DETERMINING THE PREFERENCES OF A SOCIAL GROUP STATISTICALLY AND DYNAMICALLY

Kamal Taha, PhD
Khalifa University, UAE

Abstract

In this paper, we propose a group-based Collaborative Filtering framework. The framework uses ontology-driven social networks, where nodes represent social groups. A social group is an entity that defines a group based on demographic, ethnic, cultural, religious, age, or other characteristics. In the proposed framework, query results are filtered and ranked based on the preferences of the social groups to which the user belongs. If the user belongs to social group G_x, results will be filtered based on the preferences of G_x. The framework can be used for various practical applications, such as Internet or other businesses that market preference-driven products. In the proposed framework, the preferences of a Social Group can be acquired statically from *hard-copy* published studies about the Social Group or can be acquired dynamically from Web pages that publish information about the Social Group. We describe and experimentally compare the above mentioned approaches.

Keywords: Personalized search systems, group profiling, collaborative filtering

1. Introduction:

Traditional search engines typically return identical results for the same query, independent of the user or the context. Conventional quantitative scoring functions may not adequately reflect users' preferences, since the same document may be queried by users, whose preferences differ. By analyzing search behavior, it is possible to see that many users are not able to accurately express their needs in exact query terms [Micarelli et al. 2006]. In contrast to conventional search engines, a personalized search engine [Keenoy and Levene 2005, Carmine and Antonio 2003, Weihua 2002] would return different results for the same query, depending on the user and the context. Profiles can modify the representation of the user needs before the retrieval takes place. Most personalized systems lean towards being Information Filtering (IF) systems more than being general Information Retrieval (IR) systems [Oard 2007].

Most existing personalized search systems do not consider group profiling. Group profiling can be an efficient retrieval mechanism, where a user profile is inferred from the profile of the social groups to which the user belongs. In this paper, we propose a framework that determines the preferences of Social Groups. The framework categorizes Social Groups based on demographic, ethnic, cultural, religious, age, or other characteristics. For example, people of ethnic group E_X; people who follow religion R_Y; and people who live in neighborhood N_Y can all be considered to form various social groups. In social communities, it is commonly accepted that people who are known to share a specific background are likely to have additional connected interests [Herlocker et al. 2002]. The framework can be used for various practical applications, such as Internet or other businesses that market preference-driven products. In the proposed framework, the preferences of a Social Group could be identified from either: (1) the preferences of its member users, or (2) from published studies about the social group (the availability of such data has had a significant boost with the emergence of the World Wide Web).

By crawling Web sites, the proposed framework initializes the preferences and ratings of Social Groups dynamically from Web pages that publish information about them. We proposed previously in [Taha and Elmasri 2010b] an approach that identifies the semantic relationships between XML elements in an XML document. We describe in this paper modifications we made to [Taha and Elmasri 2010b] to suit the extraction of Web content data for the sake of dynamically initializing Social Groups' preferences. We also describe modifications we made to an approach proposed in [Tang 2008] in order to initialize a Social Groups' preferences. The system generates items' scores by converting the preference data (obtained from the two approaches) into weighted web-feature and feature-item matrices.

2. Initializing the ratings of a Social Group Statically

The preferences of a Social Group can be acquired statically from *hard-copy* published studies such as:

a) Published articles and books (e.g.,[Kittler 1995; Tesoro 2001]).
b) Published studies conducted by organizations (e.g., [FAQ Archives 2008]), or specialized centers belonging to universities.

First, we need to decide on the publications to be used. The more publications used, the more accurate the results are. We need to select ones issued by reputable sources. Preferences on an item's features obtained from a hard-copy published study are represented as a publication-feature matrix M with entries f_i and P_j: feature f_i is recommended by publication P_j. The rating of publication P_j on feature f_i is the element $M(j, i)$ of matrix M. Element $M(j, i)$ is a Boolean value, where *one* denotes that publications P_j stresses the importance of feature f_i to the Social Group and *zero* otherwise. That is, the rating $M(P_j)$ of publication P_j is the j-th row of matrix M. For example, consider the following *car* preferences of the residents of neighborhood N_x (*i.e.,* Social Group N_x). N_x is a neighborhood in the State of Minnesota, USA. According to published surveys, 68% of Minnesotans prefer cars with *snow-proof* features[1], 61% prefer *fuel-efficient* cars[2], and 76% of the residents of N_x prefer *cost-efficient* cars[3]. The preferences of N_x on each of these three features will be assigned a weight of *one* in matrix M. The score of a feature is the *summation of publications' weights on it (see Equation 1).* Table 3 shows an example data set of matrix M. For example, the score of feature f_1 is the sum of the weights of publication P_2, P_3, and P_5 on feature f_1.

$$Score\ f_i = \sum_{j=1}^{|I|} M(P_j, f_i) \qquad (1)$$

We now introduce an item-feature matrix N, where element $N(j, i)$ is *one*, if item I_j contains feature f_j and *zero* otherwise. The profile $N(I_j)$ of item I_j is the j-th column of matrix N. The score of item I_j is the *summation of the normalized scores* of the features that I_j contains *(see equation 2)*

$$Score\ I_j = \sum_{\forall N(f_i, I_j)=1} score\ f_i \qquad (2)$$

3. Employing the XCDSearch Approach in [Taha and Elmasri 2010b] for Initializing the Ratings of a Social Group from Web Pages Dynamically by Crawling Web Sites

We proposed in [Taha and Elmasri 2010b] techniques called XCDSearch to build *semantic relationships* between elements in XML documents. For the sake of this work, we modified these techniques in order to build semantic relationships between Web content data *(i.e., instead of XML data)* to initialize the ratings of Social Groups. We constructed a

[1] Due to the very snowy winter in the state of *Minnesota*.
[2] Which is due, in part, to the fact that the government of *Minnesota* offers sales tax break incentive for buying fuel-efficient cars.
[3] Due to the fact that N_x is a middle-class neighborhood (e.g., [Minneapolis Census 2000]).

prototype that employs these techniques to dynamically identify the preferences of Social Groups from Web pages that publish information about them. The system will then generate items' scores *dynamically* by converting this preference data to weighted *web-feature and feature-item matrices* using equations 1 and 2. The system will use these matrices to *initialize* Social Groups' ratings. First, the system will mark up a Web page with XML tags and model the resulting document as a rooted and labeled XML tree *(e.g., Fig. 1).* A Social Group is represented as an interior node in the XML tree, and its preferences as data/leaf nodes. For example, Fig. 1 is a fragment of an XML tree modeling the content data of Web page publishing information about some Social Groups.

We first define key concepts used in the modified techniques. We use the term Ontology Label to refer to the ontological concept of a node in an XML tree. Let (*m* "is-a" m') denote that class *m* is a subclass of class m' in an Object-Oriented ontology. m' is the most general superclass (root node) of *m* in a defined ontology hierarchy. m' is called the Ontology Label of *m*. The system converts an XML tree into a tree called ontology-based tree. For example, Fig. 2 shows an ontology-based tree constructed from the XML tree in Fig. 1. An ontology-based tree is constructed as follows. First, the system removes all interior nodes that do not have children data nodes *(for example, nodes 4, 7, and 13 are removed from Fig. 1).* Then, the system replaces the remaining *interior* nodes with their Ontology Labels *(for example, nodes ethnic group(1) and sect(8) in Fig. 1 are replaced by their Ontology Label, which is GROUP as shown in Fig. 2).*

Let *a* be an interior node and *b* a data node in an ontology-based tree. Nodes *a* and *b* are *semantically related* if the paths from *a* and *b* to their *Lowest Common Ancestor* (*LCA*), not including *a* and *b*, do not contain more than one node with the same Ontology Label. The LCA of *a* and *b* is the only node that contains the same Ontology Label in the two paths to *a* and *b*. Consider that node *b* contains the preference data[4] P_i and that node *a* represents Social Group G_j. If nodes *a* and *b* are semantically related, P_i is a preference of Social Group G_j. For example, consider Fig. 2. Preference *"no pork-related products"* (node 10) belongs to religious sect R_Y (node 6) and not to ethnic group E_X (node 2), because the LCA of nodes 10 and 2 is node 1, and the path from node 1 to node 10 includes two nodes with the same Ontology Labels *(i.e., nodes 1 and 8).* Similarly, the preference *"spicy flavor"* (node 3) belongs to E_X and *not to* S_Z (node 9). Using the same techniques, both of *"spicy flavor"* and *"no pork-related products"* are preferences to religion group R_Y (node 6).

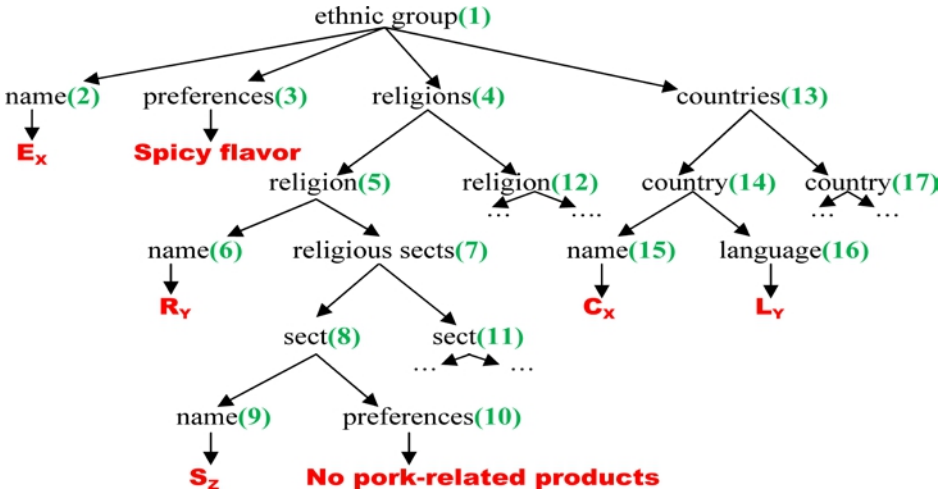

Fig. 1. A fragment of an XML tree modeling the content data of a Web page about some Social Groups. *Nodes are numbered for easy reference.*

[4] The system identifies such data via *text mining program*

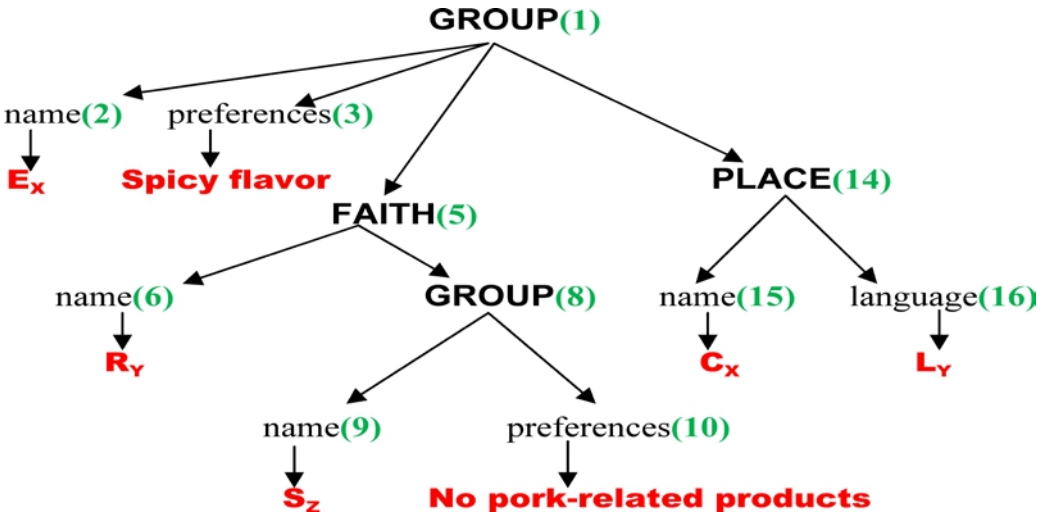

Fig. 2. Ontology-based tree constructed from the XML tree in Fig. 1

4. Employing the TopDown Approach in [Tang 2008] for Initializing the Ratings of a Social Group from Web Pages Dynamically by Crawling Web Sites

[Tang 2008] studies the effect of *topic taxonomy* on dynamic *group profiling*. A topic taxonomy consists of topic nodes. Each internal node is defined by its vertical path *(i.e., ancestor and child nodes)* and its horizontal list of attributes. To perform taxonomy adaptation, the paper proposes a top-down hierarchical traversal approach called TopDown. We constructed a prototype that employs an *adjusted* version of the TopDown approach to identify and initialize the preferences of a Social Group from Web pages publishing information about it. For each topic node *n* representing a Social Group G_x, this copy of the prototype identifies the *best neighbor* nodes of *n* that contain preference data about G_x. The TopDown approach consists of multiple iterations to search for *better hierarchies*, as follows:

1. Identification of the node to check: A list of topic nodes in the hierarchy is maintained for the search. Nodes at the upper level are given higher priority.

2. Identification of promising neighboring hierarchies concerning a node: The promising hierarchies are checked by *rolling-up nodes* to their upper level. Then, the hierarchies are checked by *pushing down* nodes to their siblings and by *merging* two sibling nodes to form a super node.

3. Identification of the best neighbor: This procedure compares all the promising neighboring hierarchies and finds the best among them.

4. Update of the current best hierarchy: The current best hierarchy is replaced with the best hierarchy just found and the list of nodes to check is updated.

Example 2: Consider that the system crawled a website publishing information about the Buddhism faith and identified the classificatory taxonomy of branches shown in Fig. 3. In the figure, p_x, p_y, p_z, and p_w are preference data. By *merging* nodes Mandalas and Shingon and *rolling up* the resulting node, and by *pushing down* node Mahayanists, the preferences of the Mahayanists can be identified as p_x, p_y, and p_z. By *pushing down* node Buddhists, it preferences can be identified as p_x, p_y, p_z, and p_w.

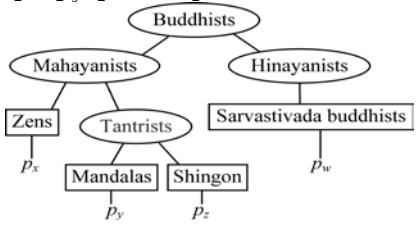

Fig. 3. Classificatory taxonomy of branches of the Buddhism faith

5. Experimental Results

5.1 Test Data for Real-User Evaluation

We asked 32 students from The University of Texas at Arlington (UTA) to evaluate and compare the four systems. The students belong to four different ethnic backgrounds and five ancestry origins. Some of them consider religion to be irrelevant and the others follow three different religions. We asked each of the students to prepare a list of 10 canned food items *ranked* based on the student's *own preferences*. We then asked this student to query our prototype systems for canned food to determine which one(s) returns ranked list of canned food *matches closely* to the one ranked by the student himself/herself.

5.2 Comparing Three Approaches for Initializing the Ratings of a Social Group

We compare in this test the three approaches described previously for *initializing* the preferences and ratings of a Social Group. These approaches are: (1) the *static initialization* from *hard-copy* published studies, (2) the *dynamic initialization* using the modified version of XCDSearch [Taha and Elmasri 2010b], and (3) the *dynamic initialization* using the modified version of TopDown [Tang 2008]. We cloned the prototype system into three identical copies, each employing one of the three approaches described above. Our objective is to determine which one of the three copies gives ranked lists of canned food *closest* to those ranked by the subject users. For the experimental dataset, we selected 18 Web sites publishing information about social groups and their preferences. For the sake of consistency, we used the same dataset for evaluating the static initialization approach also (*rather than using published hard copies*).

We ran the Web pages (*dynamically*) against each of the two copies employing the *dynamic approaches*. As for the copy employing the *static initialization* approach, we entered the preference data from the Web pages *manually* into the copy. We then measured the distance $d(\sigma_u,\sigma_s)$ between each list ranked by a resident *u* and the corresponding list ranked by one of the three copy systems *s*, using the following Euclidean distance measure.

$$d(\sigma_u,\sigma_s) = \sum_{x \in X} | \sigma_u(x) - \sigma_s(x) | \qquad (3)$$

X: Set of canned food items.

$\sigma_u \in [0,1]^{|X|}$: List of items ranked by resident u.

$\sigma_s \in [0,1]^{|X|}$: A list ranked by *one of the three copy systems*

$\sigma_u(x)$ and $\sigma_s(x)$: *position* of canned food item $x \in X$ in the lists σ_u and σ_s respectively (*a ranking of a set of n items is represented as a permutation of the integers 1, 2, . . . , n*).

Intuitively, the *static initialization approach* is expected to be more accurate than the other two approaches, since data is entered to the system *manually*. However, we aim at studying: (1) how much less accurate are the dynamic approaches than the static approach and *whether this accuracy difference is significant*, (2) whether the *practicality and convenience* of the dynamic approach makes up for its lower accuracy, in case the accuracy difference is not significant, and (3) the impact of *number of publications* on the accuracy of the three approaches. Fig. 4 shows the results. We can infer from the results the following:

(1) The static approach outperforms the two dynamic approaches as long as *the number of publications is less than about 25*.

(2) The XCDSearch's approach outperforms the TopDown approach *as long as the number of publications is greater than about 10*.

Based on the experiment results, we advocate employing the XCDSearch's approach for the sake of practicality and dynamicity, especially for recommender systems that target a rather wide range of Social Groups.

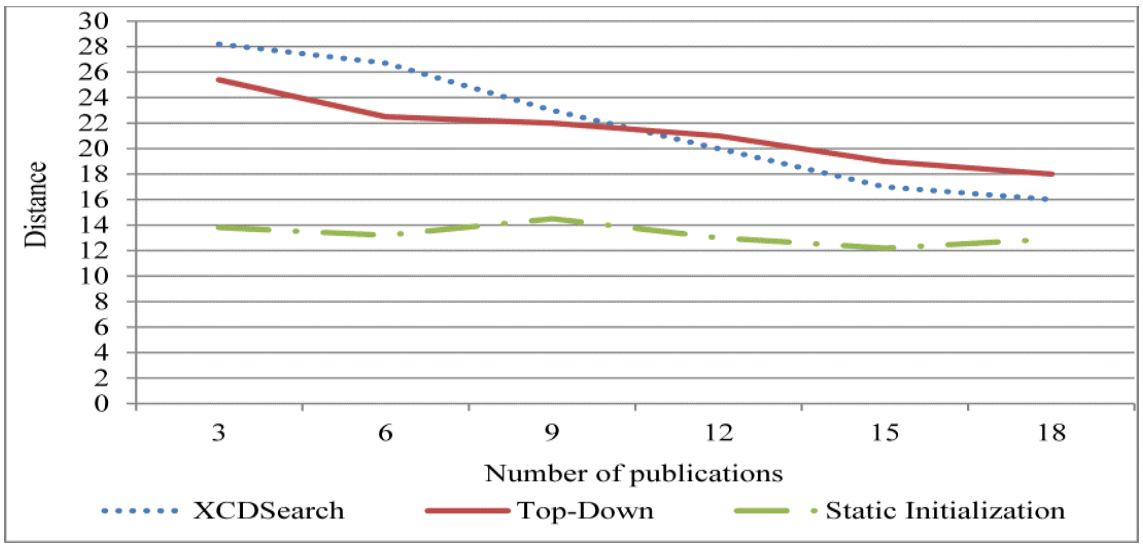

Fig. 4. Distance between the lists of items ranked by the subject users and the lists ranked by the prototypes employing the XCDSearch, TopDown, and static initialization approaches

6. Conclusion:

In this paper, we proposed a group-based Collaborative Filtering framework. The framework uses ontology-driven social networks, where nodes represent social groups. A social group is an entity that defines a group based on demographic, ethnic, cultural, religious, age, or other characteristics. The framework can be used for various practical applications, such as Internet or other businesses that market preference-driven products. In the proposed framework, the preferences of a Social Group can be acquired statically from *hard-copy* published studies about the Social Group or can be acquired dynamically from Web pages that publish information about the Social Group.

We experimentally compared the approach of determining the preferences of a Social Group statistically from published studies with the approach of determining these preferences dynamically from Web pages. Based on the experiment results, we advocate the approach of determining the preferences dynamically from Web pages for its practicality and dynamicity, especially for recommender systems that target a rather wide range of Social Groups.

References:

Carmine, C., Antonio, P. An Intelligent Search Agent System for Semantic Information Retrieval on the Internet. *WIDM*'03

Herlocker, J. L., Konstan, J. A., and Riedl, J. "Explaining Collaborative Filtering Recommendations". Proc. of the ACM Conference on Computer Supported Cooperative Work, 2002

FAQ Archives (2008). *Religion and Dietary Practices.* Available: http://www.faqs.org/nutrition/PreSma/Religion-and-Dietary-Practices.html

Keenoy, K. and Levene, M. 2005. Personalization of Web Search. *Lecture Notes in Computer Science, 3169.201-228.*

Kittler, P. 1995. *Food and culture in America: A nutrition handbook.* West Publishing, Redding.

Minneapolis Census (2000). Selected Economic Characteristics by Neighborhood, city of Minneapolis.
http://www.ci.minneapolis.mn.us/citywork/planning/Census2000/maps/economic/

Micarelli, A., Gasparetti, F., Biancalana, C. 2006. *Intelligent Search on the Internet.* Reasoning, Action and Interaction in AI Theories and Systems. LNCS, ISSN 0302-9743, Vol. 4155.

Oard, D. W. The state of the art in text filtering. *User Modeling and User-Adapting on the internet, Springer 7, 141-178.*

Taha, K., and Elmasri, R. "SPGProfile: Speak Group Profile." *Information Systems (IS)*, 2010, Elsevier, Vol. 35, No. 7, pp. 774-790

Taha, K., and Elmasri, R. "XCDSearch: An XML Context-Driven Search Engine". *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 2010, Vol. 22, No. 12, pp. 1781-1796.

Tang, L., Liu, H., Zhang, J., Agarwal, N., Salerno, J. Topic Taxonomy Adaptation for Group Profiling. ACM Transactions on Knowledge Discovery from Data, 2008, Vol. 1, No. 4.

Tesoro, E. (2001). *Religious Determinants of Food Choices.*

Weihua, L. Ontology Supported Intelligent Information Agent. *International IEEE Symposium On Intelligent Systems'02*