Adaptive Generation of Multilingual Questions and Answers from Web Content

Marta Gatius
Computer Science Department
Technical University of Catalonia
Barcelona, Spain
e-mail: gatius@cs.upc.edu

Abstract—**This article describes the adaptive generation of questions and answers from web content to assist the user when searching for information. The approach proposed integrates user models, domain-restricted knowledge representation, natural language and wrappers techniques. This proposal is based on the representation of the main concepts involved in a specific scenario as a set of attributes, and the user models as the subset of those conceptual attributes relevant for each user. The approach described has been applied to generate multilingual personalized questions and answers from web sites in two different domains: university courses and dissemination on nutrition.**

Keywords-user models; semantic knowledge representation; multilingual questions and answers

I. INTRODUCTION

The growing amount of web information, in different formats and languages, has increased the need of tools and resources that facilitate the user access to this heterogeneous data. There are already many web crawlers and directories designed to locate web sites containing information related to a user's query. However, the presentation of a set of web documents does not always satisfy the users' information goal. The presentation of a precise answer to the user's question, bellow the document level, could be very useful either the user's goal consists of obtaining a specific data, browsing (casually exploring) or getting a general idea about a subject. However, the problem of extracting the specific relevant information from web documents is not yet solved. The conventional information retrieval (IR) and information extraction technologies cannot be applied directly to web sources due to the singular characteristics they present: large volume, data in several formats (text, tables, images and video) and hyperlinked information.

Different approaches have been followed to face these challenges. Several of those approaches are focused on the structure of the web documents, such as the wrappers systems, that use delimiter-based methods to extract specific information in a particular web document. However, although there are several commercial and research wrappers systems, their scope is usually restricted, because the cost of generating a wrapper is still a problem, as described in [1].

Many approaches to face the problem of assisting the user when searching web information are based on the use of semantic knowledge that could facilitate semantic search. Semantic metadata (e.g., semantic tags) have been incorporated into many web sites to help machines understand the semantics of the pages. However, there are not yet enough web documents that contain explicit semantic information of quality. For this reason, the works focused on the use of semantic knowledge to enhance web usability can follow different approaches, such as query reformulation ([1]), information-seeking dialogues ([2]), summarization ([3]) and tag clouds ([4]).

On the other hand, the widely diverse audience of the web has also increased the amount of research on the incorporation of user models (UMs) to enhance web accessibility. As described in [5], UMs have already been incorporated in different types of web applications, such as recommended systems as well as adaptive presentation of web contents.

The work described in this article proposes the generation of personalized questions and answers from web documents to assist the user when searching for information. In order to generate the most relevant questions and answers for a particular user in a specific scenario, technologies from several areas are combined: semantic models, UM, natural language (NL) and wrappers.

This paper studies an organization of knowledge that favors its adaptability to new domains, users and languages. This organization is based on a separate and declarative representation of the different types of knowledge needed in the generation process: semantic, linguistic and UM. The main concepts involved in a specific scenario are represented as a set of attributes. The UM is represented as those conceptual attributes relevant for each user type. The attributes could also be classified according to a general syntactico-semantic taxonomy that associates an attribute with several patterns to express questions and answers about its value.

In a previous work (described in [6]), the combination of UM and domain ontologies for generating personalized questions and answers about a particular university course was studied. Current research is focused on the extension of that previous work to reduce the human intervention needed as well as to adapt the organization proposed to more complex scenarios.

In order to limit human efforts, several semantic models for representing domain concepts as a set of attributes could be used. Although ontologies provide a rich and flexible formalism, for specific scenarios, more simple representations (database models, frames and even a list of attributes) could be also appropriate, because the human effort needed for representing the domain concepts in those models is reduced.

On the other hand, the adaptation of the organization proposed to more complex scenarios involves new challenges, such as the integration of data from several web sites, in different formats. To face this new problem, a system of wrappers has been incorporated. Those wrappers automatically extract web data in different formats and represent them as values of the conceptual attributes

The next section presents an overview of relevant related work. The Section 3 describes how semantic knowledge and UM have been integrated in the approach proposed. The Section 4 describes the generation of personalized questions and answers from the semantic representation of web content. Finally, the last section draws the conclusion.

II. Related Work

As mentioned in the introduction, the large volume of web documents and their organization present new challenges to IR. In order to face these challenges, techniques from several other disciplines have been combined with traditional IR, such as semantic models, UM, NL and wrappers.

Several works propose the use of general semantic knowledge sources, such as WordNet, a general database widely used, that groups synonyms in synsets. The terms of the user's query as well as those indexing the documents can be expanded by those in the synset provided by WordNet. Thus, the term mismatch problem is reduced, that is, not only the documents containing the terms in the query are considered, but also all documents containing any of the synonyms of the terms.

However, the use of a general resource (such as WordNet) also presents a new problem, not all the synonyms of a term provided are correlated in a particular domain. There are different solutions to select only the terms correlated in a particular domain from those synonyms provided, such as the incorporation of an interface to allow the user to choose them manually (proposed in [1]), the use statistic metrics (in [4]) and use of an specific algorithm (e.g. the PageRank algorithm in [7]).

The use of domain-restricted knowledge sources do not present this problem, therefore, they are incorporated in several research works on semantic search (such as [6], [7] and [8]).

The study of user interaction for IR has also focused relevant works, such as [9], that studies several models of information-seeking dialogues. Most of the works on conversational systems are focused on a specific domain, such as the NL interface to knowledge-intensive systems described in [10].

Another active line of research for enhancing web usability is focused on the incorporation of UMs. One of the most simple UM consists on the classification of users in a small number of groups or stereotypes. Models based on stereotypes have been used since the 90's because they are simple and powerful for domains where it is easy to classify users in a small number group, based on their background or interests. Stereotypes have already been incorporated in web content adaptation ([13]), as well as in adaptive web dialogue systems ([10]).

There are already relevant works that combine NL generation and UM techniques for generating personalized dialogues in a particular domain, such as that described in [14], in the medical domain. NL and UM are also used for the adaptive presentation of dynamic hypertext in the information-systems Ilex and PEBA-II (described in [15]), that dynamically generate pages, according to the user profiles, from canned text with various types of annotations and items from a knowledge base.

III. Integration of Domain Knowledge and User Model

A. *The General Organization of Knowledge*

This paper proposes the generation of personalized questions and answers to assist the user when searching web information. This proposal is related to the interactive approach to IR, because it is based on the presentation of data as questions and answers. However, the process of developing a flexible dialogue system is much more complex and expensive than that of generating questions. The presentation of most relevant questions for a user can also be useful and the human effort required is reduced, especially when the different types of knowledge involved are represented in an appropriate form.

The approach proposed to generate personalized questions combines the use of semantic knowledge, UMs and NL techniques. This proposal is based on a separate and declarative representation of the different types of knowledge involved: semantic, linguistic and UM. This organization of knowledge facilitates its adaptation to new domains, users and languages.

Initially, a similar modular organization of knowledge was studied for a simple scenario where the user was looking for specific information on a particular university course, as described in [6]. The Prolog language was used for implementing the knowledge involved: the domain concept Course, partially described in Figure 1, as well as the syntactico-semantic classes and the lexical entries. The unification mechanism of Prolog facilitates the use of general categories augmented with features that represent particular data: concept, attribute, stereotype and language.

The work described in this paper is focused on how to extend this organization to face the challenges more complex scenarios could present, such as the integration of data from several web sites, in different formats, as well as dealing with many different types of users.

COURSE
CODE *Teacher* **description**
NUMBER OF CREDITS *Teacher/Student* **quantity**
TEACHING LANGUAGE *Student* **description**
COORDINATOR TEACHER *Teacher/Student* **agent**

Figure 1. A fragment of the concept course.

The Figure 1 shows a partial description of the main concept in the simple scenario the user searches information on a university course. For this particular scenario, two different user groups were distinguished: teachers and students. As can be seen in the Figure 1, the concept Course is described by a set of attributes. Each attribute describing the concept has been associated with the user group(s) interested on it: teachers, students or both. These attributes have also been classified according to the general syntactico-semantic taxonomy that associates an attribute with several patterns to express questions and answers about its value. From this description, the most frequently asked questions by teachers and students (together with their answers) are generated semi-automatically, and have to be manually supervised.

Although this organization of knowledge facilitates the generation of questions and answers, human intervention is needed in three steps:

1. Definition of the concepts involved in the scenario.
2. Association of each conceptual attribute with the related stereotype.
3. Association of each conceptual attribute with a syntactico-semantic class.

In the scenario the user searches for information about a university course, the human effort needed in these steps is limited. The description of a university course is usually presented in web sites in a very structured form, and its representation as a set of attributes is a simple process. Besides, usually only one web page has to be accessed, because all data about a course are located in one document. Additionally, the general description of the concept course could be reused for most university courses. The web description of those courses usually includes the data provided by the general concept course, even when it could contain additional information. Moreover, only two groups of users have to be distinguished and the domain experts (the teachers), can easily determine which information is relevant for each group.

In a more complex scenario, the human effort required is higher. The number of users groups could increase and related information could be placed in several web sites, with different contents and organization. The work described in this paper is focused on how to face those new challenges complex scenarios could present. In order to study these challenges, a new scenario has been considered: the user accesses the web to get a general idea about nutrition.

In the nutrition domain there are many different web sites with similar content but very different organization. Although there are dissemination web sites developed from public organizations, there are also many different types of sites. For this reason, users could find difficult to understand the main ideas in this domain, and personalized questions could help them to achieve this goal.

The description of the main concepts in the nutrition web sites presents several challenges. On the one hand, dissemination webs on nutrition are examples of web sites where usually information is already organized to facilitate browsing, not querying. Besides, in those web sites, data is usually presented in different formats, that can include tables, images, video and interactive tools.

On the other hand, the separation of users in groups in the nutrition domain also presents several difficulties, because there are many different types of users interested on those pages. Furthermore, several websites contain information related only to a particular type of user (e.g., those web pages about nutrition for children), while others present information related to many different users.

The rest of this section describes the approach followed to face the main challenges when representing the main domain concepts as well as the UM in the nutrition domain are faced.

B. The Representation of the Domain Concepts

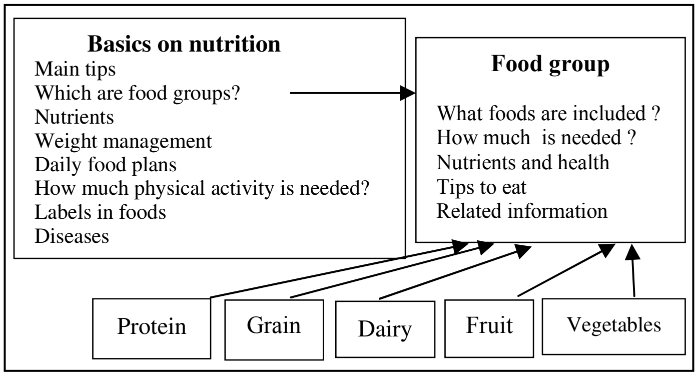

Figure 2. A fragment of the main concepts in the nutrition domain.

The representation of the main concepts in a complex domain, such as the nutrition domain, could imply the integration of information from several web sites, with different content and organization.

The Figure 2 shows a fragment of the representation of the main concepts in the nutrition domain. In this representation, concepts are described by a set of attributes which values can be text, links, tables, graphics and other concepts. For example, the value of the attribute *Main tips* is a list of tips (where each tip is represented by a sentence). The value of the attribute *Nutrients* is a list containing most relevant nutrients (each nutrient represented by a word). The value of the attribute *Weight management* is a link to a web document containing detailed information about controlling the weight. The value of the attribute *How much physical activity is needed?* is a table showing the quantity of physical activity needed by each type of user.

As shown in the Figure 2, the value of the attribute *Which are food groups?* consists of the instances of the concept *Food group*. There are five instances of this concept: *Protein, Grains, Dairy, Fruit and Vegetables*. Each of those instances is described from the set of attributes inherited by the concept *Food group*.

In order to reduce the effort of obtaining web information, a system of wrappers to facilitate the automatic extraction of specific data has also been integrated. Wrappers are especially appropriate for obtaining data that change frequently, such as dates of exams and teachers attending timetable. Wrappers are also appropriate for extracting text from tables and captions from pictures, and they are present in most web sites on nutrition.

Most wrappers use methods based on delimiters (such as HTML tags) to locate data, although linguistic knowledge can also be used. The system of wrappers described in [1], has been incorporated because it provides a specific language to reduce the effort of writing a wrapper. The data to be extracted from a particular web document have to be described in that language, and from this description the system generates automatically the wrapper.

C. User Models

The UM is represented as the set of conceptual attributes relevant for each stereotype. A particular conceptual attribute can be relevant for more than one stereotype. In that case, there are two different possibilities: the attribute may have the same value for all stereotypes and the attribute may have different values, one for each stereotype associated. For example, in the nutrition domain, most attributes describing the basic concepts on nutrition are related to all the stereotypes considered. However, while the value of several of those attributes is the same for all stereotypes, the value of the others is different for each type of user.

The conceptual attributes shown in Figure 2 are related to all types of users. Several of those attributes have the same value for all user types, such as the attributes *Which are food groups?* and *Nutrients*. There are also several attributes which value is specific for each stereotype, such as the attributes *How much physical activity is needed?* and *How much (food group) is needed*?.

In the web sites on nutrition, tables are used to present different quantities (e.g., time of physical activity, calories of a food group) for each type of user (e.g., children, women, and men). From these tables, wrappers can extract easily these data and represent them as values of the corresponding conceptual attribute. Then, the quantity related to each user type is represented separately to facilitate the generation of the personalized answer for each user. For example, the answer to the question *How much fruit is needed ?*, will be different for each type of user.

The process of defining the different groups to be considered and the information relevant for each group is not the same for all the domains.

In the nutrition domain, this information is obtained from the web contents. In most web sites, groups of people are distinguished according to their age and sex: *children, girls, boys, men* and *women*. However, in a particular web site, a more general or a more specific group can be considered, depending on the particular data presented. For example, in the description of the calories needed daily, groups are usually further subdivided considering more precise age intervals (e.g., *Children from 2 to 3*), while in the description of the exercise needed, users can be grouped in more general types (e.g., *adults, adolescents* and *children*). Furthermore, new groups can be distinguished, considering additional information, such as profession (*teachers* and *health care professional*) and/or diseases (such as *allergies* and *diabetes*).

The representation of the user models described could be extended for new scenarios in several forms. A taxonomy of user types could be included, in case the complexity of dealing with all stereotypes become high. Additionally, in several scenarios, it could also be useful to represent the relation of a stereotype and a particular conceptual attribute with a numeric value, instead of a binary value (as in the model described). A numeric value can represent more precisely how much relevant is the attribute for a particular user type.

The static UM described could also be extended to include information about a particular user, obtained dynamically. The combination of stereotypes and information about the user obtained dynamically can be very useful in several domains, such as in the educational domain, where the learning progress.

In current work, the process of determining which information is relevant for each user group in a particular scenario is done manually. However, in the nutrition domain, lexical resources could also be used to perform this selection process semi-automatically, because the key highlighted sentences often include words and expressions that could be easily related to a particular user group (e.g., *Kid-Friendly Veggies and Fruits, Be an Active Family, Men's Health*).

In any scenario, logs on users queries could also be analyzed to determine relevant attributes for each user, if the number of logs is high enough this process could be done (semi)-automatically, as described in [16].

IV. INTEGRATION OF THE LANGUAGE TECHNIQUES

The process of generating the personalized questions and answers for a particular user consists of selecting the conceptual attributes and values related to the user stereotype and presenting them in the user's language.

In order to facilitate the generation of questions and answers from the domain concepts, two different approaches for representing attributes have been considered. For those scenarios where contents in web sites are already organized as questions and key highlighted sentences, those questions and sentences can be extracted from the web documents and represented as conceptual attributes. As can be seen in the Figure 2, the attributes describing the main concepts in the nutrition domain are questions and nominal groups. Those sentences have been directly extracted from web documents.

In case the conceptual attributes are already presented as questions, the process of generating personalized questions consists on selecting the attributes associated with the user type, and presenting them in the corresponding language.

The answer to those questions will be the attribute value associated with the stereotype. Thus, the language processing is limited to translate the attribute and its value, when necessary.

In case the web data are not presented as questions and key sentences, a different approach is followed. The conceptual attributes are represented by words and nominal phrases, that can be extracted from the web document, although not necessary. Then, each attribute has to be classified according to a general syntactico-semantic taxonomy, defined in previous work, following [17]. This taxonomy relates attributes describing concepts to the linguistic structures needed for expressing question and answers about them.

The basic classes of the taxonomy are associated with the grammatical roles: *participants, being, possession, descriptions, relationships* and *related processes*. These subclasses have been further subclassified when additional information related to the linguistic realization can be considered. For example, the class *description* has been subclassified into subclasses representing time, place, manner, cause, quantity, name and type. This classification can be reusable in many languages, although the patterns associated with each class are different for each language.

The Figure 1 shows the syntactico-semantic classes associated with the attributes describing the course: the attributes *code* and *teaching_language* with the basic class *description*, *number_of_credits* with the class *quantity* (subclass of *description*) and *coordinator_ teacher* with the basic class *agent* (subclass of *participant)*.

CONCLUSION

This article describes the adaptive generation of questions and answers from web content, in several scenarios of diverse complexity. In order to facilitate the generation of the most appropriate questions and answers for a user in a particular scenario, this work proposes a general organization of the different types of knowledge involved: semantic, linguistic and UM. This organization consist of representing the main concepts involved in the scenario as a set of attributes, and the UM as the subset of those attributes, relevant for each user type. NL and wrapper techniques have been incorporated to reduce the human intervention required.

Future work could include the evaluation of the questions generated. In particular, questions and answers about several university courses could be incorporated in the web sites describing the courses, to evaluate if they really help students find the information they need.

The approach proposed could be incorporated for providing the user with friendly access to structured data, which is published more and more on the web. Additionally, the proposal described could also be incorporated in web dialogue systems assisting the user.

ACKNOWLEDGMENT

This work has been supported by the Spanish project Skater (TIN2012-38584-C06-01).

REFERENCES

[1] M. Gatius, M. Bertran, and H. Rodríguez, "Multilingual and Multimedia Information Retrieval for Web Documents," Proc. International Workshop, DEXA, 2004.

[2] M. Walker, S. Whittaker, A. Stent, P. Maloor, J. Moore, M. Johnston, G. Vasireddy. "Generation and Evaluation of User Tailored Responses in Multimodal Dialogue", Cognitive Science, 28: 811-840, 2004.

[3] R. Móro and M. Bieliková, "Personalized Text Summarization Based on Important Terms Identification". Proc. International Workshop, DEXA, 2012.M.

[4] M. Rinaldi, "Tag Clouds with Ontologies and Semantics," Proc. International Workshop, DEXA, 2012.P.

[5] P. Brusilovsky and E. Millán,"User Models for Adaptive Hypermedia and Adaptive Educational Systems," The Adaptive Web, LNCS 4321, 2007, pp. 3–53.

[6] M. Gatius, "User Models and Domain Ontologies for Generating Personalized Questions and Answers," Proc. CENTRIC 2014.

[7] B. Sajgalk and M. Bielikovaıa, "From ambiguous words to key-concept extraction," Proc. International Workshop, DEXA, 2013.

[8] W. Guo and S. Kraines,"Integrating Knowledge of City Entities Extracted from DBpedia and GeoLite into the EKOSS Failure Cases Repository to Enhance Semantic Search Capabilities," IJCISIM,Vol.3, 2011.

[9] S. Dey and S. Abraham. "User Interface For A Search Engine: A Customized and Multi-domain Approach," IJCISIM, Vol. 4, 2012.

[10] M. Gatius and M. González, "The use of Domain Ontologies for Improving the Adaptability and Collaborative Ability of a Web Dialogue System," IJCISIM, Vol 3, 2011.

[11] N. J. Belkin, C. Cool, A. Stein, and U. Thiel, "Cases, Scripts, and Information-Seeking Strategies: On the Design of interactive Information Retrieval Systems", Expert Systems. Appl., vol. 9, 1995.

[12] M. L'Abbate, I. Frommholz, U. Thiel, and E. Neuhold, "Using Case Based Retrieval Techniques for Handling anomalous Situations in Advisory Dialogues", Proc. DEXA 2004, pp. 539–548.

[13] J. Richards and V. Hanson, "Web accessibility: A broader view," In Proceedings of World Wide Web Conference, 2004, pp. 72–79.

[14] S. Williams, P. Piwek, and R. Power, "Generating monologue and dialogue to present personalised medical information to patients," Proc. of the European Workshop on Natural Language Generation,2007.

[15] M. Milosavljevic and J. Oberlander, "Dynamic hypertext catalogues: helping users to help themselves," In Proc. HYPERTEXT, ACM, 1998.

[16] L.Limam, D. Coquil, H. Kosch, and L. Brunie, "Extracting user interests from search query logs: A clustering approach," Proc. International Workshop, DEXA, 2011.J.A.

[17] J.Bateman, B. Magnini and F. Rinaldi, "The Generalized {Italian, German, English} Upper Model," Proc. Of the ECAI Conference. Workshop on Implemented Ontologies, 1994.