IMPROVED RANKING BASED RECOMMENDATION SYSTEM BY FUZZY DATA ANALYSIS

Tanmaya Bhatt, Department of Computer Science, Graphic Era University, Dehradun,India
bhatt.tanmaya@gmail.com

Abstract—***Recommender system is a personalized information filtering technique used to identify desired number of items based on interests of a user. It is based on past behavior , relations to other users , item similarity and context. A system is developed using past user ratings by applying different techniques. The techniques focus on improving the accuracy of the recommendations and other important aspects of recommendation quality, such as the diversity of recommendations. Fuzzy systems are used to solve the data analysis problems with the main aim of creating models for predictions and interpretation. In this paper we aim to improve recommendations by fuzzifying a Trip Advisor dataset.***

Keywords— ***Recommender systems, recommendation accuraacy, evaluation of recommender system, fuzzy data, fuzzy data analysis***

I. Introduction

Recommender systems have revolutionized the way people find products, information, and even other people online. These systems study and find patterns of different behavior to know what someone will prefer from among a collection of things that , that person has never experienced. The technology behind recommender systems has evolved over the past few years into a humongous collection of tools that help enable the researcher to develop effective algorithms for better accuracy of recommendation systems.

The algorithms are differentiated usually by the type of filter they have. There are different approaches to recommendation like

- Collaborative Filtering (Items recommended based only on the user's past behavior),
- Content Based (Items recommended based on item features),
- Personalized Learning To Rank (Recommendation is treated as a ranking problems),
- Demographic (Items recommended based on user features),
- Social-Recommendations(trust-based)
- Hybrid (combine any two of the above).

Traditionally, recommendation systems gather the information by explicitly asking the users or by implicitly collecting the information by their behavior. But sometimes users don’t give the correct information about them. So the alternative of this problem is collecting the information about the user preferences from their behavior which can be found by their recent activities which are available in online communities.

In this paper we report on the method by which we can create recommendation system which will work by using fuzzy data analysis. To give better recommendations we have used trip advisor data set in this paper where we have reviews of the hotels given by actual visitors to these hotels.

This paper is organized as follows. Section 2 presents the related work. Section 3 presents the methods used to create data sets. In section 4, we present the experimental results and summary. And in section 5, we give the conclusion derived from the experiment.

II. Related work

This section contains all the research work done by researchers on similar topics along with the technology and dataset used.

A. *Trip Advisor Data Set*[1]

The TripAdvisor dataset consists of around 240,000 customer-supplied reviews of 1,850 hotels. Each review is associated with a hotel and a star-rating, 1-star (most negative) to 5-star (most positive), chosen by the customer to indicate his/her evaluation. This dataset contains around 90,000 hotel reviews, in three subsets (for libsvm): the train, validation and test subsets contain approximately 76,000, 6,000 and 13,000 reviews respectively. Each of these three subsets contains a balanced number of negative (1-star and 2-star) and positive (4-star and 5-star) reviews. The dataset also includes neutral reviews (e.g. with a rating value of 3) that are used in three-class classification. For binary classification, these neutral reviews are omitted from the dataset.

Data set is mostly related to the contents of the single database table where every column represents a particular variable and the rows are related to the given member of dataset. Dataset makes a record of the values for each variable and those values are known as datum. The properties of data set are defined by various characteristics That is the number and type of variables and different statistical measures applicable.

B. *libSVM*[2]

It is an open source learning library for support vector machine written in C++ with C API with header (*.h*) files that live in the*subversion/include* directory of the source tree. It can solve C-SVM classification, nu-SVM classification, one-class-SVM, epsilon-SVM regression, and nu-SVM regression. It also provides an automatic model selectiontool for C-SVM classification.

It implements the SMO algorithm for kernelized support vector. It is easy to use software for svm classification and regression.Two steps are followed for accessing the libSVM software:-

- Train a data set for creating a model by creating class labels {1, -1} or {0, 1} depending on the implementation. If your dataset has different labels for the positive or/and negative class, make sure that this is supported or convert the class labels otherwise

- Use the model for predicting the information about testing data set.
- The sub-routines of LIBSVM contains svm train and svm predict. Svm train developes a two-class problems by decoupling a multi-class problem and one several times calls SVM train.

LIBSVM,a popular source machine learning library, was developed in National Taiwan University. Some other opensource machine learning toolkits like KNIME, GATE, Orangeand scikit-learn also use the SVM learning code from LIBSVM library. This software is free and was released under the 3- clause BSD license

C. *Clustering and Collaborative Filtering*[3]

Breese et al. used a Bayesian clustering model to cluster users based on their ratings. Their work showed mixed results; in some cases the clustering approach was competitive in terms of accuracy of the ratings and in others it performed poorly. Ungar and Foster also used a Bayesian approach to cluster users based on their preferences. Their results also showed that clustering users was not a particularly successful approach. Graph theoretic methods for clustering users based on preferences were discussed, however they do not evaluate the impact these clusters have on recommendation accuracy or quality. Finally, users were clustered using a scalable neighborhood algorithm and, once again, the clustering approach had a higher MAE than the standard collaborative filtering method.

D. *Trust and Collaborative Filtering*[3]

Social networks, and trust in particular, have been used to generate recommendations for users. In these cases, trust is used directly to generate the recommendation. This work follows from the fact that people tend to develop connections with people who have similar preferences. Trusting the opinion of another particularly speaks to this type of similarity. The applicability of this effect to recommender systems has been established in several papers. Ziegler and Lausenthat showed a correlation between trust and user similarity in an empirical study of a real online community. Using All Consuming, an online community where users rate books. The authors showed that users were significantly more similar to their trusted peers than to the population as a whole. This work was extended in which augmented the analysis of the All Consuming community and added an analysis.

The second result in used the FilmTrust system(described below) where users have stated how much they trust their friends in a social network and also rated movies. Within that community, results also showed a strong correlation between trust and similarity in movie ratings. Further work in shows that trust captures similarity in more nuanced ways, such as similarity on items with extreme ratings and large differences. Empirical results show that using trust from social networks can improve recommendations. O'Donovan and Smyth performed an analysis of how trust impacts the accuracy of recommender systems. Using the MovieLens dataset, they create trust-values by estimating how accurately a person predicted the preferences of another. Those trust values were then used in connection with a traditional collaborative filtering algorithm, and an evaluation showed significant improvement in the accuracy of the recommendations.

Massa and Bhattacharjee also conducted a study on the applicability of trust in recommender systems. Their study relied on the user ratings of products and trust ratings of other users from e-pinions as their dataset.

In the FilmTrust recommender system mentioned above, trust is used in place of the Pearson correlation coefficient to generate predictive ratings. Results showed that when the user's rating of a movie is different than the average rating, it is likely that the recommended rating will more closely reflect the user's tastes. As the magnitude of this difference increases, the benefit offered by the trust-based recommendation also increases. Moleskiing, at http://moleskiing.it , is another real system built to utilize trust ratings in a recommender system. Using a similar approach, it recommends routes to users based on information supplied by trusted peers.

E. *Fuzzy Logic*[4]

In the past few decades, fuzzy logic has been used in a wide range of problem domains. Although the fuzzy logic is relatively young theory, the areas of applications are very wide: process control, management and decision making, operations research, economies and, for this paper the most important, pattern recognition and classification. Dealing with simple 'black' and 'white' answers is no longer satisfactory enough; a degree of membership (suggested by Prof. Zadeh in 1965) became a new way of solving the problems. A fuzzy set is a set whose elements have degrees of membership. A element of a fuzzy set can be full member (100% membership) or a partial member (between 0% and 100% membership). That is, the membership value assigned to an element is no longer restricted to just two values, but can be 0, 1 or any value in-between. Mathematical function which defines the degree of an element's membership in a fuzzy set is called membership function. The natural description of problems, in linguistic terms, rather than in terms of relationships between precise numerical values is the major advantage of this theory. An idea to solve the problem of image classification in fuzzy logic manner as well as comparison of the results of supervised and fuzzy classification was the main motivation of this work. Behind this idea was also the question if the possible promising results can give the answer to the question of diminishing the influence of person dealing with supervised classification.

III. METHEDOLOGY

To ensure the effectiveness of fuzzy data we created two sets of data:

1. Percentage Data: The no of reviews were tested for accurate recommendations based on % calculated.

2. Fuzzy Data: The reviews were tested for accurate recommendations based on the fuzzy values of the data set.

A. Percentage Data

We classified the hotel review sections as *i, j, k, l, m, n, o*. Each of these values are the ratings given by the users to the hotels as depicted in the sample data given in Fig.1.To understand the data each of these sections were summed individually for each hotel such that we get *Σi, Σj, Σk, Σl ,Σm, Σn* and *Σo* for each hotel. This was done to compensate for that fact that the number of customers reviewing each hotel can differ. To calculate the percentage value each section of every hotel we used the following formula:

$$\%value=\frac{\sum x}{n\cdot 5}$$

Where n is the total no of users and x is the corresponding summation of the values ie.*i, j, k, m, n, o.*

Review Ratings			
Sleep Quality	***Service***	***Author Location***	***Author Name***
5	5	Boston	gowharr32
5	5	Madison, Wisconsin	Nancy W
4	5	Ketchikan, Alaska	Janet H
3	5	Florida	TimothyFlorida
1	1	Armstrong, BC	Karen Armstrong_BC
4	4	Kingston, Canada	Shane33333
2	5	Boise, Idaho, USA	Bnkruzn

Fig.1.Example of a part of the dataset in its original form.

The data was converted to a new form. All the given ratings were summed to acquire cumulative data as depicted in Fig. 2. Here we considered the hotel's rating to improve our results. The proposed form for the data set was to get an accurate review of the hotels which were subject to the amount of users.

Summed Ratings			
Hotel Name	***Service***	***Cleanliness***	***Sleep Quality***
BEST WESTERN PLUS Pioneer Square Hotel	991	1012	1021
Grace Inn Phoenix	117	122	123
BEST WESTERN PLUS Eagle Rock Inn	47	54	44
Comfort Inn Near Old Town Pasadena - Eagle Rock	624	609	851
Rodeway Inn & Suites Pacific Coast Highway	861	869	855
Dunes Inn - Sunset	30	36	41
BEST WESTERN PLUS Pioneer Square Hotel	991	1012	1021

Fig. 2. Example of a part of the dataset in its converted form

The data was finally converted to its % form.

% Ratings			
Hotel Name	***% - Ratings - Service***	***%- Ratings - Cleanliness***	***%- Ratings – Sleep Quality***
BEST WESTERN PLUS Pioneer Square Hotel	91.75926	92.84404	87.63948
Grace Inn Phoenix	68.82353	64.21053	55.90909
BEST WESTERN PLUS Eagle Rock Inn	72.30769	83.07692	67.69231
Comfort Inn Near Old Town Pasadena - Eagle Rock	47.63359	46.13636	64.22642
Rodeway Inn & Suites Pacific Coast Highway	93.58696	94.45652	89.5288
Dunes Inn – Sunset	60	72	63.07692
Hotel Name	% - Ratings - Service	%- Ratings - Cleanliness	%- Ratings – Sleep Quality

Fig. 3. % form of a part of the dataset

B. Fuzzifying Data using Triangular Membership Function

Zadeh introduced the membership function in his first paper on fuzzy set in 1965. Zadeh proposed his theory using the membership function with the range (0,1) working on all possible values.

There are different applications of membership function, one of those are capacities in decision theory.

Fuzzifying Data means to convert a given data into fuzzy data. A membership function characterizes the fuzzy set perfectly. A membership function has a membership value or its degree which tells about the grade of membership of the element of the function to the fuzzy set. It is used to graphically represent the fuzzy set. Such that in the graph, *x* axis is used to represents the universe of discourse, and the *y* axis is used o represent the degrees of the membership. The triangular membership function is specified by its three parameters. In which the curve is a function of a vector *x*, and depends on the scalar parameters *a, b* and *c.*

$$triangle(x;a,b,c)=\begin{cases} 0 & x\leq a \\ \frac{x-a}{b-a} & a\leq x\leq b \\ \frac{c-x}{c-b} & b\leq x\leq c \\ 0 & c\geq x \end{cases}$$

where *a* is the lower limit , *c* is the upper limit and *b* is the value.

We took each %value from above and fuzzified it. Each of these %values was subjected to the triangular membership function and was given it's equivalent fuzzy value.
It is to be remembered that the triangular membership function gives data in ranges. Thus the values closer to the peak have a value nearing 1 while the values near the lower and upper limits have values nearing 0.

IV. EMPERICAL RESULTS

We did the testing in two sets

A. Testing Set 1

We used the percentage values as training set for libSVM's svm-train member. On testing it with svm-predict using a set of test values we noted that accuracy obtained with simply the percentage value was around 34.97 to 37.47% using SMO.

B. Testing Set 2

We used the fuzzy data as training set for libSVM. We noted that the accuracy in this case was around 58.77% to 60.97%.

Rating Name	% Data Accuracy	Fuzzy Data Accuracy
Service	35.65	60.97
Cleanliness	37.47	58.77
Sleep Quality	34.97	58.97
Location	36.23	58.81
Rooms	35.75	59.22
Overall	37.01	59.1
Value	37.22	58.78

Fig. 4.1. States the results of testing on the two sets

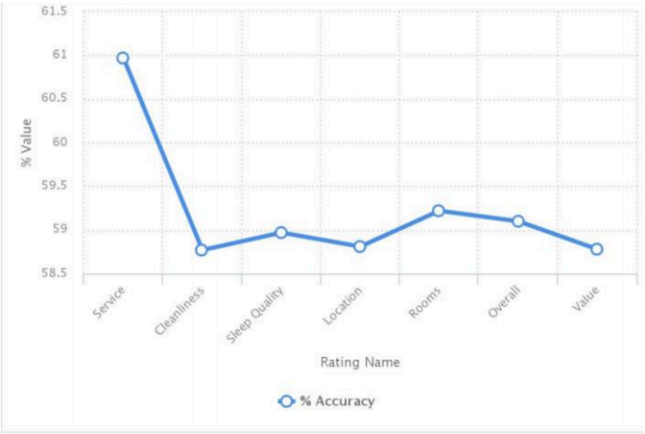

Fig.4.2 shows Accuracy for all seven aspects of % data

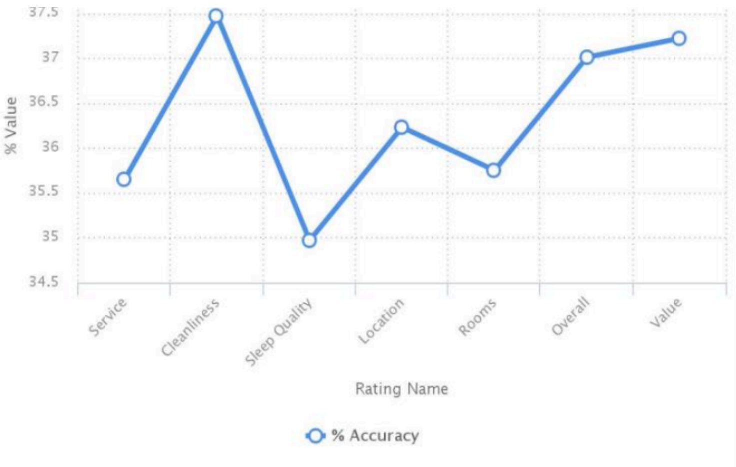

Fig.4.3. shows Accuracy for all seven aspects of Fuzzy data

V.CONCLUSION

In this paper, we study how fuzzy data analysis leads to better recommendations. We proposed an efficient form of data which leads to better recommendations for the user. Theoretical analysis is provided to guarantee the efficiency and effectiveness of our method. Extensive experimental results on real-world data sets as in this case done on the Trip Advisor data sets also confirm our theoretical findings.

REFERENCES

1. Advances in Social Media Analysis,edited by Mohamed MedhatGaber, MihaelaCocea, NirmalieWiratunga, AyseGoker
2. https://www.csie.ntu.edu.tw/~cjlin/libsvm/
3. Improving Recommendation Accuracy by Clustering Social Networks with Trust Tom DuBois Computer Science Department University of Maryland, College Park College Park, MD 20741 tdubois@cs.umd.edu Jennifer Golbeck Human-Computer Interaction Lab University of Maryland, College Park College Park, MD 20741 jgolbeck@umd.edu John Kleint Computer Science Department University of Maryland, College Park College Park, MD 20741 jk@cs.umd.edu Aravind Srinivasan Computer Science Department University of Maryland, College Park College Park, MD 20741 srin@cs.umd.edu,K. Elissa, “Title of paper if known,” unpublished.
4. IMAGE CLASSIFICATION BASED ON FUZZY LOGIC I. Nedeljkovic MapSoft Ltd, Zahumska 26 11000 Belgrade, Serbia and Montenegro igor.n@sezampro.yuY. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740-741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
5. C. Aggarwal, and P. Yu, *A Framework for Clustering*
6. Massive Text and Categorical Data Streams, SDM, (2006).
7. C. Aggarwal, Data Streams: Models and Algorithms,
8. Springer, (2007).
9. C. Aggarwal, Social Network Data Analytics, Springer,

   (2011).
10. J. Allan, R. Papka, and V. Lavrenko, *On-line new event*

    detection and tracking, SIGIR Conf., (1998).,*Document Collections*, SIGIR Conf., (1992).
11. A fuzzy logic approach to analyzing gene expression data**,** PETER J. WOOLF, YIXIN WANG
12. Multiscale event detection in social media,Xiaowen Dong1 · Dimitrios Mavroeidis2 · Francesco Calabrese3 · Pascal Frossard4
13.]FUZZY DATA MINING AND GENETIC ALGORITHMS APPLIED TO INTRUSION DETECTION,Susan M. Bridges Bridges@cs.msstate.edu Rayford B. Vaughn vaughn@cs.msstate.edu 23rd National Information Systems Security Conference October 16-19, 2000
14. Pairwise Preference Regression for Cold-start Recommendation,Seung-TaekPark,Samsung Advanced Institute of Technology,Mt. 14-1, Nongseo-dong, Giheung-gu,Yongin-si, Gyunggi-do 446-712, South Korea,park.seungtaek@gmail.com,WeiChu,Yahoo! Labs,4401 Great America, Parkway, Santa Clara, CA 95054, USA,chuwei@yahoo-inc.com
15. Proceedings of the ACM RecSys'09 Workshop on
16. Recommender Systems & the SocialWeb edited by,DietmarJannach, Werner Geyer, Jill Freyne, Sarabjot Singh Anand,
17. Casey Dugan, BamshadMobasher, Alfred Kobsa,October 25, 2009,New York, NY, USA
18. Multiscale event detection in social media, Xiaowen Dong1,Dimitrios Mavroeidis2,Francesco Calabrese3,Pascal Frossard4