A semantic web constructor system (100) supported by a plurality of web crawlers (112) in a World Wide Web upon declaring a plurality of trustworthy websites is provided, characterized in that, the system (100) includes at least one web crawler controller (110) engageable to manage the plurality of web crawlers (112), a semantic web database (116) connectable to the plurality of web crawlers (112) and a plurality of data building editors (122, 124, 126) connectable to the at least one web crawler controller (110) wherein a semantic browser (120) is further connectable to the semantic web database (116) to receive at least one natural language query from at least one user.
wo 2012/091541 A1

Published:

— with international search report (Art. 21(3))
A SEMANTIC WEB CONSTRUCTOR SYSTEM AND A METHOD THEREOF

FIELD OF INVENTION

The present invention relates to a semantic web constructor system supported by a plurality of web crawlers in a World Wide Web, upon declaring a plurality of trustworthy websites in order to query a semantic browser for a plurality of websites that are page-ranked.

BACKGROUND OF INVENTION

Most contemporary web search engines function based on keyword searches. Keyword based search engines usually pull out websites with occurrence of one or more of the keywords specified in the search. However, searching based on keywords tend to produce large numbers of hits which burdens the user to ensure relevance of results. Users end up spending large amount of time going through the resulting websites to identify a relevant document. This becomes a huge drain on the user.

In addition to the problem of having a large volume of hits, the user also needs to internalize the relevant material in order to make it applicable and utilize it for the problem that is being solved.

With the current web browser technology, a user is not able to issue a natural language query as a keyword based search engine is not able to understand the context of the natural language query and is unable to search for relevant information effectively from the web.
Furthermore, the web is made up of a large amount and highly distributed information that may come from questionable sources. Therefore, the user must be able to discern a reliable website from an unreliable one.

Most documents available on the World Wide Web are not semantically enabled for computers to interpret and to process beyond simple keyword search. Human input is required to read the document and to interpret the information. This is a very slow and inefficient process made painfully obvious with the current information explosion online.

U.S. 6,044,375 describes a method of automatically extracting metadata from a document through a neural network to generate metadata guesses including word guesses, compound guesses and document guesses along with confidence factors associated with the guesses indicating the likelihood that each of the guesses is correct. However, the described invention does not look at the information on the website itself and does not extract data from the website itself.

Therefore, there is a need for a solution in order to search for relevant data, for users who are looking for specific and relevant data without having to spend too long manually reading through long lists of results.
SUMMARY OF INVENTION

Accordingly there is provided a semantic web constructor system supported by a plurality of web crawlers in a World Wide Web upon declaring a plurality of trustworthy websites, characterized in that, the system includes at least one web crawler controller engageable to manage the plurality of web crawlers, a semantic web database connectable to the plurality of web crawlers and a plurality of data building editors connectable to the at least one web crawler controller wherein a semantic browser is further connectable to the semantic web database to receive at least one natural language query from at least one user.

There is also provided a method of constructing a semantic web supported by a plurality of web crawlers in a World Wide Web upon declaring a plurality of trustworthy websites, characterized in that, the method includes the steps of crawling the web to select unprocessed websites, selecting websites based on trustworthiness calculated by comparing a trustworthiness numeric value to a predetermined threshold value, extracting a plurality of text from the selected websites, tokenizing the extracted plurality of text, applying a predetermined set of data transformation rules to the tokenized extracted plurality of text, converting the tokenized extracted plurality of text to metadata and storing the metadata in semantic web database.

There is further provided a method of querying a semantic web supported by a plurality of web crawlers in a World Wide Web upon declaring a plurality of trustworthy websites, characterized in that, the method includes the steps of receiving a query from a user, parsing the query into an internal representation format, searching the semantic web database using the internal representation format, ranking a plurality of websites based on trustworthiness and returning the ranked plurality of websites to the user.
The present invention consists of several novel features and a combination of parts hereinafter fully described and illustrated in the accompanying description and drawings, it being understood that various changes in the details may be made without departing from the scope of the invention or sacrificing any of the advantages of the present invention.
BRIEF DESCRIPTION OF THE DRAWINGS

The present invention will be fully understood from the detailed description given herein below and the accompanying drawings which are given by way of illustration only, and thus are not limitative of the present invention, wherein:

Figure 1 is a block diagram illustrating architecture of a preferred embodiment of a semantic web constructor system;

Figure 2 is flowchart illustrating a preferred embodiment of the steps of constructing a semantic web supported by a plurality of web crawlers; and

Figure 3 is a diagram showing an example of a natural language query using the preferred embodiment of a semantic web constructor method and system.
DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

The present invention relates to a semantic web constructor system supported by a plurality of web crawlers in a World Wide Web, upon declaring a plurality of trustworthy websites in order to query a semantic browser for a plurality of websites that are pageranked. Hereinafter, this specification will describe the present invention according to the preferred embodiment of the present invention. However, it is to be understood that limiting the description to the preferred embodiment of the invention is merely to facilitate discussion of the present invention and it is envisioned that those skilled in the art may devise various modifications and equivalents without departing from the scope of the appended claims.

The following detailed description of the preferred embodiment will now be described in accordance with the attached drawings, either individually or in combination.

The present invention provides a semantic web constructor system (100) supported by a plurality of web crawlers (112) in a World Wide Web upon declaring a plurality of trustworthy websites as seen in Figure 1. The system (100) includes at least one web crawler controller (110) engage able to manage the plurality of web crawlers (112). A semantic web database (116) is further connectable to the plurality of web crawlers (112). A plurality of data building editors (122, 124, 126) is connectable to the at least one web crawler controller (110) wherein a semantic browser (120) is further connectable to receive at least one natural language query from at least one user. A trust engine (118) is connectable to the at least one web crawler controller (110).

A plurality of data building editors (122, 124, 126) further include a website list editor (122), a rule editor (124) and a concept editor (126). However, it is to be appreciated by
one skilled in the art that the plurality of data building editors may also include other editors of a similar nature in other embodiments of the system (100). In order for the system (100) to be functional, the system (100) must first be provided with a plurality of relevant concepts, a predetermined set of transformation rules and an initial set of websites where relevant information associated with the relevant concepts are found. The concept editor (126) is used to specify a plurality of relevant concepts, properties and relationships in a subject area of interest. The concept editor (126) uses external resources such as Wordnet to expand on a specified concept. The rule editor (124) is used for defining data transformation rules. A website list editor (122) is used to specify an initial set of websites where data related to the subject area of interest may be found.

An example of a web crawler (112) is a spider as used in a search engine to retrieve information from the World Wide Web. Accordingly, an example of a web crawler controller (110) is a spider controller. The web crawler controller (110) as seen in Figure 1 filters websites to ensure that trustworthy websites are processed first by the plurality of web crawlers (112). Each web crawler (112) is assigned to a different website. Websites that are to be processed by the plurality of web crawlers (112) are delegated in a balanced manner by the web crawler controller (110). The web crawler controller (110) maintains a plurality of websites that are related to the subject area of interest. Initially, the websites are provided by the user. However, the plurality of websites increases as the web crawlers (112) encounter newly linked Uniform Resource Locators (URLs) on websites that are processed.

An intermediary database (114) is connectable to the plurality of web crawlers (112) and the semantic web database (116). An example of the intermediary database (114) is a merger database.
A method of constructing a semantic web supported by a plurality of web crawlers (112) in a World Wide Web upon declaring a plurality of trustworthy websites is described herein as seen in Figure 1. The method includes crawling the World Wide Web to select unprocessed websites using a plurality of web crawlers (110) such as spiders. An expansion of the initial plurality of relevant concepts is first performed, wherein each concept provided to the concept editor (126) is expanded to include a plurality of similar concepts. This step is carried out by means of an external lexical database such as Wordnet. Upon declaring the initial set of websites, the trust engine (118) determines trustworthiness of each declared website. Websites are selected based on trustworthiness calculated by comparing a trustworthiness numeric value to a predetermined threshold value. For example, the trustworthiness value is a numeric value from 0 to 100. The trust engine (118) accepts websites with a trustworthiness value that is higher than the predetermined threshold value. A trusted list of websites is created as shown in Figure 1.

A plurality of text is extracted from the selected trusted list of websites. It is to be understood that the plurality of text may also be extracted from documents found on the World Wide Web. Further, the extracted plurality of text is tokenized and a predetermined set of data transformation rules is applied to the tokenized extracted plurality of text. The tokenized extracted plurality of text is converted to metadata, such as Resource Description Framework (RDF) data and the RDF data is stored in a semantic web database (116) as seen in Figure 2.

Identification of new Uniform Resource Locators (URLs) on documents as found on the World Wide Web is carried out continuously. Upon encountering unprocessed URLs,
the plurality of web crawlers (110) will pass the unprocessed URLs to the web crawler controller (110) for processing.

Each web crawler (110) processes allocated web pages in a non-intrusive manner.

Figure 2 shows a method performed by a web crawler (112) to process a web page. A document or web page is classified to determine relevance of each document. Sentence extraction is performed by identifying a targeted sentence from the document or web page and extracting the sentence. The extracted sentence is then tokenized in order to apply a set of predetermined data transformation rules to the tokenized extracted sentence. The tokenized sentence is then converted to metadata such as RDF data. The RDF data is then stored in a web crawler's local semantic web crawler database (201).

In an event where there are no more URLs to be assigned to any web crawlers (112), an intermediary database (114) such as a merger database retrieves RDF data collected by each local semantic web crawler database (201) and merges the RDF data with the semantic web database (116).

Upon merging all RDF data into the semantic web database, the system (100) is ready to respond to natural language or structured natural language query processing. A method of querying a semantic web supported by a plurality of web crawlers (112) in a World Wide Web upon declaring a plurality of trustworthy websites is now described as seen in Figure 1. The method includes the steps of receiving a query from a user and parsing the query into an internal representation format. A semantic browser (120) is used to receive a query from the user and a query parser is used to parse queries entered by users. The semantic web database (116) is searched using the internal representation format. The internal representation format is then passed to a semantic
search engine that performs the search. The semantic web database (116) responds with a set of answers to the query by the user as well as references to a plurality of websites where the set of answers may be found. The plurality of websites is ranked based on trustworthiness by a page-ranker. The ranked plurality of websites is then returned to the user who issued the query.

An example of the described embodiment of system (100) and method is seen Figure 3. A typical website called "Bali Thai" is identified. The plurality of web crawlers (112) then process the website and construct RDF data as seen in Figure 3. A user may issue a query as follows:

QUERY: Find me a Thai restaurant that is Halal, not too expensive, no alcohol served, near Jurong Point Shopping Centre in Singapore.

The system (100) then responds with a reference to the website entitled "Bali Thai". It is to be appreciated that it is not possible to receive meaningful results such as those seen in this embodiment in a conventional search engine by issuing a natural language query of this nature. The meaningful results in this embodiment are specific websites that contains information as searched by the user.

Each individual web crawler (112) processes data extraction and data transformation of websites locally by using concepts and rules. Further, semantic web databases that are local, are created and then merged with global semantic web databases. This architecture allows the system (100) to be inherently scalable, which is a critical requirement for creating a semantic web database for the World Wide Web. Ontology based concepts and rules are used to automatically extract data of interest from websites. It is to be understood that the usage of websites in this description includes web pages, documents, messages and other information in a text format. Formats of all
concepts and rules used are compatible with World Wide Web Consortium (W3C) ontology standards in order to be created or edited using commercially available ontology editing tools or with a specialized editor.

The described method and system can be used to transform data of interest into RDF knowledge representation to enable downstream semantic search, knowledge discovery and process automation. Therefore, the described system and method is able to transform natural language queries into appropriate internal semantic query. This produces results that directly answer the query rather than relying on users to check each search result for relevancy. The described invention can be applied, but not restricted to, create a knowledge database for any domain from any data source that is unstructured, semantically unintelligible documents to transform said documents into a computer-understandable, semantic database to perform knowledge discovery and semantic information search efficiently and accurately.
CLAIMS

1. A semantic web constructor system (100) supported by a plurality of web crawlers (112) in a World Wide Web upon declaring a plurality of trustworthy websites, characterized in that, the system (100) includes:
   i. at least one web crawler controller (110) engage able to manage the plurality of web crawlers (112);
   ii. a semantic web database (116) connectable to the plurality of web crawlers (112); and
   iii. a plurality of data building editors (122, 124, 126) connectable to the at least one web crawler controller (110)
       wherein a semantic browser (120) is further connectable to the semantic web database (116) to receive at least one natural language query from at least one user.

2. The system (100) as claimed in claim 1, wherein a trust engine (118) is connectable to the at least one web crawler controller (110).

3. The system (100) as claimed in claim 1, wherein the plurality of data building editors (122, 124, 126) further include a website list editor (122), a rule editor (124) and a concept editor (126).

4. The system (100) as claimed in claim 1, wherein the plurality of web crawlers (112) is a plurality of spiders.

5. The system (100) as claimed in claim 1, wherein the at least one web crawler controller (110) is at least one spider controller.
6. The system as claimed in claim 1, wherein an intermediary database (114) is connectable to the plurality of web crawlers (112) and the semantic web database (116).

7. A method of constructing a semantic web supported by a plurality of web crawlers (112) in a World Wide Web upon declaring a plurality of trustworthy websites, characterized in that, the method includes the steps of:
   i. crawling the web to select unprocessed websites;
   ii. selecting websites based on trustworthiness calculated by comparing a trustworthiness numeric value to a predetermined threshold value;
   iii. extracting a plurality of text from the selected websites;
   iv. tokenizing the extracted plurality of text;
   v. applying a predetermined set of data transformation rules to the tokenized extracted plurality of text;
   vi. converting the tokenized extracted plurality of text to metadata; and
   vii. storing the metadata in semantic web database (116).

8. The method as claimed in claim 8, wherein the metadata used is Resource Description Framework (RDF) data.

9. A method of querying a semantic web supported by a plurality of web crawlers (112) in a World Wide Web upon declaring a plurality of trustworthy websites, characterized in that, the method includes the steps of:
   i. receiving a query from a user;
   ii. parsing the query into an internal representation format;
   iii. searching the semantic web database (116) using the internal representation format;
iv. ranking a plurality of websites based on trustworthiness; and

v. returning the ranked plurality of websites to the user.
Figure 1
Figure 2

Determine page relevancy
Extract sentences of interest
_tokenize sentences
Convert sentences into metadata RDF

Document Classifier
Sentence Extractor
Sentence Tokenizer
Metadata Generator

Web Page
RDF Data
Semantic Web Database

Figure 3

Find me a Thai restaurant that is Halal, not too expensive, no alcohol served, near Jorong Point Shopping Centre in Singapore.
### INTERNATIONAL SEARCH REPORT

**A. CLASSIFICATION OF SUBJECT MATTER**

Int. Cl.

G06F 17/30 (2006.01)  
G06F 7/00 (2006.01)

According to International Patent Classification (IPC) or to both national classification and IPC

**B. MINIMUM CLASSIFICATION SYMBOLS**

Facsimile No. +61 2 6283 7999

F. Telephone No: +61 2 6222 3664

Form PCT/ISA/210 (second sheet) (July 2009)

**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

<table>
<thead>
<tr>
<th>Category</th>
<th>Citation of document, with indication, where appropriate, of the relevant passages</th>
<th>Relevant to claim No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>US 7603350 B1 (GUHA) 13 October 2009 Whole document particularly: Fig.1, col.6 Ins.47-54; col.5 In.6 1-col.6 In.1 1, Figs.2-4; col.10 Ins.47-5 1; Abstract; col.9 Ins.11-13.</td>
<td>1-9</td>
</tr>
<tr>
<td>P, Y</td>
<td>US 2011/0022598 A1 (PENNACCHIOTTI et al) 27 January 2011 Whole document particularly: Abstract; [0028]; [0065], Fig.1 (130), [0084]; [0022]; [0085]; [0028].</td>
<td>1-9</td>
</tr>
<tr>
<td>Y</td>
<td>US 7 117 207 B1 (KERSCHBERG et al) 03 October 2006 Whole document particularly: col.10 Ins.13-23, Fig.9, col.4 Ins.7-10; col.2 Ins.64 - col.3 In.5; col.15 Ins.50-58.</td>
<td>1-9</td>
</tr>
<tr>
<td></td>
<td>WO 2002/03 1705 A1 (SCIENCE APPLICATIONS INTERNATIONAL CORPORATION) 18 April 2002 Whole document particularly: pg.9 Ins.1-10; pg.9 Ins.14-20; Fig.6 (650), pg.16 Ins.30 - pg.17 In.2; pg.11 Ins.10 - pg.12 In.15; pg.24 Ins.7-10; 14-2 1; pg.24 Ins.10-14</td>
<td>1-9</td>
</tr>
</tbody>
</table>

[ X ] Further documents are listed in the continuation of Box C  
[ X ] See patent family annex

* Special categories of cited documents:
  
  **"A"** document defining the general state of the art which is not considered to be of particular relevance
  
  **"E"** earlier application or patent but published on or after the international filing date
  
  **"L"** document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
  
  **"O"** document referring to an oral disclosure, use, exhibition or other means
  
  **"P"** document published prior to the international filing date but later than the priority date claimed

**Date of the actual completion of the international search**

05 September 2011

**Date of mailing of the international search report**

08.09.2011

**Name and mailing address of the ISA/AU**

AUSTRALIAN PATENT OFFICE

PO BOX 200, WODEN ACT 2606, AUSTRALIA

E-mail address: pct@ipaustralia.gov.au

Facsimile No. +61 2 6283 7999

**Authorized officer**

JOHN REISNER

AUSTRALIAN PATENT OFFICE

(ISO 9001 Quality Certified Service)

Telephone No: +61 2 6222 3664

Form PCT/ISA/210 (second sheet) (July 2009)
### DOCUMENTS CONSIDERED TO BE RELEVANT

<table>
<thead>
<tr>
<th>Category*</th>
<th>Citation of document, with indication, where appropriate, of the relevant passages</th>
<th>Relevant to claim No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>US 5642502 A (DRISCOLL) 24 June 1997 Whole Document</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>US 5694592 A (DRISCOLL) 02 December 1997 Whole Document</td>
<td></td>
</tr>
</tbody>
</table>
This Annex lists the known "A" publication level patent family members relating to the patent documents cited in the above-mentioned international search report. The Australian Patent Office is in no way liable for these particulars which are merely given for the purpose of information.

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>u u 7603350</td>
<td>NONE</td>
</tr>
<tr>
<td>u s 201 1022598</td>
<td>NONE</td>
</tr>
<tr>
<td>Us 7 117207</td>
<td>NONE</td>
</tr>
<tr>
<td>wo 0231705</td>
<td>AU 89194/01 US 6678679 US 200408321 1</td>
</tr>
<tr>
<td></td>
<td>US 6954750</td>
</tr>
<tr>
<td>u s 5642502</td>
<td>US 5893092 US 6088692</td>
</tr>
<tr>
<td>u s 5694592</td>
<td>US 5576954</td>
</tr>
<tr>
<td>u s 2005010553</td>
<td>US 6970860 US 20051 14325 US 7349895</td>
</tr>
<tr>
<td></td>
<td>US 7627556</td>
</tr>
<tr>
<td>u s 2007050338</td>
<td>CA 2621031 CN 101971 172 EP 1934825</td>
</tr>
<tr>
<td></td>
<td>JP 2009510547 KR 20080043865 US 7653617</td>
</tr>
<tr>
<td></td>
<td>US 2010125564 WO 2007027469</td>
</tr>
</tbody>
</table>

Due to data integration issues this family listing may not include 10 digit Australian applications filed since May 2001.

END OF ANNEX