Enhancing User Profile Accuracy for Personalized Web Systems

Anas El-Ansari, Abderrahim Beni-Hssane, Mostafa Saadi

Abstract: Personalized Web Applications aim to improve the user's browsing experience by offering customized products and services based on his preferences and needs. A key feature of a successful personalization system is building profiles that accurately express the real interests and needs of each user. In this work, we focus on creating accurate, complete and dynamic profiles by capturing and tracking the users’ browsing activities. Moreover, we implement techniques to increase the accuracy of the retrieved user profiles by collecting more browsing data, identifying the most important concepts and removing irrelevant ones, and the number of levels from the concept hierarchy in the reference ontology that we should use to efficiently represent the users’ real interests and needs. The result is a complete, dynamic, and accurate user profile that can be used to give users better-personalized browsing experience.

Keywords: Web personalization, profile, Accuracy, Ontology.

I. INTRODUCTION

The amount of accessible information on the Web continues to grow rapidly and has far exceeded human processing capabilities [3]. This growth often prevents users from finding desired information or product. It also aggravates making correct choices in time.

This problem highlights a pressing need for intelligent personalized systems that simplify information access and content discovery taking each user’s preferences and needs into consideration.

An example of personalized Web applications that has recently become popular is the personalized Web search system. These systems provide users personalized answers about services, products, and information that they may be interested to examine or purchase. Although, building a user profile that accurately represents the user real interests and can adapt over time, is a major challenge in the field of personalization systems [8].

In this work, our main objective is to create an accurate, dynamic ontology-based user profile. We focus on implicit methods for building profiles that have the potential to adapt over time, reflecting changing user interests.

The rest of the paper is structured as follows. The following section addresses related works. Next, we present the used methods for building user profiles. Section 4 presents an improved profile structure. Section 5 discusses the experimentation results. The paper ends with a conclusion.

Revised Manuscript Received on January 22, 2020.
Anas EL-ANSARI*, Chouaib Doukkali University, Morocco. Abderrahim Beni-Hssane, Chouaib Doukkali University, Morocco. Mostafa Saadi, Univ Sultan Moulay Slimane, Morocco.

II. RELATED WORK

This research relates to some studies in the field of personalized web systems in particular. Personalized web systems are developed to help users; browse news articles (NPA [5], SIREN [10], UP-TreeRec [27]), find scientific and research papers (Pigue [14], Personalized reading system [28]), purchase favorite products (Amazon [11], eBay [17]), improve search results (Hide-n-Seek[23], Persona [25]) or even combine some of the previous tasks (Basar [26], Syskill and Webert [19]).

The type of collected data used for building user profiles also makes a difference in these systems. Most build the profile by collecting and analyzing visited Web pages. Some researchers also used other data sources (Bookmarks in Basar [26], queries and search results in [22][19] and Persona [25]). To describe a user profile, Basar [26] produce a bookmark-like list of Web pages, while Syskill and Webert [19] create a list of concepts of interest, and Persona [25] create hierarchically arranged collections of concepts or ontology.

To create a user profile we can use a variety of learning techniques such as the vector space model [20][12], genetic algorithms [16], the probabilistic model [29], or clustering [18]. Filtering and rating methods are usually applied to improve the profile, as it may contain irrelevant concepts. Some systems require users' feedback for that purpose [19][25][26]. Others adapt autonomously.

PVA (Personal View Agent) System [4] offers customized news articles based on a concept-hierarchy user profile, implicitly created from the visited Web pages. The user profile is learned here by analyzing and classifying Web pages employing the vector space model and adjusted by splitting or merging concepts in the user profile.

Another system in [9] Also learns a user interest hierarchy (UIH) based on browsed Web pages, though, it clusters main terms obtained from each web page to identify concepts representing user preferences as a hierarchy in which the parent nodes are considered common or long-term user interests, while leaf nodes are viewed as specific or short-term ones.

Authors in [12] use the vector space model to cluster browsed Web pages and bookmarks. While collecting more data, the changes in the profile are used to track interest shifts.
Our research combines aspects from several systems reviewed in this section. We first build an initial user profile, called session profile, by collecting user browsing data (URL, date, time spent on the page) using a Firefox add-on component called Meetmeimer [15]. We modified this add-on to implicitly track user-browsing behavior including all the visited web pages, timestamps, and durations. In case RSS feed is available; we use it to reduce the processing time. We use the vector space model [12, 2] to classify keyword vectors of the visited web pages to the most suitable concept in the reference ontology.

An initial profile is created with the obtained non-zero concepts (concepts associated with at least 1 visited page). To improve the profile accuracy we use other data sources; in particular, search queries, click-through data, and bookmarks. Experimentation results show that all data sources we combined in our system improve the profile accuracy.

### III. BUILDING ACCURATE USER PROFILES

Personalization and recommendation systems have emerged to help users find products, services, or information efficiently. Content-based and collaborative-filtering are the main techniques used in such systems. The first one collects user data in the form of purchased products, items ratings, etc. to calculate user-item similarities, and then suggest new recommendations for each user. The second technique tracks user browsing activities to implicitly create a profile reflecting his interests and preferences, and then propose new recommendations based on this profile.

Our work is based on the collaborative-filtering technique, used for personalization systems. In this section, we focus first on the methods for creating an ontology-based user profile from the user browsing activities, and second on the way to improve the profile accuracy.

#### A. Building initial user profile

To build a user profile, we use the browsing data mainly the visited Web pages. The system classifies each page into the most similar concept in a reference ontology. We create user profiles in a three phases’ process:

1. **Reference ontology preparation.**
2. **Collecting user-browsing data.**
3. **Classifying the visited Web pages.**

##### 1) Reference ontology preparation

Ontologies are a description of concepts and a specification of concepts' associations [7]. Personalized systems can use ontologies to handle the cold-start problem [13]. Usually, personalization systems perform poorly until they collect enough relevant data. Using ontologies as the basis of the profile helps to ease this problem since user data is matched with existing concepts and inter-concepts relations.

As a reference ontology, we made use of the Open Directory Project (ODP) ontology [1]. Preparing the reference ontology ODP means to create for each concept in this ontology, a vector with the important related terms with their weight. Each concept within the ODP ontology has a number of pages available as training data combined in a super document. This creates for each concept a super document processed to exclude stop words, and remove common suffixes with the Porter stemming technique [6].

After this process, we calculate and save vectors for each concept storing weighted important terms. Hence, all concepts are represented with n-dimensional vectors (where n is the number of distinct terms related to each concept). We calculate term weights in each concept vector using TF*IDF. In more detail, \( u_{ij} \) (the un-normalized weight of each term \( i \) in a concept \( j \)) is measured as follows:

\[
 u_{ij} = tf_{ij} \times idf_{ij} 
\]

Where:
- \( tf_{ij} = \) frequency of \( t_i \) in \( sd_j \)
- \( sd_j = \) the super document used for training concept \( j \)
- \( idf_j = \) inverse document frequency of term \( i \)

The final weight \( w_{ij} \) for each term \( i \) in a concept \( j \), is calculated as follows:

\[
 w_{ij} = \frac{u_{ij}}{\sqrt{\sum_{t_i \in d_j} (u_{ij})^2}} 
\]

##### 2) Collecting user browsing data

After the browsing session, the system collects the user’s browsing data: the visited web pages (URLs, the time when visited, Web page size...), search queries, Click-through Data, bookmarks, etc.

- **Visited Web pages and RSS feeds:**

  Based on the visited URL the system extracts either the RSS feed file if it exists or the HTML web page. The reason we first check for the RSS feed file is that generally, this file contains a simple less noisy version of the web page, which helps to reduce the noise and processing time. If it doesn’t exist the system extracts the HTML page filtering all the extracted files to remove pages too short to have any content (less than 1 KB) or even the ones a user viewed for a short time (less than 5 seconds), such as pop-ups, pages with no content of interest or the ones that were silently redirected. The visited Web pages are then processed to create a Web page vector.

- **Search queries and Click-through Data:**

  A search query is a text that a user tapes in a web search engine looking for information to satisfy his needs. It is usually a set of keywords with some optional search-directives (and, or, +, -, "", etc). Search queries are different from standard query languages, which are governed by strict syntax rules, and can be expressed in natural language. After each browsing session, the system collects a user search data including a historical query set \( \mathcal{Q}(q_1,q_2,...,q_m) \) and for each query a set of click-through documents \( \mathcal{D}(d_1,d_2,...,d_n) \). \( \mathcal{D} \) is a set of click-through documents a user choose from the list of results returned by a search engine. The couple (query, click-through documents set) is then processed same as the visited pages to create a query vector.
• Bookmarks:
  Links stored by a user to Web pages in which he has an interest. We use Bookmarks mainly to increase the corresponding concept’s weight in the user profile.

3) Classifying visited Web pages
  We process the collected browsing data the same way as described in the reference ontology preparation section. The visited pages and query vectors are created the same way as the ontology concepts vectors. Since the number of terms selected from the visited page is not constant, we only use the top 20 terms sorted by weight to represent a Web page content. In the classification process, we calculate the similarity between each visited Web page’s vector, and all the ODP concepts’ vectors (previously created and stored) applying the cosine similarity measure. For instance, the similarity between a visited Web page $p_k$ and an ontology concept $c_j$ is computed as follows:

$$\text{Sim}(c_j, p_k) = \sum_{i=0}^{n} w_{ij} \times p_{ik}$$ (4)

Where:
- $w_{ij}$: the weight of term $i$ in concept $j$
- $p_{ik}$: the weight of term $i$ in page $k$
- $n$: number of distinct terms in the vocabulary

Similarity results are then sorted, and each browsing data vector is classified into the most similar concept in the reference ontology with its weight/similarity. This system applies this process on the whole user browsing data, then accumulates weights from all data vectors matching a particular ontology concept. The result of this process is an initial user profile (Figure 1), including all classified concepts with non-zero weights.

B. Improving the profile accuracy
   To enhance user profile accuracy, we take the following measures:
   1. Examining the profile stability and determining the amount of browsing data needed to achieve stability.
   2. Rank ordering the concepts in the user profile to cut the lower-ranked ones, and produce a more accurate profile.
   3. Determine the number of ontology levels that are sufficient/required to produce an accurate user profile.
   4. Exploit the user’s feedback.

1) Profile stability
   Classifying a new page or a search query either increases the weight of an existing concept or adds a new one to the user profile. Though, even if the concepts' number increases by time, eventually the most important ones become relatively constant, indicating the user’s main preferences.

   To determine the volume of browsing data needed to achieve stability in a user profile, we measured the similarity between the classified pages and the top 50% of the user profile as the number of concepts increases over time to observe when (and if) the profile becomes stable (Figure 2).

   Users have different browsing habits, and the amount of concepts varies considerably between users, and continuously increases showing no convergence over time.

   When the amount of collected browsing data was plotted against the number of concepts in the user profiles, there was no considerable convergence. Though, when we considered only the top 50% of the profile concepts (sorted by the number of associated pages), we noticed some stability.

2) Concepts Rank Ordering
   We can sort concepts in a user profile by their weights or the number of browsing data vectors associated. Calculating the F-measure helps in deciding which sorting method produces a more accurate profile. We also determine, for each method, the non-relevant concepts' average rank.

   Ordering concepts in the user profile by importance helps to eliminate the lower-ranked ones and produce a more accurate profile.

   Next, we evaluated the profile accuracy using F-measure based on the percentage of concepts kept at different cut-offs. To set a cut-off threshold that allows for better profile accuracy, we use the percentage with the highest F-measure.

Fig. 1. Screen shot of an initial user profile.

Fig. 2. The classified pages similarity with the top concepts in the profile.
3) Ontology levels
The concept-hierarchy in the ODP ontology contains many levels. The levels used in the user profile affects accuracy. To decide the number needed, we investigated building profiles with 1 level, 2, and 3 levels in the concept-hierarchy of the reference ontology. Based on the users’ feedback, we received the following relevance judgments:

![Fig. 3. Relevance of the user profile for each ontology level.](image)

When using the first ontology level, most profiles contain a maximum of 9 concepts, reflecting users' broad interests, and providing little information about his specific preferences. Therefore most users have similar profiles while their reel interests are different.

The specificity of the profile increases as we use more levels from the reference ontology. With two levels, we get an average of 42 concepts per profile, and 76 with three.

4) Exploiting user’s feedback
To ensure the user profile’s accuracy, an initial version of the profile is presented to the user using a graphical user interface. The user is asked to give his feedback on the presented profile in order to illuminate irrelevant concepts, add new ones, increase or decrease the concept’s weight. This process is optional for the user and helps get an even accurate user profile.

IV. IMPROVED USER PROFILE STRUCTURE
Building user profiles requires observing, and learning each user’s browsing behavior after each browsing session. For reasons we discuss later in this section, we use a profile composed of three layers or sub-profiles:

- Short-term.
- Long-term.
- The archive.

After each browsing session, a list of visited concepts is prepared and analyzed to be classified as a short-term or long-term interest. User preferences change over time, as he may suddenly be interested in a topic or simply forget about another.

For example, Football fans show interest in the World Cup when this event starts, but likely turn their interest to other sports after.

The short-term layer contains the user’s recent interests. To classify topics in this category, we define a threshold for concepts’ weights. Then we classify concepts, with weights above it, as short-term, which in time, can turn long-term.

Long-term interests appear more regularly in the user profile over a longer period. For example, programming languages as a topic appear more regularly in a programmer’s profile. Thus, the long-term layer reflects the stable user's interests, while the changing ones are in the short-term.

The archive contains interests that are no longer important to the user. Concepts, that are gradually losing their weight values (importance), will eventually be sent to the archive layers based on a weight threshold. We used a three layers profile for the following reasons:

- The three layers profile help the system adapt to the user’s interest change.
- With the short and long-term layers, we separate stable from occasional interests.

V. EXPERIMENTAL RESULTS AND DISCUSSION
We have conducted a series of initial experiments using two versions of the system, after the implementation phase, to test the accuracy of the user profile based on some volunteers’ feedback.

In the first version, we use only the visited Web pages to build profiles. While in the second one we combine all user browsing data (Section 3.3.2) to get a more accurate profile. We asked subjects to browse freely for three separate sessions, and then give their feedback on the resulted user profiles by selecting irrelevant concepts in this one.

Since the entire profile construction is an information retrieval process, calculating the profile accuracy is also based on the metrics used when evaluating information retrieval systems [21].

For that purpose, we calculate precision, recall, and F-measure metrics considering all non-zero concepts in all subjects’ profiles on both versions.

![Table-1: Collected feedback data for the 1st version](image)
The following graphs represent those metrics calculated based on the users’ feedback data:

**Fig. 4.** Precision calculated for all subjects.

The precision graph shows how relevant the concepts in the profile are to the user calculated by dividing the number of relevant concepts by the total number of the user profile concepts.

**Fig. 5.** Recall calculated for all subjects.

Recall shows how complete the profile is, measured by dividing the relevant concepts’ number in the user profile by the number of relevant concepts approved by the user.

### Table- II: Collected feedback data for the 2nd version

<table>
<thead>
<tr>
<th>Subjects</th>
<th>Total number of relevant concepts in the profile</th>
<th>Total number of relevant concepts by the user</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>16</td>
<td>14</td>
<td>14</td>
<td>0.87500</td>
<td>1.00000</td>
</tr>
<tr>
<td>S2</td>
<td>12</td>
<td>11</td>
<td>12</td>
<td>0.91667</td>
<td>0.91667</td>
</tr>
<tr>
<td>S3</td>
<td>8</td>
<td>8</td>
<td>9</td>
<td>1.00000</td>
<td>0.88889</td>
</tr>
<tr>
<td>S4</td>
<td>12</td>
<td>12</td>
<td>13</td>
<td>1.00000</td>
<td>0.92308</td>
</tr>
<tr>
<td>S5</td>
<td>10</td>
<td>9</td>
<td>9</td>
<td>0.90000</td>
<td>1.00000</td>
</tr>
<tr>
<td>S6</td>
<td>14</td>
<td>12</td>
<td>13</td>
<td>0.85714</td>
<td>0.92308</td>
</tr>
<tr>
<td>S7</td>
<td>15</td>
<td>14</td>
<td>15</td>
<td>0.93333</td>
<td>0.93333</td>
</tr>
<tr>
<td>S8</td>
<td>10</td>
<td>8</td>
<td>10</td>
<td>0.80000</td>
<td>0.80000</td>
</tr>
<tr>
<td>S9</td>
<td>20</td>
<td>17</td>
<td>19</td>
<td>0.85000</td>
<td>0.89474</td>
</tr>
<tr>
<td>S10</td>
<td>18</td>
<td>17</td>
<td>18</td>
<td>0.94444</td>
<td>0.94444</td>
</tr>
</tbody>
</table>

### Fig. 6. F-measure calculated for all subjects.

F-measure combines both precision and recall; it is the harmonic mean of the previous metrics. Results show that the profile accuracy and relevance rate increases considerably when combining browsing data. However, irrelevant concepts still exist especially when the total number of concept in the profile is important.

Overall, the classifier we have built is quite accurate, and most of the concepts detected are indeed relevant. The noise caused by misclassification does not follow a particular pattern, which can be studied in future works.

## VI. CONCLUSION AND FUTURE WORK

Our work in this paper focuses on creating accurate profiles representing users’ real interests. We collect and classify users’ browsing data over time based on the ODP reference ontology. And we use techniques to increase the accuracy of the retrieved profiles by collecting more browsing data, identifying the most important concepts and removing irrelevant ones, etc. Yet, we can still make some improvements in future. For example, we need to detect and remove non-relevant concepts efficiently. We can also use other data sources such as social networks to improve the user profile.

Using the obtained profiles to improve search results is also one of our intentions. We plan to implement personalization in a natural language question-answering system that we have developed in [30], to produce personalized answers according to the user’s interests.

## REFERENCES

Enhancing User Profile Accuracy for Personalized Web Systems


AUTHORS PROFILE

ANAS EL-ANSARI received the master’s degree in software quality from ChouaibDoukkali University, El Jadida, in 2014. He is currently pursuing the Ph.D. degree on Personalized Web systems at the LAROSERI laboratory, Sciences Faculty, ChouaibDoukkali University, ElJadida, Morocco. His research interests include Web Personalization, Recommender systems, Cryptography and Semantic Web.

ABDERRAHIM BENI-HSSANE received hisPh.D. degree in computer science from Mohamed V University, Rabat, Morocco, in 1997. Since September 1994, he has been a Researcher and a Professor at the Science Faculty, ChouaibDoukkali University, ElJadida, Morocco. His research interests include performance evaluation in wireless networks, cryptography, cloud computing, and big data.

MOSTAFA SAADI receivedhis PhD. Degree from the Faculty of Sciences, ChouaibDoukkali University, ElJadida. Since 2014, he has been a Researcher and a Professor with ENSA Khouribga, Sultan MoulaySlimane University, Khouribga, Morocco. His research interests include Computer Networking, Cloud Computing and Network Security.