Extracting User Interest Center Based on a Semantic User Profile  

Khadija Mahjoubi*, Hassan Labriji, Ibtissam El achrak  
Laboratory TIM, Hassan II University, Mohammedia  
Casablanca, Morocco  

Abstract—The fundamental purpose of a personalized information retrieval system is to provide relevant results to the specific need of the user. The selection of interesting documents to a user is done on the basis of his area of interest, inferred from the information about the user or his user profile. Thus the calculation of the interest center is one of the essential elements for a relevant research. On one hand, information retrieval systems converge to a semantic representation of user profiles. On the other hand, they do not take into account the semantic links between the concepts (meaning of the word) when defining the center of interest. In this paper we use a graphical representation of the user profile based on ontology. Thus we were able to use the power of mathematical calculations of graphs to deduce the interest’s centers of the user.

Keywords—User Profile, Semantic Web, Information Retrieval System, Recommendation System, Ontology, User Profile Modeling, Similarity Measurement, Conceptual Graph...

I. INTRODUCTION

The rapid development of new information and communication technologies as well as the rise of the web, confronted us with a very large mass of heterogeneous information. The masses of accessible information have been steadily increasing, and the volumes of documents storing them are rising very rapidly. Because of this increase in the volume of information, we arrive at a paradoxical situation, never has there been so much information available, but finding in this accumulation what we are looking for, becomes more and more difficult, which leads us to both major problems: Cognitive overload and disorientation [1].

The problem of cognitive overload is related to the difficulty that can the user have to select the information corresponding to his specific needs facing a large amount of data. For example, a programmer who submits the query "java" to a search engine and which is waiting for results related to the programming language java, will have a difficulty to choose the results that interests him in a mass of results where is intermingle those related to programming and those related to the Java region in Indonesia and those related to the Java Parisian dance. The problem of disorientation is related to the fact that the user does not know which path to follow when navigating via user interface for example when browsing a website. So the problem is no longer the availability of information but the ability to select information that responds to the specific needs of the users. The conception and implementation of effective tools, allowing the user to have access only to the information he deems relevant, becomes an absolute necessity. Studies by [2] have shown that the majority of users have no idea about how the information retrieval systems work; as a result, they do not express their needs. On the other hand, users usually use only a few words (4 or 5 at most) to formulate their queries, which gives unfinished specifications on their information needs. Thus, the analysis of the interest’s center of the user reveals a particular importance. Indeed, it is
by knowing perfectly his interests that it will be possible to offer him relevant information for his research.

In this paper, we present our method of calculating the center of interest using the graph structure of the ontology as a reference in order to use the semantic link between the concepts. This article is organized as follows: First of all, we present the user profile (its modeling, its representations and its uses), then we represent the ontologies and their importance’s in some mechanisms especially in information retrieval systems and the ontology that we well use on our approach, after that we introduce our method of calculation the center of interest with its various axes, in particular the representation from the user profile to the conceptual graph ,The similarity function used between their concepts as well as the calculation method used and finally we give a conclusion with our perspectives.

II. USER PROFILE DEFINITION AND MODELING

The user profile represents a collection of personal data associated with a specific user that describes a set of attributes. These attributes can include geographic location, academic and professional experience, goals (short-term and long-term), behavior, interests, preferences, etc. [3]. A profile can be relative to a single person, or to a group of people with commonalities, such as members of a working group. Generally a user profile can be built by two methods: either by the user himself what is called explicit profile, or automatically from data from the interactions between the user and the system, in this case it is called implicit profile. This last step is the most common, since the manual entry of parameters (preferences, interests ...) by the user can be a tiring task for him and can take a lot of time to express his needs.

The user profile data is represented as needed. In general, they are stored in the knowledge base of the system as a property-value pair. Once the data is structured, they are used by the data mining algorithms to build the user profile. Several techniques of data mining can be used. [4] classify the techniques used according to the models of user profiles to build: Behavior modeling, Interests modeling and Intentions modeling

- Behavior modeling: Consists of analyzing user behaviors via browsing histories or transactions they perform on Web servers [5], in order to determine recurring navigation courses, to validate marketing strategies or to check the relevance of marketing campaigns. Among the methods of this type of analysis we cite: The Markov chains to predict future URLs that will be visited by the user [6].

- Interest modeling: this type of modeling is characterized by the definition of the centers of interests by a function fct (i) which gives the degree of interest or disinterest of a user for an item i by analyzing his previous behavior. [7]. Three approaches are widely used to build the centers of interest of the users. [8]: The direct approach, the semi-direct approach and the indirect approach.

- Intention modeling: An intention here is the purpose for which the user uses the information system. The modeling of intentions of the users is to build a model that will identify the purpose of each user of the information system. For example, customers of an e-commerce website can be divided into two groups: those who really aim to buy and those who do not aim to buy. Intention modeling is largely based on classification techniques with predefined categories [5].

The modeling of interests allows to build user profiles that are only data or information that will be reused by the appropriate mechanisms of adaptation of the information to the to produce the knowledge. These profiles are generally represented in three ways: ensemblist representation, semantic networks representation, and conceptual representation. We will discuss these types of representations in the following section.

A. User's Interest

The center of interest expresses the area of expertise of the user or its area of exploration. It can be defined by a set of keywords (concepts) as mentioned above, or a set of logical expressions (queries).

In many approaches the importance of each concept is defined by a weighting of the keywords
of the center of interest. The ontology of the domain completes the definition of the center of interest by explaining the semantics of certain terms. For example, we can explicitly define Apple ‘a computer ‘in the profile and not ‘a fruit’. The center of interest can be seen as a virtual screening which reduces the amount of information to be taken into account. We can bring the center of interest closer to the concept of Apple view. Therefore any request issued by the user will be enriched with the keywords or predicates of the requests defining the center of interest. The center of interest can be correlated with personal data and be enriched by deduction of certain information.

There are several techniques for representing the centers of interest that make the user profile.

1) The Ensemblist Representation

This approach consists of representing the user's profile by packets of weighted terms, also known by vector representation by analogy to the vector model of Salton [9] on which it is based. These packages of terms, reflecting the interest of the user can be grouped differently according to the approach taken to exploit the user's profile. We distinguish four major approaches that represent the user profile based on this kind of representation: Representation by keyword list [10], like the case of web portals: MyYahoo, InfoQuest, representation by vector of weighted terms for each center of interest [11], representation by a set of vectors of terms: weighted (or not) independent [12], representation by definition of a relation of order between the centers of interest of the profile, in this case we talk about preferences [13]. The weighting of terms is usually based on the tf.idf schema commonly used in information retrieval [14].

2) Representation by semantic networks

In order to solve the problem of polysemy of the terms inherent to the ensemblistrepresentation, a first solution consists in representing the profile by a network of weighted nodes in which each node represents a concept reflecting the center of interest of the user. This type of representation offers the double advantage of structuring and associative representation (relationships between terms) allowing considering all aspects representative of the profile.

The centers of interests are often represented by node-pair relationships in which each node contains a term derived from implicit data used to build the profile. The arcs linking the nodes are created on the basis of co-occurrences between these terms. However, the separaterepresentation of each word by nodes in the semantic network is not precise enough to decline the different meanings of the centers of interests of the user. Another possibility is to exploit external sources such as ontologies to establish links between nodes, hence the appearance of conceptual representation.

3) The conceptual representation

The representation of the profile in this approach highlights one or more semantic relations between the information of the profile. According to a general context [15], this representation offers an interesting alternative to the semantic networks approach. Indeed, current work tends to represent the profile in the form of ontology of personal concepts based on the knowledge contained in ontologies rather than building user profiles only from the implicit data collected from its interactions [16]. The conceptual representation is similar to the semantic network representation in the sense that represents the user's interests through a network of conceptual nodes. However, in the conceptual approach, the nodes correspond to abstract domains representing the interests of the user, contrary to the semantic approach where the centers of interest are represented by specific words, or a set of relevant words.

From the association of the user's interest to the concepts of the domains of the ontology, we obtain a profile represented in the form of a hierarchy of concepts. The Implicit data from user interactions is classified into these concepts and the user's interest in such concepts is recorded. (It’s the type of representation that inspires our work). Semantic representation approaches exploit ontology of reference to represent interests of the user according
to weighted concept of vectors of the ontology used.
We mention the hierarchy of concepts "Yahoo" or
the ODP as the most evidencesources used in this
type of approach.

B. Exploitation of the user profil/interest

The center of interest is widely used in many
fields and applications because of its great
importance, such as the improvement of processes
and the processing of information, as the case of
information adaptive systems where we integrate the
user profile or more specifically theirs centers of
interests in order to solve two major problems as we
have mentioned in the introduction of this article:
cognitive overload and disorientation .In information
adaptive systems there are two main types of
systems:

1) Personalized information search system

Information Retrieval (IR) refers to the set of
methods, procedures and techniques for the
acquisition, organization, storage, retrieval and
selection of information (data, texts, images
, videos). Unlike the recommendation systems, the
user expresses his need for information by a query in
an IR system. In a conventional IR system, the user's
query (weighted wordlist) is matched with the list of
indexed documents (weighted wordlist) in order to
return relevant documents corresponding to the
user's query (search engines): Google search, AOL,
Yahoo!, etc.). In a hypertext environment like the
Web, documents (Web pages) are not indexed only
on their content (keyword list), but also via
additional metrics based on the structure of links
between pages, Google's Page Rank algorithm [17]
is one of the best known in this context. The queries
expressed by users in conventional IR systems are
generally short and may contain ambiguities [18].
So the personalized IR systems come to improve the
cognitive overload problems of conventional IR
systems by integrating the interests of profile in
these mechanisms. According to [19] this integration
can be done in three main ways in the information
retrieval process:

a) Reformulation or expansion of the request:

b) Customized selection of information

It consists of integrating the user profile
parameters into the similarity function between the
user's request and each document, as the case of [21].

c) Reordering the results

It consists of using the user's profile terms to
reorder the results from a conventional IR system,
[1].

2) Recommendation systems

The recommendation systems are a specific form
of information filtering, the purpose of which is to
present the pieces of information that may be likely
to interest the user. The most common case in this
kind of system is the recommendation based on
filtering by contents using the user's interests in-
commerce sites that sell a very large variety
of products, while each customer potential is very often
interested in a very limited number of products. A
content recommendation system will compare the
profile of resources (e.g. product) with the user's
profile (interests) to select the resources

corresponding to their specific needs. The filter used
to select the relevant resources for the user is a
similarity function between the user's profile and
each resource

III. DEFINITION OF INTERESTS CENTER OF THE USER
BASED ON A SEMANTIC USER PROFILE

After defining the user's model and touch its
importance in the different phases of information
adaptive systems, defining the interest’s center is an
essential research area in this field. Indeed, the
deduction of interests of a user will allow the
information retrieval system to target the specific
need of the user by using it in the different phases.
Unfortunately, the existing methods of calculating
the center of interest do not take into account the
semantic links between the concepts, so we have
represented the user profile in graph in order to use
the power of the mathematical methods, which
allowed us to deduce the center of interest using the
semantic relationships between concepts. In this section we present our method of extraction of the interest of the user based on a semantically structured user profile. First, we present the ontologies and especially the ODP ontology, then we present the model of definition of user profiles based ontology and finally we present our method of calculation of the center of interest.

A. Ontology

Ontology refers to a set of hierarchically related semantic resources. It allows to describe the knowledge of a specific domain and to present the relations between the concepts as well as to give the missing rules and axioms to the semantic networks. The fundamental purpose of ontology is to semantically process information. There are several ontologies designed to list the content of web pages for an easier navigation by users. We cite for example Online portals such as "Yahoo", "Lycos", and "ODP" for Open Directory Project which is an open-concept hierarchies in RDF format widely adopted by many systems such as OBIWAN (Ontology Based Informong Web Agent Navigation) [22], Personae [23], and since the ODP is the largest and most complete directory of the web published and edited by experts of the domain, we use it as a source of semantic knowledge in the process of access to information. Our goal in this section is to represent each concept of the ODP by a set of terms, thus later serving the derivation of the semantic representation of the user's profile.

ODP data is available in two RDF files: the first one contains the tree structure of the ODP ontology, and the second lists the resources or web pages associated with each of the concepts. Each concept of the ODP represents an area of interest for the user and associated by experts of the domain to web pages; the latter are the most relevant in relation to the concept. The concepts are organized hierarchically where high level concepts represent general concepts and low level concepts represent specific concepts. Each concept of the ODP is represented by a title and a description describing in general the content of the associated web pages, and each web page is also associated with a title and a description describing its content. The concepts of ontology are connected by relations of different types such as "is-a", "symbolic" and "related". The links "is a" allows us to pass hierarchically from generic concepts to more specific concepts. "Symbolic" links support the multi-classification of pages in several concepts and allow the user to navigate between concepts related semantically without resorting to general concepts. "Related" links are labeled "see also" permit to point to the concepts dealing with the same theme without having web pages in common.

B. User profile as a concepts graph

We have represented the profile as a set of semantically related concepts comprising concepts of several specific levels. Indeed, the data of the ODP are represented by two files of "RDF" type (Structure.rdf and Content.rdf), the first file contains the tree structure of the ontology and the second file lists the web pages associated with each category. The goal is to represent the user profile using the same structure of the "Content.rdf" file to exploit the architecture of this ontology. We define the file "Profil.rdf" constituting the user profile containing all the knowledge concerning the user.

Formally, the structure of the graph \( G = (V, E, W) \) is constituted by a hierarchical component formed by the links of type "is-a" and a non-hierarchical component formed by links of different types predefined in the ontology of the ODP, where:

- \( V \) is a set of weighted nodes, representing the concepts used by the user,
- \( E \) is a set of arcs between the nodes of the graph \( V \),
- \( W (Ci) \) is the weight of the concept \( Ci \); it represents the number of use of this concept by the user.

1) The importance of the center of interest

Two concepts are neighboring or adjacent if they are connected by one arc, one is the father or the son of the other. The neighborhood of a concept \( c \), denoted \( V (c) \), is the set of his neighboring vertices. The importance of a concept is calculated by the following formula:

\[
Ip(c) = \sum_{a \in V(c)} poids(a) sim(a, c) \quad (1)
\]
Where \( p(w(a)) \) is the weight of the concept \( a \), and \( \text{sim}(a, c) \) is the similarity between the concepts \( a \) and \( c \).

**a) Similarity function (The choice)**

The similarity function is a function characterized by values between 0 and 1, which allows probabilistic interpretations of similarity. To choose the appropriate function for the purpose of our work we must first choose the appropriate approach to identify the measure of similarity. Generally there are four main approaches:

**The Arc Approach**

The most intuitive similarity measure of objects in ontology is their distances. These measures use the hierarchical structure of the ontology to determine the semantic similarity between the concepts. Among the works classified under this category we have:

- The measurement of Rada [24]: this measure is adopted in a semantic network and is based on the fact that we can calculate the similarity from the hierarchical "is-a" links.

- The measurement of Wu and Palmer [25]: This measure has the advantage of being fast in terms of execution time, but in some situations, we can obtain \( \text{Simwp}(L, Z) < \text{Simwp}(L, M) \), where \( Z \) is a child of \( L \) and \( M \) is a brother of \( L \), this situation is inadequate in the context of information research.

- The measurement of Labriji [26]: This measurement comes to improve the measurement of Wu and Palmer by ameliorating the similarities between the concepts located on the same hierarchy of ontology.

**The Node-based Approach**

These approaches adopt a new measure in terms of the entropic measure of information theory; it is essentially based on the content of the information to define the conceptual similarity e.g.: Resnik measurement [27].

**The Hybrid Approach**

The hybrid approach [28] is based on a model that combines between arc-based approaches (distances) and the content of information that is considered as a decision factor.

**Approach Based on the Vector Space**

In the field of information retrieval, vector space models are widely adopted. These approaches use a characteristic vector, in a dimensional space, to represent each object and calculate the similarity based on the cosine measure or the Euclidean distance. Among the approaches cited in the literature, we find: **Similarity of Jaccard, Cosine similarity**

Comparing these four approaches we have chosen the arcs-based approach and more specifically **Labriji measurement**, because a comparison was made in [26] between the measurements of "Labriji" and the most used measures "Wu and Palmer" and also "Rada" shows that the measure of "Labriji" improves the values of similarities between the concepts and seen the most realistic and powerful results it produces. The formula of Labriji measuring is given by the following function:

\[
\text{Sim}(C_1, C_2) = \frac{2 \times N}{N_1 + N_2 + 2N + N \times N_1 \times N_2}
\]

**2) Extraction of interest’s center**

The calculation of the quantity given by IP (C0) allows us to deduce the interest center of the user P:

\[
\text{Ip}(C_0) = \text{Max}\{\text{Ip}(c), c \in P\} \quad (2)
\]

Where \( P \) is the set of concepts that constitutes the user profile, Ip(C0) is a finite number that represents the concept C0.

**Application**

We want to apply our method of extraction of the center of interest of the user on the ontology of reference which one mentioned previously. We assume the following user profile:
The table below shows the deduction of neighborhoods and the importance of concepts:

<table>
<thead>
<tr>
<th>Concept</th>
<th>Weight</th>
<th>Neighborhood</th>
<th>Ip/(Concept)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AdministrativeStaff</td>
<td>5</td>
<td>ClericalStaff</td>
<td>2.66</td>
</tr>
<tr>
<td>Employee</td>
<td>3</td>
<td>Person, Faculty</td>
<td>5.68</td>
</tr>
<tr>
<td>Faculty</td>
<td>4</td>
<td>Employee, PostDoc, Professor, Lecture, ClericalStaff</td>
<td>13.54</td>
</tr>
<tr>
<td>Organization</td>
<td>5</td>
<td>Root, Department, University, Program, Institute, College, ResearchGroup</td>
<td>13.32</td>
</tr>
<tr>
<td>Work</td>
<td>4</td>
<td>Student, Research, Course</td>
<td>9.2</td>
</tr>
<tr>
<td>Person</td>
<td>4</td>
<td>Root, Student, graduateStudent, ResearchAssistant, Employee</td>
<td>10</td>
</tr>
<tr>
<td>Professor</td>
<td>4</td>
<td>AssistantProf, Faculty, Chair, VisitingProfessor, FullProfessor, Dean</td>
<td>14.97</td>
</tr>
<tr>
<td>Publication</td>
<td>4</td>
<td>Research, Software, Book, Article, Specification, Manual, unofficialPublication</td>
<td>20.21</td>
</tr>
</tbody>
</table>

So we deduce from the table that the center of interest for this profile is the concept “Publication”.

3) Etallement of a graph of concept
We define the spreading of a graph the measure that characterizes the largest distance between two nodes in a graph otherwise it is the distance that separates the furthest nodes in a graph. So it will also represent the diameter of the circle that encloses this graph. The formula of etallement:

\[
\text{Etallement} = \min \left( \text{Sim}(C_i, C_j) \right) \quad (i,j) \in \{1..N\}
\]

With:

- \( \text{Sim}(C_i, C_j) \): is the similarity between each concept with the others concepts of the graph,
- \( N \): the number of nodes (concepts) of the graph;
- \( C \): a concept of the graph of the profile \( P \);

This measure represents also the conceptual eccentricity of this concept since the eccentricity of a node is its maximum distance to all the other vertices.

4) Radius of a graph
We’re inspiring by the rule of “the similarity is a function such that its value is greater when two points are closer unlike the distance, which is a measure of dissimilarity: the closer the points are, the smaller the distance”, to define the conceptual radius of a graph by the following formula:

\[
\text{RayonG} = \frac{\text{Etallement}}{2} = \frac{\min(\text{Sim}(C_i, C_j))}{2} \quad (i,j) \in \{1..N\} \quad (4)
\]

This measure will allow us to deduce the surface of the circle that englobe this graph in order to calculate its density.

5) The Circle surface of a graph
In this situation the surface of the circle of a graph is calculated by the following formula:

\[
\text{Area_Circle} = 2 \times \Pi \times \text{RayonG} \quad (5)
\]

The surface of the circle of a graph includes all the nodes (concepts) of this graph. So its calculation enables us to deduce its conceptual density.

6) The Conceptual density of a graph
We define the formula of the Conceptual density of this graph by:

\[
\text{Conceptual_Density} = \frac{\text{Nbr}}{\text{Area_Circle}} \quad (6)
\]

With \( \text{Nbr} \): The Number of concepts of this graph.

Application( Etallement, Radius, Conceptual Density )
Let’s us consider the following graph of concepts:
Extraction of l’étalement du graphe:

The table below represents the similarity values between the concepts of this graph:
So from the results of the similarity table we have:

\[ \text{Etallement} = \min(Sim(C_i, C_j)) = 0.15 \]

- Extraction of the radius of the graph:

\[ \text{Etallement} = \min(Sim(C_i, C_j)) = 0.15 \]

- Calculation of the conceptual density of this graph:

First of all let's calculate the area of the circle of this graph:

We have:

\[ \text{Area}_{\text{Circle}} = \pi \times \text{Radius}^2 = 2 \times 3.14 \times 0.471 \times 0.471 = 0.471 \]

And we have the number of nodes of this graph is 8, so we deduce the conceptual density:

\[ \text{Conceptual Density} = \frac{\text{Nor}}{\text{Area}_{\text{Circle}}} = \frac{8}{0.471} = 16.98 \]

IV. CONCLUSION

After having presented the user profile in graph of concepts, we proposed in this article a method of semantic computation of the centers of interest. We plan to define a threshold to choose the most relevant interest centers for the user and to use this method as well as the definitions we have defined in our semantic search meta-search engine. Indeed, our method can be used in all phases of an information retrieval system including the enrichment phase of the query, the search phase or even in the indexing phase.

REFERENCES


