Proc. of The Second Intl. Conf. On Advances In Computing, Control And Networking - ACCN 2015

ISBN: 978-1-63248-073-6 doi: 10.15224/ 978-1-63248-073-6-23

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.

Sujoung Oh
Ewha Womans University
Korea

Bohyun Kim2
Ewha Womans University
Korea

Minsoo Lee
Ewha Womans University
Korea

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.

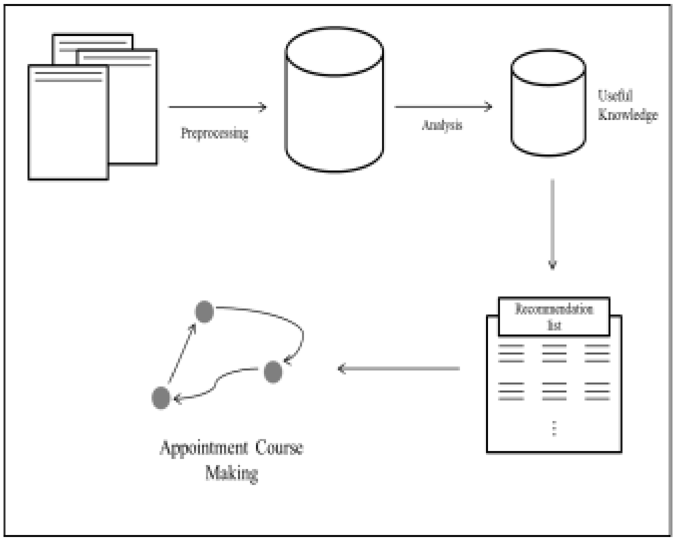
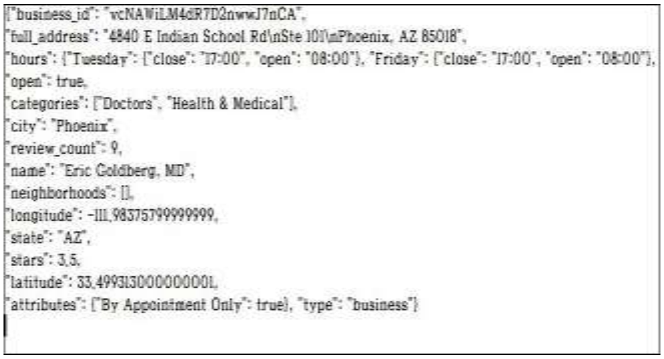


Figure 1 Outline of Proposed Method

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.



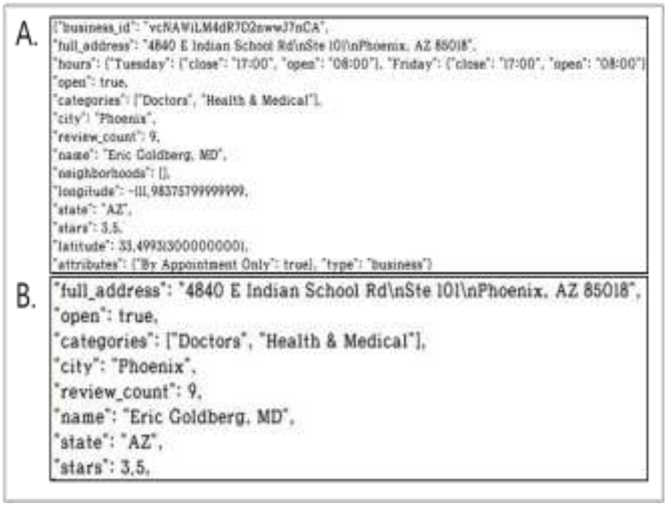
```
{"business_id": "vcNAWiLM4dR7D2nwwJ7nCA",
"full_address": "4840 E Indian School Rd\nSte 101\nPhoenix, AZ 85018",
"hours": {"Tuesday": {"close": "17:00", "open": "08:00"}, "Friday": {"close": "17:00", "open": "08:00"},
"open": true,
"categories": ["Doctors", "Health & Medical"],
"city": "Phoenix",
"review_count": 9,
"name": "Eric Goldberg, MD",
"neighborhoods": [],
"longitude": -111.98375799999999,
"state": "AZ",
"stars": 3.5,
"latitude": 33.499313000000001,
"attributes": {"By Appointment Only": true}, "type": "business"}
```

Figure 2 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.



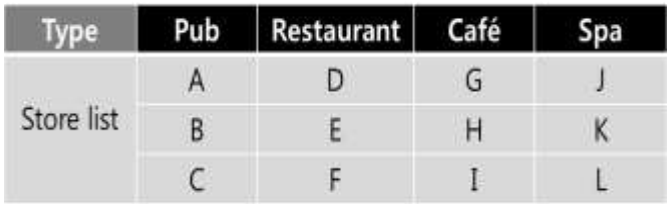
```
{"business_id": "vcNAWiLM4dR7D2nwwJ7nCA",
"full_address": "4840 E Indian School Rd\nSte 101\nPhoenix, AZ 85018",
"hours": {"Tuesday": {"close": "17:00", "open": "08:00"}, "Friday": {"close": "17:00", "open": "08:00"}
"open": true,
"categories": ["Doctors", "Health & Medical"],
"city": "Phoenix",
"review_count": 9,
"name": "Eric Goldberg, MD",
"neighborhoods": [],
"longitude": -111.98375799999999,
"state": "AZ",
"stars": 3.5,
"latitude": 33.499313000000001,
"attributes": {"By Appointment Only": true}, "type": "business"}
```

```
"full_address": "4840 E Indian School Rd\nSte 101\nPhoenix, AZ 85018",
"open": true,
"categories": ["Doctors", "Health & Medical"],
"city": "Phoenix",
"review_count": 9,
"name": "Eric Goldberg, MD",
"state": "AZ",
"stars": 3.5,
```

Figure 3 Raw Data and Pre-processed Data

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.



Type	Pub	Restaurant	Café	Spa
Store list	A	D	G	J
	B	E	H	K
	C	F	I	L

Figure 4 Example for grouped store

Figure 4 shows example for grouped store. The store A, B and C are grouped in "Pub" category because they have a same "Pub" category. Similarly, the other categories are grouped as different categories. After grouping, we analyze other data such as "stars", "location", and "review_count". Using these data, we rank stores to extract best store.

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.

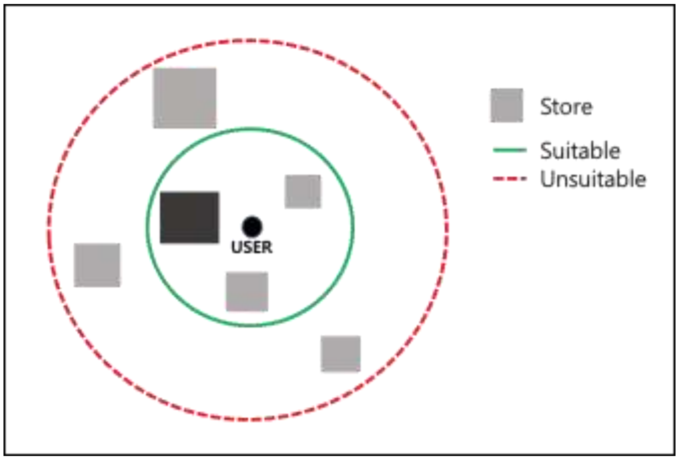


Figure 5 Recommendation Example

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.

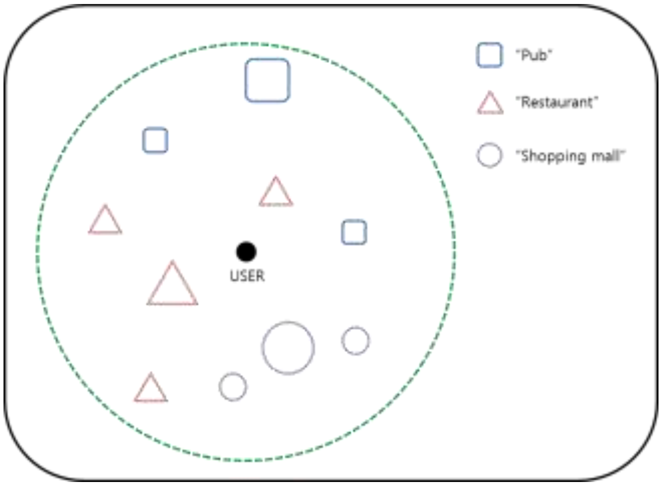
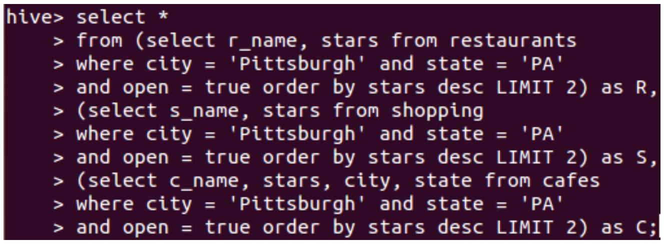


Figure 6 Example for Combinations of Appointment Places

First, we are obtained category list from user. After obtaining categories from user, we extract best store for each category. By considering location of user, we recommend best and suitable store to user for each category. Figure 6 indicates example for combination making. For each category, we calculate ranking to find best store. Using the ranking, we make a store list. Based on these list and location information, we recommend store to user. If we recommend several stores, users can choice stores they want and they can make a combination of appointment places easily.

IV. Implementation

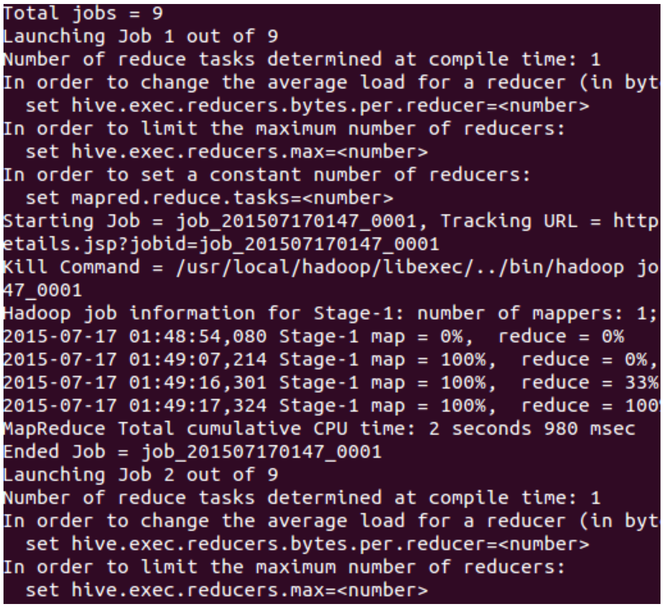
We implemented our method using Hive. The Hive provides a chance to analysis big data based on SQL (Structured Query Language). Figure 7 indicates query example to extract best stores.



```
hive> select *
    > from (select r_name, stars from restaurants
    > where city = 'Pittsburgh' and state = 'PA'
    > and open = true order by stars desc LIMIT 2) as R,
    > (select s_name, stars from shopping
    > where city = 'Pittsburgh' and state = 'PA'
    > and open = true order by stars desc LIMIT 2) as S,
    > (select c_name, stars, city, state from cafes
    > where city = 'Pittsburgh' and state = 'PA'
    > and open = true order by stars desc LIMIT 2) as C;
```

Figure 7 Query Example

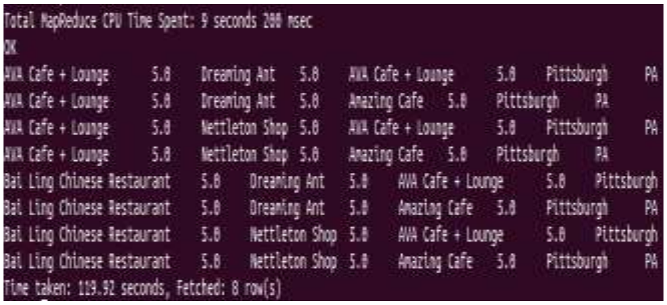
In figure 7, we extract restaurants, shopping stores, and cafes by considering stars and location for each store. In example, if a store is nearby Pittsburgh with high stars, then the store can be extracted as candidate best store.



```
Total jobs = 9
Launching Job 1 out of 9
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in byt
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapred.reduce.tasks=<number>
Starting Job = job_201507170147_0001, Tracking URL = http
etails.jsp?jobid=job_201507170147_0001
Kill Command = /usr/local/hadoop/libexec/../bin/hadoop jo
47_0001
Hadoop job information for Stage-1: number of mappers: 1;
2015-07-17 01:48:54,080 Stage-1 map = 0%,  reduce = 0%
2015-07-17 01:49:07,214 Stage-1 map = 100%,  reduce = 0%,
2015-07-17 01:49:16,301 Stage-1 map = 100%,  reduce = 33%
2015-07-17 01:49:17,324 Stage-1 map = 100%,  reduce = 100
MapReduce Total cumulative CPU time: 2 seconds 980 msec
Ended Job = job_201507170147_0001
Launching Job 2 out of 9
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in byt
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
```

Figure 8 Running status

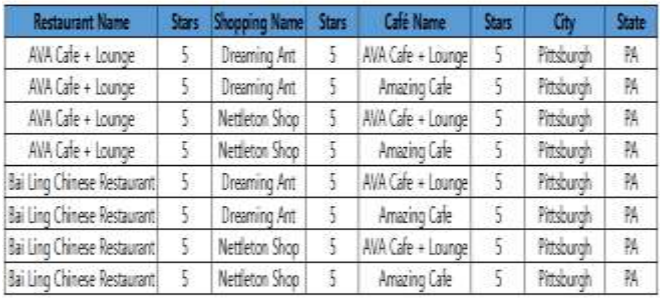
Figure 8 shows running status in Hive when executing hive query. The Hive divides jobs into two processes which include map and reduce. As shown in figure 8, we confirmed that our queries are executed successfully.



```
Total MapReduce CPU Time Spent: 9 seconds 200 msec
OK
AVA Cafe + Lounge       5.0     Dreaming Ant    5.0     AVA Cafe + Lounge       5.0     Pittsburgh      PA
AVA Cafe + Lounge       5.0     Dreaming Ant    5.0     Amazing Cafe    5.0     Pittsburgh      PA
AVA Cafe + Lounge       5.0     Nettleton Shop  5.0     AVA Cafe + Lounge       5.0     Pittsburgh      PA
AVA Cafe + Lounge       5.0     Nettleton Shop  5.0     Amazing Cafe    5.0     Pittsburgh      PA
Bai Ling Chinese Restaurant     5.0     Dreaming Ant    5.0     AVA Cafe + Lounge       5.0     Pittsburgh
Bai Ling Chinese Restaurant     5.0     Dreaming Ant    5.0     Amazing Cafe    5.0     Pittsburgh      PA
Bai Ling Chinese Restaurant     5.0     Nettleton Shop  5.0     AVA Cafe + Lounge       5.0     Pittsburgh
Bai Ling Chinese Restaurant     5.0     Nettleton Shop  5.0     Amazing Cafe    5.0     Pittsburgh      PA
Time taken: 119.92 seconds, Fetched: 8 row(s)
```

Figure 9 Result for example query

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



Restaurant Name	Stars	Shopping Name	Stars	Café Name	Stars	City	State
AVA Cafe + Lounge	5	Dreaming Ant	5	AVA Cafe + Lounge	5	Pittsburgh	PA
AVA Cafe + Lounge	5	Dreaming Ant	5	Amazing Cafe	5	Pittsburgh	PA
AVA Cafe + Lounge	5	Netleton Shop	5	AVA Cafe + Lounge	5	Pittsburgh	PA
AVA Cafe + Lounge	5	Netleton Shop	5	Amazing Cafe	5	Pittsburgh	PA
Bai Ling Chinese Restaurant	5	Dreaming Ant	5	AVA Cafe + Lounge	5	Pittsburgh	PA
Bai Ling Chinese Restaurant	5	Dreaming Ant	5	Amazing Cafe	5	Pittsburgh	PA
Bai Ling Chinese Restaurant	5	Netleton Shop	5	AVA Cafe + Lounge	5	Pittsburgh	PA
Bai Ling Chinese Restaurant	5	Netleton Shop	5	Amazing Cafe	5	Pittsburgh	PA

Figure 10 Example Result
for Combinations of Appointment Places

Proc. of The Second Intl. Conf. On Advances In Computing, Control And Networking - ACCN 2015
Copyright © Institute of Research Engineers and Doctors, USA .All rights reserved.
ISBN: 978-1-63248-073-6 doi: 10.15224/ 978-1-63248-073-6-23

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

[1] X. Xie. Potential Friend Recommendation in Online Social Network, 2010 IEEE/ACM Int'l Conference on & Int'l Conference on Cyber, Physical and Social Computing, pp. 831-835, Dec. 2010..

[2] J. Chen, W. Geyer, C. Dugan, M. Muller, I. Guy., "Make New Friends, but Keep the Old" – Recommending People on Social Networking Sites, SIGCHI Conference on Human Factors in Computing System.., pp. 201-210, 2009.

[3] G. Lekakos, P. Caravelas, A hybrid approach for movie recommendation, Multimedia Tools and Applications, Vol. 36, Issue 1-2, pp. 55-70, Jan. 2008.

[4] C. Ono, M. Kurokawa, Y. Motomura, H. Asoh, A Context-Aware Movie Preference Model Using a Bayesian Network for Recommendation and Promotion, 11th International Conference UM 2007, Vol. 4511, pp. 247-257, 2007.

[5] H. Chen, A. Chen, A music recommendation system based on music data grouping and user interests, 10th international conference on Information and knowledge management, pp. 231-238, 2001.

[6] JSON, Wikipedia, [Online] Available:
http://ko.wikipedia.org/wiki/JSON

[7] J. Venner, Pro Hadoop, Apress, 2009.

[8] USE case Study of Hive/Hadoop, Slideshare, [Online] Available: http://www.slideshare.net/evamtse/hive-user-group-presentation-from-netflix-3182010-3483386

[9] L. Chunck, Hadoop in Action, ManningPubn, 2010.

[10] Facebooks Petabyte Scale Data Warehouse using Hive and Hadoop, [Online] Available:
http://www.sfbayacm.org/wp/wp-content/uploads/2010/01/sig_2010_v21.pdf

[11] Y. He, R. Lee, Y. Huai, Z. Shao, N. Jain, X. Zhang, Z. Xu, RCFile: A Fast and Space-efficient Data Placement Structure in MapReduce-based Warehouse Systems, IEEE 27th international conference on Data Engineering, ISSN. 1063-6382, pp. 1199-1208, Apr. 2011.

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.