Efficient Grouping of Tourism Webpages Considering Ratings and Reviews

Shahnila Zaman, Sabiha Salma and Shaily Kabir
Department of Computer Science and Engineering
University of Dhaka
Dhaka, Bangladesh
nila.csedu@gmail.com, sabiha.csedu@gmail.com, shailykabir2000@yahoo.com

***Abstract*— Tourism websites consist of online user ratings and user reviews for tourist places. Data mining techniques can be applied on these opinions to automatically recommend the best tourist places to travelers. However, on one hand user ratings are subjective; on the other hand, users who write reviews may not be able to verbally express their opinions. Hence, a tourism recommender system based either on ratings or reviews will provide an inaccurate recommendation of tourist spots. In this paper, we propose a heuristic approach to group webpages considering both the user ratings and the user reviews for a tourism recommender system. Typically, the user ratings are given on a numeric scale. But, since the users write reviews in natural language, the words used in the reviews must be converted into numeric values. After the numeric conversion of the user ratings and the user reviews, we assign a weight to each page. Next, we measure similarity among the weighted pages by our proposed similarity measure and group them using a fuzzy clustering algorithm. The experimental results show that the highest rated and the highest reviewed web pages are grouped into the same cluster while considering both the user ratings and the user reviews.**

Keywords-tourism website; recommender systems; user ratings; user reviews; similarity measure.

I. INTRODUCTION

Now a days most people who wish to travel search tourism websites for information related to tourist destinations. Typically, the tourism websites consist of user ratings and user reviews in which the users freely express their opinion regarding different tourist places. Recent data have shown that travel planning has been greatly influenced by online resources. In particular, the tourism websites have large impact on tourism related decisions, and a large number of travelers search online before they decide to travel.

There exists some recommender systems for tourism websites based either on the online ratings or reviews. Moreover, these ratings are subjective, and users who make reviews may find it difficult to verbally express their opinion. Keeping them in mind, we consider that converting the user reviews to numerical values and incorporating them with the numeric ratings will be useful for determining the best tourist places in a particular location. Therefore, this paper proposes a heuristic approach to group web pages considering both the user ratings and the user reviews for a tourism recommender system.

II. RELATED WORKS

In tourism domain, the travel agents usually provide recommendations to the tourists regarding places worthwhile to visit. In this regard, Loh et al. [1] developed a tourism recommender system that assists the agents in recommending travel plans for the customers. However, Zhang et al. [2] showed that the travelers are turning to tourism websites to plan trips instead of going to travelling agents. Park et al. [3] further exhibited that prospective travelers generally feel consumer reviews to be more familiar, understandable, and trustworthy. Moreover, narrative reviews may convey rich information that cannot be captured in numerical ratings [4].

Even if the consumers use the online ratings and reviews to make decisions, they may still be overloaded with information as a large number of websites offer recommendations along with conflicting reviews. To overcome the problem, the travelers apply several heuristics to make tourism decisions where heuristics are general rules of thumb people rely on to arrive at their judgments [5]. However, these heuristics may result in less accurate decisions and biased responses.

Instead of depending on the heuristics, a tourism recommender system is needed so that the travelers can easily and efficiently discover the best tourist places to visit. Senecal and Nantel [6] demonstrated that the online recommender systems are the most influential source than any traditional recommendation sources such as ‘human expert’ and ‘other consumer report’ in the consumers' product choice processes. In addition, Cheong and Morrison [7] described the importance of user-generated content (UGC), produced from the feature comments and reviews of the users about brands and products, on the consumer’s future online purchase. Meeting up the need, Zhang et al. [8] developed the Informed Recommender specifically for the tourism domain where it collects relevant information from the user reviews to make recommendations. However, none of the recommender systems developed so far considers both the user ratings and the user reviews to create the recommendations.

III. PROPOSED APPROACH

Travel and tourism websites usually consist of the user ratings and/or the user reviews where tourists can submit their opinions about the tourist destinations. The tourists' ratings, often scaled from 1 to 5 stars, offer an idea of how good a tourist spot is. However, this arrangement of the ratings limits the users' ability to fully express their opinion about a place. Since the ratings are subjective, different users may rate the same spot differently even if they feel the same way about a tourist place. Besides, considering only the user reviews for recommendations is also problematic because people might use words in a wrong context or they might use very few words to describe a place. Moreover, they may not be able to verbally express their views accurately due to the language barriers. All of these limitations may affect the overall performance of the recommender systems.

Our proposed approach incorporates both the user ratings and the user reviews for grouping the tourism webpages to improve the reliability and the accuracy of the recommendations. The proposed approach involves the following four major steps:

1. Collection and conversion of user ratings and reviews: Collect the user ratings and reviews for each web page. Convert the ratings to the numeric values on a scale of 1 to M, and the review words to the numeric values on a scale of 1 to N, where M and N are positive integers.

2. Page-weight generation: Each webpage is assigned a weight based on their numeric ratings and review values.

3. Similarity calculation: Apply our proposed similarity measure in order to calculate the alikeness among the weighted pages.

TABLE I. EXTRACTED CATEGORY-1 WORDLIST

10-star words			
Amazing	awesome	best	Brilliant
Breathtaking	excellent	fantastic	fabulous
Gorgeous	incredible	loved	lovely
Magical	magnificent	marvelous	masterpiece
Outstanding	perfect	spectacular	Stunning
Superb	terrific	wonderful	unique
Great	love	phenomenal	greatest
Incredibly	brilliantly	breathtakingly	excellently
Gorgeously	magnificently	marvelously	perfectly
Spectacularly	superbly	terrifically	wonderfully
Amazingly	amazed	coolest	thrilled
Remarkably	serenest	miraculously	dazzling

TABLE II. EXTRACTED CATEGORY-2 WORDLIST

7-star words			
Good	enjoy	Enjoyed	Enjoyable
Nice	exciting	Fun	entertaining
Entertain	entertained	Friendly	interesting
Liked	peaceful	Informative	safe
Cool	recommend	Huge	large
Clean	comfortable	Spacious	refreshing
Better	fresh	Pleasant	impressive
Reasonable	fine	Relaxing	Relax
Beautiful	convenient	Helpful	inexpensive
Appropriate	charming	Favorite	satisfactory
Pleasantly	impressed	Organized	relaxed

4. Group the webpages: Utilize fuzzy-C means clustering algorithm for generating an offline page model.

A. *Collection of words from the user reviews*

Generally, the user ratings for a particular tourism webpage are given in a range of one star to five stars, where one star represents the worst and the five stars shows the best. Thus, the ratings can easily be changed to a scale of 1 to 5. However, the conversion of a verbal review to a numeric scale is rather difficult as various users may use different words to express the same feelings. Fortunately, while analyzing the users' reviews it reveals that most people tend to use the same set of words to articulate the tourist places. For an example, when a place fascinates the tourists, they usually use words-'*great*', '*excellent*', '*best*' to describe that place. In contrast, for the detested places, they frequently use '*bad*', '*terrible*', and etc. In the similar manner, the users who fairly choose a place but are not entirely charmed, they may express as being '*good*' or '*fun*'. On the other hand, the average state of disliking may be expressed by the words- '*boring*' or '*disappointing*'. We mine all of these words from the user reviews to enrich our extracted review words database. Moreover, we include past, present and future tenses of some commonly used words. In addition, we consider the words such as '*masterpiece*', '*informative*' and etc., commonly used to describe some specific attractions like '*Museums*' and '*Art Galleries*'.

B. *Conversion of words into numerals*

We categorize all the extracted words into four groups. *Category-1* includes the words which people use to state for the loving tourist place. However, *Category-2* consists of the words describing their liking for the place. *Category-3* comprises the words employed by the people for the boring place, while *Category-4* covers the words used to present the hatred. Next, we assign each category of words into a numeric value based on a scale of 1 to 10. *Category-1* is given the highest value of 10 for representing the best, whereas

Category-2 is assigned a value of 7 for showing goodness. In the similar manner, *Category-3* is set to a value of 4 and *Category-4* is put the lowest value of 1 for presenting the worst. Table I, II, III and IV present the wordlists extracted from the user reviews.

TABLE III. EXTRACTED CATEGORY-3 WORDLIST

4-star words			
Okay	Noisy	Small	crowded
Crowd	Rude	Lazy	boring
Bored	disappointing	disappointed	plain
Expensive	waste	difficult	criticism
Criticized	flaws	negative	unfortunately
Hard	overpriced	poor	ordinary
Slow	irritating	irritated	average
Bad	annoyed	annoying	annoy
Downside	bland	waiting	wait
Obstructed	loud	unfriendly	rudely

C. *Calculation of page weight*

We convert the user star ratings for each tourism page into numeric values on the scale of 1 to 5. Besides, the extracted words from the page review belonging to the four categories are assigned numeric values on the scale of 1 to 10. For example, a 3-star attraction *'Manhattan Skyline'* owns a review like "every street and park in Manhattan provides a *fantastic* photograph. The skyline is *unique* and *gorgeous* Manhattan Skyline". Therefore, this attraction is assigned a numeric rating of 3 and a numeric review of 10. Finally, we calculate a weight for each page based on the numeric ratings and the numeric reviews.

TABLE IV. EXTRACTED CATEGORY-4 WORDLIST

1-star words			
Awful	worst	terrible	Avoid
Rubbish	dull	frustrating	Painful
Pain	nightmare	gruesome	Beware
Horrible	dirty	ugly	Horribly
Avoided	nightmarish	hated	Foul

D. *Proposed similarity measure*

After assigning the weights to the pages, we calculate the alikeness among the pages. Equation (1) presents our proposed similarity measure for discovering the alikeness among two attractions P and Q.

$$Similarity(P,Q) = \frac{W_{min}(review) + W_{min}(rating)}{W_{max}(review) + W_{max}(rating)} \qquad (1)$$

Where, $= \frac{\sum w_i v_i}{\sum w_i}$, w_i is the total no. of extracted words in each category C_i (or the total no. of users giving the rating R_i) and v_i is the numeric value assigned to C_i or R_i.

Tables V, VI, and VII depict the similarity calculation among three attractions *'Times Square(TS)'*, *'Broadway Theatre(BT)' and 'Governor's Island(GI)'* using (1).

TABLE V. EXTRACTED CATEGORIES FOR TS, BT AND GI

Categories	**Numerals**	**No. of Words**		
		TS	***BT***	***GI***
Category-1	10	74	87	15
Category-2	07	47	48	10
Category-3	04	20	9	50
Category-4	01	02	01	10
Total Assigned Weight		8.05	8.57	5.06

TABLE VI. EXTRACTED RATINGS FOR TS, BT AND GI

Ratings	**Numerals**	**No. of Users**		
		TS	***BT***	***GI***
5-star	05	40	25	15
4-star	04	50	103	45
3-star	03	84	29	12
2-star	02	19	01	10
`1-star	01	02	15	05
Total Assigned Weight		3.55	3.54	3.63

TABLE VII. SIMILARITY RESULTS BETWEEN TS, BT AND GI

	TS	***BT***	***GI***
TS	1.0	0.96	0.74
BT	0.96	1.0	0.71
GI	0.74	0.71	1.0

Since the travel and tourism websites consist of thousands of tourist destinations, some places may have high ratings but low review comments while a number of others may possess low ratings and high reviews. With such fuzziness of the ratings and reviews, we reasonably apply the fuzzy c-means algorithm [9] to our similarity results in order to group the tourism pages into a number of clusters with various memberships.

IV. EXPERIMNTS AND RESULTS

For the experiments, we used a set of 600 webpages from the website *"www.tripadvisor.com"*, the largest travel website on the Internet with more than 335,000 tourist destinations in 34 different countries. We considered three countries with two hundreds of tourist places for each country. Since each place is both rated and reviewed by many users, we chose only those places having at least hundred users' comments and ratings. Thus, we included at least 600,000 ratings and reviews in our database. All experiments were performed on an Intel(R) Xeon(R) 3400 series based workstation running at 2.67 GHz with 4GB RAM.

We evaluated the performance of our proposed page grouping using the following cluster validation indices:

- Davies-Bouldin Index
- Xie-Beni Index
- Kwon Index.

The Davies-Bouldin index is a cluster validation used for evaluating the quality of the clusters. It is defined using (2).

$$DB = \frac{1}{n_c}\sum_{i=1}^{n_c} R_i, \text{ where } R_i = (R_{ij}), i = 1 \dots n_c \quad (2)$$

Besides, the Xie-Beni index is defined as follows:

$$S = \frac{\sum_{i=1}^{c}\sum_{k=1}^{n} \mu_{i,k}^2 \|v_i - x_k\|^2}{n\, min_{i\neq j}(\|v_i - v_j\|^2)} \quad (3)$$

In addition, the Kwon index is based on the Xie-Beni index, and is defined as:

$$v_K(U,V;X) = \frac{\left(\sum_{k=1}^{n}\sum_{i=1}^{c} u_{i,k}^2 \|x_k - v_i\|^2 + \frac{1}{c}\sum_{i=1}^{c}\|v_i - \bar{x}\|^2\right)}{min_{i\neq j}(\|v_i - v_j\|^2)} \quad (4)$$

All of these validation indices present better clustering of pages for small index values. We applied the following features for finding the overall performance of our proposed similarity measure and also the grouping of pages with our approach.

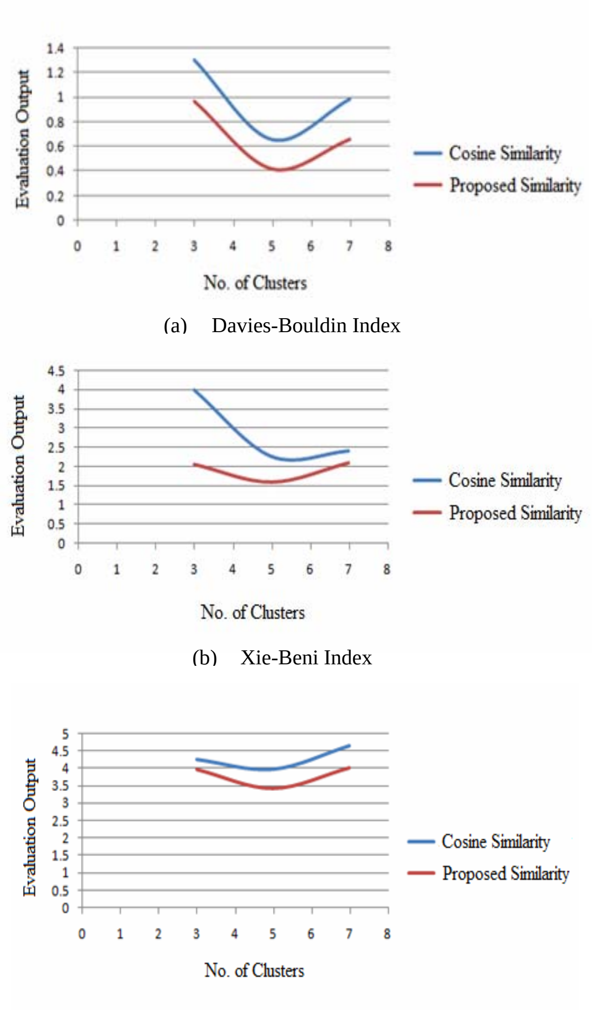
- Based on the similarity equations
- Based on the approach (ratings, reviews, or both ratings and Based reviews) giving the grouping of the highest rated and the highest reviewed pages together

A. Performance analysis with respect to similarity measure

To compare the grouping of pages using our proposed similarity measure, we carried out the same experiment with other well-known measure, namely, the Cosine similarity. The performance results of these similarity measures based on all three validation indices are shown in Fig. 1. It is already mentioned that the lower the index/evaluation output the better the clustering is. From Fig. 1, it is easy observed that our proposed similarity gives small outcome in case of all indices. Thus, our similarity shows improved grouping of tourism pages as compared to the Cosine similarity.

B. Performance evaluation taking ratings,reviews and both of ratings and reviews

Table VIII, IX and X present the results of the first ten webpages collected from the best grouping of five clusters using only ratings, only reviews, and both ratings and reviews respectively with our similarity measure for the country '*America*'. From Table X, it is easily observed that taking both



(a) Davies-Bouldin Index

(b) Xie-Beni Index

(c) Kwon Index

Figure 1. Evaluation output with proposed similarity and cosine similarity

TABLE VIII. TEN TOP PAGES OF THE BEST GROUING WITH MEMBERSHIPS WHILE TAKING ONLY THE HIGH-RATED PAGES

Webpage	***Membership***	***Rating***	***Review***
Greek town	0.9777	4	3.979
The Late Show	0.9576	4	2.405
National Geographic Museum	0.9473	4	7.594
Blue Gold Fleet	0.9368	4	3.518
Untouchable Tours	0.9166	4	5.474
Bunker Hill Monument	0.9054	4	7.431
Improve Asylum	0.8931	4	5.109

Corcoran Gallery of Art	0.8742	4	6.511
Merchandise Mart	0.8612	4	7.436
Loeb Boathouse	0.8512	4	5.558

the user-reviews and the user-ratings groups the highest-rated and the highest-reviewed web pages together. However, from Tables VIII we can find that only high-rated pages are grouped together where low-reviewed pages are present. In the similar manner, from Tables IX we can find that only high-reviewed pages are grouped together where low-rated pages are present. When the URL's of these ten webpages are checked in '*TripAdvisor*', we found that the web pages grouped by our similarity measure are also the top-ranked and the top-reviewed pages on the website. Therefore, from the experimental results, we can state the following points:

- Our proposed similarity measure gives better grouping of pages as compared to Cosine similarity.
- Incorporation of the ratings and the reviews in our similarity equation helps in grouping the highest-rated and the highest-reviewed webpages together.

TABLE IX. TEN TOP PAGES OF THE BEST GROUING WITH MEMBERSHIPS TAKING ONLY THE HIGH-REVIEWED PAGES

Webpage	***Membership***	***Review***	***Rating***
Gray Line	0.9911	7.604	3.5
Frog Pond	0.9907	7.601	3.5
Little Tokyo	0.9799	7.616	3.5
National Geographic Museum	0.5196	7.594	3.5
Bunker Hill Bridge	0.9575	7.623	4
Greek town	0.9427	7.579	3
Alcatraz Cruises LLC	0.9226	7.632	3
Old State House	0.9122	7.632	3
Bank of America Theatre	0.9049	7.642	4
Discovery Children's Museum	0.8936	7.567	2

TABLE X. TEN TOP PAGES OF THE BEST GROUING WITH MEMBERSHIPS TAKING THE HIGH-RATED AND HIGH-REVIEWED PAGES

Webpage	***Membership***	***Rating***	***Review***
Wall Street Walks	0.9896	4.5	7.899
Moorea Beach Pool	0.951	4.5	7.861
Copley Square	0.9458	4.5	7.923
George's Island	0.9447	4.5	8.021
Albert Einstein Memorial	0.9329	4.5	7.885
Lincoln Park Conservatory	0.9301	4.5	7.827
Chicago History Museum	0.9293	4.5	7.736
USS Pampanito	0.9179	4.5	7.707
Grant Park	0.9141	4	7.936
San Francisco Whale Tours	0.9041	3	7.192

V. CONCLUSION

Considering the limitations of utilizing either the ratings or the reviews for recommendation in the tourism domain, this paper has proposed a heuristic approach which collects both the user ratings and the user reviews of the pages, assigned weights to the pages based on the ratings and the reviews, calculates alikeness among them using the proposed similarity measure, and finally groups them by applying a fuzzy clustering algorithm. The experimental results show that our similarity measure considering both the ratings and reviews assembles the highest-rated and the highest-reviewed webpages in the same cluster. This offline page grouping substantiates high accuracy and greater reliability in online recommendations.

As we didn't consider users' expressions of negation (e.g., didn't) and adverbial phrases (e.g., very much), in the future work we will duly consider them. We hope to extend our offline heuristic approach to an online recommender system for making the recommendations for the users.

REFERENCES

[1] S.Loh, F.Lorenzi, R.Saldana, and D. Licthnow, "A tourism recommender system based on collaboration and text analysis," Information Technology and Tourism, Vol. 6, 2004.

[2] L.Zhang, B.Pan, W.Smith, and X. Li, "An exploratory study of travelers' use of online reviews and recommendations: A qualitative approach," Journal of Information Technology and Tourism, Vol.11, No.2, pp. 157-167, 2009.

[3] D.Park, J.Lee, and I.Han, "The effect of online consumer reviews on consumer purchasing intention: The moderating role of involvement," International Journal of Electronic Commerce,Vol.11, No.4, pp. 125–148, 2007.

[4] P. A.Pavlou and A. Dimoka, "The nature and role of feedback text comments in online marketplaces: Implications for trust building, price premiums and seller differentiation," Information Systems Research, Vol.17, No.4, pp. 392–414, Dec. 2006.

[5] Payne et al., "Adaptive strategy selection in decision making," Journal of Experimental Psychology: Learning, Memory, and Cognition, Vol.14, No.3, pp. 534-552, Jul. 1988.

[6] S.Senecal and J. Nantel, "The influence of online product recommendations on consumers' online choices," Journal of Retailing, Vol.80, pp. 159-169, Jun 2004.

[7] H. Jun Cheong and M. A. Morrison, "Consumers' reliance on product information and recommendations found in UGC", Interactive Advertisement, Vol.8(2), 2008.

[8] S. Aciar, D. Zhang, S. Simoff, J.Debenham, "Informed Recommender: Basing recommendations on consumer product reviews," IEEE Computer Society, Vol.22, No.3, pp. 39-47, Jun. 2007.

[9] J.C.Bezdek, R.Ehrlich, and W.Full, "FCM: The Fuzzy C-Means Clustering Algorithm," Computers & Geosciences, Vol.10, pp. 191-203,1984.

AUTHORS PROFILE

Shahnila Zaman received B.Sc (Hons.) degree in Computer Science and Engineering from University of Dhaka, Bangladesh in 2014. She is currently pursuing M.Sc. degree in Computer Science and Engineering in the University of Dhaka. Her field of interest includes data mining, web mining and database management systems.

Sabiha Salma received B.Sc (Hons.) degree in Computer Science and Engineering from University of Dhaka, Bangladesh in 2014. She is currently continuing M.Sc. degree in Computer Science and Engineering in the University of Dhaka. Her field of interest includes data mining, web mining, and web design and development.

Shaily Kabir received B.Sc (Hons.) and M.S degree in Computer Science and Engineering from the University of Dhaka, Bangladesh and also M.Comp.Sc degree in Computer Science from Concordia University, Canada in 2012. Currently she is working as an assistant professor in the Department of Computer Science and Engineering, University of Dhaka. Her research interests include computer networks and network security, data and web mining, and database management systems.

.