Subject Recommendation Using Ontology for Computer Science ACM Curricula

M. M. Lotfy\textsuperscript{1,*}, A. A. Salama\textsuperscript{1}, H. A. El-Ghareeb\textsuperscript{2}, M. A. El-dosuky\textsuperscript{2}

\textsuperscript{1} Department of Mathematics and Computer Science, Faculty of Science, Port Said University, Egypt.
\textsuperscript{2} Department of Information Systems, Faculty of Computers and Information Sciences, Mansoura University, Egypt.

Received: 01 November 2013; Accepted: 04 January 2014

Abstract

Recommender systems are needed to find subject items of one’s interest. We review recommender systems and recommendation methods. We propose a subject personalization framework based on adaptive hypermedia for Computer Science ACM Curricula. We extend Hermes framework with subject recommendation functionality. We combine TF-IDF term extraction method with cosine similarity measure. Specialization and standard subject database are incorporated into the knowledgebase. Based on the performed evaluation, we conclude that semantic recommender systems in general outperform traditional recommenders systems with respect to accuracy, precision, and recall, and that the proposed recommender has a better F-measure than existing semantic recommenders.

Keywords: Ontology; Semantics-Based recommendation; Heuristics

1. Introduction

Recommender systems are needed to find subject items of one’s interest. Challenges in building recommender systems can be classified as those concerning the user, and those concerning the algorithms used [1]. Different models are proposed [2] to deal with the missing or incorrect data from subject recording measurements. Other challenges have a trade-off between them such as the perfect databases size and the cold-start problem. The cold-start problem can be solved by using information about the user’s previous experience to calculate similarity measures to recommend new files [3]. Challenges about user compliance can benefit from many suggested strategies[4].

* Corresponding author.

\textit{E-mail address:} dev.magdy2011@gmail.com (M. M. Lotfy).
Section 2 reviews the previous attempts in building subject recommenders and recommendation approaches. Section 3 presents our solution and the evaluation of the proposed framework. We conclude in Section 4 with plans for future work.

2. Previous Work

The State-of-the-Art for building recommender systems is shown in a recent paper [5]. Many successful recommendation systems are proposed for news ([6], [7]), nutrition planning ([8], [9], [10], [11], [12]). The personalized recommendation technique in recommender systems, one of the most important tools of personal service in websites, makes great significance in Internet marketing activities of e-Commerce([13], [14]).

Recommendation system is a significant part of e–learning systems for personalization and recommendation of appropriate materials to the learner [15]. In the context of e-learning recommender systems, the users with greater knowledge have greater weight in the calculation of the recommendations than the users with less knowledge [16]. Recommender system can be used for predicting student performance [17].

There are four types of recommender approaches: content-based, semantics-based, collaborative filtering, and hybrid [22], but we restrict our discussion to the first two only. Content-based recommenders make use of Term Frequency-Inverse Document Frequency (TF-IDF)[18] and cosine similarity to compare the similarity between documents. Semantics is concerned only with concepts, and employing approaches such as concept equivalence [7], binary cosine[7], Jaccard [19], and semantic relatedness [20]. Next section shows how these approaches can be implemented.

3. Proposed Framework

The proposed framework is shown in fig. 1.

![Figure 1. The proposed framework](image-url)
The first step is to take the raw description directly from the user or from his profile. Stop words are removed, followed by stemming words back to the root and removing punctuation and converting to lower case. The next stage is to match the description or the output of the rule to the domain ontology. User profile is constructed by calculating TF-IDF values for each term. We determine the term frequency \( f_{i,j} \) for a term \( t_i \) within an recipe \( a_j \):

\[
f_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}
\]

(1)

dividing \( n_{i,j} \), the number of occurrences of term \( t_i \) in file by \( a_j \), the total number of terms in the document. Then the inverse document frequency (IDF):

\[
idf_j = \log \frac{|A|}{|\{a : t_i \in a\}|}
\]

(2)

dividing the total number of subject items by the number of subject items containing term \( t_i \).

The final value is computed by multiplying TF and IDF:

\[
tf\text{idf}_{i,j} = f_j \times idf_j
\]

(3)

Semantic measures benefit from the ontology that is defined by a set of concepts:

\[
C = \{c_1, c_2, c_3, \ldots , c_n\}
\]

(4)

The subject file can be defined by a set of \( p \) concepts:

\[
A = \{a_1^n, a_2^n, a_3^n, \ldots , a_p^n\}
\]

(5)

The user profile, \( U \), consists of \( q \) concepts found in the subject items read by the user:

\[
U = \{u_1^n, u_2^n, u_3^n, \ldots , u_q^n\}
\]

(6)

The similarity between a subject file and the user profile can be computed by:

\[
Similarity(U,A) = \begin{cases} 1 & \text{if } |U \cap A| > 0 \\ 0 & \text{otherwise} \end{cases}
\]

(7)

We can employ binary cosine to compute the similarity:

\[
B(U,A) = \frac{|U \cap A|}{|U||A|}
\]

(8)

by dividing the number of concepts in the intersection of the user profile and the unread subject file by the product of the number of concepts in respectively \( U \) and \( A \).

Similarly, Jaccard computes the similarity between two sets of concepts:

\[
J(U,A) = \frac{|U \cap A|}{|U \cup A|}
\]

(9)
Semantic neighborhood of $c_i$ is all concepts directly related to $c_i$ including $c_i$:

$$N(c_i) = \{c'_1, c'_2, c'_3, \ldots, c'_n\}$$  \hspace{1cm} (10)

A subject item $a_k$, which consists of $m$ concepts is described as the following set:

$$A_k = \{c^k_1, c^k_2, c^k_3, \ldots, c^k_m\}$$  \hspace{1cm} (11)

To compare two new items $n_i$ and $n_j$, a vector can be created:

$$V_i = (\{c'_1, w'_1\}, \ldots, \{c'_p, w'_p\}) \quad l \in \{i, j\}$$  \hspace{1cm} (12)

where $w_i$ is the weight of $c_i$. The similarity between subject items $a_i$ and $a_j$ is:

$$\text{SemRel}(a_i, a_j) = \cos(V_i, V_j) = \frac{V_i \cdot V_j}{\|V_i\| \cdot \|V_j\|} \in [0,1]$$  \hspace{1cm} (13)

The proposed framework is implemented in PHP. It allows the user to formulate queries and execute them to retrieve relevant subject items. We use the approach applied to adaptive hypermedia [21] and Hermes framework[6]. We extend Hermes with subject recommendation functionality. It utilizes OWL[23] for representing the ontology.

Performed tests are based on a subject corpus of 145 courses extracted from the Computer Science ACM Curricula [24] as shown in Table 1.

<table>
<thead>
<tr>
<th>Group</th>
<th>No. of courses</th>
<th>Group</th>
<th>No. of courses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithms</td>
<td>11</td>
<td>Artificial intelligence</td>
<td>11</td>
</tr>
<tr>
<td>ComputerArchitecture</td>
<td>10</td>
<td>Networking</td>
<td>9</td>
</tr>
<tr>
<td>Computation</td>
<td>3</td>
<td>Operating Systems</td>
<td>14</td>
</tr>
<tr>
<td>Discrete math</td>
<td>6</td>
<td>Problem formulation</td>
<td>8</td>
</tr>
<tr>
<td>Graphic</td>
<td>12</td>
<td>Programming languages</td>
<td>11</td>
</tr>
<tr>
<td>HCI</td>
<td>10</td>
<td>Software engineering</td>
<td>14</td>
</tr>
<tr>
<td>Information Models</td>
<td>15</td>
<td>Special topics</td>
<td>11</td>
</tr>
</tbody>
</table>

We have used 5 users with different but well-defined interests in our experiments. An example of a user interest is “algorithms”. Each user has manually rated the subject items as relevant or non-relevant for his interest. For each user we split the subject items corpus in two different sets: 60% of the subject items are the training set and 40% of the subject items are the test set. Recommenders compute the similarity between the subject items and previously
computed user profile. If the computed similarity value is higher than a predefined cut-off value the subject item is recommended and ignored otherwise.

Evaluating the recommenders is done by measuring accuracy, precision, recall, specificity, and F-measure. This is done by calculating a confusion matrix for each user. Fig. 2 shows the results of the evaluations.

The best recommenders for accuracy is the proposed framework, for precision is the proposed framework, for recall is binary cosine, for specificity are TF-IDF, Jaccard, and the proposed framework, and for F-measure is the proposed framework. The proposed algorithm scores well on accuracy as it makes relatively small amount of errors for both recommended subject as well as discarded subject items. For precision, the proposed algorithm scores the best for precision as most recommended subject items are relevant. The good results for recall obtained by the concept equivalence are due to the optimistic nature of the algorithm: any subject item which involves previously viewed concepts is recommended. TF-IDF, Jaccard, and the proposed framework score well on specificity as these algorithms do not recommend most of the non-relevant subject items.

4. Conclusion and Future Work

We propose a subject personalization framework based on adaptive hypermedia for Computer Science ACM Curricula. We extend Hermes framework with subject recommendation functionality. We combine TF-IDF term extraction method with cosine similarity measure. Specialization and standard subject database are incorporated into the knowledgebase. Performed tests are based on a subject corpus of 145 courses extracted from the Computer Science ACM Curricula.

Based on the performed evaluation, we conclude that semantic recommender systems in general outperform traditional recommenders systems with respect to accuracy, precision, and
recall, and that the proposed recommender has a better F-measure than existing semantic recommenders.

In the future we plan to extend the querying language by defining its grammar, and applying it for extracting deep knowledge from subject ontology.

Another possible research direction relates to the advanced traditional weighting schemes that other than TF-IDF such as logarithmic TF functions [25]. Another research direction is the considered similarity function. We would like to evaluate alternatives for cosine similarity as Ln(lt+l) [26] which seem to remove some of the cosine similarity bias favoring long documents over short documents.

Also, we would like to incorporate other services such as opinion mining [27]. As additional further work we would like to consider other types of available subjects in the ACM Curricula.

References


