

# On the Use of Ontologies for an Optimal Representation and Exploration of the Web

Nicolas Guelfi, Cédric Pruski  
Laboratory for Advanced Software Systems  
Faculty of Sciences, Technology and Communication  
University of Luxembourg  
6, rue Coudenhove-Kalergi  
Luxembourg-Kirchberg, Luxembourg  
{nicolas.guelfi, cedric.pruski}@uni.lu



Journal of Digital  
Information Management

**ABSTRACT:** *The use and definition of ontology for the representation and the exploration of knowledge are critical issues for approaches dealing with information retrieval. In this paper, we propose a new ontology-based approach for improving the quality, in terms of relevance, of the results obtained when searching documents on the Internet. This is done by a coherent integration of ontologies, Web data and query languages. We propose new data structures built upon ontologies: the WPGraph and the W<sup>3</sup>Graph which allow Web data to be modelled. We also discuss the use of ontologies for an efficient exploration of the knowledge contained in our conceptual structures using ASK, a specific query language introduced in this paper. An experimental validation of our approach is proposed through a prototype supporting our innovative framework.*

## Categories and Subject Descriptors

D.3.3 [Language Constructs and Features]; H.2.3 [Data Description Languages]

## General Terms

Ontology, Web graph, Web data

**Keywords:** Web search, Query Language, Ontology, Conceptual Structures, Graphs

Received 20 Feb. 2006; Revised 7 June 2006; Accepted 16 June 2006

## 1. Introduction

Since the introduction of the World Wide Web (WWW) by Tim Berners-Lee in the early Nineties, and more recently the definition of the Semantic Web [1] and its promising results, the interest for concepts and tools that can improve representation and retrieval of Web information is increasing. With the popularity of the WWW and an ever increasing number of Web pages, the need for tools and concepts that allow the retrieval of Web data in a powerful and relevant way is of utmost importance. Only Web search engines like Google<sup>1</sup> or Yahoo<sup>2</sup> allow users to search the Web and find these pages. As search engines index a huge amount of data, it becomes obvious that during a search many unwanted pages slip-in among and pollute the most relevant results proposed by the engines. This point can be easily highlighted, by entering a common query in Google. For instance, if a user is looking to gain information about books and journal on the subject of graph theory trees, the query “publications on trees” delivers results from totally different areas of interest. Among the results one can find documents related to botanic, but also information about genealogy or even graph theory and this force the user to spend too much time skimming the results before reaching the relevant information. This can be avoided if the initial query is built properly.

The formula for a good query, a query that will give results satisfying the user, requires a rigorous selection of keywords [11]. The role of keywords in existing Web search engines is to characterize the context of a query, called domain. The domain is simply the area of interest. Ongoing research<sup>3</sup> in the design of Web search engines attempts to integrate a users' behaviour by analyzing the already visited pages in order to establish their research domain. There is

need for a new search method that will effectively describe the search domain. We believe that the use of tools, such as ontologies [5], could partly solve this problem. Ontologies would be used to model the domain of research that the user has in mind, which in turn will facilitate the elaboration of future queries. Ultimately, search engines would give optimal results by integrating the vocabulary contained in the ontology to filter Web pages before displaying them to end-users. Such an approach would reduce drastically the number of keywords to enter.

Users must have a good knowledge of what information exists on the Web and how it is stored. For instance, assume that you are looking for the technical specifications of the Samsung TV with the following reference WS32Z308. The query “+Samsung +WS32Z308” would prevalently be entered in Google. However, among the most relevant results, the engine returns only web sites that allow the purchase of the requested product. This is mainly due to the use of PageRank algorithms [4] that are part of search engines and which use hyperlinks to determine the relevance of a page. Thus, in order to obtain the right page, the user must transform the initial query as follows, “+Samsung +WS32Z308 +group”. In this case, the appropriate pages for the technical specifications of the Samsung TV, model WS32Z308, is returned. Presently, users have to know specific keywords that figure on the wanted web page in order to build a query. A well-adapted organization of the Web that takes into account content of Web pages would make it possible on one hand to reach the desired web pages quickly, and on the other it would facilitate the construction of good queries. As a result, the global speed of Web document retrieval will increase.

Previous queries also illustrate, via the appearance of the ‘+’ symbol, the role played by the query language. Although the user interface aspect is important, as it must provide a user-friendly way to build queries, the most important aspect is the foundation of the language. Most of the existing languages devoted to the Web [40, 41, 9, 10] rely on formal, well-defined logic based semantics, from first-order logic to description logics. This has the advantage to make the query unambiguous and therefore verifiable on structures representing the Web, thus setting the scene for effective and powerful search engine design. However, with the intention of being usable, a query language must be defined in such a way that it offers the features that users are expecting and returns the desired results.

In this paper, we present an original approach that aims at improving the relevance of the returned results when searching the Web. The proposed approach is based first on the definition of new conceptual structures: the WPGraph and the W<sup>3</sup>Graph. Inspired from John Sowa's conceptual graphs [3], these graphs are built according to a prescribed ontology and allow the data of the Web to be represented in a more intelligent and intuitive way. Second, we define ASK, a new query language that is designed to extract the wanted knowledge from our conceptual structures.

The remainder of this paper is organized as follows. Section 2 presents state-of-the-art Web search and ontology, as well as the approach we advocate. We introduce in Section 3 the WPGraph and the W<sup>3</sup>Graph: our new conceptual structures. Section 4 is devoted to the presentation of the ASK query language. Section 5 presents the prototype of our framework. Finally, the last section wraps up with our concluding remarks and future work.

## 2. Web Exploration and Ontology: state-of-the-art

Conceptually, ontology and Web search are different. However, since the emergence of the Semantic Web and the need for adapted knowledge management structures able to treat huge amount of

<sup>1</sup> <http://www.google.com>

<sup>2</sup> <http://www.yahoo.com>

<sup>3</sup> <http://www.google.com/psearch>

data, ontologies are being used more often in the development of Web search techniques. To illustrate this convergence, we need to study the existing work in the field of Web search and ontology.

## 2.1 Web Exploration

Although it's a crucial point for Net surfers, techniques for Web document retrieval are not numerous. Without knowing the URL of the desired Web page, the user has no option but to use a Web search engine in order to find the desired information. The problem of document retrieval on the Internet is summarized in Figure 1. Berners-Lee's vision of the World Wide Web [14] inherently requires research work in the area of Web search to become increasingly user-oriented. Hence, the first important component of the architecture presented in Figure 1 is the user interface. It is an important component since it should facilitate both query construction and result readings. Interfaces allowing the construction of queries with natural language seem a promising field of research [15, 19, 13]. While recent work in the domain of user interface seems to reveal a new tendency that uses multi-media for displaying search results on the screen. In this context, Dziadosz and Chandrasekar [12] suggest an approach using thumbnail previews for helping users select Web pages.

The second relevant aspect appearing on Figure 1 deals with queries: their elaboration, interpretation by Web search applications and their refinement to get more relevant results. Several approaches for query elaboration are based on implicit query expansion techniques [20, 22, 47, 6]. Added terms are useful to characterize the domain of the query. Other approaches use machine learning techniques [21] or agent [23] to learn user interests and to apprehend their behaviour as well as possible in order to improve retrieval of information on the Web by making the most of the harvested data. The interpretation of the query, which differs from one engine to another [17, 18], is very important for Net surfers. They must have a well-established idea how the engine will interpret queries in order to choose the right words to elaborate queries. To touch a broad range of Web documents, it is also possible to use the technique recommended by the Semantic Web [1] community. This consists of enriching not the queries but the Web pages themselves in meta-data, using an adapted language [28, 29], thus offering broader provisions for the verification of queries by search engines. This technique strives to be more precise in the returned results since it has the ability to answer precise questions like "Who is the actual president of the USA?" Such queries would give very random results with a common Google-like query. Query refinement is the third important step dealing with query. This iterative process consists in expanding or reducing the initial query by adding or deleting keywords to specify the domain of the query, and thus obtains more accurate results. However, since users are in charge of the refinement process for most search engines, the new expanded query could lead to totally irrelevant results depending on users' experience. As a result, new approaches

strive to provide better support to users and existing approaches aim at automating the refinement process by introducing statistics [2] or by analyzing query logs [49], or even by using tools such as thesaurus or ontology [48, 50].

The last point highlighted by Figure 1 is the problem raised by the structure of the Web. In most of the approaches [4, 25, 16], the structure of the Web is build upon hyperlinks pointing from a page to another. But, Web structure can also be defined from other characteristics like its content [26, 27], which allows both surfing and searching the Web to be done in a more intuitive way.

## 2.2 Ontology

The term *ontology* has its origin in philosophy and denotes the science of "what is." Very early on, ontology became a central notion in sciences and presently it is becoming more in vogue in various fields. In zoology, botany or chemistry, scientists use ontology like taxonomy to describe a given field of knowledge. Since the beginning of the 80s, John McCarthy [30] gave the word ontology another dimension by introducing the notion of "how it exists." With this synthesis, the term ontology has become a must in information science fields and particularly in artificial intelligence and knowledge management. Several formal definitions for ontology were proposed until Tom Gruber [5] stated that "An ontology is an explicit specification of a conceptualization." Gruber's definition became the *de facto* standard and is used in various fields. It is in this context that ontology or ontological methods have been implemented in the field of database development [31, 32] and also in Software Engineering for the writing of software [33] (e.g. through class models using the Object-Oriented paradigm). Additionally, ontology-based methods have been applied to solve many problems related to information retrieval in contexts ranging from medicine to the Internet with its Semantic Web paradigm [1]. The latter aims at using ontology as a tool for taming the immense diversity of sources from which Internet content is derived, and here, even a small dose of ontological regimentation may provide significant benefits to both producers and consumers of online information. Other domains are also affected by the use of ontology. Among them, one finds linguistics, where ontologies are used for text parsing and words disambiguation [7, 34]. For now, research efforts strive toward ontology use to support business areas like banking [36]. The notion of ontology seems to be promising, but specialists agree that in IT systems ontologies suffer from a lack of formalization and are often not suitable to the domain they are supposed to represent.

### 2.2.1 Formalization of Ontologies

Tremendous obstacles stand in the way of ontology building and integration. These obstacles are analogous to, let's say, the task of establishing a common metaontology of world history, which would require a neutral and common framework for all descriptions of

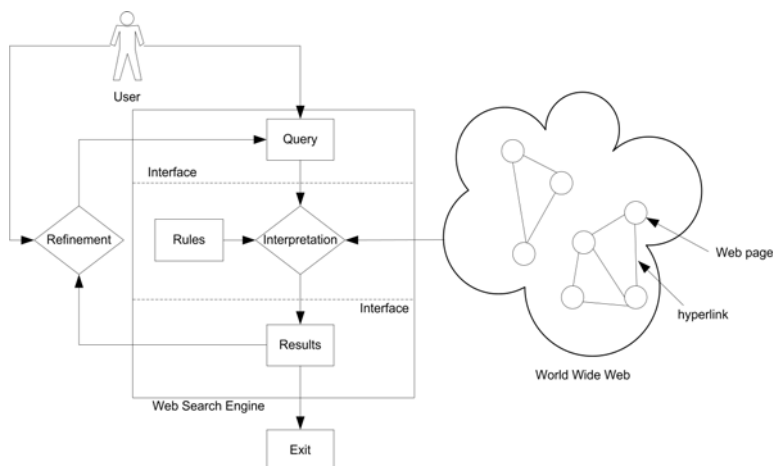


Figure 1. Web Search Process

historical facts, which would, in turn, require that all events, legal and political systems, rights, beliefs, powers, and so forth, be comprehended within a single perspicuous list of categories. In addition to the problem of extension are the difficulties that arise at the level of adoption. To be widely accepted, ontology must be neutral between different data communities. As experience has shown, a formidable trade-off is required between this constraint of neutrality and the necessity that ontology be maximally wide-ranging and expressively powerful. One way to address this problem is to consider, on the one hand, the formal aspect of ontology: the ontology of part and whole, of identity and difference, of dependence and independence. On the other hand particular domain specific or regional ontology needs to be considered, for example, ontologies of geography, or medicine, or ecology. The relation between formal and domain-specific ontologies is in some respects analogous to that between pure and applied mathematics. Just as many sciences use mathematics, domain-specific ontologies should ideally share, at their foundation, the same robust and widely accepted top-level ontology.

Despite these problems, several formalisms have been developed to formalize ontologies. All the existing formalisms are based on first-order logic and are motivated by the same goal in developing an expressive, flexible, computer (and human)-readable medium for exchanging knowledge bases. Knowledge Interchange Format (KIF) [39] has been the first formal language developed in the context of knowledge management. The language has three essential features:

- Standard set-theoretical semantics
- Logical comprehensiveness, which means that it has all the expressive power of the first-order predicate calculus
- Ability to support the representation of representations, or representation of knowledge about knowledge

On the basis of KIF, Gruber has developed a more serviceable language for ontology representation known as Ontolingua [38], designed to serve as a *lingua franca* for those involved in building ontologies. Ontolingua is built on the basis of KIF 3.0, but has a very distinctive purpose. Where KIF is conceived as an interface between knowledge representation systems, Ontolingua is intended as an interface between ontologies. It provides an environment and a set of software tools designed to enable heterogeneous ontologies to be brought together on a common platform via translation into a single language. Finally, the last family of language used to formalize ontologies is the one utilizing description logic [37]. DAML+OIL [29], which is the most popular used language and has the goal of exploiting the power and flexibility of XML as a framework for the construction of specialist ontologies (XML is the universal format for structured documents and data on the World Wide Web) and part of a standardization effort in the attempt to create the Ontology Web Language (OWL) [29].

### 2.2.2 Suitability of Ontologies

There are many problems with the definition of ontology as a specification of a conceptualization. There are different kinds of specification (in English, or in KIF, or in first-order predicate logic) all

of which might describe what we can intuitively recognize as the same ontology. We can simply speak about good and bad conceptualizations. The former reflecting what actually exists in reality, the latter resting on ontological error; the former illustrated by a conceptualization of types of slime mold, the latter by a conceptualization of types of evil spirits. Conceptualizations are set-theoretic objects. They are built out of two sorts of components: a universe of discourse (objects 'hypothesized to exist in the real world') and a set of properties, relations and functions. Good conceptualization can be defined as, conceptualizations whose universe of discourse consists only of existing objects (we will need to make similar restrictions on the associated properties, functions and relations in a more careful treatment). Bad ones are all conceptualizations that do not satisfy this condition. If, then, there are not only good but also bad (objectless) conceptualizations, it follows that only certain ontologies, as specifications of conceptualizations, can be true of some corresponding domain of reality, while others are such that there is simply no corresponding domain of reality for them to be true of. But recall, that information systems ontology is a pragmatic enterprise. It starts with conceptualizations and from there goes to the description of corresponding domains of objects, which are nothing more than models describing a point of view. What is most important is that all of the mentioned examples are treated (or recognized) equally by the ontological engineer. In a typical case, customers will specify the universe of discourse and, for the purpose of the ontological engineer, the client is always right. This is why the ontological engineer aims for adequacy for the client-defined domain. The main focus is on reusability of application domain knowledge in such a way as to accelerate the development of appropriate software systems in each new application context. The goal as we have seen is not truth-relative to an independently existing domain of reality, but merely, at best, truth-relative to a conceptualization.

### 2.3 The O<sup>3</sup> Approach

According to the existing work presented above, the context of a query and Web structure play an important part in the information retrieval process. We present below a conceptual summary of the O<sup>3</sup> approach that is discussed in detail within the remaining sections of this paper along with some notes on related work. The main idea of the process for document retrieval on the Web, Figure 2, rises from our observations and experiments concerning the handling of tools for document retrieval. The use of search engines is very often ineffective, as users do not correctly apprehend the structure of the Web and ways of building good queries. This lack naturally appears as a result of the structure and the behavior of search engines and leads to results really far from those awaited by the user, obliging the user to rework the request. Therefore, our approach is based on the use of ontology for restructuring the Web while using the content of Web pages (more intuitive for users) and hyperlinks, and the enrichment of queries so as to directly target the domain of the desired results. This is done through the development of new conceptual structures, the W<sup>3</sup>Graph and the WPGraph, that aim at representing the Web and its Web pages respectively. In addition, we propose ASK, a new query language for the extraction of the desired knowledge from the proposed graphs.

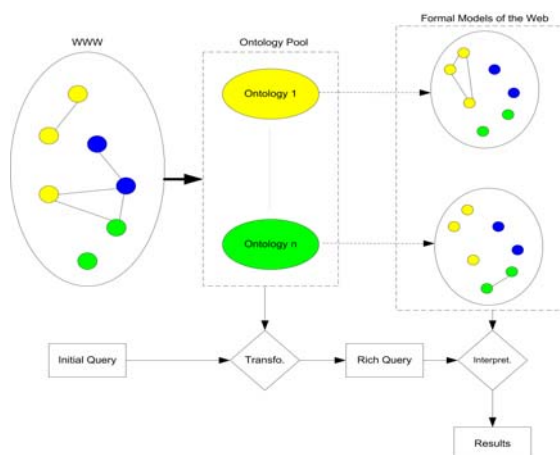


Figure 2. The O<sup>3</sup> Approach

### 3. Ontological Structure for Representing Web Data

Ontology describes in a generic way the knowledge specific to a given field by offering a consensual comprehension. Ontology can be seen as a graph where nodes denote concepts and arcs represent the existed relations between concepts. Thus we define an ontology as follows:

**Definition 1:** Let  $\Sigma$  be a finite alphabet. We define an ontology  $O$  as a triple  $(C, R, E)$  where:

1.  $C = \{c | c \in \Sigma^+\}$  is the set of concepts<sup>4</sup>
2.  $R = \{\perp, \text{Synonymy}, \text{Antonymy}, \text{Hyperonymy}, \text{Hyponymy}, \text{Meronymy}, \text{Holonymy}\}$  is a set of relations
3.  $E = \{(c1, c2, r) \mid c1, c2 \in C, r \in R\}$  is the set of arcs

Concepts are defined as being words of the natural language, whereas relations between concepts are those proposed by Fellbaum [7] in the definition of WordNet and which have been formalized in [42]. But, although being formalized, relations used to build ontologies are themselves problematic because of their too general definition and their rather imprecise semantics. In fact, to be even more precise in the description of a knowledge domain, we must consider several degrees in a relation. For example, we must make the distinction between the relation of type member/collection (tree/forest) or material/object (steel/car), although these kinds of relations are all meronymies. In the context of Web data modeling, ontology, though being extremely subjective and very context dependent, is a very powerful tool that will help to represent, at various levels, the data of the Web as well as the Web itself. In the following section we will illustrate how ontologies can be used to represent, in an intelligent way, the content of the Web by using both the data and the structure of a Web page.

In our approach, ontologies are currently built manually as it is done in most of approaches that use ontologies for information retrieval [47, 51]. However, ontology construction and maintenance should be made semi-automatically. We would support a construction of the ontology using pre-defined ontology patterns and composition operators over ontologies (union, intersection ...). Nevertheless, the basic ontological pattern may probably be built manually by domain experts. Our approach for the construction of the basic ontologies would be, in the spirit of supervised machine learning, to start with informal requirements, design and construct a first version of the ontology, evaluate the ontology on a set of queries sample against the pre-defined expected results. Then the analysis of the results would infer ontology modification thus reaching incrementally a fix point.

#### 3.1 WPGraph for Representing Data of Web Pages

We introduce here the WPGraph, a new conceptual structure for the representation of the content of a Web page. This structure is directly inspired from Sowa's conceptual graphs [3], with the only difference that the relations binding the concepts do not appear directly on the graph but are integrated in a metric allowing to measure the distance, in a semantic way, between the concepts.

Unlike Semantic Web approaches [41, 40] that use only part of the data and especially meta-data, added to HTML code, we use the whole data contained in a Web page to construct its associated WPGraph. We use the Web page data, including meta-data, as well as structural information (type, size and color of the fonts ...), and the kind of objects contained in HTML documents (audio, video, image...).

**Definition 2** Let  $\Sigma$  be a finite alphabet. We define GD the WPGraph associated with a Web page as a tuple  $(V, E, T, \varphi, \rho, \rho_e)$  where:

1.  $V = \{x | x \in \Sigma^+\}$  is the set of vertices
2.  $E = \{\{u, v\} \mid u, v \in V\}$  is the set of edges
3.  $T = \{\perp, \text{img}, \text{vid}, \text{snd}, \text{fil}\}$  is the set of document types
4.  $\varphi: V \rightarrow \mathbb{R}^+$  is a vertices labeling function
5.  $\rho: V \rightarrow \mathbb{R}^+$  is a weight function for the vertices

<sup>4</sup> In our case  $\Sigma$  denotes natural language words whereas  $\Sigma^*$  denotes all the words built upon  $\Sigma$

6.  $\rho_e: E \rightarrow \mathbb{R}^+$  is a weight function for the edges

Thus, set  $V$  represent the important concepts that appear on the Web page. Based on Jansen's study [44] we have decided to take only nouns into account and we plan to integrate also adjectives. Set  $E$  contains the edges that represent potential relations between the concepts according to a given ontology. Function  $\varphi$  gives the type of the concept contained in a vertex as well as a description. Our study [35] shows that we could consider only the five types (file, video, image, audio and  $\perp$ )<sup>5</sup> figuring in set  $T$ . The two other functions  $\rho_e$  and  $\rho_v$  measure respectively the importance of a concept in a given Web page and the distance between two concepts according to a given ontology.

The construction of such a graph is done through a syntactic and semantic analysis of the content of a Web page. The first investigation will lead to the construction of sets  $V$  and  $E$  and functions  $\rho_e$  and  $\rho_v$ . The function is computed first by taking text statistics into account, to know, the frequency of a word, the distance separating two occurrences of the same words, to this end we need to introduce formally the definition of the frequency of a word. This is done through definition 3.

**Definition 3** We define the frequency of a word  $m$  in a text  $T_x$  as being the ratio of all occurrence of  $m$  in  $T_x$ .

$$Fq(m, T_x) = \begin{cases} \frac{n}{|T_x|} & \text{if } n \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Where  $n$  denotes the occurrence of  $m$  in  $T_x$  and  $|T_x|$  is the number of words in  $T_x$ .

The second investigation is made via the analysis of typographic features offered by the HTML language. This is done based on the specifications of the HTML language<sup>6</sup> and, more precisely, on the study of HTML tags to determine what are the interesting objects and terms figuring on the Web page. Therefore we propose (see definition 4) to take the font type, size, color, and style into account in the definition of the  $\rho_v$  function.

**Definition 4** Let  $m$  and  $n$  be two words of a text  $T_x$ .

1. Let FT the function that permits to detect if there is a change in the type font

$$FT: T_x \times T_x \rightarrow \{0, 1\}$$

$$(m, n) \rightarrow \begin{cases} 1 & \text{if } \text{Font}(m) = \text{Font}(n) \\ 0 & \text{otherwise} \end{cases}$$

Where  $\text{Font}$  return the type of the font.

2. FS measure the importance of the font size of a word. Assume that  $\text{Size}$  return the size (in point) of a word in a given text

$$FS: T_x \rightarrow [0, 6] \cap \mathbb{N}$$

$$m \rightarrow \begin{cases} 0, & \text{if } \text{Size}(m) < 8 \\ 1, & \text{if } 8 \leq \text{Size}(m) < 10 \\ 2, & \text{if } 10 \leq \text{Size}(m) < 12 \\ 3, & \text{if } 12 \leq \text{Size}(m) < 14 \\ 4, & \text{if } 14 \leq \text{Size}(m) < 16 \\ 5, & \text{if } 16 \leq \text{Size}(m) < 18 \\ 6, & \text{if } \text{Size}(m) \geq 18 \end{cases}$$

3. FC detect if two words  $m$  and  $n$  have the same font colour. Assume that  $\text{Colour}$  return the font colour of a word

$$FC: T_x \times T_x \rightarrow \{0, 1\}$$

$$(m, n) \rightarrow \begin{cases} 1 & \text{if } \text{Colour}(m) \neq \text{Colour}(n) \\ 0 & \text{otherwise} \end{cases}$$

<sup>5</sup>  $\perp$  means a string of characters

<sup>6</sup> <http://www.w3.org/TR/html401>

$$FE : Tx \rightarrow \{0,1,2,3\}$$

$$m \rightarrow \text{Underline}(m) + \text{Bold}(m) + \text{Italic}(m)$$

Consequently, definition 3 and 4 lead to the definition of  $\rho$ .

**Definition 5** Let  $\rho$  be the weight function for WPGraph vertices,  $Tx$  a text and  $m$  a word of  $Tx$ .

$$\rho : V \rightarrow IR +$$

$$m \rightarrow Fq(m,Tx)+FT(m,Succ(m,Tx))+FS(m)+FC(m,Succ(m,Tx))+FE(m)$$

Assume that  $Succ(m,Tx)$  returns the word next to  $m$  in  $Tx$ .

The semantic analysis of a Web page will complete the construction of the WPGraph by implementing ontologies. These will be used for two different things. The first one concerns the construction of the  $\rho$  function. This function allows to measure the distance (in a semantic way) of two concepts of a WPGraph. It is build according Hirst and St-Onge metric [43] by taking notably the maximum length of the shortest path between two concepts in a given ontology and the homogeneity of the relations binding the concepts of the path into account. The second concerns the use of ontologies to disambiguate words sense. It is widely accepted that words have sense only in a particular context this is why knowledge contained in an ontology combined to the text figuring on a Web page will allow the system to make the difference between for example a board in the sense of a *board of director* and a board in the sense a *wood board*.

**Example 1:** Let's have a small example to illustrate both the construction of a WPGraph and the impact of the ontology on it. Consider the Web page Figure 3 and the two ontologies figure 4.



Figure 3. A Web Page

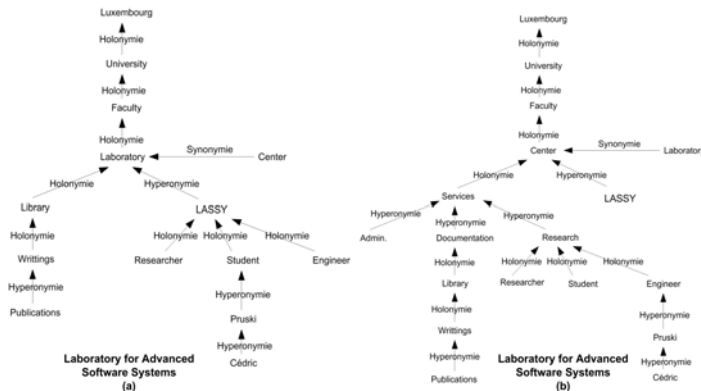


Figure 4. Ontologies

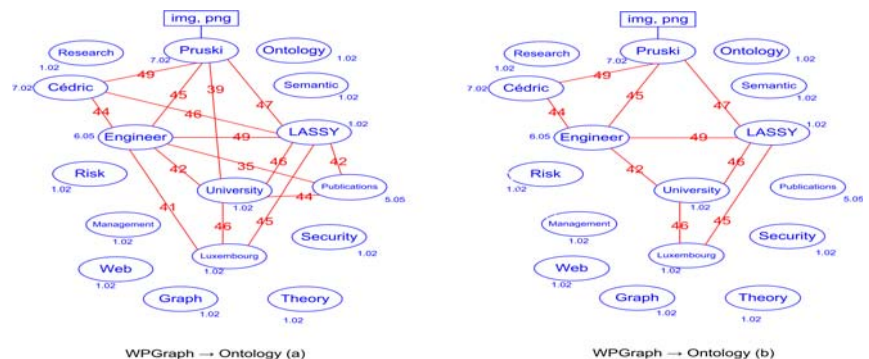


Figure 5. WPGraphs

Observe that the two ontologies describe the same domain (our laboratory the LASSY), but with more or less precision. Figure 5 shows the corresponding WPGraphs.

First observation concerns the obtained WPGraph. As illustrated by our example the WPGraph is strongly ontology dependent. Actually, the set of edges which is built according Hirst-StOnge metric contains more elements if the analyzed Web page is close to the used ontology of the domain. This means that the more connex the WPGraph is, the more interesting the Web page it is supposed to represent can be according to the knowledge of the domain users have.

Second, coefficients appearing on graphs are obtained according to the two weight functions defined above. For instance, the term 'publications' appear twice on the Web page, moreover, one occurrence is in bold, coloured font therefore  $\rho$  (publications) =

$$\frac{2}{40} + 0 + 4 + 1 + 1 = 6.05$$

Concerning  $\rho$  which is directly linked to the ontology, the coefficient is obtained according to Hirst-St-Onge metric. Observe that  $\rho$  (publications, LASSY) = 42 in WPGraph (a) because there is a path whose length is greater than 5 in ontology (a) which is not the case in ontology (b) and therefore  $\rho$  (publications, LASSY) = 0 (i.e. there is no edge between 'publication' and 'LASSY') on WPGraph (b).

### 3.2 The $W^3$ Graph for Modeling the Web

The WPGraph, presented in the previous section, is a first step toward the building of the  $W^3$  Graph aiming at model the whole Web (or a subset of it). However, unlike classical approaches using graphs to represent the Web [25, 16], we define a graph whose edges do not represent a hyperlink between two pages, but are built on the basis of potential semantic relation between the content of pages. Moreover, the nodes of such a graph consist of both a WPGraph associated with a Web page and the URL of that Web page.

**Definition 6** Let  $W$  be a set of Web pages, we define  $WS$  as being the set of WPGraph. Each element of  $WS$  correspond to a unique element of  $W$ .

**Definition 7** Let  $\Sigma$  be a finite alphabet and  $W$  a set of web pages, we define  $GW$  a  $W^3$  Graph as a triple  $(S, A, \rho_{GW})$  where:

1.  $S = \{(ws,url) | ws \in WS, url \in \Sigma^*\}$  is a set of vertices of the graph (url stand for the address of the ws web page)
2.  $A = \{(x,y) | x,y \in S\}$  is a set of edges
3.  $\rho_{GW} : WS \times WS \rightarrow \mathbb{R}^+$  is a weight function for the edge.

The  $\rho_{GW}$  function is also based on the Hirst-StOnge metric [43]. The idea is to measure the distance between two Web pages by comparing their content. Thus we use this metric to compare the main concept of two WPGraphs, the two second and so on and finally we make the mean of all the distances we obtained. This is what we usually do in order to compare the elements of two sets. Thus we first need a function that has the ability to arrange the concepts according their importance in the Web page:

**Definition 8** Let  $WP = (V, R, T, \Phi, \rho, \rho)$  be a WPGraph, we define:

$$ordwp: 1..|V| \rightarrow V$$

$$\forall (i,j) \in (1..|V|)^2: (i < j \Rightarrow \rho(ordwp(i)) > \rho(ordwp(j)))$$

Where  $|V|$  denotes the cardinality of  $V$ . Then we can define the  $\rho_{GW}$  as follows:

**Definition 9** Let  $G = (VG, EG, T, \Phi_G, \rho_G, \rho_G)$  and  $H = (VH, EH, T, \Phi_H, \rho_H, \rho_H)$  be two WPGraphs. We define:

$$\rho_{GW}(G,H) = \sum_{i=1}^k \frac{\rho(\rho_G(ordG(i)), \rho_H(ordH(i)))}{k}$$

with  $k = \min(|VG|, |VH|)$

In this third section, we have proposed new conceptual structures based on ontology to represent the Web, but now we need to define powerful tools to benefit the best from these structures. In the following section we introduce ASK, a new query language, tailored to our graphs.

#### 4. Ontology-based Query Language for Exploring the Web

##### 4.1 Query Languages for the Web

Query languages for the Web have attracted a lot of attention recently, one can find W3QL [9] that focuses on extensibility, WebSQL [10] that provides a formal semantics and introduce a notion of locality. More generally, SQL-like languages [46] give users the illusion that the Web is structured as a database and provide him the way to query it. On another side, there are languages based on first-order logic. These languages are mostly those proposed by Web search engine and offer a set of possibilities to build queries quickly. Among these possibilities we find Boolean operators (for example AND represented by + by Google), but also specific keywords allowing user to target a specific kind of document. A more detailed survey of query languages can be found in [35].

##### 4.2 ASK: A Language for Querying WPGraph and W<sup>3</sup>Graph

We propose a query language allowing to extract the interesting part of the W<sup>3</sup>Graph. Our language is based on the study of the Web search engines' query languages [4], on linguistic analysis [44] and on Semantic Web languages [45] study. Based on the survey carried out by Jansen [44], we have decided to take only nouns into account in the development of ASK. The grammar of the ASK language is as follow:

$$Gr = (XGr, VGr, expr\_list, RGr)$$

Where  $XGr$  denotes the terminal symbols,  $VGr$  the non-terminal symbols,  $expr\_list$  the initial symbol and  $RGr$  the production rules.

$$XGr = \{(), [], {}, !, |, \&, \#, img, snd, fil, empty, vid\}$$

$$VGr = \{expr\_list, expr\_part, expr, type, X, R\}$$

$$RGr = \{expr\_list = expr\_list expr\_part$$

$$| expr\_part$$

$$expr\_part = expr$$

$$expr = X$$

$$| expr[R]$$

$$| expr \{type\}$$

$$| (expr)$$

$$| ! expr$$

$$| expr \& expr$$

$$| expr \# expr$$

$$| expr expr$$

$$Type = img$$

$$| snd$$

$$| empty$$

$$| vid$$

$$| fil$$

$$X = 'a'..'z'$$

$$| 'A'..'Z'$$

$$R = '0'..'9'$$

$$| '0'..'9'.'0'..'9'\}$$

As every Web search engines' query language, ASK has the ability to support Boolean operators (! for NOT, # for XOR, & for AND, | for OR), it also takes the type of the desired document into account as well as the importance of the concept in the Web pages (with respect to set T and function  $\rho$  presented in section 3.1). Here are some basic sample queries built with ASK.

1. `dessert#cheese` : Web pages containing the word cheese or the word dessert but not both
2. `(tree | plant)[2.1][img]` : Web pages containing images of a plant or a tree with  $\rho(tree) \geq 2,1$  and  $\rho(plant) \geq 2,1$
3. `(flower | plant)&!tree` : Web pages containing the flower or plant nouns but not tree

Our language, though basic, makes it possible to take advantages of the WPGraph and W<sup>3</sup>Graph, because it has been conceived in order to take all the characteristics of these structures into account. Moreover, ASK has been designed to facilitate the use of the ontology defined in section 3 for queries expansion.

##### 4.3 Using Ontologies for Enriching ASK Query

Ontology-based query expansion is not a new technique. Navigli [47] has already proposed an approach using such methods which consist in using the concepts and the relations of a given ontology and adding them in a query. This must be done in a very rigorous way, because it can have disastrous effect on the result and also on the interpretation of a query. In the framework of data mining, ontologies are represented using the formalism given in section 3. Such formalism is useful to facilitate the management of the concepts as objects, to establish a hierarchy, to compare their properties with respect to the relations that are binding them and to make the navigation amongst ontology elements easier. Such an approach can improve the relevance of results when searching the Web. The ontology has the properties to sharpen a system based on a traditional indexation process by increasing the chances to formulate a query from terms or descriptors that represent the need of information as well as possible. In other words, the ontology contains constraints to put in the query that has to be validated on a given class of models. This process has several advantages:

- Reduce the silence as regards the whole set of restored documents while resting on words not explicitly figuring in the query. To this end, queries are expanded starting from terms and relations of an ontology which are bound to those of the initial query.
- Reduce the noise with respect to the whole set of restored documents. The idea consists in using ontology to make word sense disambiguation, and thus keeping in the query only words expressing the need of information as well as possible.
- Reduce the noise with respect to the whole set of restored documents. The idea consists in using ontology to make word sense disambiguation, and thus keeping in the query only words expressing the need of information as well as possible.

We have defined several rules for utilizing ontology to expand ASK queries one of them is as follows the other ones are referenced in [35]. Assume that  $O=(C,R,E)$  is an ontology, if a query is composed of only one word  $\omega$  we can enrich the query by all the synonyms of  $\omega$  in  $O$ . Thus:

$$Enrich(\omega, 0) = \omega \& (\bigcup_{\omega \in W} \omega)$$

$$W = \{\omega_i | \forall i = 1, \dots, n (\omega, \omega_i, \text{Synonymy}) \in E \vee (\omega, \omega_i, \text{Holonymy}) \in E\}$$

The proposed rules are useful to shortcut the refinement phase shown figure 1. Actually, by formulating directly the appropriate query, user will not be obliged to refine it to obtain the desired relevant results. This is a first step toward optimality because it spares much time to users.

### 5. Prototyping the $O^3$ Approach

To strengthen the validity of the  $O^3$  approach, we have developed a prototype that implements the theoretical concepts presented in the previous sections. We also propose an optimal algorithm for the search of Web documents based on our approach and whose appropriate generalization is as follows:

- $O^3$  ( $W^3$  Graph web, Query  $r$ , Ontology  $O$ )
1. URLSet  $\leftarrow \emptyset$
  2.  $r' \leftarrow Enrich(r, O)$
  3. Walk(web,  $r'$ , URLSet)
  4. Return URLSet

Basically, the idea behind this algorithm is to enrich the initial query, elaborated by the user, with the vocabulary contained in the ontology according to the rules presented in section 4.3 and then to validate the query on the WPGraph and  $W^3$  Graph. However, to optimize the development of our prototype we decide to formalize the concepts presented in this paper using first-order logic<sup>7</sup>. Nevertheless, the exploration of the  $W^3$  Graph through the Walk procedure can be optimized by considering the  $\rho_{GW}$  function and, for example, by exploring nodes whose edges have a strong weight first. Moreover, as  $\rho_e$  function takes into account the properties of a given ontology, to construct the WPGraph associated with a given web page, the obtained WPGraph will be as dense (by considering the edges weight average) as it is closed to the ontology used to model the domain of research. Consequently, results can be sorted according to decreasing density of the corresponding WPGraphs so users will be able to consult Web pages which should interest them the most, first.

#### 5.1 Logic-based Unified Framework

The main goal of this effort is to offer a formal framework of our concepts for handling them in a non ambiguous way and to be able to rapidly prototype our approach. We first define the logic structure of our conceptual structures (ontologies and graphs) and then we focus on the semantic of the ASK language. We will then be able to verify the logic formulae obtained from ASK queries on the logic structure of our graphs, hence obtaining the desired URL list.

##### 5.1.1 Logical Formalization of the Conceptual Structures

The logic formalism of our graphs respects the formalism given section 3. Let  $O=(C,R,E)$  be an ontology, we define:

$$Ont = \langle Dont, Cont, Eont \rangle$$

as being the logic structure of an ontology with  $Dont = \Sigma^+ \cup R$  the domain which is the union of words of the natural language and the 7 semantic relation names of  $R$  and two relations,  $Cont$ ,  $Eont$  where:

$$Cont \subseteq Dont: Cont = \{c | c \in C\}$$

$$Eont \subseteq Dont : Eont = \{(c1, c2, r) | c1, c2 \in C \wedge r \in R \wedge (c1, c2, r) \in E\}$$

$Cont$  is a generic relation allowing to uniformly cover the set of concepts of an ontology (i.e.  $Cont(c)$  if  $c \in C$ ). In the same spirit, the  $Eont$  relation is defined to cover the relations that bind concepts of an ontology (i.e.  $Eont(x, y, synonymy)$  if  $x, y, \in C$  and  $Synonymy \in R$  and  $(x, y, synonymy) \in E$ ).

<sup>7</sup> We have prioritized the use of first order logic instead of graph theory for instance because we use PROLOG for the development of our prototype.

We need then to formalize in the same way the WPGraphs and the  $W^3$  Graph, what is done hereafter. Let  $GD$  be a WPGraph  $GD = (V, E, T, \varphi, \rho^v, \rho^e)$  we define the logic structure:

$$SLGGD = \langle DGD, VGD, EGD, \varphi GD, \rho^v GD, \rho^e GD \rangle$$

Let  $DGD = \Sigma^+ \cup T \cup IR^+$  be the domain (i.e the union of words of natural language, the type of objects contains in a Web page and the set of positive real numbers) we define the 5 relations  $VGD$ ,  $EGD$ ,  $\varphi GD$ ,  $\rho^v GD$ ,  $\rho^e GD$  as follows:

$$VGD \subseteq DGD: VGD = \{v | v \in V\}$$

$$EGD \subseteq DGD^2: EGD = \{(v1, v2) | v1, v2 \in V \wedge \{v1, v2\} \in E \vee \{v2, v1\} \in E\}$$

$$\varphi GD \subseteq DGD^3: \forall v, l, t \in DGD \varphi(v) = (l, t) \Rightarrow (v, l, t) \in \varphi GD$$

$$\rho^v GD \subseteq DGD^2: \forall v, r \in DGD \rho^v(v) = r \Rightarrow (v, r) \in \rho^v GD$$

$$\rho^e GD \subseteq DGD^3: \forall v1, v2, r \in DGD \rho^e(\{v1, v2\}) = r \Rightarrow (v1, v2, r) \in \rho^e GD$$

$VGD$  covers the whole set of vertices of a WPGraph,  $EGD$  covers the set of edges of a WPGraph (without taking the orientation of the edges into account).  $\varphi GD$  covers the set of label and type that can have a vertex (for instance on example 1,  $\varphi GD(Pruski, png, img) = true$  because  $\varphi(Pruski) = (png, img)$ ). Finally,  $\rho^v GD$  and  $\rho^e GD$  are two relations that are build using the weight functions of a WPGraph.

Lastly, let  $GW = (S, A, \rho_{GW})$  be a  $W^3$  Graph, we define the logic structure

$$SLGGPW = \langle DGPW, SGPW, AGPW, \rho_{GPW} \rangle$$

with  $DGPW = \Sigma^+ \cup IR^+$  the domain composed of natural language words and positive real numbers and the three relations:

$$SGPW \subseteq DGPW^2: SGPW = \{(u, v) \in S\}$$

$$AGPW \subseteq DGPW^4: AGPW = \{(S1, S2) | S1, S2 \in S \wedge \{S1, S2\} \in A \vee \{S2, S1\} \in A\}$$

$$\rho_{GPW} \subseteq DGPW^5: \forall s1, s2, r \in DGPW, \rho_{GPW}(\{s1, s2\}) = r \Rightarrow (s1, s2, r) \in \rho_{GPW}$$

$SGPW$  covers the set of vertices that compose a  $W^3$  Graph,  $AGPW$  covers the set of edges and  $\rho_{GPW}$  correspond to the weight function define for the  $W^3$  Graph.

#### 5.1.2 Logical Formalization of ASK Queries

Since the formalization of our structures is clearly established we can focus on that of the ASK queries. So that the queries can be interpreted on our graphs, they must respect their formalism. We give the formalization of the queries one can elaborate with the ASK language, this can be seen as being the semantics of the language. In this part we assume that  $WP = \langle DGD, VGD, EGD, \varphi GD, \rho^v GD, \rho^e GD \rangle, url$  is a Web page (i.e. a vertex of a  $W^3$  Graph).

ASK Expressions	Equivalent in First-order Logic
1. $\omega$	$\exists v \in VGD \ v = \omega \wedge \varphi GD(\omega) = (e, l) \wedge \rho_{GD}(\omega) \geq 0$
2. $\omega\{r\}$	$\exists v \in VGD \ \exists x \in \Sigma^+ \subseteq DGD \ \exists t \in T \subseteq DGD \ v = \omega \wedge \varphi GD(\omega) = (x, t) \wedge \rho_{GD}(\omega) \geq r$
3. $\omega[r]$	$\exists v \in VGD \ v = \omega \wedge \varphi GD(\omega) = (e, l) \wedge \rho_{GD}(\omega) \geq r$
4. $\omega\{r\}\{t\}$	$\exists v \in VGD \ \exists x \in \Sigma^+ \subseteq DGD \ \exists t \in T \subseteq DGD \ v = \omega \wedge \varphi GD(\omega) = (x, t) \wedge \rho_{GD}(\omega) \geq r$
5. $! \omega$	$\forall v \in VGD \ v \neq \omega$
6. $\omega_1 \& \omega_2$	$\exists v_1, v_2 \in VGD \ v_1 = \omega_1 \wedge v_2 = \omega_2 \wedge \varphi GD(\omega_1) = (e, l) \wedge \varphi GD(\omega_2) = (e, l) \wedge \rho_{GD}(\omega_1) \geq 0 \wedge \rho_{GD}(\omega_2) \geq 0$
7. $\omega_1 \# \omega_2$	$\exists v_1, v_2 \in VGD \ (v_1 = \omega_1 \wedge \varphi GD(\omega_1) = (e, l) \wedge \rho_{GD}(\omega_1) \geq 0) \vee (v_2 = \omega_2 \wedge \varphi GD(\omega_2) = (e, l) \wedge \rho_{GD}(\omega_2) \geq 0)$
8. $\omega_1 \# \omega_2$	$\exists v_1, v_2 \in VGD \ (v_1 = \omega_1 \wedge v_2 \neq \omega_2 \wedge \varphi GD(\omega_1) = (e, l) \wedge \rho_{GD}(\omega_1) \geq 0) \vee (v_2 = \omega_2 \wedge v_1 \neq \omega_1 \wedge \varphi GD(\omega_2) = (e, l) \wedge \rho_{GD}(\omega_2) \geq 0)$
9. $(\omega_1 \& \omega_2)[r]\{t\}$	$\exists v_1, v_2 \in VGD \ \exists x \in \Sigma^+ \ \exists t \in T \subseteq DGD \ v_1 = \omega_1 \wedge v_2 = \omega_2 \wedge \varphi GD(\omega_1) = (x, t) \wedge \varphi GD(\omega_2) = (x, t) \wedge \rho_{GD}(\omega_1) \geq r \wedge \rho_{GD}(\omega_2) \geq r$
10. $(\omega_1 \# \omega_2)[r]\{t\}$	$\exists v_1, v_2 \in VGD \ \exists x, y \in \Sigma^+ \ \exists t \in T \subseteq DGD \ (v_1 = \omega_1 \wedge \varphi GD(\omega_1) = (x, t) \wedge \rho_{GD}(\omega_1) \geq r) \vee (v_2 = \omega_2 \wedge \varphi GD(\omega_2) = (y, t) \wedge \rho_{GD}(\omega_2) \geq r)$
11. $(\omega_1 \# \omega_2)[r]\{t\}$	$\exists v_1, v_2 \in VGD \ \exists x, y \in \Sigma^+ \ \exists t \in T \subseteq DGD \ (v_1 = \omega_1 \wedge v_2 \neq \omega_2 \wedge \varphi GD(\omega_1) = (x, t) \wedge \rho_{GD}(\omega_1) \geq r) \vee (v_2 = \omega_2 \wedge v_1 \neq \omega_1 \wedge \varphi GD(\omega_2) = (y, t) \wedge \rho_{GD}(\omega_2) \geq r)$
12. $(\omega_1 \text{ op } \omega_2)[r]\{t\}$	$\Leftrightarrow \begin{cases} (\omega_1[r]\{t\} \text{ op } \omega_2)[r]\{t\} \\ (\omega_1 \text{ op } \omega_2[r]\{t\})[r]\{t\} \\ (\omega_1[r]\{t\} \text{ op } \omega_2[r]\{t\})[r]\{t\} \end{cases}$

<sup>8</sup> op is one of the three Boolean operators  $\&$ ,  $|$ ,  $\#$

From a more pragmatic point of view query 10 means that we are looking for a WPGraph that contains two vertex (i.e.  $\exists v1, v2 \in VGD$ ) that correspond to the concepts  $\omega_1$  or  $\omega_2$ . The two concepts must be of  $t$  type and have an importance of at least  $r$  in the WPGraph (i.e. the rest of the formula).

For instance, the query (tree | plant)[2.1]{img} given section 4.2 has the following first-order logic correspondence:

$$\exists v1, v2 \in VGD \exists x, y \in \Sigma^* (v1 = tree \wedge \rho_{GDD}(tree) = (x, img) \wedge \rho_{GDD}(tree) \geq 2.1) \vee (v2 = plant \wedge \rho_{GDD}(plant) = (y, img) \wedge \rho_{GDD}(plant) \geq 2.1)$$

It means that we are looking for web pages containing images of plant or tree where plant and tree have a weight greater than 2.1.

of the above tools and committed the obtained results in table 1 hereafter. Since the construction of WPGraphs and  $W^3$  Graph is not automatic, we have selected randomly a set of web pages among the results returned by the various tools we want to compare and have transformed them manually into WPGraph and  $W^3$  Graph in order to make our prototype effective, the same query has then been processed by the prototype.

Although being only at the prototyping phase, our tool already gives promising results. The aspect on which our approach stands out from the other is the relevance of the returned results (precision<sup>13</sup> column). This is mainly due to a combination of the concepts proposed in our framework:

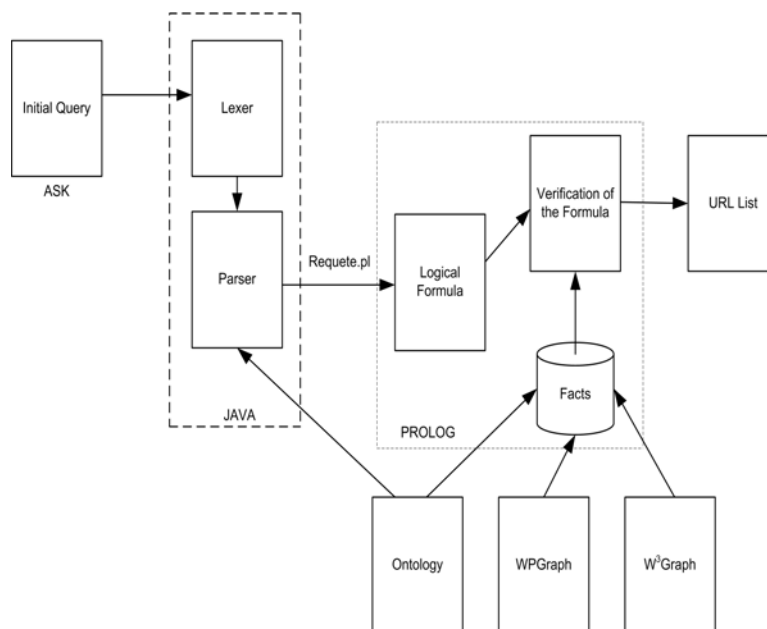


Figure 6. Prototype General Schema

## 5.2 Tool Architecture

The proposed prototype (see Figure 6) (also available through the Web <http://se2c.uni.lu/tiki/tiki-index.php?page=TargetTool>) is made up of three parts. The first one consists in a JAVA lexer and parser, allowing to treat ASK queries, in order to enrich them with words taken from an ontology and to transform them into a formula interpretable on the logic structures defined above. The second part is a PROLOG-based<sup>9</sup> engine able to verify the emitted formula on the structures contained in the base of facts which is the third part of our tool. More details about this prototype including source code and screenshots can be found in our technical report [35].

## 5.3 Case Study

In order to strengthen our approach, we have developed a basic case study using our prototype and have confronted the obtained results to those obtained using classic Web search techniques. We decided to compare our approach to Google, the well-known search engine, Clusty<sup>10</sup>, a metasearch engine that use clustering techniques, Swoogle<sup>11</sup>, the search engine for the Semantic Web and lastly Metacrawler<sup>12</sup> which is a metasearch engine widely used. For this case study assume that we are looking for pages containing publications about trees in the sense of the graph theory. Thus, we have submitted a simple query, to know “publications on trees” to all

- First, the use of ontology to characterize the research domain. However, this result can be still be improved by performing concepts lemmatization (considering singular and plural) for instance as it is done in Google or Metacrawler or to make the ontology much finer.
- Second, the query language that facilitate both the construction of good queries and their expansion using the vocabulary contained in an ontology
- Third, the adequacy of graph structure to represent and explore the Web.

On the other hand, the complexity in time as well as the user interface problems, which are not optimal with respect to other approaches, will be of course corrected for future releases of our tool.

## 6. Conclusion

In this paper we have presented a new ontology-based approach for improving Web information retrieval by proposing new conceptual structures to model the Web and a query language to extract Web information from them. This study has shown that the use of ontology for Web scale applications will increase and will be the subject of many research projects. An important weakness to note is the current difficulty in designing ontologies. Ontologies are not designed easily and, moreover, there are no quality measures already in place for ontologies. Thus, in order to make our approach optimal, we will need to resolve this problem. Providing a unified precise, rigorous

<sup>13</sup> The precision is the ratio of relevant returned pages to quantity of returned documents.

<sup>9</sup> <http://www.swi-prolog.org>

<sup>10</sup> <http://www.clusty.com>

<sup>11</sup> <http://swoogle.umbc.edu/>

<sup>12</sup> <http://www.metacrawler.com/>

	Speed	Query language	Types of retrieved documents	Retrieved documents	Precision	Interface
Google	Excellent (0.06s)	Basic boolean operators, lemmization	Web page, images, pdf, ps, doc, xls ...	Huge (34 100 000)	11% intelligent pagerank sorting	Web page, URL + description + file type
O <sup>3</sup>	N.A. since semi-automatic	Richer boolean operators, query expansion	url, files, images, video, audio	Depending on W3Graph size	75%	only URL list
Clusty	Good (2.3 s)	Poor, -negation and "" phrases	Web page, images, pdf, ps	Little (251)	8% + clustering	Web page, URL+ description + file type + images
Swoogle	Good (2.123s)	only terms	Semantic Web files	Very little(7)	(irrelevant)	Web page, URL+ description + file type
Meta crawler	Good (1.5 s)	Basic boolean operators, lemmization	Web page, images, audio, video, pdf	Little (95)	3% + kind of pagerank sorting	Web page, URL+ description + file type

Table 1. Results Comparison

and homogeneous framework for this would be positive and this has been partially illustrated in our paper. Several ongoing projects like ALT INFO<sup>14</sup> or GOL<sup>15</sup> try to lessen the impact of these weaknesses. The proposed conceptual structures (WPGraph and W<sup>3</sup>Graph) have shown the adequacy of graph structures to Web exploration and representation. Nevertheless, these structures must be studied and enhanced, including the development of adequate metrics in order to cope with the complexity of the Web. We also plan to develop tools for end-users that would have advanced user interfaces to facilitate both the selection of the ontology and the elaboration of effective queries. We will make more rigorous statistical studies that compare our approach to popular ones.

## References

- [1] Berners-Lee, T., Hendler, J., Lassila, O (2001). The Semantic Web. Scientific American.
- [2] Cui, H., Wen, J.-R., Nie, J.-Y., Ma, W.-Y (2002). Probabilistic query expansion using query logs. In Proceedings of the 11th international conference on World Wide Web, p. 325-332, Honolulu, Hawaii, USA. ACM Press.
- [3] Sowa, J.F (1984) Conceptual Structures - Information Processing in Mind and Machine. Addison Wesley.
- [4] Page, L., Brin, S (1998). The anatomy of a large-scale hypertextual web search engine. Proceedings of the Seventh International World-Wide Web Conference.
- [5] Gruber, T.R (1993) A translation approach to portable ontology specifications. *Knowl. Acquis.*, 5 (2), 199-220.
- [6] Cui, H., Wen, J.-R., Nie, J.-Y., Ma, W.-Y (2002). Probabilistic query expansion using query logs. In: Proceedings of the 11th international conference on World Wide Web, 325-332, Honolulu, Hawaii, USA. ACM Press.
- [7] Fellbaum, C.D (1998). WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA.
- [8] Introna, L., Nissenbaum, H (2000). Defining the web: The politics of search engines. *Computer*, 33 (1) 54-62

[9] Konopnicki, D. Shmueli., O (1995). W3QL: A query system for the world-wide web. Proceedings of the 21th International Conference on Very Large Data Bases, 54-65, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[10] Mendelzon, A.O., Mihaila, G.A., Milo, T (1997). Querying the World Wide Web. *International Journal on Digital Libraries* 1(1) 54-67.

[11] Silverstein, C., Marais, H., Henzinger, M., Moricz, M (1999). Analysis of a very large web search engine query log. *SIGIR Forum*, 33 (1) 6-12.

[12] Dziadosz, S., Chandrasekar, R (2002). Do thumbnail previews help users make better relevance decisions about web search results? In: *SIGIR*, 365-366.

[13] Liu, H., Lieberman, H., Selker, T (2002). Goose: A goal-oriented search engine with commonsense. In: AH '02: Proceedings of the Second International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, 253-263, London: Springer-Verlag.

[14] Berners-Lee, T., Cailliau, R., Groff, J.F, Pollermann, B (1992). World-wide web: The information universe. *Electronic Networking: Research, Applications and Policy*, 1 (2) 74-82.

[15] Stenmark, D (1999). To search is great, to find is greater: a study of visualisation tools for the web. <http://w3.informatik.gu.se/dixi/publ/mdi.htm>.

[16] Huang, X., Lai, W (2003). Identification of clusters in the web graph based on link topology. In: Seventh International Database Engineering and Applications Symposium (IDEAS'03), p. 123.

[17] Chiang, R. H. L., Chua, C. E. H., Storey, V. C (2001). A smart web query method for semantic retrieval of web data. *Data Knowl. Eng.*, 38 (1) 63-84.

[18] Lawrence, S., Giles, C.L (2000). Accessibility of information on the web. *Intelligence*, 11 (1) 32-39.

[19] Jackson, P., Moulinier, I (2003). Briefly noted: natural language processing for online applications: Text retrieval, extraction, and categorization. *Comput. Linguist.*, 29 (3) 510-511.

[20] Lawrence, S (2000). Context in web search. *IEEE Data Engineering Bulletin*, 23 (3) 25-32.

<sup>14</sup> <http://www.loa-cnr.it/altinfo>

<sup>15</sup> <http://www.ontology.uni-leipzig.de/>

- [21] Webb, G.I., Pazzani, M.J., Billsus, D (2001). Machine learning for user modeling. *User Modeling and User-Adapted Interaction* 11 (1-2), 19-29.
- [22] Budzik, J., Hammond, K.J (2000). User interaction with everyday applications as context for just-in-time information access. Proceedings of the International Conference on Intelligent User Interfaces, New Orleans, Louisiana. ACM Press.
- [23] Bauer, T., Leake, D.B (2001). Wordsieve: A method for real-time context extraction. *Lecture Notes in Computer Science*, 2116:30.
- [24] Haveliwala, T.H (2003). Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE Trans. Knowl. Data Eng.*, 15 (4) 784-796.
- [25] Broder, A.Z., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., Wiener, J.L (2000). Graph structure in the web. *Computer Networks*, 33 (1-6) 309-320.
- [26] Bharat, K., Henzinger, M.R (1998). Improved algorithms for topic distillation in a hyperlinked environment. Proceedings of the 21<sup>st</sup> annual international ACM SIGIR conference on Research and development in information retrieval, 104-111. ACM Press.
- [27] Flake, G. W., Lawrence, S., Giles, C. L., Coetzee, F (2002). Self-organization of the web and identification of communities. *IEEE Computer*, 35 (3) 66-71.
- [28] Klyne, G., Carroll, J.J (2004) Resource description framework (rdf): Concepts and abstract syntax. W3C Recommendation.
- [29] McGuinness, D.L., van Harmelen, F (2004). Owl web ontology language overview. W3C Recommendation.
- [30] McCarthy, J (1980). Circumscription-a form of non-monotonic reasoning. *Artificial Intelligence*, 13, 27-39.
- [31] Hakimpour, F., Geppert, A (2005). Resolution of semantic heterogeneity in database schema integration using formal ontologies. *Inf. Tech. And Management*, 6 (1), 97-122.
- [32] Ouzzani, M., Benattallah, B., Bouguettaya, A (2000). Ontological approach for information discovery in internet databases. *Distrib. Parallel Databases*, 8 (3), 367-392.
- [33] Wang, X., Chan, C.W., Hamilton, H.J (2002). Design of knowledge-based systems with the ontology-domain-system approach. Proceedings of the 14th international conference on Software engineering and knowledge engineering, 233-236, Ischia, Italy. ACM Press.
- [34] Fragos, K., Maistros, Y., Skourlas, C (2003). Word sense disambiguation using wordnet relations. *In: Proceeding of the 1st Balkan Conference in Informatics*, Thessaloniki, Greece.
- [35] Pruski, C (2005). An optimal algorithm for the interpretation of first-order logic formulae on the web based on graph semantics. TR-SE2C-05-06, Software Engineering Competence Center - University of Luxembourg, Luxembourg-Kirchberg, Luxembourg.
- [36] Labrou, Y (2002). Agents and ontologies for e-business. *Knowl. Eng. Rev.*, 17 (1) 81-85.
- [37] Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P.F (2003). *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press.
- [38] Gruber, T.R (1993). Towards principles for the design of ontologies used for knowledge sharing. *In: N. Guarino and R. Poli, editors, Formal Ontology in Conceptual Analysis and Knowledge Representation*, Deventer, The Netherlands. Kluwer Academic Publishers.
- [39] Genesereth, M.R., Fikes, R.E (1992). Kif version 3.0. reference manual. Technical Report Logic-92-1, Stanford, Stanford University.
- [40] van Harmelen, F., Fensel, D. (1999). Practical knowledge representation for the web. Proceedings of the IJCAI'99 Workshop on Intelligent Information Integration.
- [41] Decker, S., Fensel, D., van Harmelen, F., Horrocks, I., Melnik, S., Klein, M., Broekstra, J (2000). Knowledge representation on the web. In F. Baader, editor, *International Workshop on Description Logics (DL'00)*.
- [42] Priss, U. (1998). The formalization of wordnet by methods of relational concept analysis. *In: C. Fellbaum, editor, WordNet: An Electronic Lexical Database and Some of its Applications*, 179-196. MIT press.
- [43] Hirst, G., St-Onge, D (1998). Lexical chains as representation of context for the detection and correction malapropisms. *In: C. Fellbaum, editor, WordNet: An electronic lexical database and some of its applications*, 305-332, Cambridge, MA. The MIT Press.
- [44] Jansen, B.J., Spink, A., Pfaff, A (2000) Linguistic aspects of web queries. *In: American Society of Information Science*, Chicago.
- [45] Broekstra, J., Fluit, C., van Harmelen, F (2000). The state of the art on representation and query languages for semistructured data. Technical Report On-To-Knowledge EU-IST-1999-10132 deliverable 8, Administrator Nederland b.v..
- [46] Spertus, E., Stein, L.A (2000). Squeal: Structured queries on the web. *In: Ninth International World-Wide Web Conference*, Amsterdam.
- [47] Navigli, R., Velardi, P (2003). An analysis of ontology-based query expansion strategies. *In: Proceeding of the Workshop on Adaptive Text Extraction and Mining, Cavtat-Dubrovnik (Croatia)*.
- [48] Carmel, D., Farchi, E., Petruschka, Y., Soffer, A (2002). Automatic query refinement using lexical affinities with maximal information gain. *In: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, p. 283-290, Tampere, Finland. ACM Press. 16.
- [49] Anick, P (2003). Using terminological feedback for web search refinement: a log-based study. *In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, p. 88-95, Toronto, Canada. ACM Press.
- [50] Stojanovic, N., Studer, R., Stojanovic, L (2004). An approach for step-by-step query refinement in the ontology-based information retrieval. *In: IEEE/WIC/ACM International Conference on Web Intelligence (WI'04)*, p. 36-43.
- [51] Paralic, J. Kostial, I (2003). Ontology-based information retrieval., *In: Proc. of the 14th International Conference on Information and Intelligent systems*, p. 23-28, Varazdin, Croatia



Nicolas Guelfi is Professor at the Faculty of Sciences, Technology and Communications of the University of Luxembourg. His main research activities concern the engineering and evolution of reliable and secure distributed and mobile systems based on semi-formal methods and transformations. He is a leading member of the Laboratory for Advanced Software Systems (LASSY). He has made significant contributions on software engineering methods and tools for distributed systems. He has been involved in three European ESPRIT BRA projects. He is a member of the executive committee of the ERCIM consortium and is chair of the working group on rapid integration of software engineering techniques (RISE) where he developed collaborations with the W3C consortium and with the Semantic Web working group.



Cédric Pruski is a Ph.D. student in the LASSY of the University of Luxembourg and in the IASI team of the LRI at Paris-Sud University XI (France). He received his Master's degree from the university of Nancy I (France) in September 2005. He also worked for two years as a R&D engineer in the SE2C group at the University of Luxembourg on research projects that dealt with security and trust management. His research interests include the use of ontologies for information management, ontologies evolution and Web information retrieval.