IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 2, Ver. IX (Mar-Apr. 2014), PP 11-15
www.iosrjournals.org

An Automated Approach for Job Scheduling and Work Flow Mining

[1]P. Rijisha , [2]V. Venkatesh
PG Student, Assistant Professor
[1,2]Department of CSE, SVS College Of Engineering, Tamilnadu.

Abstract: *Now a day's work allotment in a software firm become more important and cumbersome. The main objective of this concept is to reduce the work of the software developers in a software company and work allocation. Here the work will be allotted for the each person automatically by splitting the modules into tasks. When developing a big project more confusion and more discussions will occur, and some clauses will arise between modules while splitting into tasks and who the person to develop the particular task And performance evolution will not be in accurate, which confuse with the actual work allotment by the HR. So that still companies are struggling to book multiple software projects at a same time. So that still companies are yielding less works from their employees. In order to reduce these problems, implementing work flow mining, some of them are Group movement pattern, Heterogeneous project details, Ranking model adaptation SVM (RA-SVM), Domain specific search, Distributed clustering etc.*

I. Introduction

An Automated Approach for Job Scheduling and Work Flow Mining" which is developed to reduce the work of the software developers in a software company and work allocation.

In order to reduce problems in workflow mining, proposed system is introducing. Here all the modules will be given as input and it will automatically indicate how many persons will need to complete this project, and how much time will be taken to complete a particular module. So one can avoid all confusions by using these methods.

The proposed system use Group movement domain specific search Existing object tracking applications focus on finding the moving patterns of a single object or all objects. In contrast, propose a distributed mining algorithm that identifies a group of objects with similar movement patterns. This information is important in some biological research domains, such as the study of animal's social behavior and wildlife migration.

The proposed algorithm comprises a local mining phase and a cluster ensemble phase. In the local mining phase, the algorithm finds movement patterns based on local trajectories. Then, based on the derived patterns, propose a new similarity measure to compute the similarity of moving objects and identify the local group relationships.

To address the energy conservation issue in esthete-constrained environments, the algorithm only transmits the local grouping results to the sink node for further ensemble. In the cluster ensemble phase, the algorithm combines the local grouping results to derive the group relationships from a global view. Then further leverage the mining results to track moving objects efficiently. The results of experiments show that the proposed mining algorithm achieves good grouping quality, and the mining technique helps reduce the energy consumption by reducing the amount of data to be transmitted.

The proposed system use Movement Pattern Mining. The temporal-and-spatial correlations and the regularity in the trajectory data sets of moving objects are often modeled as sequential patterns for use in data mining. Agrawal and Srikant first defined the sequential pattern mining problem and proposed an Apriori-like algorithm to mine frequent sequential patterns. Free Span is proposed by Han et al, which is an FP-growth-based algorithm that addresses the sequential pattern mining problem by considering the pattern-projection method. For handling the uncertainty in trajectories of mobile objects, Yang and Hu developed a new match measure and proposed Traj Pattern to mine sequential patterns from imprecise trajectories. Moreover, a number of research works have been elaborated upon mining traversal patterns for various applications.

For example, Chen et al. proposed the FS and SS algorithms for mining path traversal patterns in a Web environment while Peng and Chen proposed an incremental algorithm to mine user moving patterns for data allocation in a mobile computing system. But, sequential patterns or path traversal patterns do not provide sufficient information for location prediction or clustering. The reasons are as follows:

First, for sequential pattern mining or path traversal pattern mining extract frequent patterns of all objects, meaningful Second, a sequential pattern or a traversal pattern carries no time information between consecutive items, so they cannot provide accurate in-formation for location prediction when time is concerned. Third, sequential patterns are not full representative to individual trajectories because a sequential pattern does not contain the information about the number of times it occurs in each individual trajectory.

To discover significant patterns for location prediction, Morzy proposed Apriori Traj and Traj-Prefix Span to mine frequent trajectories, where consecutive items of a frequent trajectory are also adjacent in the original trajectory data.

II. Existing System

System use trajectory clustering based on objects' movement behavior has attracted more attention. For example, Li et al. employ Moving Micro clusters (MMC) to discover and maintain a cluster of moving objects online. Meanwhile, Lee et al. proposed trajectory clustering to discover popular movement paths. Clustering similar trajectory sequences to discover group relationships is closely related to the problem. Wang et al. transform the location sequences into a transaction-like data on users and based on which to obtain a valid group. But, the proposed AGP and VG-growth algorithms are Apriori like or FP-growth-based algorithms that suffer from high computing cost and memory demand. Here apply a density-based clustering algorithm to the trajectory clustering problem based on the average euclidean distance of two trajectories.

The proposed work in which discover group information based on the proportion of the time a group of users stay close together or the average euclidean distance of the entire trajectories may not reveal the local group relationships, which are required for many applications.

Still most of the software companies are facing much more problems in the work allotments. Even though the company is having sufficient numbers of employee and even thou they are trained well there is always a problem in work allotment strategy. This leads to facing the company into many problems like could not able to deliver the project at a particular time, More numbers of error occurrence, More maintenance, complexion or much more effort for a simple works and etc, In the existing system there may some solutions are available but one cannot solve all these problem in one solution.

Drawbacks Of Existing System

- Software cannot be in user friendly so that it could not solve much problem in the work allotments.
- There should be always an expert needed to solve these problems or handing this domain to get a prior output.
- Manual operations will be more, even most of the process in manual mode means one could not obtain a accurate result
- Difficult to understand the output of the domain.
- Time consumption will be more in order to producing an output for a simple workflow process.

The existing system primarily focusing on the Trajectory Clustering in Job scheduling algorithm, which makes to user to do more manual works.

For example Grade Allocation will not allow in this model, this is because; manual input data will not generate automated values or grids. Here the Job scheduling algorithm is implemented for work allotment purpose, but here no availability of the user, so that the resthece data will be affected. No training data will be available here for the future reusable works, and no neural network implemented for data training.

III. Proposed System

Distributed clustering is an important research topic. Most of the approaches proposed in the literature focus on seeking a combination of multiple clustering results to achieve better clustering quality, stability, and scalability. Here introduced and formulated the clustering ensemble problem to a hyper-graph partitioning problem, and proposed CSPA, HGPA, and MCLA to compute the best K-partition of the graph. And also presented a probabilistic model to combine cluster ensembles by utilizing information theoretic measures. Also combine multiple runs of the K-means algorithm with random initializations and random numbers to obtain the final consensus partition. Apply random projection to the high dimensional data and cluster the reduced data by using EM for a single run of clustering. The Collective PCA technique is applied to reduce the vector dimension for distributed clustering of high dimensional heterogeneous data. But, since the trajectory data set is composed of sequential data, one of the challenges addressed in the paper is the similarity comparison of location sequences. The data types of the above works are most integer vector or categorical data, and the related issues are thereby different from this. In addition, previous works that require a predator mined the clustering or ensemble algorithms are not suitable for the applications. Besides, although the local grouping results in a vector

of integers, each of which represents the mapping between an object and its belonging group, dimension reduction like Collective PCA is unnecessary in the case will be identified using domain specific search

The future works over came all the problems in the existing system, so that this implementation will satisfy the end user and purpose will solved in too short period of time. Here the implementation design will be in the combination of Job Scheduling model in architecture design using neural networks.

Grades will be available here, so that the work load of the HR will be reduced and the accuracy will be maintained.3 types of job scheduling will be used namely Long term scheduling, medium term scheduling, short term scheduling. By using these scheduling methods data accuracy will be increased. By using the neural networks training data can be trained for the future use, so that reusability can be increased up to 80 percentage. In the proposed system here overcame all the drawbacks in the existing system.

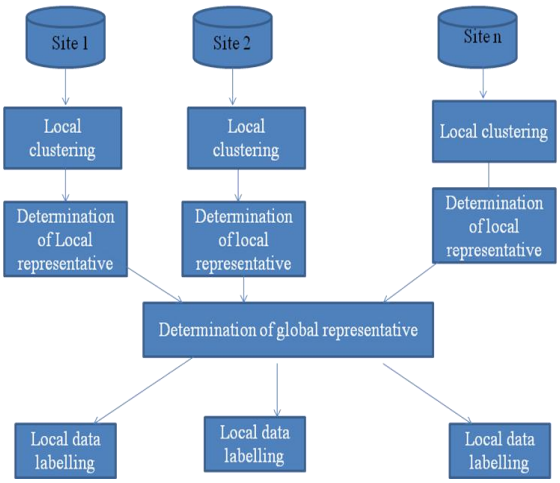
Likewise develop a dynamic application which can able to perform a multi tasking process at a short period of time, so that one can able to yield more output from this application. Work allotment will be done from the initial stage of the employee, that means, Whenever an employee is joining in this company he or she will be comes in the work flow category. So that employee will be monitored right from the beginning of the work. Here the employee work status will be reported right from the first of his work. After the allocation is over the project details will be allotted to the employee according to the experience. Here an intelligent system is used for the wok allotment for the admin. According the performance of the employee the work can be allotted.

Recently, various domain-specific search engines emerge, which are restricted to specific topicalities or document formats, and vertical to the broad-based search. Simply applying the ranking model trained for the broad-based search to the verticals cannot achieve a sound performance due to the domain differences, while building different ranking models for each domain is both laborious for labeling sufficient training samples and time-consuming or the training process. In the proposed system in which address the above difficulties, investigate two problems: (1) whether can adapt the ranking model learned for existing Web page search or verticals, to the new domain, so that the amount of labeled data and the training cost is reduced, while the performance requirement is still satisfied; and (2) how to adapt the ranking model from auxiliary domains to a new target domain.

The second problem from the regularization framework and an algorithm called ranking adaptation SVM is proposed. Our algorithm is flexible enough, which needs only the prediction from the existing ranking model, rather than the internal representation of the model or the data from auxiliary domains. The first problem is addressed by the proposed ranking adaptability measurement, which quantitatively estimates if an existing ranking model can be adapted to the new domain.

Advantages Of Proposed System

- User Friendly application which can be used for both the employee work track out as well as the project maintenance.
- Even a low level employee can able to operate this domain in order to fetch the work details from the admin.
- Here most of the process is automated well so that manual work will be mostly avoided here.



IV. Algorithm Used In Proposed System

a) Confidence of an Association Rule using distributed clustering.

When we search for association rules, we do not want just any association rules, but good association rules. To measure the quality of association rules, the inventors of the apriori algorithm, introduced the confidence of a rule. The confidence of an association rule R = A and B -> C is the support of the set of all items that appear in the rule (here: the support of S = { A, B, C }) divided by the support of the antecedent (also called if-part or body) of the rule (here X = { A, B)). That is,

conf(R) = (supp({A, B, C}) / supp({A, B})) *100%.

More intuitively, the confidence of a rule is the number of cases in which the rule is correct relative to the number of cases in which it is applicable. For example, let R = wine and bread -> cheese. If a customer buys wine and bread, then the rule is applicable and it says that he/she can be expected to buy cheese. If he/she does not buy wine or does not buy bread or buys neither, then the rule is not applicable and thus does not say anything about this customer.

If the rule is applicable, it says that the customer can be expected to buy cheese. But he/she may or may not buy cheese, that is, the rule may or may not be correct. Naturally, we are interested in how good the prediction of the rule is, that is, how often its prediction that the customer buys cheese is correct. The rule confidence measures this: it states the percentage of cases in which the rule is correct. It states this percentage relative to the number of cases in which the antecedent holds, since these are the cases in which the rule makes a prediction that can be true or false. If the antecedent does not hold, then the rule does not make any prediction, so these cases are excluded.

Rules are reported as association rules if their confidence reaches or exceeds a given lower limit. That is, we look for rules that have a high probability of being true: we look for good rules, which make correct predictions. My apriori program always uses a minimum confidence to select association rules. The default value for the minimum confidence is 80%. This value can be changed with the option -c.

In addition to the rule confidence my apriori program lets you select from several other rule evaluation measures, which are explained below, but it will also use rule confidence. If you want to rely entirely on some other measure, you can do so by setting the minimal rule confidence to zero.

b) Support of an Association Rule

The support of association rules may cause some confusion, because I use this term in a different way than do. For them, the support of an association rule A and B -> C is the support of the set S = { A, B, C }. This may be fine if rule confidence is the only evaluation measure, but it causes problems if some other measure is used. For these other measures it is often much more appropriate to call the support of the antecedent of the association rule, that is, the support of X = { A, B } in the example above, the support of the association rule.

The difference can also be stated in the following way: the support of the rule is the number of cases in which the rule is correct, whereas for me the support of a rule is the number of cases in which it is applicable, although in some of these cases it may be false.

One reason for this choice, as already mentioned, is that the definition of does not work well for evaluation measures other than rule confidence. This is explained in more detail below. Another reason is that I prefer the support of a rule to say something about the statistical support of a rule and its confidence, that is, from how many cases the confidence is computed in order to express how well founded the statement about the confidence

Maybe an example will make this clearer. Suppose you have a die which you suspect to be biased. To test this hypothesis, you throw the die, say, a thousand times. 307 times the 6 turns up. Hence you assume that the die is actually biased, since the relative frequency is about 30% although for an unbiased die it should be around 17%. Now, what is the statistical support of this statement, that is, on how many experiments does it rest? Obviously it rests on all 1000 experiments and not only on the 307 experiments in which the 6 turned up. This is so, simply because you had to do 1000 experiments to find out that the relative frequency is around 30%, and not only the 307 in which a 6 turned up.

Or suppose you are doing an opinion poll to find out about the acceptance of a certain political party, maybe with the usual question you ask 2000 persons, of which 857 say that they would vote for the party you are interested in. What is the support of the assertion that this party would get around 43% of all votes? It is the size of your sample, that is, all 2000 persons, and not only the 857 that answered in the positive. Again you had to ask all 2000 people to find out about the percentage of 43%. Of course, you could have asked fewer people, say, 100, of which, say, 43 said that they would vote for the party, but then your statement would be less reliable, because it is less supported. The number of votes for the party could also be 40% or 50%, because of some random influences. Such deviations are much less likely, if you asked 2000 persons, since then the random influences can be expected to cancel out.

The rule support can be used to filter association rules by stating a lower bound for the support of a rule. This is equivalent to saying that you are interested only in such rules that have a large enough statistical basis. The default value for this support limit is 10%. It can be changed with the option -s. Note that the argument, if positive, is interpreted as a percentage. If, however, the given argument is negative, it is interpreted as an absolute number rather than a percentage.

The minimum support is combined with the minimum confidence to filter association rules. That is, my apriori program generates only association rules, the confidence of which is greater than or equal to the minimum confidence and the support of which is greater than or equal to the minimum support.

Despite the above arguments in favor of my definition of the support of an association rule, a rule support compatibility mode is available. With the option -o the original rule support definition can be selected. In this case the support of an association rule is the support of the set with all items in the antecedent and the consequent of the association rule, that is, the support of an association rule as defined.

V. Conclusion

Thus the proposed work for each employee of an organization that is registered is allocated based on their performance and the job can be completed in faster time. Thus the proposed system can be used for finding the best team leader to do a particular project in a short span of time and allocating work automatically for employee by finding the right person for the right work by the team leader based on the information that is being updated continuously by the admin so as to reduce the work and time of the software developers in a software company.

VI. Future Work

This system provide a successful way of allocating the job for each employee in a company, the number of employee may not be limited now. It can be extended as allowing admin to input the size of organization and to specify how to make change from mid scale to big scale, as the organization grows. And the employee working in a project will be relieved only after that project is completed. But there may be cases in which employee need to be relieved even when he/she is working in a project, so this project can be extended to provide the same. This can be done by assigning his/her work to the employee who is not allotted a project.

References

[1]. LinjunYang, Member, IEEE, Chao Xu, Xian-Sheng Hua,"Ranking Model Adaptation for Domain specific Search", IEEE Transaction on Knowledge and Data Engineering,Volume:24,Issue:4,2012.
[2]. Dik Lun Lee , Wang-Chien Lee,"PMSE: personalized mobile search engine", Browse Journals & Magazines,Knowledge and Data Engineerin, Volume:25 Issue:4, 2013.
[3]. Lam, Wai, Tsang, Ivor W. Wong, Tak-Lam, "Discovering low- rank shared concept space for adapting text mining model", Journals & Magazines, Pattern Analysis and Machine, Volume:35 Issue:6,2013.
[4]. Xi Li ; Dick, A. ; van den Hengel, A, "Boosting Object Retrieval With Group Quires", Journals & Magazines,Signal Processing Letters, Volume:19 Issue:11,2012.
[5]. Cheolkon Jung ; Key Lab. of Intell, Perception & Image Understanding of Minist. of Educ. of China, Xidian Univ., Xi"an, China ; Jiao, L.C. ; Yanbo Shen Ensemble ranking svm for learning rank. Machine Learning for Signal Processing (MLSP), 2011.
[6]. Freund, R. Iyer, R. E. Schapire, Y. Singer, and G. Dietterich. An efficient boosting algorithm for combining preferences. Jthenal of Machine Learning Research, 4:933–969, 2010.
[7]. M. Belkin,P. Niyogi, andV. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. Jthenal of Machine Learning Research, 7(Nov):2399–2434, 2010.
[8]. R. Herbrich, T. Graepel, and K. Obermayer. Large margin rank boundaries for ordinal regression. In Advances in Large Margin Classifiers, 2010.
[9]. Geng, L. Yang, C. Xu, and X.-S. Hua. Ranking model adaptation for domain-specific search. In CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management, pages 197–206, 2009.
[10]. Geng, L. Yang, C. Xu, and X.-S. Hua. Ranking model adaptation for domain-specific search. In CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management, pages 197–206, 2009.