Theoretical Approach of Search of Missing Values in Data Set using Data Mining

Ajay Singh Mavai
Department of Computer Science & Engineering
LNCT Bhopal (M.P)

Sadhna K. Mishra,Ph.D
Department of Computer Science & Engineering
LNCT Bhopal (M.P)

ABSTRACT
The uncontrollable expansion of the information over the internet creates a critical job to discover which information is supportive or useful for a particular user. This paper proposes a new filtering technique using rough set and clustering technique to seek out the nearest neighbor, So that the user can get the right choice for the collection of the objects which are appropriate for them. In this paper, we are going to find out missing values in data set by using data mining specifically with the help of filtering technique that uses a membership function which is a unique suggesting approach by combining the marking, trait, likes and features of user's information about items. Our approach moves towards the state of the art suggesting system, and reduce the recorded problems. Filtering technique suggest items by taking in to order of the taste of users, under the supposition that users will be attracted by a particular item that users alike to them have rated highly. To our best information, this is the unique study of integrating traits and likes of user's information converted into a missing value for the development of suggesting manager..

Keywords
Suggesting system, Missing value, Social Media, Clustering

1. INTRODUCTION
Social Media (SM) is a cluster of Internet-based applications that enhanced on the idea and technology of Web 2.0, and permit the arrangement and exchange of User Generated Content. SM sites can also be referred to as web-based services that permit those to generate a public/semi-public report within a domain such that they can communicatively bond with a list of other users within the system. SM is an essential source of learning of attitudes, reactions, subjectivity, opinions, approaches, estimation, influences, examinations, feelings, borne out in text, reviews, blogs, conversations, news, remarks, or some other manuscripts. Before the arrival of SM, the homepages was popularly used in the late 1990s which made it probable for common internet users to share information [1]. Nevertheless, the behavior on today's SM appears to have altered the World Wide Web into its intended original creation. SM platforms allow rapid information swap between users regardless of the location. SM can be referred to as group of human relations where communication and relationships form nodes and edges; the nodes consist of entities and the relationships between them build up the edges. Several organizations, individuals and even government of countries pursue the activities on SM in order to acquire data on how their viewers act in response to postings that concern them. SM permits the effective compilation of huge amount of data and this gives increase to major computational challenges. Nevertheless the well organized mining of the information retrieved from these huge amounts of data helps to find out helpful knowledge of supreme significance in various aspects like marketing, banking, government and defense. Again efficiently mined SM data can be used as judgment supportive device by different entities that make use of SM contents for various purposes. SM sites are usually known for information distribution, attitude/sentiment expression, and product reviews. News alerts, breaking news, political debates and government policy are also discussed and analyzed on SM sites [2]. However, while some views on SM support users and other entities to make valuable decisions, some are mere assertions and hence misleading. User's attitude/sentiments on SM such as Twitter, Face book, YouTube and Yahoo are mostly positive, negative or neutral (neutral being usually considered as no view on a particular subject). Online view can be revealed using conventional methods but this is conversely not enough considering the huge amount of information produced on all SM sites.

2. LITERATURE REVIEW
Clustering is a useful technique of data mining for finding of data allocation and samples in the core data. The main idea and target behind the clustering is to determine both the concentrated and thin sections in a data set. Data clustering has been considered in the statistics, machine learning, and database domain with various emphases. The earlier approaches do not adequately consider the facts that the data set can be too large to fit in the main memory, here we are also using a membership function theory which is a tool of sets and relations and associations for learning ambiguity, imprecision, vagueness, and uncertainty in data analysis [3].

There are some filtering techniques that recommend a item to users are demographic filtering, content based filtering and collaborative filtering. Each one has different effectiveness and accuracy about recommendation according to the applying areas and the activity level. For that reason, it is essential to build up a system that merges and join the uniqueness of each filtering technique rather than the one dependent on a individual filtering technique, if we talk about Demographic filtering technique analyzes users characteristics by demographic elements or items based on their personal attributes such as user's sex, age, height, weight job etc, and then suggests the items to the user's in different perspectives. It is suitable for objective marketing or preliminary implementation of structure, which is not enough for preference information. Content-based filtering technique filters all the information from the users based on the textual information enclosed in items, under the assumption that users will like similar items to the ones they liked before. Analysis of data is not complicated and the recommendation results are reflected easily. Items that are considered sufficiently similar to the user profile are suggested to the user. However, it is difficult to learn dynamic changes of the users. Collaborative filtering is the method, which suggests items based on the evaluation and taste of users and other users having similar preferences, under the supposition of

those users will be attracted in items that users similar to them have rated highly. This technique has higher precision in case of determining unseen preference patterns but there are a small number of difficulties in preliminary evaluation and personal examine [5].

Finally it may be concluded that, there are two strongly probable problems with the above suggesting techniques. One is the scalability, which is how rapidly a suggesting technique can generate a suggestion, and the second is to improve the quality of the suggestion for a user, so we need an especial hybrid technique which may be appropriate for particular case based suggesting system that predicts the user's favorites and suggests to them with help of choosing items, following and analyzing the user's activity patterns. It can analyze the user's preferences robotically without user's input. Nevertheless, there are few issues to analyze, if the user visits the site for first time or there aren't any commercial activities, but finally, we can get the missing value with help of rough set function which is describe broadly in next section.

3. PROPOSED MODEL

In this section, we demonstrate the proposed model by considering a data set study in which we recommend the missing values by applying membership function and using membership function. In existing models the values in given data set were fully indiscernible relation [1]. In our data set values are almost indiscernible. For example if we have one data set in which the ratings of movies is given by different users. Here we can see that in given data set for user2 item 5 value is missing; here by comparison by other users we can predict the value of user2 for item5. So by data set User1 user2 user3 will come in one group and user3 and user4 will come in another group. So as we can say that {user1, user2, usre3} and {user4, user5}, so we can say that for user 2 item5 value will be 4. Here in the case all values are in full indiscernible relation. But if values are not in full indiscernible relation then how we will predict the values in that case for that we are proposed model.

Table 1. Data Set

	Item i1	Item i2	Item i3	Item i4	Item i5
User 1	7	3	8	2	4
User 2	7	3	8	2	?
User 3	7	3	8	2	4
User 4	5	7	9	1	3
User 5	5	7	9	1	3

Here we are demonstrating proposing model by taking example of giving data set. In Table 2 given below, we consider a few users those are giving values for items. The membership function has been adjusted in such a manner that their values should lie in [0, 1] and these functions must also be symmetric. We define a relation $S(r_i, r_j)$ in order to identify the almost indiscernibility among the objects r_i and r_j, where,

$$s(ri, rj) = 1 - \frac{|v_{ri} - v_{rj}|}{2(v_{ri} + v_{rj})} \quad (1)$$

The membership function has been adjusted in such a manner that their values should lie in [0, 1] and these functions must also be symmetric

4. RESULT ANALYSIS

In this section, we discuss in detail the subsequent steps of the preprocess architecture design for the empirical study taken under consideration. A target dataset for analysis as shown in Table is considered. We have designed relations based on the attributes and computed the almost similarity between them. The relation identifies the almost indiscernibility among the objects. This result induces the equivalence classes. We obtain categorical classes on imposing order relation on this classification. The fuzzy proximity relations S_i, i = 1,2,3,4,5,6 corresponding to the items, item1, item2, item3, item4 and item5 is given in

Table 2. Data Set

	Item 1	**Item 2**	**Item 3**	**Item 4**	**Item 5**
User1	4.1	1.8	10.8	10.3	2.7
User2	1.7	5.5	4.2	6.0	?
User3	5.8	3.7	7.1	2.9	8.0
User4	2.0	2.7	4.5	5.8	5.4
User5	4.0	10.0	10.6	1.5	2.4
User6	1.9	5.9	4.8	4.1	5.2

Now on consider Almost similarity of 90%. It is observed that $S_1(r_1,r_1) = 1$; $S_1(r_1,r_5) = 0.98$; $S_1(r_1,r_6) = 0.99$; $S_1(r_2,r_2) = 1$; $S_1(r_2,r_4) = 0.91$; $S_1(r_2,r_7) = 0.94$; $S_1(r_3,r_3) =1$; $S_1(r_3,r_8) = 0.99$; $S_1(r_4,r_4) = 1$; $S_1(r_4,r_7) = .97$; $S_1(r_5,r_5) = 1$; $S_1(r_5,r_6) = .99$. Thus, the users r_1, r_5, r_6 are α – Identical. Similarly r_2, r_4, r_7 are α – Identical and r_3, r_8 are α – Identical. Therefore we get,

$R/S_1 = \{\{r_1,r_5\}, \{r_2,r_4,r_6\}, \{r_3\}\}$

Therefore, the values of the Item on 1 are classified into three categories can be shown by any integer like Similarly, the different equivalence classes obtained corresponding to the items 2, 3, 4, 5 are given below.

$R/S_2 = \{\{r_1\},\{r_2,r_6\},\{r_3\},\{r_4\},\{r_6\}\}$

$R/S_3 = \{\{r_1,r_5\}, \{r_2,r_4, r_6\},\{r_3\}\}$

$R/S_4 = \{\{r_1\},\{r_2,r_4\},\{r_3\},\{r_5\},\{r_6\}\}$

$R/S_5 = \{\{r_1,r_5\},\{r_3\},\{r_4, r_6\}\}$

Table 3. Result

	Item 1	Item 2	Item 3	Item 4	Item 5
USER1	4.1	1.8	10.8	10.3	2.7
USER2	1.7	5.5	4.2	6.0	**5.3**
USER3	5.8	3.7	7.1	2.9	8.0
USER4	2.0	2.7	4.5	5.8	5.4
USER5	4.0	10.0	10.6	1.5	2.4
USER6	1.9	5.9	4.8	4.1	5.2

So we get missing value for user2 for item 5 is 5.3.

5. CONCLUSION

Explore SM data especially attitude/sentiments articulated by SM users with data mining techniques has verify effective and helpful think about the research carried out so far in the field. This is so because of the capability data mining possess in handling noisy, huge, dynamic and vibrant data. Many authors have come up with a number of algorithms that can be used to mine the attitude of online users of the SM. Bulky number of works reviewed mainly utilized Support Vector Machine (SVM), Naive Bayes and Maximum Entropy. While some authors considered other data mining techniques like Association Rule Mining, Decision Tree, KNN and Neural Network, these techniques have not gained popularity as much as SVM, Naïve Bayes and Maximum Entropy. Nevertheless their information have been useful for trade-off interpretability reason. It is expected that future work will make use of both currently used and yet-to-be-explored data mining techniques to delve deeper into mining the ever increasing online data generated daily on SM. The results of the investigations are expected to assist different entities in retrieving vital information on SM and consequently using this information as decision support tools that helps to suggest missing values.

6. REFERENCES

[1] Bowman, Adedoyin-Olowe, M., Gaber, M., Stahl, F.: A Methodology for Temporal Analysis of Evolving Concepts in Twitter. In: Proceedings of the 2013 ICAISC, International Conference on Artificial Intelligence and Soft Computing. 2013.

[2] Andreas M. Kaplan, Michael Haenlein: Users of the world, unite! The challenges and opportunities of SM, Business Horizons, Volume 53, Issue 1, January–February 2010, Pages 59-68, ISSN 0007-6813, http://dx.doi.org/10.1016/j.bushor.2009.09.003.

[3] Aggarwal, N., Liu, H.: Blogosphere: Research Issues, Tools, Applications. ACM SIGKDD Explorations. Vol. 10, issue 1, 20, 2008.

[4] Dong, W. 2006. "Influence Modeling of Complex Stochastic Processes." July. Master's Thesis in Media Arts and Sciences.

[5] V. A. Balasubramaniyan, A. Maheswaran, V. Mahalingam, M. Ahamad, and H. Venkateswaran. A crow or a blackbird?: Using true social network and tweeting behavior to detect malicious entities in Twitter. 2010.

[6] A. L. Barabasi. The origin of bursts and heavy tails in human dynamics. Nature, 435(7039):207{211, 2005.

[7] C. M. Bishop. Pattern Recognition and Machine Learning. Springer, 2006.

[8] J. Ben Schafer, Dan Frankowski, Jon Herlocker, Shilad Sen. Collaborative Filtering Recommender System. 2007, 4321:291-324.

[9] Funakoshi , Kaname. A Content Based Collaborative Recommender system with Detailed Use of Evaluations, 2000, 1:253-256.

[10] Yao Y Y. Information tables with neighborhood semantics. In: Data Mining and Knowledge Discovery-Theory, Tools, and Technology (Dasarathy B V. Ed.), Society for Optical Engineering, Bellingham, Washington, 2000, 2: 108~116.