

A Unified Framework For Extracting Similar Patterns From Uncertain Data Set

Ankur Jain

M. E.

Affiliation: RGPV Bhopal (M.P.)

Abstract

Tourism is a joyful travelling concept by which one meets to known or unknown places or location, communities, culture, language, peoples, art object, ancient things. Many negative impacts of tourism occur when the person is totally unknown to travelling place or he/ she travels first time the touring place. Some of the consequences will arise at the time of paying travelling cost. Generally it happens there where the person is unknown about travelling cost per KM. Many time Bus, Car or other vehicle fare increased by the vehicle owner for tourist. This is biggest disadvantages of tour. Other disadvantage are time wasting, more hotel cost, insufficient facilities, unfair fare, misguide, long journey, unclean rest house, loot etc. Due to all this, we need a proper standard website which can capable to remove all limitation of tour and provide transparent tour in low cost with cover all tourist places in better and proper way. This paper presents an artificial intelligent approach for extract the optimal tour information according to requirement. Raw travel information is stored in data ware house and according to chosen of three cities by traveler our tool design a circular cluster that contain all basic information with connection of one city to another city with proper manner that one can travel more places in less time and less expense of money for the required tour. Here selection of all three cities is uncertain.

This tool is implementation of "K-Medoids Clustering Algorithm". Key to our algorithms are exploiting detailed source models, using different filtering ideas to find the tour that fulfill the requirement of user on the basis of three cities selected by user. Good scaling properties are obtained using Artificial Intelligence technique. The tool is "AUTO TOUR DESIGN" that helps the traveller in most significant way and approaches to that places where traveller want to travel in minimum cost with losing less time and enjoying more without taking tension of costing, unknown place phobia, looting fear, unavailability of hotel's room, less facilities and knowledge weakness about distances.

KEYWORDS: Data Mining.

Introduction

K-medoid is a classical partitioning technique of clustering that clusters the data set of n objects into k number of clusters [1, 3]. This k : the number of clusters required is to be given by user. This algorithm operates on the principle of minimizing the sum of dissimilarities between each object and its corresponding reference point. The algorithm randomly chooses the k objects in dataset D as initial representative objects called medoids. A medoid can be defined as the object of a cluster, whose average dissimilarity to all the objects in the cluster is minimal i.e. it is a most centrally located point in the given data set. Then for all objects in the dataset, it assigns each object to the nearest cluster depending upon the object's distance to the cluster medoid. After every assignment of a data object to particular cluster the new medoid is decided.

With the increase in Information Technology, the size of the databases created by the organizations due to the availability of low-cost storage and the evolution in the data capturing technologies is also increasing. These organization sectors include retail, petroleum, telecommunications, utilities, manufacturing, transportation, credit cards, insurance, banking and many others, extracting the valuable data, it necessary to explore the databases completely and efficiently. Knowledge discovery in databases (KDD) helps to identifying precious information in such huge databases. This valuable information can help the decision maker to make accurate future decisions. KDD applications deliver measurable benefits, including reduced cost of doing business, enhanced profitability, and improved quality of service. Therefore Knowledge Discovery in Databases has become one of the most active and exciting research areas in the database community.

The Aspects of Auto Tour

Basically all the following researches and data are based on the experience of all the tours taken earlier. All research is nothing but it is the huge collections of combine experience [4] of cities, tour places, distances, that is already taken earlier in the form of different ways, which works like as Fig. 1.



Fig. 1

Artificial intelligence Approach

It is based on fuzzy logic in which artificial intelligence tool also works. This technology is based on fuzzy logic and artificial intelligence. Here artificial intelligence help to maintain the discipline the calculation of finding nearest tour place with minimum expenses.

With combination of both approaches that are fuzzy logic and artificial intelligence we prepare a unique formula for finding nearest tour places with a sequence that tour is enjoyable with less expense and less time consume.

Data warehouse Approach

It is a central repository of data which is created by integrating data from one or more disparate sources. Data warehouses store current as well as historical data and are used for creating trending reports for senior management reporting such as annual and quarterly comparisons.

The integrated data source systems may be considered to be a part of a distributed operational data store layer. Data federation methods or data virtualization methods may be used to access the distributed integrated source data systems to consolidate and aggregate data directly into the data warehouse database tables. This integrated data warehouse architecture supports the drill down from the aggregate data of the data warehouse to the transactional data of the integrated

source data systems.

In our research the collection or summarization of tour after making cluster of cities worked as a DATAWARE HOUSE.

In our research the collection or city information with related tour places as shown Fig. 2.

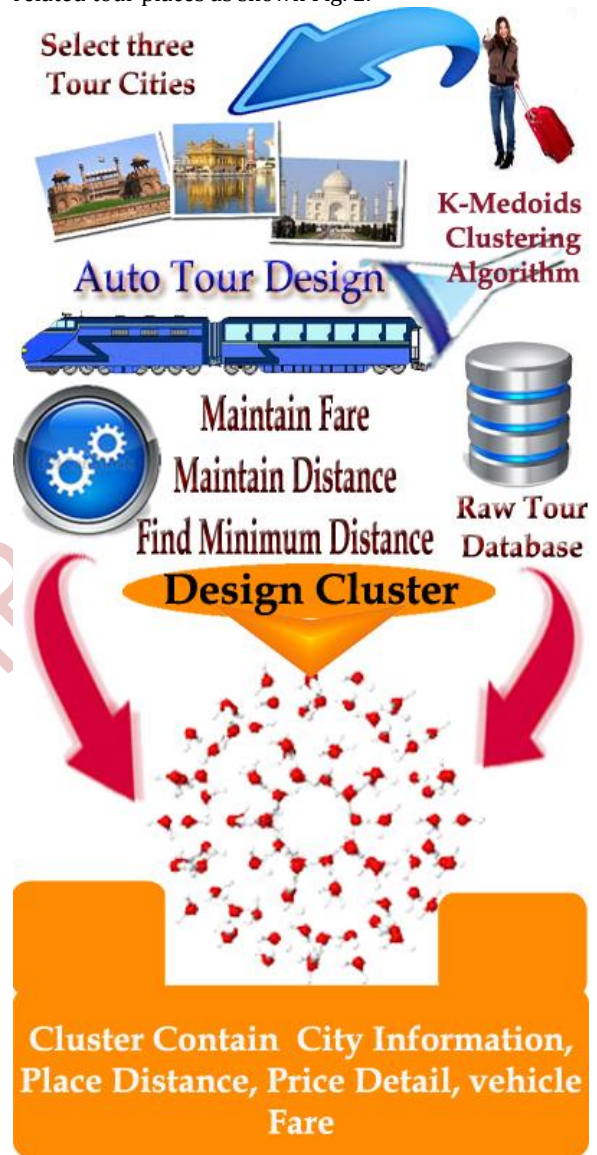


Fig. 2

Data mining Approach

Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets involving methods at the intersection of

artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data preprocessing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.

Mining knowledge from large amounts of spatial data is known as spatial data mining. It becomes a highly demanding field because huge amounts of spatial data have been collected in various applications ranging from geo-spatial data to bio-medical knowledge. The database can be clustered in many ways depending on the clustering algorithm employed, parameter settings used, and other factors. Multiple clustering can be combined so that the final partitioning of data provides better clustering. In this paper, an efficient k-medoids clustering algorithm has been proposed.

The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining). This usually involves using database techniques such as spatial indices. By this approach our tool work more efficient and the value of this tool is increased by this approach [5]. As you see the Fig. 3 that shows the data mining approach more clearly defined in Fig 3.

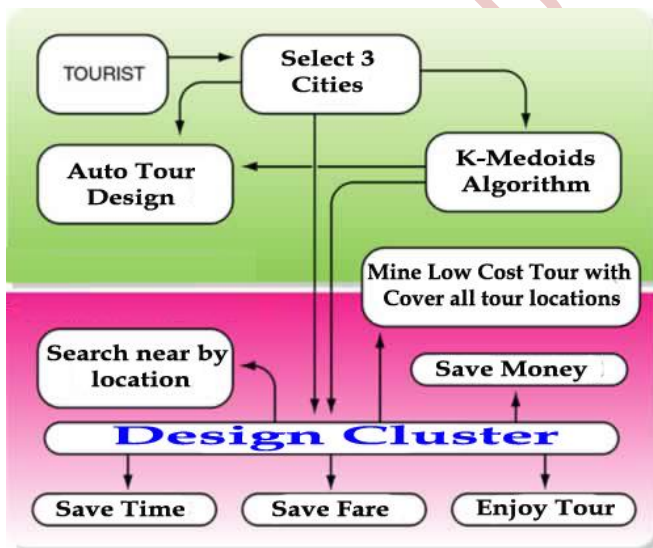


Fig. 3

K- Medoids Clustering algorithm

K-medoid is a classical partitioning technique of clustering that clusters the data set of n objects into k number of clusters [1, 3]. This k : the number of clusters required is to be given by user. This algorithm operates on the principle of minimizing the sum of dissimilarities between each object and its corresponding reference point. The algorithm randomly chooses the k objects in dataset D as initial representative objects called medoids. A medoid can be defined as the object of a cluster, whose average dissimilarity to all the objects in the cluster is minimal i.e. it is a most centrally located point in the given data set. Then for all objects in the dataset, it assigns each object to the nearest cluster depending upon the object's distance to the cluster medoid. After every assignment of a data object to particular cluster the new medoid is decided.

1) *Input*

- k : the number of clusters.
- D : a data set containing n objects.

2) *Output*

- A set of k clusters.

3) *Algorithm*

1. Randomly choose k objects in D as the initial representative objects;
2. for all objects in the data set D
 - a. Find the cluster C which is nearest to object i by using the dissimilarity measure;
 - b. assign object i to cluster C ;
 - c. set the member object in cluster C having minimum intra cluster variance as new centroid of C

3. Display statistics of clusters obtained.

Reason behind the Adopted Aspect

We are in generation of Artificial Intelligence where everything can be automated operated [6]. In this field we work one step more that is automated text arrangement that is automatically arrange the text in different subjects without wasting the time in arranging in their respective subject manually.

The adopted aspect automatically break down the single notes into different parts according to criteria so by this proposed help line user concentrate only writing their work in place of managing their work.

The proposed model helps student, notes maker, news manager, speech designer and many more persons for managing their written work more refine manner and save their time in converting their rough notes into fair notes[7].

In whole world the theme that is "Save Paper, Save Tree, Save Environment, Save World", we try to discover India that if Indian people prepare notes digitally then no need to maintain the notes subject wise. Our tool is very efficient to understand or recognize that for which subject it is written.

These are all basic reasons for which we announce our tool which is very helpful for managing the notes in future with saving precious time.

Auto Tour: Deployment Strategy

Basically all the following researches and data are based on the experience of all the tour around the world.

- ❖ A vast research has to be conducted on the cities and facts related to those cities.
- ❖ Facts and statistics related to the tour are researched.
- ❖ Data warehouse will be build for the cities and relation between them. A huge amount of sample has to be collected from various cities.
- ❖ Comparison of cities is required for tour design.
- ❖ Data mining techniques are used to fetch accurate result after comparison the cities.
- ❖ Analysis will be done on the basis of the facts related and the comparison in the cities

AI tools for calculation the tour program among those three cities of city matching.

Steps of k-Medoids Clustering Algorithm Implementation

The first step in the analysis is to abstract over the input that is three cities taken from user, in order to comparison of cities where cities are a part of DATAWARE HOUSE.

The second step is to abstract over the implementation, to find particular tour after comparison of preferred cities to database cities that is actual exist for tour making process.

Conclusion

Proposed system is more efficient and more reliable for making a tour on the basis of three cities. With reference the cities it is automatically show the path, payments, cities, travel condition.

This tool is very helpful for those travellers who are regularly make tour and want to get knowledge of any unknown places for required tour.

REFERENCES

- [1] S. Abiteboul, P. Kanellakis, and G. Grahne, "On the Representation and Querying of Sets of Possible Worlds," Proc. ACM SIGMOD, 1987.
- [2] Managing and Mining Uncertain Data, C. Aggarwal, ed. Springer, 2009.
- [3] P. Andritsos, A. Fuxman, and R.J. Miller, "Clean Answers over Dirty Databases: A Probabilistic Approach," Proc. 22nd IEEE Int'l Conf. Data Eng. (ICDE), 2006.
- [4] L. Antova, C. Koch, and D. Olteanu, "From Complete to Incomplete Information and Back," Proc. ACM SIGMOD, 2007.
- [5] L. Antova, C. Koch, and D. Olteanu, "10⁶ Worlds and Beyond: Efficient Representation and Processing of Incomplete Information," Proc. 23rd IEEE Int'l Conf. Data Eng. (ICDE), 2007.
- [6] L. Antova, T. Jansen, C. Koch, and D. Olteanu, "Fast and Simple Relational Processing of Uncertain Data," Proc. 24th IEEE Int'l Conf. Data Eng. (ICDE), 2008.
- [7] C.C. Aggarwal and P.S. Yu, "Outlier Detection with Uncertain Data," Proc. SIAM Int'l Conf. Data Mining (SDM), 2008.
- [8] C.C. Aggarwal, "On Unifying Privacy and Uncertain Data Models," Proc. 24th IEEE Int'l Conf. Data Eng. (ICDE), 2008.
- [9] C.C. Aggarwal, "On Density Based Transformations for Uncertain Data Mining," Proc. 23rd IEEE Int'l Conf. Data Eng. (ICDE), 2007.
- [10] C.C. Aggarwal and P.S. Yu, "A Framework for Clustering Uncertain Data Streams," Proc. 24th IEEE Int'l Conf. Data Eng. (ICDE), 2008.
- [11] C.C. Aggarwal, J. Han, J. Wang, and P.S. Yu, "A Framework for Clustering Evolving Data Streams," Proc. 29th Int'l Conf. Very Large Data Bases (VLDB), 2003.
- [12] M. Arenas, L. Bertossi, and J. Chomicki, "Consistent Query Answers in Inconsistent Databases," Proc. 18th ACM Symp. Principles of Database Systems (PODS), 1999.
- [13] M. Ankerst, M.M. Breunig, H.-P. Kriegel, and J. Sander, "OPTICS: Ordering Points to Identify the Clustering Structure," Proc. ACM SIGMOD, 1999.