Identifying Frequent Point of Interest using Geotagged Photos

Dr. V. R. Kavitha[1] & Ms. S. Keerthipriya[2]

[1]Department of Computer Applications, PSG College of Arts and Science, Coimbatore.

[2]Department of Computer Science,

[1]kavitha.vr@grd.edu.in, [2]selvakeerthi17@gmail.com

Abstract: *The Tourist Recommendations has been a growing interest and it is identified based on the Relevant Point of Interest (POIs) to form a Personalized Trip. It is using Location Based Social Networks database. Most frequently visited places are identified as person's point of interest. The aim of this paper is to identify and analyze the main tourist attraction places in a city. In this work Association Rule Mining Algorithm is used to analyze the frequently visited locations along with the visitor's information. R programming language is used for analysis. The dataset flickr is used to identify the places where most frequently visited places based on geospatial information. The current work read the flickr dataset. The dataset is divided into different segments and individual segments are analyzed. Final result is generated by combining the results of the individual segments and it will tell you the places where people will gather frequently.*

Keywords: location based social network, trip recommendations, and spatiotemporal data mining

1. INTRODUCTION

This research work discusses about the point of interest of a tourist and also gives an idea about the places visited often by the tourists. It is difficult for a tourist to get to know about the places of visit in a city. The behavior and semantic annotation can be identified using Point of Interest (POI).The advancement in the web-based and mobile based technologies give space for the users to access social media which in turn provide more data for analysis. Social media are user generated online contents that allow the user to interact and communicate with their friends, colleagues, and families. In net photo can be shared using popular photo sharing websites like fickr. Flickr is the most popular photo sharing websites. The association rule is the most popular and well known approach for discovering interesting relationship between variables in large databases. The *arules* package available in R provides methods for creating and manipulating input dataset. It also analyze the resultant item sets and rules. The R Programming is a statistical computing language useful for analyzing statistical problem and all other related Data mining Algorithms. The proposed algorithm can be used to mine frequent item sets in a particular city.

2. LITERATURE SURVEY

Xiaoting Wang has identified the trip recommendation with the relevant point of interest (POIs), and also selecting the best time of day to visit the POIs. They propose the personalized crowd-aware trip recommendation (PersCT) algorithm to recommend personalized trips that avoids the most crowded times of the POIs. This algorithm evaluates using foot traffic information from a real life pedestrian sensor dataset and user travel histories with the PersCT using real life datasets from Flickr geo-tagged photos. Finally by PersCT method it display the overall travel histories information.[1].

Ickjai Lee mines point of interest and their associations using data mining techniques such as clustering and association rule mining to mine area of attraction, and their association patterns. For analyzing purpose the data are collected from flickr in the area of Queensland, Australia. They are the most popular tourist destinations. The authors propose a PoI association mining framework for geotagged photos with a combination of two popular data mining techniques, clustering and association rule mining. It particularly focuses on Queensland one of the hottest tourist destinations in Australia [2].

The author discussed in this paper using point of interest mining. The people behavior can be identified using their interesting photo taking patterns. Ickjai Lee analyzes geotagged photos from flickr for Queensland and Australia. Using the people's photo-taking PoI the main interesting places are used for decision making by local businesses, policy makers, and travelers[3].

Thanh-Hieu Bui discusses the Point of Interest(POI) mining from a collection of geotagged photos in combination with proper semantic annotation using additional POI information from high coverage external POI databases. The POI mining refers to the processes of pattern recognition that is clustering, extraction of information and semantic annotation. The author proposes a novel POI mining framework by using two-level clustering, the random walk and constrained clustering. Experimental results were generated based on two datasets of geotagged flickr photos of two cities in California, USA. Finally with comparing the existing approaches, they propose a novel clustering framework based on proper semantic annotation using additional POI information from a high coverage external POI databases [4].

Guochen Cai, et al, discussed to identify the most Regions of Interest (RoI) and frequent trajectory patterns of the photo-takers using the Trajectory Pattern Mining Algorithm with the dataset taken from Flickr. The software is used in the paper is JAVA programming language. Trajectory Pattern Mining Algorithm (TPM) requires three users provided input parameters. For example to find out the region of interest and frequent patterns using the flickr dataset, which contains different spatial regions: Australia Queensland. Using TPM Algorithm the photos taken from above mentioned regions were analyzed and are able to identify major cities, tourist locations, and expected tourist routes without any external user specification[5].

Yan-Tao Zheng, el al, says that the region of attraction means, a place of interest where sequence of visits by the tourists. Using GPS tagged photos it is easy to identify the most regions of attraction. The typical travel path and Region of Attraction (RoA) using GPS tagged photos, GPS-tagged photos downloaded from Internet contains the local travel information stored in a database. It builds a statistically reliable database of travel paths, and mine a list of regions of attraction (RoA). The analysis of the travel path database is done using Entropy based Mobility Measure and Z test. Using DBSCAN density-based clustering algorithm can generate the regions of attractions (RoA) and able to identify the four major cities, it includes San Francisco, New York City, Paris and London[6].

3. IMPLEMENTATION DETAILS

Point of interest is identified as a particular city where the user is frequently visiting. After analysing the literature the present study concentrates on association rule mining algorithm using R package. The R package arules presented in this paper provides, creating and manipulating input dataset and analyzing the resultant item sets and rules. The dataset used for the analyzes is a flickr dataset which comprises of a set of users and their visits to various places in a city, with a total of 3975 tours and 17,087 visits. The user-POI visits are determined based on geo-tagged YFCC100M Flickr photos. This dataset contains the information like Userid, Photoid, photo taken date, PoiId, poi theme, SeqId, latitude, longitude data.

Using Association Rule Mining the frequent place can be identified. The association rule is the most popular and well researched approach for discovering interesting relationship between variables in large a databases. R is a powerful language and environment for statistical computing and graphics. R is open source and widely adopted by statisticians, biostatisticians, and geneticists. There is a huge wealth of existing libraries which can be directly in builded in our code.

Association rules mining has been a popular mining approach for discovering positive associations. Given a set I = {I1, I2, …, Ik} of items (k-itemsets) in a transactional database D. Each transaction T ∈ D is a subset of I. It is often called frequent pattern mining since it discovers frequent patterns. A k-itemset is frequent if its' frequency is greater than or equal to a user specified threshold. An association rule is an expression in the form of X => Y ($X \cap Y = \varphi$) where X is called antecedent and Y is called consequent.

Association rules mining involves two estimates: support and confidence. The support of an itemset X ∈ D is the number of transactions in D that have X in them. That is, it is the probability X and Y being in the dataset. The confidence of an association rule X => Y is the conditional probability of having Y contained in a transaction given that X is in that transaction. Support = probability($X \cup Y$), confidence = probability ($X \cup Y$)/probability (X).

For the current analysis support value is 0.01, and confidence value is 0.5 for finding Association between items.

The arules package is used for Mining Association Rules and Frequent Itemsets with R Programming language. The arules package for R provides the infrastructure for representing, manipulating and analyzing transaction data and patterns (frequent itemsets and association rules). Mining Associations with Apriori is used to mine frequent itemsets, association rules or association hyperedges. The Apriori algorithm employs level-wise search for frequent itemsets.

Table 3.1. User Visit Data

photoID	userID	date/time taken	poiID	poiTheme	poiFreq	seqID
12344732513	10089564	1/27/2014	120	Leisure/R	357	1
12344760173	10089564	1/27/2014	188	Place of W	225	1
12321530175	10089564	2/5/2014	190	Transport	351	2
12321735773	10089564	2/5/2014	120	Leisure/R	357	2
12322029584	10089564	2/5/2014	120	Leisure/R	357	2
12322044814	10089564	2/5/2014	120	Leisure/R	357	2
11768725924	10188434	12/13/2013	97	Office	148	3
11769094426	10188434	12/13/2013	97	Office	148	3
11768756124	10188434	12/13/2013	189	Place Of A	765	3
11768597463	10188434	12/13/2013	189	Place Of A	765	3
11768762064	10188434	12/13/2013	189	Place Of A	765	3
11768766564	10188434	12/13/2013	189	Place Of A	765	3
11768359575	10188434	12/13/2013	189	Place Of A	765	3
11768616573	10188434	12/13/2013	189	Place Of A	765	3
11768619193	10188434	12/13/2013	189	Place Of A	765	3
11768378535	10188434	12/13/2013	189	Place Of A	765	3
11768382115	10188434	12/13/2013	189	Place Of A	765	3
11768636943	10188434	12/13/2013	189	Place Of A	765	3
11768643743	10188434	12/13/2013	189	Place Of A	765	3
909465863	10195518	6/3/2007	57	Place Of A	58	4
1145062626	10195518	8/15/2007	6	Place of W	67	5
1144222893	10195518	8/15/2007	6	Place of W	67	5
1144228037	10195518	8/15/2007	241	Place Of A	23	5

The table 3.1 depicts the user visit data that contains the relevant data such that photoid, userid, date/time taken, poiid, poitheme, poifrequency, sequenceid.

Table 3.2. Geospatial Data

poiID	theme	subTheme	poiName	lat	long
1	Transport	Railway St	Flemingtc	-37.7882	144.9393
2	Mixed Use	Retail/Off	Council H	-37.8143	144.9666
3	Place Of A	Library	The Melb	-37.8149	144.9673
4	Leisure/R	Informal (	Carlton Ga	-37.8061	144.9713
5	Place of W	Church	St Francis	-37.8119	144.9624
6	Place of W	Church	Wesley Ch	-37.8102	144.9682
7	Place of W	Church	St Augusti	-37.817	144.9549
8	Place of W	Church	St James (	-37.8101	144.9525
9	Place of W	Church	St Mary's /	-37.8032	144.9538
10	Place of W	Church	Romanian	-37.8052	144.967
11	Place of W	Church	Welsh Pre	-37.8104	144.9599
12	Place of W	Church	Church of	-37.8105	144.9639
13	Place of W	Church	Scots Chu	-37.8146	144.9686
14	Place of W	Church	St Michae	-37.8144	144.9692
15	Place of W	Church	Greek Ort	-37.8088	144.9783
16	Place of W	Church	St Peter's	-37.8097	144.9753
17	Place of W	Church	Lutheran T	-37.811	144.9757
18	Place of W	Church	Holy Trinit	-37.8141	144.9832
19	Place of W	Church	St Johns L	-37.8209	144.9671
20	Place of W	Church	North Mel	-37.8036	144.9477
21	Place of W	Church	Melbourn	-37.8114	144.9847
22	Place of W	Church	All Nation	-37.7959	144.969
23	Place of W	Church	Our Lady (	-37.8026	144.9693
24	Place of W	Church	St Michae	-37.7941	144.9454

The table 3.2 depicts the geospatial data that contains the relevant information that is photoid, theme, subtheme, poiname, latitude and longitude data.

4. ALGORITHM

In this paper the Apriori Algorithm is used for identifying frequent Point of Interest (POI). Apriori is an algorithm for frequent item set mining and association rule for learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets. As long as those item sets appear sufficiently often in the database. The frequent item sets determined by the Apriori can be used to determine association rules which will highlight general trends in the database. This has applications in the domains such as market basket analysis. The dataset contains 17,088 number of information in that data we split them into individual group of 300.

Using the Apriori Algorithm the frequent item set of the transaction is identified so easily. It can also identify the frequent Point of Interest (POI).

For doing the Analyses the dataset used in the work is flickr. It contains Geotagged photos which is having the information like spatio-temporal information such as latitude and longitude.

The input for the analyzes is userid and poitheme data. Finally the is generated as a chart and userid and poi as theme value.

Steps involved in the Algorithm

Step 1: Read the data for the execution of an algorithm. To read a file in R first reads the files into data frame that it creates called data. Then header=TRUE specifies that this data includes a header row and sep=”,” specifies that the data is separated by commas.

Step 2: Split the entire dataset into the specified size files for execution

Step 3: View the file in R by selecting the attributes.

Step 4: Use the split file for analyzes. The apriori algorithm the data file name with the support value and confidence value is used for the analyses.

Step 5: Then the Final Result will be Stored in Separate Files.

Output for the algorithm

The five charts display the execution of an algorithm with the splitting of 300 data elements.

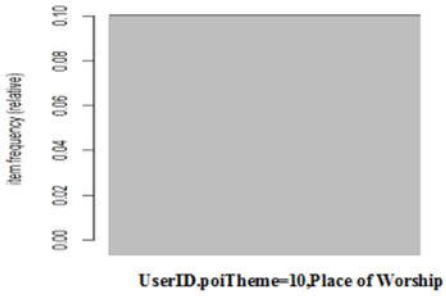

Figure3.1.1. Place of Worship

In this Figure 3.1.1 represent the higher value of data with the userid and poitheme.The value of userid is 10 and the poi theme is Place of Worship and the frequency value is 10.

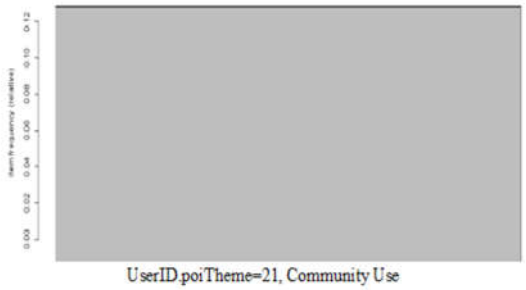

Figure 3.1.2. Community Use

In this Figure 3.1.2 represent the higher value of data with the userid and poitheme. The value of userid is 21 and the poi theme is Community use and the frequency value is 13.

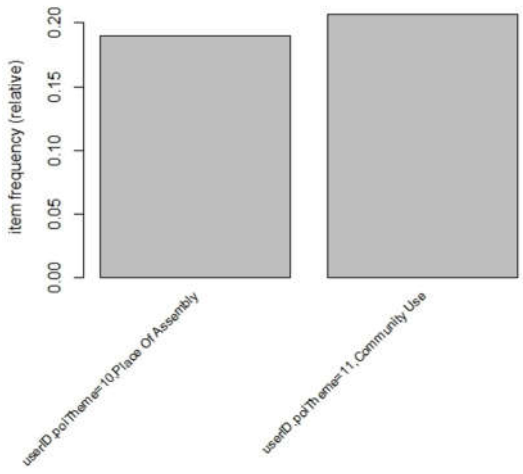

Figure 3.1.3. Community Use and Place of Assembly

In this Figure 3.1.3 represent the higher value of data with the userid and poitheme.The value of userid is 10 and the poi theme is Place of Assembly, and then another value is higher userid is 11 and the poi theme is Community use. Then the frequency value of both is 18 and 20.

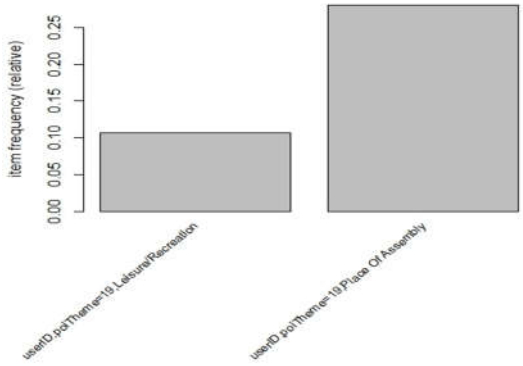

Figure 3.1.4. Place of Assembly

In Figure 3.1.4 represent the higher value of data with the userid and poitheme.The value of userid is 19 and the poi theme is Leisure Recreation, and then another value is higher userid is 19 and the poi theme is place of Assembly. Then the frequency value of both is 10 and 25.

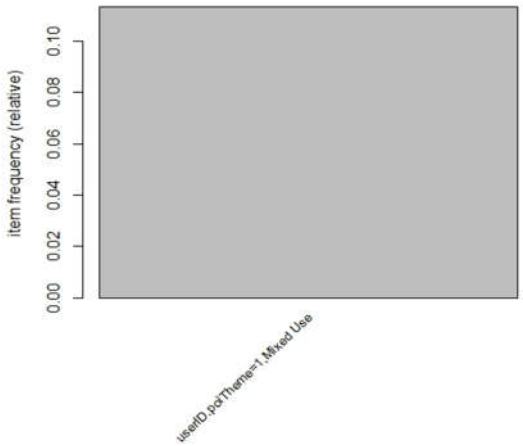

Figure 3.1.5. Mixed Use

In this figure 3.1.5 represent the higher value of data with the userid and poitheme. The value of userid is 1 and the poi theme is Mixed Use. Then the frequency value 14.

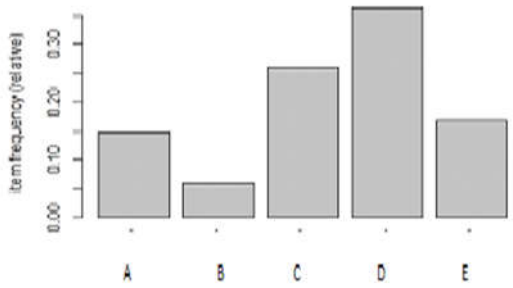

Figure 3.1.6. Place of Assembly

The value of a,b,c,d,e is

- A= userid.poitheme=1,mixed use
- B= userid.poitheme=10, Place of Assembly
- C=userid.poitheme=11, community use
- D= userid.poitheme=19, Place of Assembly
- E= userid.poitheme=21,community use

Finally when combining the five individual chart data and then execute them in the algorithm it displays the below chart. The most frequent Point of Interest is Place of Assembly. Result shows the place of Assembly is most frequency of tourist people visited again and again.

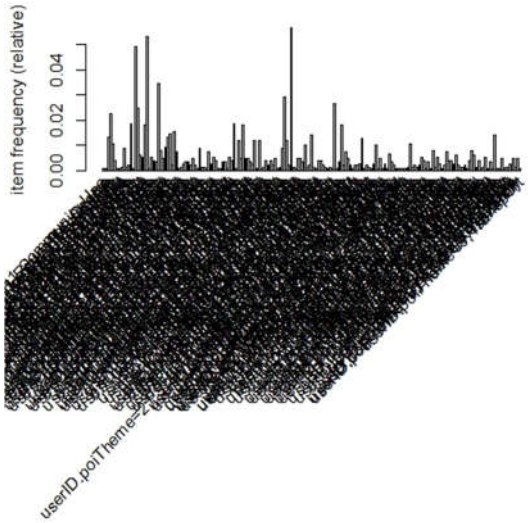

Figure 3.1.7. Overall Data Figure

On the whole, 17,088 data elements were analysed, Figure 3.1.7 shows that it is not giving the clear information. That is the reason the entire file is split into affordable size and each sub file is given for analysis. The results were also combined; it shows better output with more clarity.

Our Algorithm gives advantages while we are considering following points:

- Distributing Computing Support.
- Output is same for comprising splitted information and whole they process.
- Viewability and Clarity will be more when we are splitting the input and do the analyzes.
- Comparing to the other figures it will be clarity.

5. CONCLUSION

The geotagged photos are very useful for gathering the spatial data information. The user takes photos when he is visiting place. It might be a place where he visits frequently. The photos taken by him was posted on the flickr data base. It is one of the popular social media photo sharing sites. If the entire dataset is analysed with frequency of visits by a person, it is not giving clear information. The same file if it is split into sub files and the result is having more clarity. The photos contain information like userid and the destination location where the photo is taken. After the analysis it is identified that Place of Assembly is the place where more number of people visited with more frequency.

6. REFERENCES

[1] *Xiaoting Wang, Christopher Leckie, Jeffery Chan, Kwan Hui Lim and Tharshan Vaithianathan."Improving Personalized Trip Recommendation to Avoid Crowds Using Pedestrian Sensor Data". Proceedings of the 25th ACM International Conference on Information and Knowledge Management (CIKM'16),* ***(2016),*** *pp. 25-34.*

[2] *Ickjai Lee, Guochen Cai, Kyungmi Lee" Mining Points-of-Interest Association Rules from Geo-tagged Photos", 46th Hawaii International Conference on System Sciences,* ***(2013)****.*

[3] *Ickjai Lee, Guochen Cai, Kyungmi Lee "Points-of-Interest Mining from People's Photo-Taking Behavior", 46th Hawaii International Conference on System Sciences,* ***(2013)****.*

[4] *Thanh-Hieu Bui and Seong-Bae Park "Point of interest mining with proper semantic annotation", November 2016, New York 2016.*

[5] *Guochen Cai, Chihiro Hio, Luke Bermingham, Kyungmi Lee, Ickjai Lee,: Mining Frequent Trajectory Patterns and Regions-of-Interest from Flickr Photos, Proceedings of 47th Hawaii International Conference on System Science, Australia* ***(2014)****.*

[6] *Yan-Tao Zheng, Yiqun Li, Zheng-Jun Zha, and Tat-Seng Chua, Mining Travel Patterns from GPS-Tagged Photo , Dept of Computer Science , National University of Singapore,****(2011)****.*

[7] *Y-T. Zheng, Z-J. Zha, and T-S. Chua, "Mining Travel Patterns from Geotagged Photos", ACM Transactions on Intelligent Systems and Technology, Vol. 3 (3): 56,* ***(2012)****.*

[8] *Y. Yang, Z. Gong, and L. Hou, "Identifying Points of Interest by Self-tuning Clustering", SIGIR,* ***(2011)****, pp.883-892.*

[9] *D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg, "Mapping the World's Photos", In Proceedings of the 18th International Conference on World Wide Web, ACM, New York,NY,* ***(2009)****, pp.761-770.*

[10] *R. Agrawal, and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases", In Proceedings of the 20th International Conference on Very large Data Bases,* ***(1994)****, pp. 487-499.*

[11] *P.C. Wong, P. Whitney, and J. Thomas, "Visualizing Association Rules for Text Mining", In IEEE Symposium on Information Visualization, San Francisco, CA, (1999), pp. 120-123.*

[12] *A. Savasere, E. Omiecinski, and S. Navathe. An efficient algorithm for mining association rules in large databases. Proceedings of the 21st International Conference on Very large Database,* ***(1995)****.*

[13] *Anurag Choubey, Ravindra Patel, J.L. Rana, "A Survey of Efficient Algorithms and New Approach for Fast Discovery of Frequent itemset for Association Rule Mining", IJSCE, ISSN: 2231-2307, vol.1, issue 2,* ***(2011)****May.*