tech(CS) from JNRV University. He has 11 years of teaching experience in india and 2 years of teaching experience in Malaysia for UG & PG IT courses. Presently he is persuing M.Tech (Computer Science Engineering) from NOVA College of Engineering & Technology, affiliated to JNTUK University, Kakinada.