Recommending Combinations of Appointment Places from Hive based Big Data

Sujoung Oh, Bohyun Kim, Minsoo Lee

Abstract—Since the development of web, we can use various type data such as movie, music, and social network data. These data is very useful to make recommendation system. For this reason, several recommendation system studies used the web data to construct outstanding system. In this paper, we proposed a method to recommend combinations of appointment places based on web data. Our system generates ranking to extract best store, and use location data to recommend suitable store. By considering these two main features, the proposed system recommend best and suitable store to user to recommend combinations of appointment places. Our system provides a chance to design combinations of appointment places with high-quality easily.

Keywords—Combinations of appointment places, Hive, Recommendation system, Yelp

I. Introduction

Since the development of web, various data are accumulated in database. These data is very useful to extract knowledge in several areas. If the data is movie data such as score, genre, and etc., then we can use the data as movie recommendation analysis information. Likewise, several areas which include music, social network, and restaurant use web data to extract useful information. The recommend system based on the web data is very valuable to user because a user can avoid boring movies or bad restaurant by considering several features in web data. For this reason, several studies related to recommendation system are conducted. In this paper, we proposed a method to recommend combinations of appointment places making using web data. Using web data, we generate ranking for various store types such as “pub”, “shopping mall”, and ”spa”. We also consider location information to recommend suitable store.

Although a restaurant is very famous and superior if the location is too far to visit, the restaurant should have excluded in recommendation list. To consider the problem, we use location data as one of the recommendation feature. Using these two crucial features which include ranking and location, we recommend best and suitable stores to user. Furthermore, our system can cover several store types. Therefore our system can be used as combinations making method. For example, if a user wants to visit several place (pub, spa, and shopping mall) then our system can recommend best stores for each category by considering ranking and location. The proposed system is designed based on HIVE to cover big data. To implement this system, we used YELP data set which includes several store type data.

II. Related Works

By development of Internet, people can save and record large amount of data in the Web. In proportion to the increase in the data amount, various studies are being conducted related to the web data analysis. In particular, research on the recommendation system based on an evaluation of the various users has been actively conducted. Typically there are many recommendation systems such as friend recommendation system on social networks, movie recommendations, music recommendations, etc.

A. Social Network Friend Recommendation

Social Network recommendation system is a system that recommend user’s friend to user. In other words, the system is still friendships, but not to recommend a list of users that guess would be the future friendships. [1] designed a potential friend recommender system in social network of biology field to show the effectiveness of proposed framework. [2] suggested four recommender algorithms in enterprise social networking site using survey and field study.

B. Movie Recommendation

Movie recommendation refers to the user that the user still unwatched movie worth recommending the movie to be interesting to see. The major methods in recommendation systems are collaborative and content-based filtering. [3] proposed a hybrid approach based on content-based and collaborated filtering, implemented a movie recommendation system. [4] developed a model that contained consideration of
users’ context in addition to users; personality and multiple applications such as recommendation and promotion.

C. Music Recommendation

Music recommendation is sillier to movie recommendation system. This system refers to a service provided by the music list. Recommended genre, such as pop or jazz and also recommended the title song from substantial research have been researched. [5] designed the Music Recommendation System to provide a personalized service of music recommendation. And the content-based, collaborative and statistics-based recommendation methods are proposed, which are based on the favorite degrees of the users to the music groups.

D. JSON

JSON (Java Script Object Notation) is the data expression method when human exchanged data from internet. JSON is made up of text, so people and machines can read and write easily. It is independent of the programming language and platform thus it is good for the exchange of objects between different systems. It can be use directly to eval command in Java Script. This is because JSON adopted Java Script syntax. This characteristic is benefit in the Web environment using Java Script frequently. However, practically when people use eval command it is susceptible to inflow the infection from external. Most modern Web browsers such as Mozilla Firefox 3.5, Internet Explorer 8, Opera 10.5, Safari and Google Chrome included the function only for JSON parser, therefore using this function is more safe and rapid way.[6]

E. Apache Hive

Apache Hive is Data Warehouse infra structure built on top of Hadoop.[7] It supports data summarize, query and analysis. Primary it created in Facebook but currently it used in company such as Netflix and they develop it.[8] Apache Hive analyzes huge data sets stored in the data storage system, such as Apache HDFS and Apache HBase. It supports SQL language called HiveQL and also supports all function of MapReduce. And it provides index contained bitmap index for executing query faster.[9] By default, Hive is stored in Apache Derby database embedded metadata. It provides the option to use a different server / client database, such as MySQL.[10] And Hive support Text file, Sequence file, ORC and RC file currently. Apache Hive also includes Hive-Metastore. It contains statistics and schemas that are useful in data exploration and query optimization.[11]

III. Approach for Combinations of Appointment Places

In this section, we introduce our proposed method to recommend combinations of appointment places. The proposed method is consisted of five steps. First, we gathered data for various stores such as restaurant, pub, shopping mall and etc. In the next step, the gathered data is filtered based on categories. After processing the data, we analyze the data to make a ranking for each store. The ranking is used to recommend store. Finally, using original data and ranking, we recommend combinations of appointment places to user. Figure 1 indicates outline of proposed method.

After preprocessing the gathered data, we analyze the data to make useful knowledge. Based on the knowledge, we construct recommendation list for each category. The category means characteristic of store such as restaurant, shopping mall, and spa. Using the recommendation list, the user can make a combination of appointment places easily.

A. Data Collection

To collect various store data, we used yelp data set. The yelp data set includes information for store such as address, hours, categories, city and etc. These data can be used to extract best store list among the all of store. Particularly, the “stars” is very useful to evaluate the store. Figure 2 shows yelp data example.

The number of yelp data set is 61,184, and the number of attribute is more than 500,000. As shown in figure 2, the data
is consisted as Json file format. To construct database, we converted Json file format into csv file format.

B. Data Preprocessing

To exclude useless data, we conduct pre-processing for yelp data. Among the store data, a few data is useless to infer best store such as “longitude”, “latitude”, “business_id”, and “neighborhoods”. On the contrary, several data is useful to infer best store such as “stars”, “review_count”, “city”, and “open”. For this reason, we selected essential data to infer best store. Figure 3 present raw data and pre-processed data.

D. Recommendation list extraction

In this step, we analyze store data for each type. Using “stars” data, we rank stores and recommend high ranking store. Additionally, by considering “location” information, we recommend stores to user because the location is very important to user to visit store. Although the store has very high ranking compared to other similar store, if the location is too far from the user, the user cannot visit the store. For this reason, we used “location” data as useful recommendation feature. After analyzing the store data, our data has two crucial attribute to recommend store. The one of the types is “ranking” which can be used to recommend best store by considering opinion of other visitors. The other is “location” which can be used to recommend suitable store to user by considering location of users. Using these two attributes, we recommend best store which is nearby user.

C. Data Analysis

In this step, we first categorized store in detail based on the rule. The rule is that if “A” store has pub as category and “B” store also has pub as category, they are grouped as “Pub”. Using the rule, all stores are grouped for each class.

<table>
<thead>
<tr>
<th>Type</th>
<th>Pub</th>
<th>Restaurant</th>
<th>Café</th>
<th>Spa</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>D</td>
<td>G</td>
<td>J</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>E</td>
<td>H</td>
<td>K</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>F</td>
<td>I</td>
<td>L</td>
<td></td>
</tr>
</tbody>
</table>

Figure 4 Example for grouped store

Figure 5 shows example for recommendation. In figure 5, the square indicates store and the size of square means ranking. The line colors indicate suitability to visit store by user. As shown in figure 5, we recommend Black Square to user because the square is larger than others which are located in inner green line. One of the Gray Squares is larger than Black Square; however the square is located between inner green line and outer dotted red line. For this reason, the largest square is considered as unsuitable store to user. Using two values, we recommend best and suitable store to user.

E. Recommending Combinations of Appointment Places

We generate combinations of appointment places based on categories. For example, if a user want to 3 categories which include “Pub”, “Restaurant”, and “shopping mall” as appointment place, then we recommend best store for each category.
First, we obtained the category list from the user. After obtaining the categories from the user, we extract the best store for each category. By considering the location of the user, we recommend the best and suitable store to the user for each category. Figure 6 indicates an example for combination making. For each category, we calculate ranking to find the best store. Using the ranking, we make a store list. Based on these lists and location information, we recommend the store to the user. If we recommend several stores, users can choose stores they want and they can make a combination of appointment places easily.

IV. Implementation

We implemented our method using Hive. The Hive provides a chance to analyze big data based on SQL (Structured Query Language). Figure 7 indicates query examples to extract the best stores.

Figure 7 Query Example

In figure 7, we extract restaurants, shopping stores, and cafes by considering the stars and location for each store. In example, if a store is nearby Pittsburgh with high stars, then the store can be extracted as a candidate best store.

Figure 8 Running status

The Hive divides jobs into two processes, which include map and reduce. As shown in figure 8, our queries are executed successfully.

Figure 9 Result for example query

Figure 9 presents results for the hive query to extract the best stores. The “AVA cafe” and “Amazing Cafe” are extracted as best stores for the café category. Likewise, the proposed method recommends the best stores for each category.

<table>
<thead>
<tr>
<th>Restaurant Name</th>
<th>Shopping Name</th>
<th>Stars</th>
<th>Cafe Name</th>
<th>Stars</th>
<th>City</th>
<th>State</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVA Cafe Lounge</td>
<td>Dreaming Art</td>
<td>5.0</td>
<td>AVA Cafe + Lounge</td>
<td>5.0</td>
<td>Pittsburgh</td>
<td>PA</td>
</tr>
<tr>
<td>AVA Cafe Lounge</td>
<td>Dreaming Art</td>
<td>5.0</td>
<td>Amazing Cafe</td>
<td>5.0</td>
<td>Pittsburgh</td>
<td>PA</td>
</tr>
<tr>
<td>AVA Cafe Lounge</td>
<td>Nettleton Shop</td>
<td>5.0</td>
<td>AVA Cafe + Lounge</td>
<td>5.0</td>
<td>Pittsburgh</td>
<td>PA</td>
</tr>
<tr>
<td>AVA Cafe + Lounge</td>
<td>Nettleton Shop</td>
<td>5.0</td>
<td>Amazing Cafe</td>
<td>5.0</td>
<td>Pittsburgh</td>
<td>PA</td>
</tr>
<tr>
<td>Bai Ling Chinese Restaurant</td>
<td>Dreaming Art</td>
<td>5.0</td>
<td>AVA Cafe + Lounge</td>
<td>5.0</td>
<td>Pittsburgh</td>
<td>PA</td>
</tr>
<tr>
<td>Bai Ling Chinese Restaurant</td>
<td>Dreaming Art</td>
<td>5.0</td>
<td>Amazing Cafe</td>
<td>5.0</td>
<td>Pittsburgh</td>
<td>PA</td>
</tr>
<tr>
<td>Bai Ling Chinese Restaurant</td>
<td>Nettleton Shop</td>
<td>5.0</td>
<td>AVA Cafe + Lounge</td>
<td>5.0</td>
<td>Pittsburgh</td>
<td>PA</td>
</tr>
<tr>
<td>Bai Ling Chinese Restaurant</td>
<td>Nettleton Shop</td>
<td>5.0</td>
<td>Amazing Cafe</td>
<td>5.0</td>
<td>Pittsburgh</td>
<td>PA</td>
</tr>
</tbody>
</table>

Figure 10 Example Result for Combinations of Appointment Places
Our proposed method recommends best stores for various categories such as shopping mall, café, restaurant, and etc. Figure 10 shows recommend list for three categories which are selected by user. In our scenario, a user selected three categories and Pittsburgh as location. By considering user needs, our system recommended best stores. Furthermore, using these stores, the system provides a chance to make an appointment course making. For this reason, user can make a plan easily using only two features.

v. Conclusion

We proposed a method to make an appointment courses by considering stars and location. Using the stars, we can recommend best stores to user. We also can propose suitable stores by considering location. In this paper, we implemented our system based on Hive and Yelp store data. The proposed method extracted best stores for each category by considering stars and location successfully.

In future works, we will use review count to consider the number of person who visit the store. The review count may be used to check reliability for stars. For example, in case of 5 stars store with 1 review count, we cannot trust the stars because the starts are considered by only one person. On the other hand, in case of 5 stars with 1000 review counts, we can recommend the store as best store with high reliability. For this reason, we will use various data to propose more suitable method to extract best stores in future works.

Acknowledgment

This work was supported by the Technology Innovation Program (or Industrial Strategic technology development program, 10046955) funded by the Ministry of Trade, Industry & Energy (MI, Korea).

References


Sujoung Oh is an MS student at the department of computer science and engineering of Ewha womans university in Seoul, Korea. Her research interests include data mining, data warehouse, Web information infrastructures, and stream data processing.

Bohyun Kim is an MS student at the department of computer science and engineering of Ewha womans university in Seoul, Korea. Her research interests include data mining, data warehouse, Web information infrastructures, and stream data processing.

Minsoo Lee is an associate professor at the department of computer science and engineering of Ewha womans university in Seoul, Korea, since March 2002. He received his Ph.D degree from the University of Florida, and his Master, and Bachelor at the department of computer science and engineering of Seoul National University, in 2000, 1995, 1992, respectively. He worked for LG Electronics from July 1995 to July 1996. And worked for Oracle Corporation in the US as a Senior Member of Technical Staff. During this period, he worked on developing business intelligence tools. His research interests include data mining, data warehouse, Web information infrastructures, and stream data processing.