2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology

An Enhanced Similarity Measure for Utilizing Site Structure in Web Personalization Systems

Shaghayegh Sahebi
University of Tehran
Karegar Ave.
Tehran, Iran
P.O. Box: 14395-515
s.sahebi@ece.ut.ac.ir

Farhad Oroumchian
Wollongong University of Dubai
Dubai, UAE
P.O. Box: 20183
oroumchian@acm.org

Ramtin Khosravi
University of Tehran
Karegar Ave.
Tehran, Iran
P.O. Box: 14395-515
rkhosravi@ece.ut.ac.ir

Abstract

The need for recommendation systems to ease user navigations has become evident by growth of information on the Web. There exist many approaches of learning for Web usage-based recommendation systems. In hybrid recommendation systems, other knowledge resources, like content, semantics, and hyperlink structure of the Web site, have been utilized to enhance usage-based personalization systems. In this study, we introduce a new structure-based similarity measure for user sessions. We also apply two clustering algorithms on this similarity measure to compare it to cosine and another structure-based similarity measures. Our experiments exhibit that adding structure information, leveraging the proposed similarity measure, enhances the quality of recommendations in both methods.

1. Introduction

With the rapid growth of Web, personalization systems have been the subject of many researches. A Web personalization system is defined as any system that tailors the Web experience for a particular user/a group of users [4]. Many web mining techniques have been used in web personalization systems to discover usage patterns from Web data such as clustering techniques, association rule mining, and click pattern analysis.

Nevertheless, pure usage-based personalization systems do not utilize the domain semantics and structural knowledge so they cannot recommend complicated objects, consisted of semantic attributes, similar to each other. As a result, hybrid recommendation systems have been emerged. Examples of hybrid systems using Web sites content are [2], [5], and [3]. As an instance of using linkage structure information in a usage-based personalization system, we can name Nakagawa and Mobasher's work [6], which switched between different recommendation algorithms based on the degree of connectivity in the site and the current location of the user within the site. Nasraoui et al. [7] also used the hierarchical linkage structure of site as an implicit concept hierarchy to be exploited in computing the similarity between pages.

In this study, we propose a similarity measure for visiting sessions of users, which is based on both usage data and linkage structure of the Web site in Section 3. This work is based on [7] and tries to enhance its similarity measure. We use an agglomerative hierarchical clustering and Relational Fuzzy Subtractive Clustering (RFSC) [8] algorithms on usage data of the DePaul University CS department in Section 6 to compare this similarity measure with the proposed measure of [7] and cosine similarity measure. These algorithms are described in Section 2. Based on the results in Section 6, we can conclude that adding structural information as a concept hierarchy, utilizing the proposed similarity measure, improves the quality of recommendations in both applied methods.

2. Applied Methods

We have applied Agglomerative Hierarchical Clustering (AHC) and Relational Fuzzy Subtractive Clustering (RFSC) algorithms to Web usage data. The AHC algorithm works by grouping data objects into a tree of clusters in a bottom-up (merging) fashion. It starts by placing each object in its own cluster and then merges these atomic objects into larger clusters, according to some criterion, until all of the objects are in a single cluster. We utilized the average distance criterion based on its less sensitivity to noise and correlation of the distances between data objects and the linking of objects in the cluster tree.

The RFSC algorithm [8] works based on the distances

978-0-7695-3496-1/08 $25.00 © 2008 IEEE
DOI 10.1109/WIIAT.2008.270

between all data points and is less sensitive to noises for not needing the fuzzy partition condition. In this algorithm, we consider every data point as a potential cluster center, choose the maximum potential point greater than an accept ratio (ϵ) as a cluster center, and update other potentials iteratively. If the potential of a data point is less than a reject ratio ($\bar{\epsilon}$), it will never be chosen as a cluster center.

3. Proposed Similarity Measure

The clustering algorithms utilize a similarity measure to gain the similarity between the data points. In Web usage mining, the cosine similarity measure between sessions is very popular. There have been some efforts to leverage other information sources, like Web site hyperlink structure, in addition to usage data in data mining for personalization. In [7], Nasraoui et al have proposed a new similarity measure based on link structure of a Web site to enhance the quality of recommendations. From now on, we call this similarity measure "the basic similarity measure".

In this measure, a user session is modeled as following: a unique number $j \in \{1, ..., N_U\}$ is assigned to each URL in the site, where N_U is the number of URLs and the i^{th} user session is modeled in a N_U-dimensional vector space as stated in Equation 1. We call this model "the Binary View Model".

$$S_j^i = \begin{cases} 1 & \text{if the user has accessed the } j^{th} \text{ URL;} \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Based on this, the first similarity measure between two user sessions A and B is:

$$S_{1,AB} = \frac{A \cdot B^t}{\|A\|\|B\|} = \frac{\sum_i A_i B_i}{(\sum_i A_i)^{0.5}(\sum_i B_i)^{0.5}} \quad (2)$$

For computing the basic structure-based similarity measure, the entire Web site is modeled as a tree each of its nodes representing a URL. In this tree, a node is another node's parent if the latter's URL is hierarchically located under the former's URL in a directory-like structure. A syntactic similarity between the i^{th} and j^{th} URL is then calculated based on Equation 3 in which P_i is the path from the root node (main page) to the page i and $|P_i|$ is the length of this path. Using the similarity measure between URLs as a matrix S_u, the syntactic similarity between sessions A and B is calculated by Equation 4.

$$S_u(i,j) = \min(1, \frac{|P_i \cap P_j|}{\max(1, \max(|P_i|, |P_j|) - 1)}) \quad (3)$$

$$S_{2,AB} = \frac{\sum_i \sum_j A_i B_j S_u(i,j)}{\sum_i A_i \sum_i B_i} \quad (4)$$

To obtain the basic similarity, a maximum of $S_{1,AB}$ and $S_{2,AB}$ is chosen in Equation 5 and the basic dissimilarity measure is obtained by Equation 6.

$$S_{AB} = \max(S_{1,AB}, S_{2,AB}) \quad (5)$$

$$d_s^2(A,B) = (1 - S_{AB})^2 \quad (6)$$

To be able to exploit the defined measures in non-binary view modeling of user session, considering the visit duration of each page instead of just zeros and ones, we can use Equation 7 and 8. This dissimilarity measure works fine for the binary view modeling of user sessions: $d_s^2(K,K) = 0$, $d_s^2(K,L) \geq 0$, $d_s^2(L,K) = d_s^2(K,L)$. But some enhancements could be considered for this similarity measure. We can eliminate calculating both cosine and structure-based similarities and just calculate a combination of them. In this way, we will also get rid of an extra maximization and an extra quadrating. The structure-based S_2 measure can not be used itself due to some problems. The first problem is that, sometimes $S_{2,KK} \neq 0$ and even $S_{2,KK}$ may be different for different values of K. The cosine similarity between two objects always scales between zero and one, in which one denotes the most similarity and zero indicates the least similarity between two sessions. In $S_{2,AB}$, there is no such an scale. In some cases even $S_{2,AB} > S_{2,AA}$. It is mainly due to the non-normalized denominators of Equations 4 and 8 with respect to their numerators. Another problem of this measure is that, quadrating the final similarity measure, to obtain the dissimilarity, makes the dissimilarity scales very small. On the other hand, by getting deeper in the URL tree, the concepts of the URLs get narrower, so the sibling URLs get closer to each other. As a result, the similarity between two sibling URLs is expected to grow by getting deeper in the URL tree. But the problem is, in the S_u measure, the similarity between two sibling URLs always equals to one which does not seem to be correct.

$$S_{1,AB} = \frac{A \cdot B^t}{\|A\|\|B\|} = \frac{\sum_i A_i B_i}{\sqrt{\sum_i A_i^2}\sqrt{\sum_i B_i^2}} \quad (7)$$

$$S_{2,AB} = \frac{\sum_i \sum_j A_i B_j S_u(i,j)}{\sqrt{\sum_i A_i^2}\sqrt{\sum_i B_i^2}} \quad (8)$$

To resolve the stated problems in the basic similarity measure, we have proposed a variation of this measure which we call it "the enhanced similarity measure". If we consider the tree modeling of Web site hyperlink structure, we can define the similarity between two URLs as below:

$$S'_u(i,j) = \min(1, \frac{|P_i \cap P_j| - 1}{\max(1, \max(|P_i|, |P_j|) - 1)}) \quad (9)$$

Now if we consider the matrix S' as the similarity matrix between different URLs, the similarity between two user sessions is defined by Equation 10 and the dissimilarity measure is obtained by Equation 11.

$$ES_{AB} = \frac{A \cdot S'_u \cdot B^t}{(A \cdot S'_u \cdot A^t)^{0.5}(B \cdot S'_u \cdot B^t)^{0.5}} \quad (10)$$

$$d_{s'}(A,B) = 1 - ES_{AB} \quad (11)$$

In our enhanced similarity measure, we are sure about the scaling of the ES_{AB} by normalizing the measure in Equation 11. The enhanced similarity is always in range $[0,1]$, $ES_{AB} = 0$ denotes the least similarity and $ES_{AB} = 1$ indicates the maximum similarity between two sessions. On the other hand, always $ES_{KK} = 1$ and as a result $ES_{KK} > ES_{KL}$. The similarity between two siblings (S'_u) in the URL tree also increases by growing the depth of the tree and narrowing the subject of the pages. We also do not need to have an optimistic maximum aggregation and an extra quadrating. Both of these structure-based similarity measures violate the "triangular inequality", which means in some cases $ES_{KL} \geq ES_{KM} + ES_{ML}$ and $S_{KL} \geq S_{KM} + S_{ML}$.

4. Recommendation in Different Algorithms

To recommend items (pages) in AHC algorithm, we first find the best cluster for each evaluation data point (x^j, a vector representing the j^{th} user visit duration on all pages) by calculating the distance between these data points with cluster centers (μ_k) in Equation 12. Then, we sort the pages in the best cluster (ω_k) based on the sum of durations of user views on those pages to find the most important pages of each cluster by Equation 13.

$$BestCl(x^j) = \arg\min_k d(x^j, \mu_k) \quad (12)$$

$$ImportantPages(\omega_k) = Sort(\sum_{x^j \in \omega_k} x^j) \quad (13)$$

We recommend most important pages of the assigned cluster which the user has not seen yet.

$$RecommSet(x^j) = ImportantPages(BestCl(x^j)) \quad (14)$$

For recommendations in RFSC algorithm, we calculate the distance between evaluation data points and cluster centers and then the fuzzy membership matrix (U) for the evaluation data using 15. We sort the clusters based on their degree of importance for each data point which means the ascending order of membership degrees in each row. For each session (x^j), the number of pages recommended from each cluster (k) is determined by the membership degrees of each session to each cluster. The constant α is a limit on maximum number of recommendations for a session. For each cluster, we calculate a weighted sum, by multiplication of membership matrix with the visit duration of each page in each session as described in the following pseudo code, to recommend the important pages of the cluster (ω_k).

$$ImportantCls(x^j) = Sort(Uj) \quad (15)$$

$$RecomNumber = \frac{U_{j,k}}{\sum_k U_{j,k}} \alpha \quad (16)$$

$$ImportantPages(\omega_k) = Sort(\sum_j U_{j,k} x^j) \quad (17)$$

Assuming:
$ImportantCls(x^j) = [c1, ..., cn]$ and
$ImportantPages(\omega_k) = [p1, ..., pm]$,
for $a = 1$ to n do
  for $b = 1$ to $RecomNumber$ do
    recommend $ImportantPages(\omega_k)[b]$
    if not visited by user

5. Data and Measures

In this study, we utilized the usage data of the DePaul University (http://cs.depaul.edu). In this data set, sessions of 13745 users on 683 pages of CTI web site of the DePaul University for a two week period have formed a 13745×683 matrix. Each member of this matrix shows the visit duration of each user on each page.

We applied some of the measures suggested in [1] and additional measures, taken from information retrieval literature, to evaluate the quality and goodness of recommendations. These measures are:

- Hit Ratio (HR): Percentage of hits with respect to number of the sessions. If a recommended page is actually requested later in the session, we declare a hit.
- Recall (Re): Percentage of hits with respect to number of pages in unvisited part of user session.
- Precision (Pr): Percentage of hits with respect to the number of recommendations for each session.
- F-Score (FS): A proportion of precision and recall which is taken from Information Retrieval literature:

$$FScore = \frac{(Recall \times Precision) \times 2}{(Recall + Precision)} \quad (18)$$

- Prediction Strength (PS): Average number of recommendations made for a page.
- Recommendation Quality (RQ): Average rank of the first hit in recommendations.
- Prediction Coverage (PC): Percentage of train pages which were recommended to users.

To be ideal, both recall, precision and so F-Score should be one. It occurs when all the unvisited pages of user session and no other pages are recommended. It is also better to have a higher hit ratio and lower recommendation quality (RQ).

Table 1. Goodness Measures of Recommendations with Different Similarity Measures

Model	Similarity	No of Clusters	HR(%)	Re(%)	Pr(%)	FS(%)	PS(%)	RQ	PC(%)
AHC	cosine	50	73.19	45.44	8.24	13.95	11	6.02	37.19
AHC	S_{AB}	10	58.93	33.27	6.03	10.21	11	5.14	3.95
AHC	ES_{AB}	50	74.65	47.43	8.60	14.56	11	5.83	48.46
RFSC	cosine	54 ($\epsilon = 0.001, \bar{\epsilon} = 0.501$)	59.99	38.43	6.49	11.11	11.81	3.89	8.78
RFSC	S_{AB}	16 ($\epsilon = 0.005, \bar{\epsilon} = 0.505$)	30.08	21.31	6.30	9.72	5.92	1.03	2.12
RFSC	ES_{AB}	62 ($\epsilon = 0.001, \bar{\epsilon} = 0.501$)	41.54	23.29	9.16	13.15	5.07	1.05	18.89

6. Experimental Results

To compare the cosine similarity, the basic similarity for non-binary view model, and our enhanced similarity measure, we applied the AHC and RFSC clustering algorithms on the described data set. The algorithms are developed in MATLAB [9] and the results are shown in Table 1.

To gain a proper result in RFSC method, we increased accept and reject ratios from 0.01 to 0.99 with 0.001 step size. In this method, utilizing a point to point similarity, the best number of clusters is determined automatically. We repeated the AHC algorithm with different number of clusters, applying different similarity measures. This hierarchical clustering algorithm, though simple, will neither revokes the merge actions done previously nor performs object swapping between clusters which may lead to low quality clusters. It only considers the similarity to average in a cluster so it is more susceptible to variations with respect to RFSC.

Although the number of recommendations was limited to 11 in both algorithms, the prediction strength measure has a higher value for two basic and enhanced similarity measures utilizing the AHC algorithm which results in better precision and weaker recall and hit ratio with respect to applying the cosine measure. Recommendation ranks are better with the basic measure. However, considering the F-Score value, we can see that the enhanced similarity measure outperforms both cosine and basic one. The cosine measure also outperformed the basic similarity measure. It may be due to the optimistic maximum aggregation in Equation 5 or the large denominators of Equations 4 and 8 which makes smaller similarities.

7. Conclusion and Future Work

In this study, we proposed a linkage structure-based similarity measure based on [7], compared it to cosine and basic hyperlink structure-based similarity measures, using different clustering algorithms in Web personalization systems. The proposed measure eased the calculation of the basic similarity measure. It also improved the scaling of that measure, corrected some problems in structure-based part of that, and outperformed it in our experiments.

For future work, enhancing this similarity measure, so that it will not violate the triangular inequality, is important. Besides, more precise results are needed for an ideal recommendation system. As a consequence, it is valuable to embed the context of Web site pages or semantic information of them in recommendation process.

References

[1] A. Bose, K. Beemanapalli, J. Sirvastava, and S. Sahar. Incorporating concept hierarchies into usage mining based recommendations. *ACM*, 1:444–448, 2006.

[2] M. Eirinaki, M. Vazirgiannis, and I. Varlamis. Sewep: Using site semantics and a taxonomy to enhance the web personalization process. *In Proceedings of the 9th SIGKDD International Conference on Data Mining and Knowledge Discovery (KDD03)*, 2:99–108, 2003.

[3] P. Kearney, S. Anand, and M. Shapcott. Employing a domain ontology to gain insights into user behaviour. *In: Proceedings of the 3rd Workshop on Intelligent Techniques for Web Personalization, at IJCAI 2005*, 2005.

[4] B. Mobasher, R. Cooley, and J. Srivastava. Automatic personalization based of web usage mining. *ACM*, 2:782–800, 2000.

[5] B. Mobasher, H. Dai, T. L. Y. Sun, and J. Zhu. Integrating web usage and content mining for more effective personalization. *In E-Commerce andWeb Technologies: Proceedings of the EC-WEB 2000 Conference. Lecture Notes in Computer Science (LNCS) 1875, Springer*, pages 165–176, 2000.

[6] M. Nakagawa and B. Mobasher. A hybrid web personalization model based on site connectivity. *In: Proceedings of the WebKDD 2003Workshop, at the ACM-SIGKDD Conference on Knowledge Discovery in Databases (KDD2003)*, 2003.

[7] O. Nasraoui, R. Krishnapuram, A. Joshi, and T. Kamdar. Automatic web user profiling and personalization using robust fuzzy relational clustering. *In Segovia, J., Szczepaniak, P., Niedzwiedzinski, M., eds.: Studies in Fuzziness and Soft Computing*, 105:233–261, 2002.

[8] B. S. Suryavanshi, N. Shin, and S. P. Mudur. An efficient technique for mining usage profiles using relational fuzzy subtractive clustering. *In Proc. WZH'05*, 2005.

[9] MATLAB Software. http://www.mathworks.com.