ASWGF! Towards an Intelligent Solution for the Deep Semantic Web Problem

Mohamed A. Khattab, Yasser Hassan and Mohamad Abo El Nasr*

Department of Mathematics & Computer Science, Faculty of Science, Alexandria University, Egypt,
*Head of Computer Engineering Department, Arab Academy for Science, Technology and Maritime, Egypt
khattab@alex-sci.edu.eg, y.fouad@alex-sci.edu.eg, *m.aboelnasr@gmail.com

Abstract — Deep Semantic Web is a new era of Web problems correlated to the release of Semantic Web search engines. This research considers the problem of deep Semantic Web as a challenge for the future of the Web in order to achieve the dream of Semantic Web connecting everything. This paper addresses the problem of Semantic Web as being a solution that needs a solution and announces the term “Deep Semantic Web” in an attempt to motivate the researchers to take into consideration such an invisible problem and find a solution for it. Our contribution is trying to provide a state of the art solution through proposing an intelligent Semantic approach for Web search engines towards a homogeneous Semantic Web. The innovative idea of The Automatic Semantic Web Generator Framework – ASWGF – is embodied in the implementation and integration process with the current Web servers once. The approach considers the integration process as the spark off the model whereas the web servers will be the heart of the framework, we are not going to bet on the webmasters anymore in the success or the failure of the model.


1. Introduction

The largest known resource for information is the World Wide Web as people all over the world use the Web to find all sorts of information. The Web is inherently distributed and the data on the Web is stored in numerous sources and those sources have central points that provide search capabilities which are the Web search engines [1]. Traditionally, crawlers have only targeted a portion of the Web called the publicly indexable Web (PIW) [2]. This refers to the set of pages reachable purely by following hypertext links, ignoring the unknown domains and hosts which are called “invisible Web”. Jill Ellsworth [3] used the term “invisible Web” in 1994 to refer to the Web sites that are not registered in any search engine. Frank Garcia [4] used the term “invisible Web” in 1996 to refer to the Web sites that are possibly reasonably designed, but the designers did not bother to register them in any of the Web search engines; therefore, no one can find them. The first use of the specific term “Deep Web” occurred 2001 in a study by Michael Bergman [5] in which Bergman avoided the term “invisible Web”. Over the years, search engines have discovered and learned how to track many factors to help them decide which pages are relevant to a given search. All such factors have to do with either where the word appears, or the structure of the Web. Search Engine Optimization (SEO) considers how search engines work, what people search for, the actual search terms typed into search engines and which search engines are preferred by the targeted audience. Optimizing Web sites may involve editing its content – HTML – and the associated coding to increase both its relevance to specific keywords and to remove barriers to the indexing activities of the search engines. Site owners recognize the value of having their Web sites highly ranked and visible in search engine results. Semantic Web crawler differs from a traditional Web crawler in the format of the source material it is traversing, and the means of specifying links between information resources [6]. This paper announces a new term; “Deep Semantic Web” which refers to Web sites that are traditionally optimized and registered through the Web search engines and will not be indexed through Semantic Web search engines; refers to the definition of term “Deep Web.” After the announcement

1
of Semantic Web search engines as alternatives to the traditional Web search engines, we can define the term “Deep Semantic Web” to refer to the surface Web from the point of view of Semantic Web search engines. Currently, there becomes a heterogeneity of structures between Web sites and Web search engines: as traditional HTML based Web sites, content management system (CMS) based Web sites optimized for traditional Web search engines or optimized for Semantic Web search engines, which Web sites being submitted to which Web search engine. Even Semantic based search engines and Semantic optimized Web sites use different approaches and ontologies which lead to the lack of optimization, and the Semantic dream is not achieved.

2. Intelligence in Semantic Web

The Semantic Web is a mesh of information linked up in such a way as to be easily processed by machines on a global scale. Think of it as being an efficient way of representing data on the World Wide Web, or as a globally linked database; to do for data what the Web did for documents. Turning the Web into a database, and everyone's data becomes a part of it. The data becomes part of the worldwide database. The Semantic Web will give the ability to move from data record to data record, like going from Web page to Web page. Shortly, it is a service that aims at “tieing it all together” in order to make sense of the Web. The social graph is about connecting people while the Semantic Web is about connecting things. As illustrated in figure 1, it is not just connecting things but connecting everything so the intelligence is in the connections [7]. This makes the social graph just a small piece of the Semantic Web which revolves around connecting more kinds of things together. The connections will improve searching and advertising.

As the Web swells with more and more data, the predominant way of sifting through all of that data—keyword search—will one day break down in its ability to deliver the exact information. In fact, some argue that keyword search already delivers diminishing returns. However, anyone frustrated by the sense that it takes longer to find something on traditional Web search engines today than it did a year ago knows there is some truth in his argument. Keyword search is okay unless the information explosion continues. The amount of information on the Web keeps growing exponentially to accommodate all those seekers, and they feel compelled to put their own personal and social information onto the Web. At a certain point, with billions and billions of Web pages to sift through, just keyword search will not cut it anymore. Figure 2 illustrates the irrelative relation between the amount of searched data and the productivity of search across the Web generations.

![Figure 1 The Semantic graph connects everything](image1)

![Figure 2 Irrelative relations between the amount of data and the productivity of search](image2)
Tim Berners-Lee – inventor of the Web – was the Semantic Web [8] to automate the process of collection and integration of information, but the dream is not yet realized.

3. Semantic Web Crawler vs. Web Crawler

The Web crawler is a program or an automated script that browses the World Wide Web in a methodical and automated manner and is known as "Web spider" or "Web robot". Other less frequently used names for Web crawlers are "ants", "automatic indexers", "bots", and "worms" [9]. The process of browsing the World Wide Web is called "Web crawling" or "spidering". Many sites, in particular search engines, use spidering as a means of providing up-to-date data. Web crawlers are mainly used to create a copy of all the visited pages for later processing by a search engine that will index the downloaded pages to provide fast searches. Crawlers can also be used for automating maintenance tasks on the Web sites, such as: checking links or validating HTML code. Also, crawlers can be used to gather specific types of information from Web pages, such as harvesting e-mail addresses.

Figure 3 Typical high level architecture of a Web crawler

Web crawlers are a central part of Web search engines, and details on their algorithms and architecture are kept as business secrets. When crawler designs are published, there is often an important lack of details that prevents others from reproducing the work. A Web crawler is one type of bot, or software agent and the typical high level architecture of Web crawlers is shown in figure 3 involving a scheduler and a multi-threaded downloader [10].

In concept a Semantic Web crawler differs from a traditional Web crawler in only two regards: the format of the source material it is traversing, and the means of specifying links between information resources. Whereas a traditional Web crawler operates on HTML documents, linked using HTML anchors, a Semantic Web crawler – scutter – operates on RDF metadata with linking implemented using the rdfls. However, in practice, the aggregation and processing of Semantic Web content by a scutter differs significantly from that of a normal Web crawler. A normal crawler must only contend with extracting text from, possibly invalid, content and subsequent link extraction, whereas a Semantic Web crawler must carry out additional processing tasks: merging of information resources via inverse-functional-properties; tracking provenance of data; harvesting schemas and ontologies in addition to source data; extraction of embedded metadata (e.g. EXIF, XMP), etc [11] provides an excellent introduction to these issues and summarizes implementation experience gained whilst constructing a scutter. For example one of the Semantic Web crawlers is Slug [12], an open source, configurable and modular scutter. The emphasis of the framework is on modularity and ease of use rather than speed (hence the name!), although performance is currently acceptable.

Figure 4 Summary of slug architecture

The UML diagram provided in figure 4 summarizes the key components and relationships present in the Slug framework. Central to the design of the framework are variations of the master-slave and producer-consumer design patterns. After presenting the two common Web crawler types, we need to think about the result of a Semantic Web crawler visiting a traditionally published...
Web sites and a Web crawler visiting a Semantically published Web sites.

4. Approaches for Semantic Web Publishing

Semantic publishing on the Web or Semantic Web publishing refers to publishing information on the Web as documents accompanied by Semantic markups [13]. Semantic publication is intended to provide a way for computers to understand the structure and even the meaning of the published information, making the search for information and the data integration more efficient. In the Semantic Web, information published on the Web is accompanied by metadata describing the published information; thus providing a Semantic context. There are many approaches for Semantic Web publishing that apply different trends and outputs to obtain a Semantic context as figure 5 shows.

![Figure 5 Approaches for Semantic Web publishing](image)

**Data objects** – Publish information as data objects using Semantic Web languages like RDF [14] and OWL [15].

**Embedded metadata** – Embed formal metadata in documents using new markup languages like RDFa [16] and Microformats [17].

**Dynamic Semantic publishing** – A new technical term introduced by the online team of BBC [18]. It describes the idea of utilizing linked data technology to automate the aggregation and publication of interrelated content objects and provides rich content relationships as figure 6 illustrates a possible, generalized approach of dynamic Semantic publishing [19].

The content objects are multi-dimensional entry points to a site and provide a much better and more user-engaging way to explore content than the usual flat archives pages, which normally do not have dimensions beyond date, tag, and author. The variety of approaches for Semantic Web publishing results in different outputs based on different ontology which lead to the inconsistency of Web contents and difficulty for Semantic Web search engines to index such contents.

![Figure 6 Generalized approach of dynamic Semantic publishing](image)

5. Semantic Web Search Engines

Unlike traditional Web search engines, Semantic Web search engine stores Semantic information about Web resources and is able to solve complex queries, in addition to considering the context where the Web resource is targeted [20]. Using a Semantic Web search engine ensures more relevant results based on the ability to understand the definition of the word or the term that is being searched for, rather than on numbers. Semantic Web search engines are able to understand the context in which the words are being used, resulting in smart, relevant results which may seriously contribute to the development of electronic business applications since it is based on strong theory and widely accepted standards. After the announcement of releasing Semantic Web search engines as alternatives to traditional Web search engines, there is a list of the top seven Semantic Web search engines with a short description about each one to get started via Semantic Web searching [21].

The release of such Semantic Web search engines results in a heterogeneous Web contents composed of different architectural-based Web sites – HTML/code-based or CMS-based, and different knowledge-based Web search engines – traditional or Semantic – in addition to different ontologies and methodologies even in the same type according to approaches for Semantic Web publishing as illustrated in figure 7.
The search engine optimization process for newly designed or developed Web sites to be submitted or indexed through a Web search engine should be done for traditional Web search engines or for Semantic Web search engines. If the Semantic will be the target, which ontology should be utilized to be indexed as there are different applied ontologies? And which type of Web crawler will traverse and index the contents of such Web sites? And how to do so? The heterogeneity of the Web contents leads to one or all of the following:

- Lack of optimization.
- Heterogeneous Web content.
- Inconsistence Semantic Web model.
- Deep Semantic Web.

The Semantic Web results in a solution needs for a solution.

6. The Overview of the Proposed Approach

The problem of deep Semantic Web is considered a challenge for the future of the Web in order to achieve the dream of Semantic Web connecting everything. We are trying here to provide a state of the art solution by proposing an intelligent Semantic approach for the Web search engines towards a homogeneous Semantic Web.

6.1 The Aspects of the Proposed Approach

The proposed approach will maintain three aspects as illustrated in figure 8. The first aspect considers ontology unification from the side of the Web search engine to guarantee the speed and the simplicity of the indexing process for Web sites. Such aspect will guarantee the homogeneity of Web sites structures according to a predefined ontology. The Web search engine will no longer need to analyze the contents of the Web site; thus saves the time for more enhanced processes. The second aspect is concerned with the automation of Semantic annotation process through the implementation of the proposed approach as an Automatic Semantic Web Generator Framework – ASWGF. The third aspect is time-saving by involving the Web servers in the Semantic annotation process and guaranteeing the perfection of such process according to the proposed approach and the implemented framework.

6.2 The Automatic Semantic Web Generator Framework – ASWGF

Webmasters – as long as we are talking about Web problems or Web solutions always there is a spot on webmasters as they must be involved in any keyword-based optimization process or Semantic annotation process which always lead to the lack of solution or burden the Web search engines as must be the source and the provider for the solution. The innovative idea of The Automatic Semantic Web Generator Framework – ASWGF – is embodied in the implementation and integration process with the current Web servers once. The approach considers the integration process as the spark off the model whereas the web servers will be the heart of the framework, we are not going to bet on the webmasters anymore in the success or the failure of the model.

The start of any Web chain is by registering a domain name then hosting a certain Web site over such registered domain and our Semantic Web generator claims such a starting point and operates through a sequence of processing phases as follows:

Phase i (crawling & analyzing) – based on the hosting Web server and by applying the Host List Protocol model [22] listing the hosted domains, crawling and analyzing
their contents regardless to the content type – HTML based or CMS based.

**Phase ii** (decision & Semantic annotation) – after the analysis process the model – based on reasoning – will specify the target domain and decide which Semantic repository is suitable for such crawled and analyzed content to use it and complete the Semantic annotation process.

**Phase iii** (Semantic Web generation) – the avenues during this phase may be in one of two directions:

1. Send the resulting Semantic content to the Web search engines to handle the received information according to the pre-agreed ontology.
2. Building an intermediate repository (core of new Semantic Web search engine) to be accessed through end users / search engines according to the policy agreed upon.

![Diagram of the ASWGF Framework](image)

**Figure 9** Members of the ASWGF Framework

The state of the art in the proposed approach can be summarized as follows:

- **Automating** the Semantic Web generation process, not just the Semantic annotation process [23].
- **Unifying** the processing phases regardless to the underlying system of the Website content – HTML-based CMS-based or other Web script-based.
- **Minimizing** the processing time – single processing time – as all Web servers will work at the same time instead of sending crawlers of Web search engines to crawl the Web site contents searching for traditionally or Semantically optimized content.

7. **Experimental Results**

One of the objectives of our model is to unify the processing phases regardless to the underlying system of the Website content HTML-based, CMS-based or other Web script-based. The experiment demonstrates the capability of the framework extracting such heterogeneous content during preserving the reliability of the system according to the accuracy rate factor. The environment for the experiment was a random sample of the World Wide Web hosted Websites classified into four main categories according to their platform including the operating system, the underlying server daemons and the development environment which was an average of five values for each environment. Accuracy rate and error rate performed on the data series obtained through the experiment.

![Graph showing distribution of results accuracy relative to the underlying development environment](image)

**Figure 10** Distribution of the results accuracy relative to the underlying development environment

According to figure 10 the best accuracy value is ranging from 98% to 100% for all underlying development environment while the worst accuracy value is ranging from 75% to 84% for ASPX underlying development environment and such for two cases of the five averages.

The evaluation of the obtained accuracy rates performed on the data series obtained through the experiment leads to the fact that the framework successfully extracted such heterogeneous content during preserving the reliability of the system in terms of the accuracy rate.

8. **Conclusion and Future Work**

Semantic Web is a solution that needs a solution. The announcement of Semantic Web search engines as alternatives to traditional Web search engines affects the surface Web as moving slowly towards being deep Semantic Web. The release of Semantic Web results in a heterogeneous Web contents composed of different architectures based Web sites – HTML/code-based or
CMS-based – and different knowledge based Web search engines – traditional or Semantic – in addition to different ontologies and applied approaches. The problem of Web heterogeneity is a challenge and a threat to the future of the Web as the Semantic Web is to the future. The heterogeneity of Web contents leads to the lack of optimization, heterogeneous Web, inconsistent Semantic Web model and deep Semantic Web. We are trying here to provide a state of the art solution through proposing an intelligent based approach results in an automatic Semantic Web Generator Framework – ASWGF – towards a homogenous Semantic Web which is the vision of the future for the Web. The evaluation of the obtained accuracy rates performed on the data series obtained through the experiment leads to the fact that the framework successfully extracted such heterogeneous content during preserving the reliability of the system in terms of the accuracy rate.

References