Using Capture-Recapture Method for Web Intelligence

Siddharth Gupta[1], Narina Thakur[2]
[1,2]Department of Computer Science, Bharati Vidyapeeth's College of Engineering, New Delhi, India
(siddharth_bvcoe@hotmail.com, narinat@gmail.com)

Abstract – **Semantic Web is, without a doubt, gaining momentum in both industry and academia. With the emergence of Semantic Web framework the naïve approach of searching information on the syntactic web is cliché. Why do we think that the web would be improved, if it understood the meaning of its contents? Doesn't it understand it now? Google is very good at correcting typing mistakes, figuring out what I "meant" when I miss-typed a query or suggesting keyword to expand our search query. In this paper we redefine the idea of searching optimised information on the web using Capture-Recapture method. This paper also explains an optimised semantic searching of keywords and detailed explanation and simulation on Ontology of Indian Universities.**

Keywords **- Semantic Query, Ontology, XML, Keyword Searching, Optimisation, Capture-Recapture Method**

I. INTRODUCTION

The Semantic Web is an extension of the World Wide Web with new technologies and standards that enable interpretation and processing of data and useful information for extraction by a computer. In 1989 Tim Barners-Lee [3] invented the World Wide Web and marked the advent of Web 1.0 in which all the web pages were published with no user interaction. Web 2.0 based on equal user participation, collaboration; inter personal connectivity, interconnected applications [6]. Some of the applications are blogging, Facebook, Flickr, MySpace, Google, and YouTube where both producer and consumers can interact with each other [3]. The major drawback of Web 2.0 is its lack of interpretability between machines. Because of the lack of metadata and knowledge management crisis, powerful and complex algorithms are required by the search engines in order to parse the keywords requested by the user. The future web, semantic web is based on the principal of interoperability between machines and giving them power to think [5], aims at attaching metadata, specifying relations between web resources and knowledge management, in order to process and integrate data by the users. Hence semantic web is a web of databases and not of documents, queried by SPARQL. RDF attaches metadata specify relations between the resources based on XML. Ontologies are another major semantic web technology built above RDF aims at providing strong semantics and vocabulary [6]. They link the data on the basis of logical reasoning, common vocabulary and analytical thinking. This paper elucidates searching of information based on semantic understanding of words with simulation of Ontology of Indian Universities using protégé is a tool. An algorithm based on semantic searching of information is exemplified by an example. The main idea behind it is to collectively manage all the related information of Indian universities on a single platform which aggregates the information related to conferences, inter-college competitions, exchange of ideas, interaction among the students hence knowledge management and sharing on a single platform.

Our proposal in this paper redefines this approach of semantic searching using Capture-Recapture method. Using this method we can enhance the current searching techniques and form an intelligent web by assigning priority to the web pages based on the high probability.

Here capture-recapture method is used which is one of the methods for estimating the size of wildlife populations. An initial sample of the population in question is caught; its individuals marked and then released back into the wild, and a note taken of the number released. These marked individuals are allowed to become randomly dispersed throughout the population and then a second sample is taken. Its size and the number in it of marked, and hence recaptured, individuals is noted.The size of the population estimated on the principle that the proportion marked in the second sample equals the proportion of marked individuals in the population as a whole, so that

$$a/d = c/b \quad (1)$$

where a is the number marked and released into the population, b is the size of the second catch, c is the number recaptured in the second catch, and d is the size of the population as a whole, such that
Estimate of population size (d) = a.b/c

II. BACKGROUND STUDY

The semantic web which is in the early stages of development, has not reached the stage of pervasive web of distributed knowledge and searching based on intelligence. Success will be attained when a dynamic synergism can be created between people and a sufficient number of infrastructure systems and tools for the semantic web in analogy with those for the original web. Numerous parallel research works has been done in order to efficiently retrieve data from the web. Ontologies are another major semantic web technology built above RDF aims at providing strong semantics and vocabulary [5] [6]. They link the data on the basis of logical reasoning, common vocabulary and analytical thinking. The semantic web architecture is explained as follows. The first layer URI/IRI (Uniform Resource Identifier / International Resource Identifier) is a string of standardized form in order to uniquely identify resources/documents. URL (Uniform Resource Locator) is a subset of URI. IRI is an internationally accepted form. XML layer with the XML namespace and XML schema ensures that there is common syntax used in the semantic web. XML is the key for platform independence and exchange data using a common language.
The core format for representation of data in semantic web is RDF which is based on triples (subject – predicate – object)

-

978-1-4244-5967-4/10/$26.00 ©2010 IEEE

and forms a graph pertaining to the given data in the same form [7]. It is the grammar of the document whereas XML is the words understood by machines. OWL (Web ontology Language) uses description logics and provides strong semantics. For querying RDF and OWL Ontologies, SPARQL is used which acts as a query tool similar to SQL [8]. The correct logics with the rules implied proves the ontologies which then intertwined with trusted inputs to yield trusted outputs. Digital signatures and cryptography is applied in order to maintain security while information exchange.

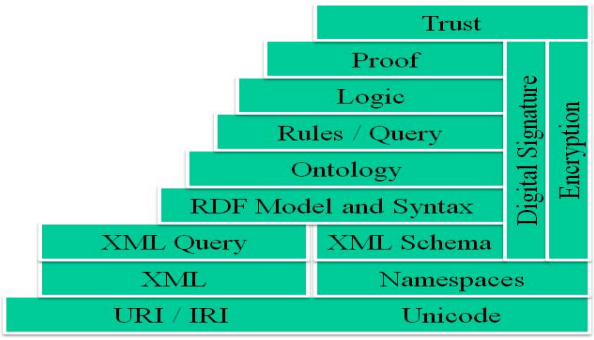

Fig 1. Semantic web architecture [7].

RDF

Resource Descriptive Framework (RDF) is the backbone of semantic web, which is the standard model for data interchange on the web. It is linked with the other resources via URI which uniquely identifies the documents. It works on the XML technology hence making data in each database understood by each other in XML form and hence platform independent. The RDF model is based on the idea of triple. It breaks down the data into three parts namely resource, property and property-value which are nothing but subject, predicate and object in simple grammar [7] [8].

OWL

OWL is a stronger language as compared to RDF and has much greater machine interpretability. Thomas Gruber defines it as "explicit specification of conceptualisation" [9], which are the relationships that concepts, objects and other entities hold with each other. According to Barry smith "the ontology is a classification of entities, represented by nodes in a hierarchal tree" [10]. The ontologies are derived form the real world conceptualisation shared by humans as a knowldege base and implemented in digital described through machine readable languages such as OWL, XML.

SPARQL

SPARQL (Simple Protocol and RDF Query Language) is similar to SQL and used to access native graph based RDF stores and extract data from the traditional databases, hence yielding perfect results.

Our previous work [1] exemplified intelligent expansion of query of searching related text information. The idea was to let user types a search query K. Then by looking at the texts found by means of K, other words related to K can be determined. Example if we type "Indian" as our search query then the computer will give us all the texts on the web containing the word "Indian" in their keysets. Now these keysets will also have other words related to "Indian". As the keywords of the same set are interrelated to each other, he proposed an algorithm to search related text information by searching those which have keywords in the keyword sets of the texts we already have [1].

The proposed algorithm in [1] is explained with the help of an example:

Input: Search: = {x1, x_2, x_3, x_4…. X_n}

Output: Set O: = {Keyword sets I, keywords P}

Procedure:

Let Set I: = Φ, Set P: = Φ

For every keyword set $Search_i$ in web database

If Search ∩ $Search_i \neq \Phi$ then add $Search_i$ to I;

Let I: = {$search_1$, $search_2$, $search_3$… $search_n$}

P: = $search_1$ ∪ $search_1$ ∪ $search_1$ ∪ …. $search_n$

C: = count number of keywords in Set P

Output: Keyword set I and keywords P

III. ONTOLOGY OF INDIAN UNIVERSITIES

Using protégé [12], which is a free, open source platform for constructing, visualizing and manipulating domain models and knowledge based applications with ontologies we have simulated ontology of Indian Universities. It implies that if different Indian colleges websites providing information about various events, services, information are published and shared under same ontology in terms they all use, then computer agents can share and aggregate information from these websites and hence provide a more viable solution to the user

query about Indian universities/ colleges. The domains of the ontology defined can be reused by other ontology, thus integrating several existing ontologies under one large ontology describing a large domain.

The name of my ontology representing unique URI:

http://www.indianuniversities/ourontology1.owl

Defining Classes:

Taking reference of Delhi University [2], we define three main classes namely Universities, Courses and States. The University class has been further subclassed into the Indian Universities along with their academics and student. In reference to our previous work [1] we have thoroughly expanded our ontology. The instances of these classes will be defined later.

The courses class contains a list of all the courses as its subclasses and states contain a list of states.

All the classes are subclassed under one class <owl:Thing> which is the root of all the classes.

The assertions are made using assertion browser and new instances of the classes can be created with the details according to the properties defined.

The following subclasses were made and its instances using instance browser, which shows a column wise creation of assertions of the classes:

TABLE 1
CLASS DISTRIBUTION

Sno	Column-Wise Distribution	
	Base Classes	Derived Classes
1	Univerisities	IP University, Delhi University, IIT
2	Courses	Commerce, Science, IT, Arts, Engineering, Law, Diploma, Language
3	States	New Delhi, Mumbai, Chennai, Kolkata
4	Academia	Colleges, Calender, Centres, Department, Faculties
5	Student	Examination, Alumni, Placement, Admission

The figure 2 shows the classes and the individual tree of the classes and its instances in a radial layout. Hence we can see the instances of the Universities class and has a subclass Colleges which has its own set of instances and how they are interconnected with the State and Courses classes all under one root <owl:Thing>.

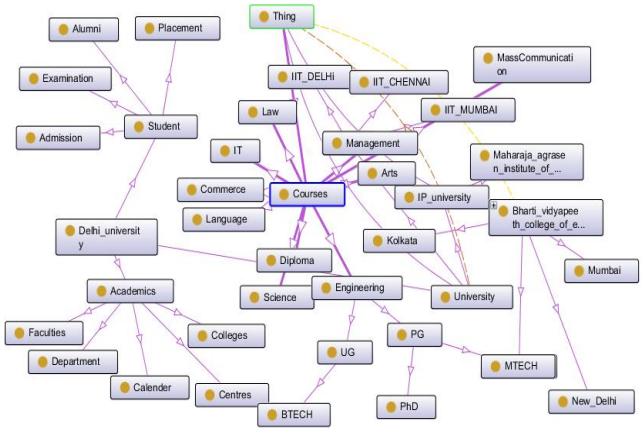

Fig 2. Radial layout

The results are hence plotted on the graphs using OWL Viz tool and following simulations were done which shows the interconnection between the classes and the domain defined. The fig. 3. Shows the class hierarchy which shows the base class as <owl:Thing> and all the classes subclassed under it.

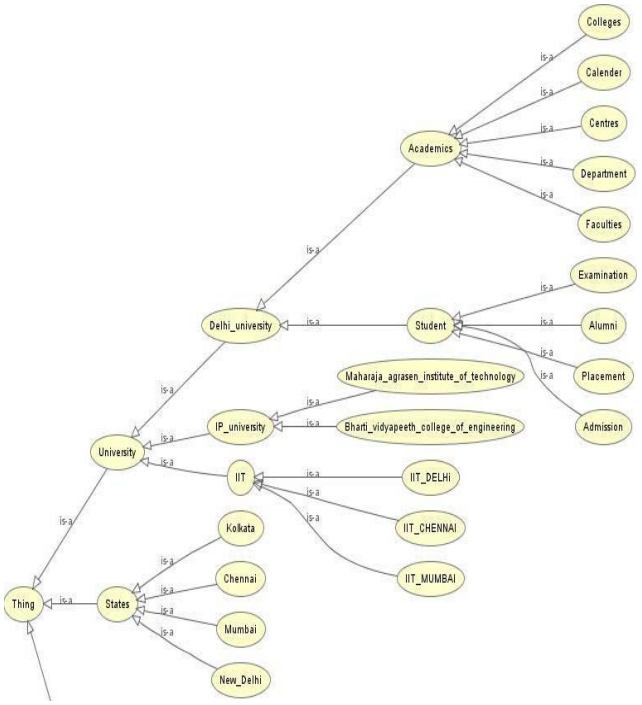

Fig 3(a)

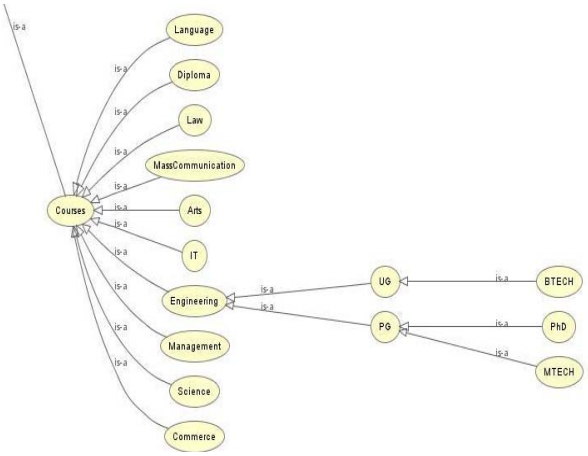

Fig 3(b)

Now taking an example of our search query:

Search Query= {"Indian Universities"}

Let Keyword sets obtained were :

$Search_1$ = {"Indian", "Courses", "IEEE", "Universities"}

$Search_2$ = {"Indian", "Conferences", "Universities"}

$Search_3$ = {"Universities", "Placements", "IEEE", "Indian"}

$Search_4$ = {"Indian", "States", "IEEE", "Universities"}

$Search_5$ = {"Indian", "Courses", "States", "Universities"}

Now we have a set P in which union of all the keysets are taken

P = { "Indian", "Universities", "Courses", "IEEE", "Conferences", "Placements", "States" }

Now as we see Indian Universities occur in every search hence the comparison is made with reference to these two keywords.

P= {Φ, Φ, 2, 3, 1, 1 ,1}

According to the observation probability of IEEE is more than other keywords hence, it is given higher priority. Though computer doesn't understand the meaning, but we can assign priorities in order to yield semantic search as keywords with higher priority occurs in greater number of keysets and implies that it is closely related to the search query.

Now, we can optimise search query by using Capture-Recapture method which is used for estimating the size of wildlife populations. One of the parameters, denoted by 'N' in, is the size of the population. We show that the estimate for the parameter 'N' that is obtained from the capture-recapture method is the value of the parameter 'N' that makes the observed data "more likely" than any other possible values of 'N'. Thus, the capture-recapture method produces the maximum likelihood estimate of the population size parameter 'N'.

Let's start with an example. In order to estimate the number of the web pages/ links which contain the keywords (as in search query "Indian Universities") in the web, a total of w= 250 keyset are found/captured and tagged and then released.

After allowing sufficient time for the tagged keywords, a sample of n =150 web pages/ links were found/caught which contain the query keyword. It was found that y= 16 web pages/ links were tagged. Estimate the size/ number of the web pages/ links in the web.

Let 'N' be the size of the web pages/ links which contain the keywords in the web. The population proportion of the tagged web pages/ links which contain the keywords is 16. The sample proportion of the tagged web pages/ links which contain the keywords is (y/n). In the capture-recapture method, the population proportion and the sample proportion are set equalled. Then we solve for 'N' [13].

$$\frac{w}{N} = \frac{y}{n} \Rightarrow N = \frac{wn}{y} = \frac{250(150)}{16} = 2343.75 = 2343 \quad (2)$$

Now, after w=250 web pages/ links which contain the keywords were captured, tagged and released, the population is separated into two distinct classes, tagged and non-tagged. When a sample of n = 150 web pages/ links which were selected without replacement, we let 'Y' be the number of web pages/ links which contain the keywords in the sample that were tagged. The distribution of 'Y' is the Hypergeometric distribution. The following is the probability function of 'Y' in [13].

$$P[Y = y] = \frac{\binom{w}{y}\binom{N-w}{n-y}}{\binom{N}{n}} \quad (3)$$

Note that (wn / y) is the estimate from the capture-recapture method. It is also an upper bound for the population size N such that the probability P(N) is greater than P(N-1). This

implies that the maximum likelihood estimate of N is achieved when the estimate is $\hat{N} = \frac{wn}{y}$. (4)

As an illustration, we compute the probabilities [13]

$$P(N) = \frac{\binom{250}{16}\binom{N-250}{150-16}}{\binom{N}{150}} \quad (5)$$

for several values of N above and below N=2343 . The following matrix illustrates that the maximum likelihood is achieved at N=2343.

$$\begin{pmatrix} N & P(N) \\ 2340 & 0.1084918 \\ 2341 & 0.1084929 \\ 2342 & 0.1084935 \\ 2343 & 0.1084938 \\ 2344 & 0.1084937 \\ 2345 & 0.1084933 \\ 2346 & 0.1084924 \end{pmatrix} \quad (6)$$

IV. FUTURE SCOPE

1. The current ontology can be reused and integrated in a large domain, hence expanding the knowledge base.
2. Information of all the universities and their corresponding colleges will be assorted under one ontology hence interoperability as well as interpretability.
3. The web will be a collection of databases with a common vocabulary for exchanging and interpreting information between machines.

V. CONCLUSION

This paper focuses on how the future web might look like and the interaction between the sources of information to yield perfect and real time results with unique power of intelligence by interpreting the best possible solution for the query. Ontologies are the bases of semantic web and can be further expanded. An example detailed ontology was simulated using protégé and results were analysed. We have also addressed Capture-Recapture method for estimating the Key sets probabilities in Semantic Searching Technique. This web searching technique by assigning priority to the web pages based on the high probabilities of the Search Keysets.

REFERENCES

[1] Siddharth Gupta and Narina Thakur *"Semantic Query Optimisation with Ontology Simulation", International Journal of Web and Semantic Technology, Vol. 1, No.4, ISSN: 0975-9026(Online) , 2010, pp 1-10.*

[2] DU's official website: http://www.du.ac.in/index.php?id=4

[3] Dean Allemang "*Rule-based intelligence in the Semantic Web",* proc. Second International Conference on Rules, and Rule Markup Languages of the Semantic Web, IEEE Computer Society, Athens GA, Nov 2006, ISBN: 0-7695-2652-7, pp. 83-88.

[4] Ora Lassila, James Hendler, *Embracing "Web 3.0" ,in IEEE Internet Computing Magazine, ISSN: 1089-7801, IEEE computer society, 2007, pp. 90-93,*

[5] Radha Guha, "*Towards the Intelligent Web Systems"*, proc. First International Conference on Computational Intelligence, Communications Systems and Networks, , Coimbator(India), IEEE 2009, pp. 459-463

[6] LI yuan, ZENG jianqiu, "*Web 3.0: A real personal web!" Beijing China , proc 3rd International Conference on Next Generation Mobile Applications, Services and Technologies.,* ISBN: 978-0-7695-3786-3, *IEEE 2009., pp. 125-128*

[7] *A Semantic Web primer* by Grigoris Antoniou and Frank Ven Harmelen, ISBN: 978-0-262-01242-3, MIT PRESS 2008.

[8] *XML Databases and Semantic Web* by Bhavani Thuraisingham, ISBN : 0-8493-1031-8, published 2002

[9] *Programming the Semantic Web* by Toby Segaran, Colin Evans, Jamie Taylor, ISBN: 978-0596-15381-6, O'REILLY 2009

[10] OWL:*http://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html*

[11] *A guide to Future of XML, Web Services and Knowledge Management* , by Michael C.Daconta, Leo J. Obrst, Kevin T.Smith, ISBN : 0-471-43257-1, published in 2003

[12] Protégé : *http://protégé.stanford.edu/download*

[13] Yashihiro Tohma et al "*The Estimation of Parameters of the Hypergeometric Distribution and its Application to Software Reliability Growth Model "in IEEE transactions on Software Engineering Vol 17, No. 5,1991* ISSN : 0098-5589 *,pp. 483-489*