УДК 004.9

PROSPECTS FOR APPLYING THE CONCEPT OF THE SEMANTIC WEB ANALYSIS FOR EXISTING SITES

V.V. Zosimov

Mykolaiv V.O. Suhomlynsky National University

zosimovvv@bk.ru

Одним з найважливіших напрямків підвищення ефективності пошуку інформації в Інтернеті є активне застосування концепції семантичного веб. В статті проведено аналіз перспектив використання концепції семантичного веб для аналізу текстових даних в мережі Інтернет. Розглянуто проблеми її застосування до вже створених сайтів та шляхи їх вирішення.

Ключові слова: семантика, Інтернет, пошук інформації, аналіз даних, сайт, пошуковий агент, онтологія, обробка та зберігання інформації, структури даних..

One of the most important ways of increasing information retrieval efficiency on the Internet is the active application of semantic web concept. The article analyzes the prospects of usage the semantic web concept for text data analysis on the Internet. The problems of its application to the already established sites and ways to solve them.

Keywords: semantics, Internet, information retrieval, data analysis, site search agent, ontology, information processing and storage, data structures.

Одним из важнейших направлений повышения эффективности поиска информации в Интернете является активное применение концепции семантического веб. В статье проведен анализ перспектив использования концепции семантического веб для анализа текстовых данных в сети Интернет. Рассмотрены проблемы ее применения к уже созданным сайтам и пути их решения.

Ключевые слова: семантика, Интернет, поиск информации, анализ данных, сайт, поисковый агент, онтология, обработка и хранение информации, структуры данных.

Introduction. Modern Internet is dynamically developing. In early 2012, there were 330 million sites. By the end of the year its number has increased more than 2 times and reached 743 million. Number of working sites in January 2014 amounted to 861.4 million. Total number of sites currently stands at more than a billion. [1] (Fig 1,2)

Sites number rapid growth has led to the fact that the developers of search engines are faced with several problems. "Dimension" problem one of the principals.

Standard information search and processing methods that used by modern search engines are losing their effectiveness with such a large amount of the required documents. User receives millions of web pages as a response for search request. Search results priority order is determined by search engines ranking algorithms which do not consider any user's preferences, or context in which the search query is consume. According to Internet live statistics [2] 31.9% of users are browsing only first results page, and then follow to one of the sites. 23%, are browsing two pages before make a choice. 16.1% browse only first three links, and do not look at the other results.

These data show that almost half (48%) of Internet users are following link from the first page and do not browse even the second. Respectively, they browse maximum ten results from hundreds thousands or even millions search results found by entered keywords. If user does not find information on the first page, he changes the search query. And there is no guarantee that the first page contains information which is more complete and meaning appropriate to the user's needs.

About half (45.9%) of the respondents admitted that almost immediately found right products and services. Thirds of respondents (33.3%) can not find required data in three cases out of four. 13.3% of users receive necessary information only in half of the cases. Finally, 5.1% are right only a quarter of cases and 2.3% do not understand how is it possible to find something using search engines.

In consequence of this there is an urgent need to develop new information retrieval models, which can choose the most appropriate documents within the meaning of user's needs among found by keywords.

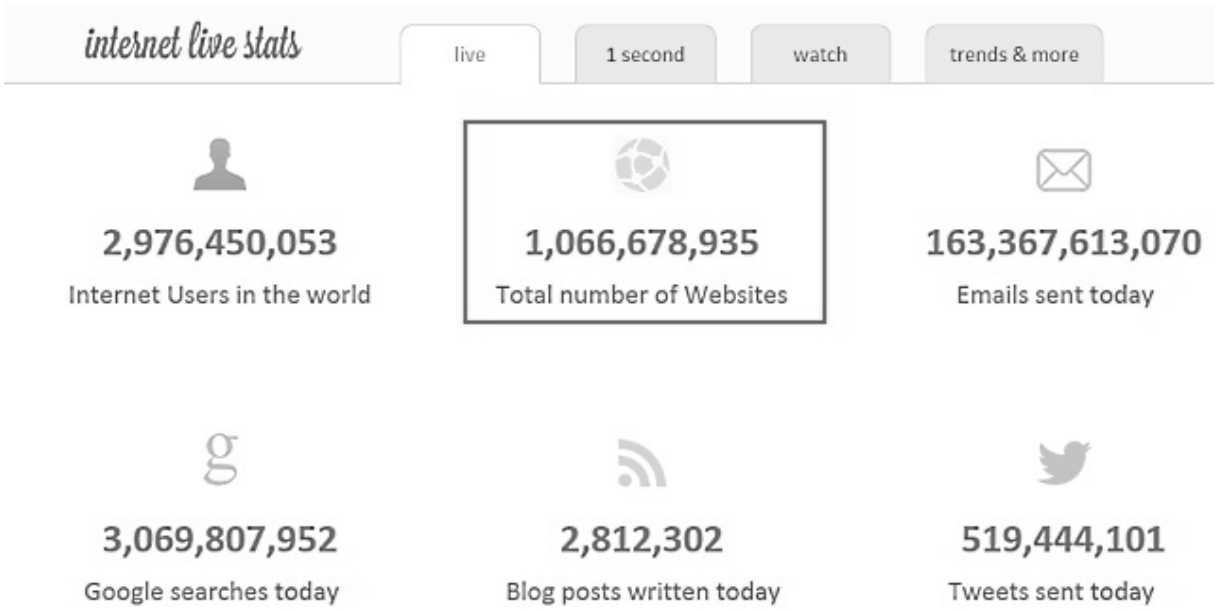

Fig 1. Live Internet statistic [3]

Tim Berners Lee proposed concept of the Semantic Web as a new stage in the Internet development, which would allow the machines easier understand and process contained in the website information due to the of semantic site markup. [2] According to the developers, this approach will help to eliminate the main drawback of modern research - the "dimension" problem, as well as to make information retrieval on the Internet more comfortable for the user. A number of search engines based on search agents were developed under this concept. These systems give the user results relevant not only to their keywords, but taking into account pages semantic content and additional search parameters input by the user.

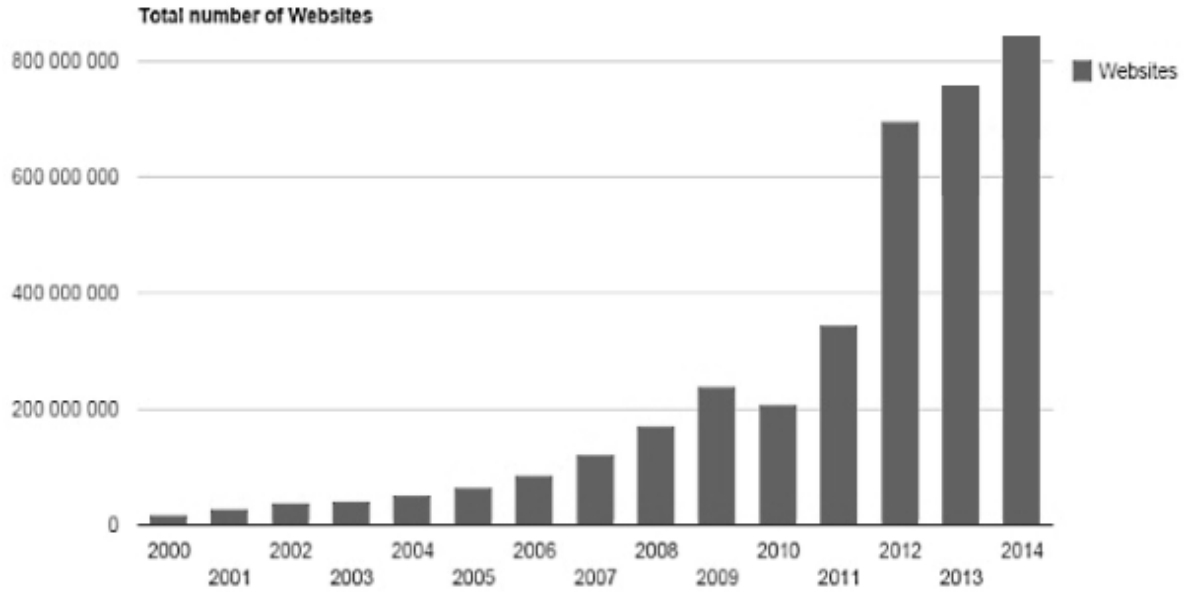

Fig 2. Web sites number growing schedule [3]

1. Semantic web concept

Semantic Web - is a public global semantic network that is formed on the World Wide Web basis through the information standardization in the machine processing suitable form.[4]

In a typical World Wide Web based on HTML-pages, the information lies in the text pages and is intended for human reading and understanding. The Semantic Web consists of machine-readable elements - semantic network nodes, buttress on ontology. Due to this, programs-clients can directly get statements "object - type relationship - the other object" and determine logical conclusions for them. Semantic Web works in parallel with the conventional World Wide Web and on its basis, using the HTTP protocol and resource identifiers URI. [5]

Semantic Web is a symbiosis of the two directions, the first of which covers data representation languages. The main languages are Extensible Markup Language XML [6] and Resource Description Framework RDF [7]. There are also a number of other formats, but XML and RDF offer more opportunities, so they are recommended by W3C

The second, conceptual direction carries a theoretical understanding of the domain model. Such models of domains in terms of the Semantic Web are called ontologies. February 10, 2004, the W3C adopted and published a web ontology language OWL [8] specification.

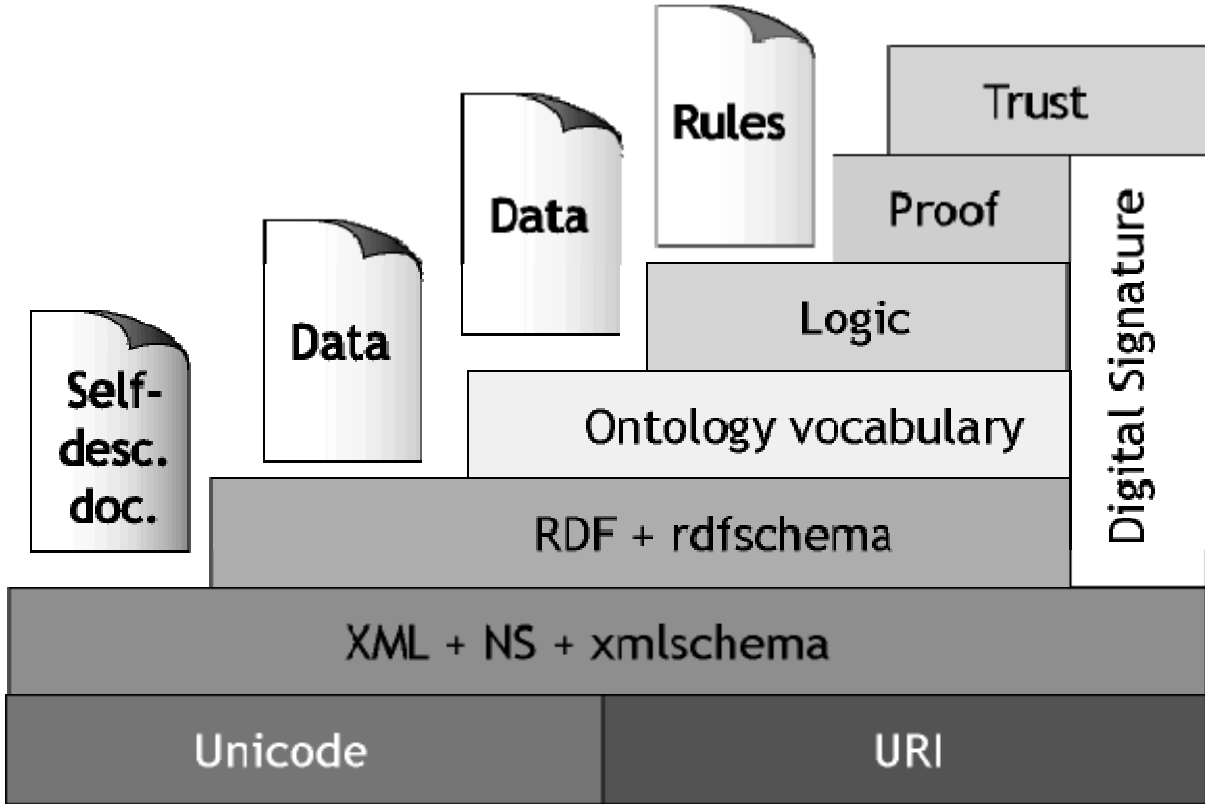

Fig 3. Semantic Web concepts Stack

The Semantic Web project involves the creation of system with "artificial intelligence" elements, which would allow special applications to look for high quality relevant information, as well as share information to each other over the Internet. At the same time the ontology language OWL is a decisive component of intellectualization, the basis for the semantic networks construction. It should be noted that the semantic networks theory has been focused on the artificial intelligence problems, such as machine translation systems. Knowledge in the semantic networks theory were represented as nodes connected by arcs, each of which specifies the relationship type. The Semantic Web is essentially a realization of the artificial intelligence idea, but the term is not very popular because of the large number of failed projects in this area, so the "semantic web" concept is today alertness. However, web ontology essentially a real knowledge base, one of the artificial intelligence conceptual foundations.

In the future, the correct website execution will allow interacting services with based on the semantic relationships analysis between concepts and objects available in the network, provide the user information that will satisfy his needs

2. Applying the Semantic Web concept to existing sites

Despite the active development of the Semantic Web concept, there are number of problems that do not allow it to be used to implement a wide range of users:

1. The Semantic Web concept assumes that the data analysis and the most optimal results selection will make machine automatically. Search agents show the

person a few most relevant results on the basis of semantic analysis. And if the search agents algorithms are not determined for users preferences, then he can not find appropriate results and other results, among which there is one that would satisfy users preferences will be filtered by machine. Thus, rather than to facilitate search task for the user application of this concept severely restricts his choice. This brings us back to the common search engines problem. Search agents usually have the opportunity to extended search settings, but not all users are willing to spend time learning new interfaces. Therefore it is necessary to develop the most simple and user-friendly interfaces of user interaction with the search agent. Otherwise, a long and complex setting search parameters may scare the user.

2. Semantic Web concept imposes special requirements on sites development standards. But most developers do not follow all the sites requirements for web pages creation. This happens for several reasons:

1) Lack of time to develop the project, the time limits set by the customer do not allow make high-quality semantic markup.

2) For today a very small percentage of Internet users are searching information using search agents. The bulk of the users prefer simple and familiar search tools. So sites customers do not see the point in spending time and money on semantic markup.

3) The company, which provides services do not make site semantic markup due to unprofitability or developers unwillingness to learn new technology.

3. The most significant drawback is the absence of effective methods and technologies for automatic semantic annotation of existing web pages. Considering the Internet growth rate, the number of active sites up to date is more than one billion. Bring all these web resources to the semantic web standards by hand without the use of automatic machine methods is an impossible task. User deliberately narrows the search to about 5% of all contained on the web pages using semantic search. If websites without semantic markup contains user necessary information he does not get access to it using a semantic search.

It follows that the Semantic Web concept can not properly function without the development automatic semantic annotation methods for existing web pages. Development of such methods will allow the concept of the semantic web to cover a large part of information resources located on the Internet, and continue to develop into a complete, user-friendly and widely used tool of web data mining and information retrieval.

Conclusions. Semantic web concept became an opportunity for the Internet to avoid the crisis of large dimension. But despite all the advantages of the concept it has is one major drawback. A huge number of existing sites created without semantic markup can not fully participate in the intellectual retrieval. Huge number of newly created sites are also done without usage of new technologies.

Therefore, semantic search will only work with the newly created site made using all semantic web rules.

This suggests need to search for new tools that enable the automatic semantic markup of existing sites and to simplify the creation of semantic markup for new sites.

Література

1. http://www.netcraft.com/ - report for the number of working sites on the Internet.

2. http://iprospect.com - statistics service for the viewed by user search results pages

3. http://internetlivestats.com/ - real time statistics service for the Internet services status.

4. http://wikipedia.org/ - the semantic web definition, architecture, implementation problems

5. http://www.w3.org/ - description of the semantic web concept, syntax, data definition language.

6. www.w3.org/TR/NOTE-xml-ql/ - query Language XML (Extensible Markup Language)

7. http://www.w3.org/TR/rdf-tuturial - Manual for the RDF (Resource Description Framework).

8. http://www.w3.org/TR/owl-ref/ - Manual for the OWL (web ontology language).