Personalized Information Retrieval Approach

Myriam Hadjouni[1], Mohamed Ramzi Haddad[1], Hajer Baazaoui[1], Marie-Aude Aufaure[2], and Henda Ben Ghezala[1]

[1] Laboratory RIADI-GDL, National School of Computer Sciences, University of Manouba, 2010 la Manouba, Tunisia
{myriam.hadjouni, hajer.baazaouizghal, henda.benghezala}@riadi.rnu.tn
haddad.medramzi@gmail.com
[2] MAS Laboratory, SAP BusinessObjects Chair, Ecole Centrale Paris, Grande Voie des Vignes, 92 295 Chatenay-Malabry, France
marie-aude.aufaure@ecp.fr

Abstract. Web information system personalization is an emergent research field with the objective to facilitate the use and the control of Web content. This paper presents a personalized information retrieval approach based on end user modelling. The proposed approach personalizes data retrieval using implicit user information and interests measurements. As the data manipulated is expressed by attributes and values, we define several similarity measures. These measurements consider both semantic and spatial user contexts. The approach personalizes Web content and especially spatial information focusing on its spatial semantic aspects.

Keywords: Web personalization, spatial Web personalization, user modeling

1 Introduction

The volume of information available on Web information systems is growing continuously. Browsing this content becomes a tedious task given the presentation of data that does not meet user's aims and needs. To satisfy user needs, personalization is an appropriate solution to provide an adaptive and intelligent Human-Computer-Interaction (H.C.I) and to improve the information systems usability. Moreover, Web resources increasingly contain geo-referenced entities generally associated with a geographical location [1]. Geographic information systems suffer from a lack of data quality since the presented data is highly complex and diverse. Manipulated objects have a very rich semantic and the degrees of user's interests towards them vary depending on context and on personal tastes. Existing personalization approaches try to determine user preferences to help him while exploring information. However, these approaches have some drawbacks and generally neglect semantic and spatial aspect of the manipulated data. The semantic Web provides an appropriate background to define and to describe information systems' content in order to take into account its different semantics and to efficiently understand it.

This paper presents a personalized information retrieval approach based on end user modeling. The proposed approach personalizes data retrieval using implicit user information and interests measurements. We start, in the next section with related works presentation and discussion. We then present our architecture including user and data modeling approaches and the similarity measures used to increase the quality of the personalization process and the measures used to deduce user's interest. These measures help us to construct a user's network based on our proposed system. Section 4 presents this network. Section 5 concludes our work and gives some perspectives.

2 Related Works

Generally, personalization methodologies are divided into two complementary processes which are (1) the user information collection, used to describe the user interests and (2) the inference of the gathered data to predict the closest content to the user expectation. In the first case, user profiles can be used to enrich queries and to sort results at the user interface level [2]. Or, in other techniques, they are used to infer relationships like the social-based filtering [3] and the collaborative filtering [4]. For the second process, extraction of information on users' navigations from system log files can be used [5]. Some information retrieval techniques are based on user contextual information extraction [6]. Information semantics are also used to enrich the personalization process; queries can be enriched by adding new properties from the available domain ontologies [7]. The user modeling based on ontology can be coupled with dynamic update of user profile using results of information-filtering and Web usage mining techniques.

With the evolution of the internet, Web resources increasingly contain geo-referenced entities generally associated with a geographical location [1]. Statistics collected through search engines show that spatial information is pervasive on the Web and that many queries contain spatial specifications, but it is more difficult to find relevant resources responding to query including a spatial component [8]. The spatial information personalization should consider spatial properties and relationships found in Web documents. Design of spatial Web applications requires at least three components: (1) a user model and associated user preference elicitation mechanisms and (2) a personalization engine combining spatial and semantic criteria and (3) a user interface enriched with spatial components [9]. The spatial Web personalization requires the representation of user features, particularly those relevant to the spatial domain. [10] explores semantic similarity and spatial proximity measures as well as relevance ranking functions on the behalf of the user. Semantic similarity is the evaluation of semantic links existing between two concepts [11]. [12] introduced a classification algorithm for measuring spatial proximity between two regions. Another aspect of spatial Web personalization techniques concerns interactive adaptive map generation and visualization. These techniques are concerned with Web maps adaptation according to user's needs [13]. In a prototype applied to maritime navigation, [14] categorizes the users according to their geographical context.

The presented personalization approaches have contributed to the improvement of information systems use. However and despite their widespread use, these approaches have weaknesses and limitations. In fact, several approaches, like the collaborative ones, present the same recommendations for all users within the same cluster. Thus, they do not consider some specific users preferences when they represent a minority in a given group. Content based approaches facilitate items retrieval by proposing some alternatives and recommending similar items to the one that the user is visiting. However it focuses only on the user's actual and temporary needs and can't highlight the items that are related to the current query results. Other approaches try to determinate the interests of each user but they are limited by their items model that doesn't describe the differences between items properties. This lack of semantic description of the items decreases the quality of personalization since similarities and dissimilarities between items can't be measured accurately. In addition, in most personalization approaches, the spatial aspect is not taken into consideration, which requires an adaptation of those approaches to be relevant while applied to spatial information. These limitations explain the importance given to hybrid approaches. The hybridization of existing approaches is presented as an alternative that would improve the quality of personalized systems [15].

3 Personalized Information Retrieval System

We present in this section a Personalized Information Retrieval approach for Web information systems. Our proposition is based on a dynamic and iterative construction of a multidimensional user model. This multidimensional approach is used to represent and describe the user towards different dimensions. The model proposed (noted M_u) is composed of 4 dimensions: user profile, spatial model, graphic model and textual model. If we consider U as the set of users, a user $u \in U$ will have as model M_u:

$$M_u = D_t \cup D_s \cup D_n \cup P \qquad (1)$$

Where:
- D_t = keywords employed by the users for their textual search
- D_s = spatial positions of the users
- D_n = entities visited by the users
- P = user profile.

This model is part of the user modeling level of our system which is composed of the following ones: user level, application level, user modeling level, information retrieval level and storage level [16]. In this paper, we only consider the modeling level which is also concerned with the exploration and the identification of semantic and spatial relationships between entities extracted from log files or from real-time navigation.

Users can be known or unknown. A known user has an account and is individually identified by the system while an unknown user does not possess proper profile. So, the user modelling level is also dealing with:

- – For unknown users: (1) the construction of user real-time profile and (2) the mapping of this current user into the closest prototype.
- – For known users: (1) the activation of the concerned user model and (2) the update of his real-time profile.

The model is based on an implicit interaction with the user: implicit because the user isn't directly asked to give opinion, and interactive because we use the navigation to measure its interest to a given entity. These measurements are based on:

- The similarities that could exist between attributes and entities of interest.
- The deduction of user interests from all its navigation.
- The calculation of the pertinence of a supposed spatial move. In fact, in our approach, we also consider that if a user in interested on a spatial zone, he aims to move to.

Next sections present the measurements cited above.

3.1. Similarity measures

The similarity measures used in our system are attributes values similarity and entity similarity.

Attributes values similarity These attributes can be simple attributes like numeric values, Booleans and strings. Otherwise, attributes can be composite such as numeric intervals and sets, or strings sets. We define for each attribute type a similarity measure:

- Numeric values: These values are discrete, so it is necessary to have a similarity function to divide their variation intervals into sets of similar values such as for hotel prices or rooms number. Each value can be considered to be similar to neighbourhood's values. The neighbourhood's width increases when the value is high and vice versa. We define neighbourhood of similar values as:

$$Sim(\alpha) = \left\{ \beta, \beta \in \left[\alpha - \left(\frac{\alpha}{\varepsilon} \right), \alpha + \left(\frac{\alpha}{\varepsilon} \right) \right] \right\} \quad (2)$$

Where ε defines the neighbourhood's width.

Semantic distance between α and $\beta \in [a,b]$ is:

$$D_{sem}(\alpha, \beta) = \frac{|\alpha - \beta|}{|a - b|} \quad (3)$$

And the similarity degree between them is:

$$D_{sim}(\alpha, \beta) = 1 - D_{sem}(\alpha, \beta) \quad (4)$$

With these three equations, we will have the similarity as:

$$Sim(\alpha) = \left\{ \beta, D_{sim}(\alpha, \beta) \geq 1 - \frac{\left(2\alpha / \varepsilon\right)}{b - a} \right\} \quad (5)$$

– Numeric intervals: The interval attributes may have common values. These values are included within the similarity measure. If we consider A and B as two numeric intervals, the similarity degree between them is:

$$D_{sim}(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{6}$$

– Sets: as for intervals, groups or lists can have common values. Here, we need to define the well known intersection and union operators to include our similarity. Let's consider $Sim(F_i,a_i)$ the set of all values that are similar to the value a_i of the attribute F_i. Intersection and union operators taking into account the similarity applied to the sets A considered as reference, and B are:

Definition 1

$$A \cap_s B = \{a_i \in A, \quad \exists b_j \in B \quad tq \quad b_j \in Sim(F_i, a_i)\}$$

$$A \cup_s B = A \cup \{b_i \in B, \quad \neg\exists a_j \in A \quad tq \quad b_j \in Sim(F_i, a_i)\}$$

So, with A and B as two sets of numeric values, the degree of similarity becomes:

$$D_{sim}(A,B) = \frac{card\{A \cap_s B\}}{card\{A \cup_s B\}} \tag{7}$$

Thus, A and B are considered similar if $D_{sim}(A,B) \geq \varepsilon$.

ε is the similarity threshold to be determined in relation to the attribute semantic.

– Other types: On the cases of the other types of attributes like strings or Boolean the similarity operator is equality.

Entities similarity The similarity measure between two entities corresponds to the aggregation of their attributes similarity degrees. If x and y are two entities of the same concept, γF_i is the attribute's importance coefficient. α_i and β_i are the attribute values submitted by x and y. Semantic distance between them is:

$$D_{sem}(x,y) = \sum_{i=1}^{n} \gamma F_i * D_{sem}(\alpha_i, \beta_i) \tag{8}$$

With n: number of attributes describing x and y which depends on the concept to which they belong.

Having ε_i as the similarity entry of the attribute F_i, we consider that x and y are similar if:

$$D_{sem}(x,y) \leq \sum_{i=1}^{n} \gamma F_i * \varepsilon_i \tag{9}$$

So the set of similar entities to x is:

$$Sim(x) = \left\{ D_{sim}(x,y) \geq 1 - \sum_{i=1}^{n} \gamma F_i * \varepsilon_i \right\} \tag{10}$$

For building the user model, we start with the profile construction, taking interests into consideration. This information is deduced by measuring user interests related to textual, spatial and navigational dimensions. Next section presents the approach used to build the user profile using information from navigational dimension.

3.2. User interests

The concept of implicit information gathering adopted is as follow: when a user touches information, he is implicitly voting for it. Based on this assumption, we introduce measures that aim to deduce user interests. In the following, *u* represents a user; an item visited by *u* is noted as x_i and a concept (accommodation, restaurant, etc…) to which x_i belongs is noted *C*.

Interests indicators The user profile, in our context, is also used to describe user interest towards the visited items. We use the following parameters:

- Visits frequency: Represents the number of visits to a given item. This number can reflect the user interest for an item or a group of items and can be used to determine the most interesting ones.
- Visits duration: In addition to the visits rate, the visit duration is also significant as users spend more time reading information perceived as relevant [1].
- Explicit items rating: User's behaviours do not always reflect his real interest: a user can visit an item description more than once or spend much time on reading it just to satisfy curiosity. To avoid such situation, the user can explicitly evaluate his interest. This evaluation is then used to adjust the previous measures notes.

Visited entities interest The main measures used here to deduce the user interest are the frequency and the duration of visits. To do this, we use all saved user transactions and their duration. If the user gives an explicit evaluation on a visited entity, we can refine these measures. The interest degree I_e of a *u* towards a visited item *x*, considering only the semantic aspect of *x*, is:

$$I_e(u,x) = \frac{1}{2} * \left(\frac{d_v(u,x)}{D_v(u)} + \frac{n_v(u,x)}{N_v(u)} \right) * V(U,x) \quad (11)$$

Where

$d_v(u,x)$: Duration of the user visits to *x*.

$D_v(u)$: Total duration of the user visits.

$n_v(u,x)$: Number of times the user *u* visited x.

$N_v(u)$: Total number of user visits *u*.

$V(u,x)$: Average rate assigned by *u* to *x*. This note is between 0 (not interesting) and 1 (Very interesting). The default value is 0.5 if the user did not rate the item.

Thus, the quality of preferences and interests elicitation process increases when the user evaluates explicitly visited entities. Taking into account ratings is important when a user spends a long time visiting an entity that is not interesting for him or when he visits frequently an item that he is not planning to visit. So, rating this entity would reduce the influence of the duration and the frequency of those visits.

Concepts interest Spatial entities are classified into several predefined concepts with regard to the considered domain. Each spatial concept is defined by several attributes which are used to describe entities belonging to it. In order to make a personalized and relevant classification of concepts we measure the user interest towards spatial/semantic entities belonging to a considered concept. With $x_i \in C$, the interest degree of *u* towards *C* is:

$$I_c(u,c) = \sum_{i=1}^{n} I_e\left(u, x_i\right) \tag{12}$$

Where

n is the number of items x_i visited by *u*.

$I_e(u,x_i)$: The interest degree of *u* towards x_i.

Features values interest After presenting the metric for classifying concepts by interests' degrees, we aim to order entities within the same concept according to their relevance. This is done with correlation to concepts, i.e. the attribute's interest degree value is calculated by considering the interest to entities with similar value.
Consider an attribute $F_i \in C$ and v_k one possible value of F_i and $X(F_i,Sim(F_i,v_k))$ the set of all entities visited by the user and which have a similar value to v_k. $Sim(F_i,v_k)$ is the set of F_i values that are similar to v_k. The similarity operators depend on the attribute type and semantic. The interest value of *u* the value v_k is:

$$I_f(u,F_i,v_k) = \sum_{x_j \in \left(F_i, Sim\left(F_i, v_k\right)\right)} I_e\left(u, x_j\right) \tag{13}$$

The next step is to normalize this interest value to be independent of the elements number.

$$I_v(U,F_i,v_k) = \frac{I_f(U,F_i,v_k)}{\sum_{v_j \in V\left(F_i\right)} I_f(U,F_i,v_i)} \tag{14}$$

Where $V(F_i)$ is the set of all possible attribute F_i values.

Interest prediction As already mentioned, user preferences are introduced by interests degrees on spatial/semantic concepts and their attributes values are deduced from user navigations. We use the previous metrics to predict the user's interest degree for the non-visited items. So, we assume that:

- The user's interest degree towards a concept increases with the number of visits to entities of this concept.
- The more preferred entity values are, the more this entity is likely to be visited.
- The concepts attributes have different importance degrees. So we can distinguish between the critical attributes and those that do not influence the user choices.

This measure is used to predict the most relevant content the user is searching for. Moreover, it contributes to refine results of textual search mainly based on the use of concepts and attributes that will define the domain ontology. We aim to deduce this ontology from previous (and current) search and navigations and not to explicitly involve the user. To achieve these points, we present the following formula, which predicts the degree of user satisfaction for one entity *x* depending on its concept *C* and its attributes values:

$$S(x,u) = I_c(u,C) * \sum_{i=1}^{n} \alpha_{F_i} * I_v(u,F_i,v) \quad (15)$$

Where
n is the number of the concept attributes *C*
v is the value of the attributes *Fi* presented by the entity *x*.
Ic(u,C) is the interest degree of u on the concept *C*.
$I_v(U,F_i,v)$ is the interest degree of *u* on the value *v* of F_i.

4 Construction of the user results

Having as objective the amelioration of users' results, we iterate the construction of his model. In the end of each iteration, we determine the data pertinence and use this value in the next iteration. With this hypothesis, two types of information are the basis of personalization: the location of user at the moment of his search and noted L_u and the location of his area of interest corresponding to the target search zone noted Z_i. The area of interest is deduced from the position on a map in the user interface or from the search keywords. The results provided to users are constructed on the basis of three points: (a) the path needed to reach the area of interest Z_i from the user location when searching L_u, (b) the arrival point and (c) the semantic and spatial distance between a departure point and the objects of interest.

The area of interest path. The estimation of the arrival path over the area of interest is a quality criterion. In fact, the taken path has an impact on the arrival point into the area of interest and therefore affects the location of objects of interest displayed to the user. In this context, we use a qualitative evaluation based on an explicit weighting of criteria such as cost or type of way. Then, the user can add its own choice value. The minimum criteria that we have taken are the type of road, its cost and distance travelled. The weights are made of values between 0 and 1, values which are given in the form of notes by the user. The path evaluation equation is:

$$EvalPath = \alpha.Type * \beta.Cost * \lambda Dis\tan ce \quad (16)$$

With α, β and λ ∈]0,1]

The arrival estimation. The move of the user from its search location to its area of interest can provide us two types of information: (1) the arrival point information deduced from the path taken when move and (2) a location marker that is the collapse point of the user, we called it departure point. This point may be the subway station closest to the object of interest or the user residence. The distance between the arrival and the departure points is equal to 1 when these two points are joined. The hypotheses are:
– Existence of an arrival point in Z_i.

- If user moves to *Zi*, he has a departure point and a path: *d(arrival point, departure point)*
- The evaluation is based on the path and on the semantic of the searched object:

$$Eval\left(ArrPt\,, Path\right) = D_{SemSpa}\left(ArrPt\,, DepPt\right) * Sem\left(SearchObje\;ct\right) \tag{17}$$

Semantic and spatial distances. These distances are the key features to include the spatial aspect into the information filtering process. Semantic and spatial distances between the departure point *DepPt* and the interest objects $object_i$ are evaluated as shown below:

$$DsemSpa\left(ArrPt\,, DepPt\right) = \sum_{i=1}^{n} \mathrm{Dspa(objecti, DepPt)} * \mathrm{Dsem(objecti, DepPt)} \tag{18}$$

Where:
n is the number of the interesting objects in the area Z_i.
$Dspa(object_i,DepPt)$ is the spatial distance between $object_i$ and *DepPt*.
$Dsem(object_i,DepPt)$ is the semantic distance between $object_i$ and *DepPt*.
In the end of every iteration, these three measures are applied to ameliorate the next one. Taking into account the area of interest path will explicitly implicate the user, while the arrival estimation and the semantic and spatial distances calculation is implicitly done.

5 Conclusion and Future Work

Web personalization attracts rising research efforts to facilitate Web information retrieval and navigation. Generally, Web personalization and user modeling approaches do not focus on the spatial aspect of the information and the constraints it implies. In this paper, we have presented some background knowledge on existing Web and spatial Web personalization systems.

We have then introduced our proposition for a personalized information retrieval approach which is mainly based on the end user modeling. The user model is multidimensional and it is built with iterations that use estimations calculated from implicit user's navigation information.

Next step of our work is to add to this personalization process (1) interactions between the users' models and (2) as a consequence, the construction of a network of models. In fact, as we consider similarities and distance between the search concepts and entities within this process, we assume that adding also similarities measurements between the users' models could be helpful in the personalization process.

References

1. Winter, S., Tomko, M.: Translating the Web Semantics of Georeferences. In Taniar, D.; Rahayu, W. (Eds.), Web Semantics and Ontology, pp. 297-333, Idea Publishing (2006)
2. Koutrika, G., Ioannidis, Y.: A Unified User-profile Framework for Query Disambiguation and Personalization. In: Workshop on New Technologies for Personalized Information Access, held in conjunction with the 10th International Conference on User Modeling, pp. 44–53 (2005)
3. Mladenic D.: Text-learning and Related Intelligent Agents: a Survey. IEEE Intelligent Systems, 14(4):44–54 (1999)
4. Goldberg, K., Roeder, T., Gupta, D., Perkins, C.: Eigentaste, a Constant Time Collaborative Filtering Algorithm. Information Retrieval Journal, 4(2):133–151, (2001)
5. Paulakis, S., Lampos, C., Eirinaki, M., Vazirgiannis, M.: Sewep: A Web Mining System Supporting Semantic Personalization. In: 15th European Conference on Machine Learning and 8th European Conference on Principles and Practice of Knowledge Discovery in Databases, LNCS, vol. 3202, pp. 552-554, Springer (2004)
6. Jones, G.J.F., Brown, P.J.: Context-Aware Retrieval for Ubiquitous Computing Environments. Invited paper in Mobile and Ubiquitous Information Access, LNCS, vol. 2954, pp. 227–243. Springer (2004)
7. Messai, N., Devignes, M.D., Napoli, A., Smaïl-Tabbone, M.: Méthode Sémantique pour la Classification et l'Interrogation des Sources de Données Génomiques. Revue des Nouvelles Technologies de l'Information, Extraction des connaissances : Etat et Perspectives, pp.43-47 (2005)
8. Yang, Y., Aufaure, M.A., Claramunt, C.: Towards a DL-Based Semantic User Model for Web Personalization. In: Third International Conference on Autonomic and Autonomous Systems, pp. 61-61. IEEE Computer Society (2007)
9. Kuhn, W.: Handling Data Spatially: Spatializing User Interfaces. In: 7th International Symposium on Spatial Data Handling, Advances in GIS Research II, vol. 2, pp.13B.1-13B.23, IGU (1996)
10. Yang, Y.: Towards Spatial Web Personalization. PhD thesis, Ecole Nationale Supérieure d'Arts et Métiers (2006)
11. Resnik, P.: Semantic Similarity in a Taxonomy: an Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. Journal of Artificial Intelligence Research, 11:95–130 (1999)
12. Larson, R.R., Frontiera, P.: Spatial Ranking Methods for Geographic Information Retrieval in Digital Libraries. In: Heery, R., Lyon, L., (eds.) ECDL. LNCS, vol. 3232, pp. 45–56. Springer (2004)
13. Maceachren, A.M., Kraak, M.J.: Research Challenges in Geovisualization. Cartography and Geographic. Information Science, 28(1):3–12 (2001)
14. Petit, M., Ray, C., Claramunt, C.: A User Context Approach for Adaptive and Distributed GIS. In: 10th International Conference on Geographic Information Science. Springer-Verlag, LNGC, pp.121-133 (2007)
15. Burke, R.: Hybrid Recommender Systems: Survey and Experiments. User Modeling and User-Adapted Interaction, 12(4):331–370 (2002)
16. Hadjouni, M., Baazaoui, H., Aufaure, M.A., Claramunt, C., Ben Ghezala, H.: Towards a Personalized Spatial Web Architecture. In: Workshop Semantic Web meets Geopatial Applications, held in conjunction with the AGILE International Conference on Geographic Information Science (2008)
17. Kelly, D., Belkin, N.J.: Reading Time, Scrolling and Interaction: Exploring Implicit Sources of User Preferences for Relevant Feedback. In: 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 408–409, ACM (2001)