P: ISSN NO.: 2394-0344

E: ISSN NO.: 2455-0817

Remarking

Vol-II * Issue-VII* December- 2015

Quest for Common-Sense Information Retrieval

Sandeep Kaushik
Assistant Professor,
Deptt. of I.T.,
L.N. Mishra College of Business Management,
Muzaffarpur

A.N. Jha
Ex-Faculty,
Deptt. of Maths,
L.N.T.College,
B.R.A. Bihar University,
Muzaffarpur

Ajay Kumar
Assistant Professor,
Deptt. of Management,
L.N. Mishra College of Business Management,
Muzaffarpur

Suket Jha
Deptt. of I.T.,
L.N. Mishra College of Business Management,
Muzaffarpur

Abstract

In last few decades the web has grown almost out of proportion. The expansion in the web has also expanded the bases of information and data. But with this growth it has enhanced the uncertainty of correct information retrieval also. Now these days we have become highly dependent on this monster of our creation for very simpler or critical needs of our day to day life. Creation of various information bases with global network succeeded in sharing information and data globally. Searching methods employed traditionally is now failing to search the correct information for correct search. The failure is due to various reasons. One of the greatest reasons for such failure is the absence of common sense schemes in information search. In this paper we are presenting the technological quest for common sense information retrieval.

Keywords: Semantic Web, Ontology, RDF, XML, Information Retrieval.

Introduction

A common sense is better defined as kind of sense which is most uncommon to the masses in general. When searched the meaning of common sense on google it first says that it is a noun and means a “good sense and sound judgement in practical matters”. Further it explains this as sensibleness, native wit, prudence, discernment, acumen, sharpness, wisdom, insight, and intuition and many more meanings. All this searching of meaning gives a chance to smile meaningfully. A very spontaneous urge simply flashes in a highly demanding mind that “why not it uses any dictionary for synonyms of the key words used in searching process of any kind”. At once it creates a fancy because it itself lacks this common sense of searching with synonyms in its searching of information or any data. A common user always believes that the information provided by these conventional search engines are perfect and true because it is from the widely used web. But the reality is just reverse. It simply scans through the available information on their resources and whatever looks similar, it presents that as result. All this is highly absurd and highly dangerous also for critical situations. It has been proven time and again that these conventional search engines don’t have the ability to understand a query with a common sense approach.

Problem Definition

The world is a busy place. It is not only complicated but also richly diverse, diverse in its heritage and history. Our ways of describing and understanding made it more complex due to differences in language and culture. We also know that the world is interconnected and interdependent. Effects of one change can be seen into each subtle may be with different meanings. World is constantly changing, so is our understanding of what exists in the world and how it affects and is affected by everything else. But knowledge is very similar to even in this nature of the world. We find that knowledge phenomenon can be found anywhere and everywhere. Knowledge is captured and codified in forms. It may be structured, semi-structured and unstructured. The structure of knowledge evolves with the incorporation of more information. Knowledge is nothing but a particular meaning and understanding of things with a particular context.

In the beginning what we used with our machines we called them data but with needs of the time by using some sophisticated designs of their storage we converted them into information providing such databases whom ultimately we called information systems. These information systems when matured become the knowledge hubs. Knowledge is always an element of meanings and understanding. Whenever we search anything on these information stations we actually search meaningfully. The search for meaning based information on our creations is the quest of common sense information retrieval.

An information retrieval seems to be very effortless task for common people either using the local database or the globally distributed database. In today’s era, information is growing at a fast pace and so as the database, and in such situation retrieving an appropriate data which seems easy for the common people is actually not so easy. Retrieving relevant data as per the user query depend on many critical points such as

1. How data are arranged in database?
2. Which indexing technique is being used?
3. How query is being processed?
4. Which searching/matching technique is being used?
5. And which ranking technique is being used?

The Quest Begins

It is a long saga of information retrieval needs. Before we go further it is pertinent to have a glance over the mess that all we have created to feed this monster similar to dinosaurs. One of the beliefs is that those giant creatures have ended due to their huge size and hunger. They have eaten all the resources around then and made them so bulky that they were unable to move for next course. Finally they started to eat them self's and ended their species forever. We have also made this present web almost similarly monstrous which is eating all the resources whatever we created for them. Let us have a look over the menu which we presented to feed the needs of web hunger.

Web has been created mostly much unplanned and much unorganised. It was either not co-ordinated also. Everything that we used with its creation was just to solve the then present problem with its functioning. Sometimes it feels that we have created even more tools than what we are getting the return from this. Some of the important tools and methods we employed in webbing all that we have can be listed as this. Let us have a brief of these.

Unicode

First among all is Unicode. Unicode is the basic universal number for every character, which works on multiple platforms. It is the basic notation that allows a single software, text or single character to be transported to other parts without corruption and re-engineering. By using Unicode to represent character and string data, we actually enable universal data exchange capabilities.

Uniform Resource Identifier

The Uniform Resource Identifier is termed as URI, which is a basic syntax for strings that is used to identify a resource. This term is used to addresses and names of objects or resources for web. A resource is any physical or abstract things in which each item has an identifier. The URI consists of two types namely 1.URL is the unique address for a file that is accessible on the web. The URL contains the name of the protocol to be used to access the file resource, a domain name that identifies a particular machine on the web, and a pathname (a hierarchical description that specifies the location of a file in that machine) 2. Uniform Resource Names (URN) gives a universal and persistent name about a resource in its namespace. This namespace guides the syntax of URN. It is used to identify resources on the web, every resource on the web should be uniquely identifiable.

Hyper Text Mark-up Language

HTML is the standard markup language used to create web pages. It is the most basic building block of a web page. It is a cornerstone technology used for interpreting and composing text, images and other material into visual web pages. It determines the content of a webpage, but not its functionality.HTML adds markup to Standard English. Hyper Text refers to links that connect web pages to one another, making the web what it is today. HTML is the language that describes the structure and the content of a web document. Using HTML static as well as dynamic web sites can be created.

Extensible Mark-up Language

The Extensible Markup Language (XML) is a W3C- recommended general-purpose markup language that defines a set of rules for encoding documents in a format which is both human readable and machine readable. XML is evolved from simplified subset of Standard Generalized Markup Language (SGML). Its main task is to facilitate the sharing of data across different information systems, particularly systems connected via the Internet. The design goals of XML are simplicity, generality and usability across the web. XML is the simplest way to send the document across the web to its specific format. It allows users to edit or modify it and again transfer it. Scientifically, XML is built upon Unicode characters and URI's.

XML Schema

An XML Schema describes the structure of an XML document. It is a document definition language that enables you to develop XML documents into a Specific vocabulary and a specific hierarchical structure. XML Schema is different to database schema, which defines the column names and data types in database tables. XML Schema allows the validation of instances to ensure the accuracy of field values and document structure at the time of creation. XML Schema secure data communication as the sender can describe data in a way that the receiver will understand.

Resource Description Framework

Resource Description Framework (RDF) is a standard model for data interchange on the web. RDF is world wide web consortium (W3C) standards designed as a metadata data model. The RDF model is often called a triple because it has three parts: subject, predicate, object.

Subject

This is the resource that is being described by the ensuing predicate and object.

Predicate

This is a function from individuals to truth-values with an parity based on the number of arguments it has.

Object

This is either a resource referred to by the predicate or a literal value.

The final outcome is known as statement. This is the combination of the three elements, subject, predicate, and object.

RDF provides interoperability between applications that exchange machine-understandable information on the Web. It defines the relationship between the resources on the web. RDF defines a simple, yet powerful model for describing resources. RDF extends the linking structure of the web to use URI to name the relationship between things. Using this simple model, it allows structured and semi-structured data to be mixed, exposed, and shared across different applications.

RDF Schema

The RDFS or RDF Schema is a knowledge representation language, providing basic elements for the description of vocabularies, intended to structure RDF resources. The RDF Schema layer is located above the XML layer, which provides more functions and capabilities than in XML schema. The data model of RDF schema allows creating classes of data. A class is defined as group of things with common characteristics. An object in the RDF schema is the instance of the class.

Now after all these discussions return to the issue of databases which are the kingpin of all the schemas. With all these tools and schemas, the main issue is what database should be used. The database is the ironic element in information retrieval, so special attention must be given to it. Earlier the database used was on the flat tabular basis, which was a first move in the organising of data, but it was unable to remove data redundancy. To overcome this bottleneck **Relational database** come into existence in 1970.It was invented by E.F.Codd.

In relational database, all data are stored in tables as the previous one but the striking point about this database is making relationships between each and every entity. Relational database overcome the problem of data redundancy using normalization. RDBMS supports for large database and is mostly used by all the information retrieval systems. But still it has many disadvantages

1. RDBMS still lag behind in providing efficient and effective integrated support for meaning based text searching within fields because of its rigidity. Most of the today's search engine relies on RDBMS and so if we search 'victory' and in table 'win' is defined then it will return 'no such result found'. It shows the lack of common sense also.
2. It represents everything in 2-D tabular form that's why it often provides poor support for storage of complex objects.
3. Until recently, it provides no support for complex data objects such as images, voice recognition.

The disadvantages of RDBMS was partially tried to be solved in Artificial Intelligence (AI). Apple's Siri which recognises speech and communicate with user and also perform actions as per the demand of user is made using AI. The artificial intelligence came into existence with a hope to give us machines capable of genuine human-level intelligence but till now it still seems quite an illusion.

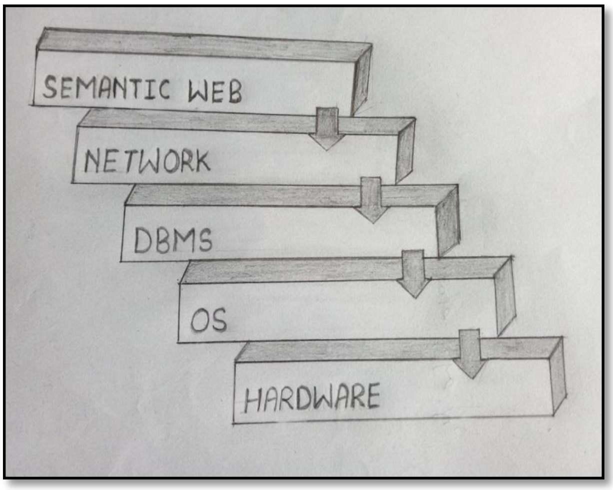
Sometimes it happens in the history of the mankind that for a time being we take very staunch moral ground for accepting any new technology. But with time when it gets settled, we accept it as part of our life. AI had also suffered from moral human cry in its beginning stage. It was said that when a machine will start working like human then it will become a hot potato for whole human species, as the importance of human will reduced by machines. Now AI has got settled in some of its proposed functional areas like voice recognition, it has become a part of our day to day life. But still the rigidity of the database remains as usual.

The context and meaning of the request made by the user is not analyzed to the full extent. So here the need arises for such a technique which enhances machines with such powers that enable them to respond for the queries using their common sense. In this quest for common sense information retrieval another technique which emerged is Semantic web**.**

Semantic Web

Semantic web is an extension of current web that allows the meaning of information to be precisely described in terms of well defined vocabularies that are understood by machines. This term was coined by Tim Berners-Lee. Due to its semantic searching which overcome the bottleneck of keyword searching, it is often called as next generation's web and web of linked data that can be processed by machines. The semantic web proposal is in fact again giving a chance to the practicing of artificial intelligence to prove that even in web searching technology a human intelligence alike understanding may be embedded. To implement of many of the claims that were given at the origin of artificial intelligence. The web is returning to the traditional grounds of artificial intelligence in order to solve its own problems. The first incarnation of the semantic web was meant to address this problem by encouraging the creation of ontology in the database so that machines can understand each other to fulfil the needs of the user by providing the right information.

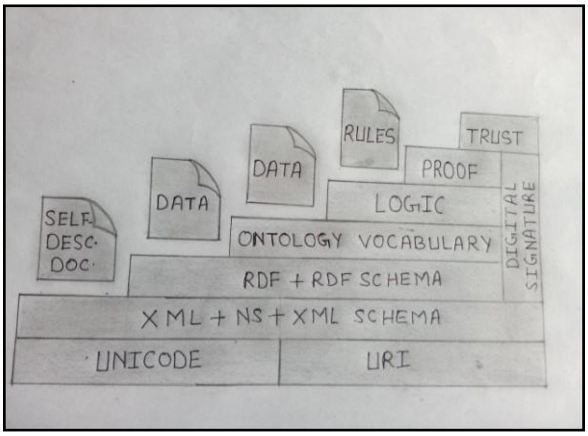
Semantic Web Design



Finding meaning is the main task of the semantic web. In order to achieve this objective, multiple lexicons with several representational structures are needed. They are illustrated in a simple structure.

Tim Berners

Lee explained them as a very complex format, which has following Layers.



The semantic web is an effort to build a new architecture to enhance the formal understanding of semantics by the available machines. At present mostly the web the content is made suitable for machine only, which should be actually suitable only for human. The efforts for semanticist will enable the intelligent machines to understand the web content, and produce an intelligent response to expected situations. In order to achieve that objective present effort can be divided into three parts.

1. Knowledge representation as XML, XML schema and RDF, RDF schema.
2. Ontology as ontological vocabulary.
3. Rules as facilitators'.

Semantic web is proposed to design as such that acknowledges the users query dynamically. The important property of the semantic web includes meaning understanding and machine useable metadata. Semantic web can be designed to learn on itself also. Learning may be based on semantic querying and navigation through different contents of the web and also by understanding the actual intension of the user. The semantic web can be exploited as a very suitable platform for implementing a self learning system, because it contains all means for learning like ontological development, ontology based annotation for learning, and their composition. According to Tim Berners Lee many versions of semantic web architecture can be designed to suit the purpose. Such versions may describe the languages needed for data interoperability between applications in the form of layering architecture.

Semantic web provides services to the upper layer where it needs a perfect abstraction for perfect functionality. But initial designs suffer from several deficiencies, to avoid these deficiencies a new design is proposed. In this design an extra layer is embedded which is called Rules. In semantic web design the challenge for the system engineers is to implement the proper integration of different layers.

Ontology

Ontology provides a mean for creating vocabulary of certain area and to define it formally. The term ontology can be defined as an explicit specification of conceptualization. The conceptualization means modelling certain domain and the ontology is used to describe important concepts of this domain, so it is the specification of this conceptualization.

Ontology is used at a stage where the vocabularies related to a specific domain are to be defined. It provides the capability to make analysis on the relationships between the different vocabularies to discover problems such as the existence of two vocabularies of the same meaning. In this stage, the relationships between vocabularies of a specific domain are created in hierarchal form by using the inheritance and classes concepts. Languages such as OWL (Web Ontology Language) which may be considered as a syntactic extension for RDF/RDFS, are provided at this stage. The main layer of semantic web architecture is ontology vocabulary, which typically consists of hierarchical distribution of important concepts in a domain, along with descriptions of the properties of each concept. Ontology play a pivotal role in the semantic web by providing a source of shared and precisely defined terms that can be used in metadata.

Ontology needs the support of languages like OWL. OWL is intended to be used when the information contained in documents needs to be processed by applications, as opposed to situations where the content only needs to be presented to humans. OWL can be used to explicitly represent the meaning of terms in vocabularies and the relationships between those terms. This representation of terms and their interrelationships is called ontology. OWL has more facilities for expressing meaning and semantics than XML, RDF, and RDF-S, and thus OWL goes beyond these languages in its ability to represent machine interpretable content on the Web. The OWL has been designed to meet the requirements of RDF, RDFS, XML Schema.

Conclusion

The aim this article was to present all that trials and triumphs that can shape the future of the web. Semantic web ontology seems to achieve a common and shared knowledge that can be used by people and different systems. Ontology play an important role in achieving interoperability across organizations and on the Semantic Web, because they capture domain knowledge and their role is to create semantics explicitly in a meaningful way and providing the basis for meaningful understanding between parties. Ontologism has become a popular research topic in many communities. We have described a reliable and an efficient system, which suggests the user all the effective details to know about an educational domain. It is reliable because though it can be inputted with synonymous words and misspell, it retrieves the similar result and does not provide an irrelevant results. it saves the user's inconvenience to move on to more pages to search for the more result. The system can be further refined of with more words in the search interface which can yield more filtration of the query result. The system can be better used with more performance indicators which can better model user requirements.

References

1. T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web", Scientific Am, May 2001.
2. D. Fensel, F.V. Harmelen, I. Horrocks, D.L. McGuinness, and P.F. Patel-Schneider, "OIL: An Ontology Infrastructure for the Semantic Web", Presented at IEEE Intelligent Systems, 2001.
3. R. Neches, R. Fikes, T.W. Finin, T.R. Gruber, R.S. Patil, T.E. Senator, and W.R. Swartout, "Enabling Technology for Knowledge Sharing", Presented at AI Magazine, 1991.
4. N.F. Noy, "Semantic Integration: A Survey Of Ontology-Based Approaches", Presented at SIGMOD Record, 2004.
5. Y. Qu, W. Hu, and G. Cheng, “Constructing virtual documents for ontology matching,” inProc. 15th Int. World Wide Web Conf., 2006.
6. http://www.semanticweb.org
7. http://www.w3.org
8. http:// Wikipedia.org