INTELLIGENT RETRIEVAL
Open Journal

Review

Moving Forward with Smart Web-Based Information Retrieval

Vandana Dhingra[1*]; Komal K. Bhatia[2]
[1]*School of Engineering and Technology, Apeejay Stya University, Gurgaon, India*
[2]*Associate Professor, CSE YMCA, UST, Faridabad, India*

[*]***Corresponding author***
Vandana Dhingra
School of Engineering and Technology, Apeejay Stya University, Gurgaon, India

Article information
Received: March 31st, 2023; **Revised:** April 30th, 2023; **Accepted:** May 15th, 2023; **Published:** July 22nd, 2023

Cite this article
Dhingra V, Bhatia KK. Moving forward with smart web-based information retrieval. 2023; 1(1). doi: https://doi.org/10.70705/ppp.ir.2023.v01.i01.pp33-36

ABSTRACT

As a repository of knowledge, the potential of the World Wide Web is almost limitless. Nevertheless, the difficulty of computers processing and integrating this data effectively, coupled with the fact that the WWW already connects over 15 billion sites, means that this promise remains mostly unrealized. One major challenge is the retrieval of useful information from the web. Any time someone does a search on the online, it leads them to hundreds of links that provide information on a certain subject; this is because the amount of information available on the World Wide online (WWW) has grown exponentially with the popularity of the internet, making it more difficult to extract important data. It is up to the user to choose how to extract valuable information. Finding out what is missing from current search methods and why the web needs an intelligent information retrieval system are the main focuses of this study. Metadata, RDF, URI, XML, XMLS, Triples, and Ontologies are some of the technologies and ideas that are presented in this article that are necessary for the realization of the semantic web.

Keywords
Semantic web; Intelligent retrieval; Triples; Ontologies; RDF; URI; XML; SWD.

INTRODUCTION

As far as human knowledge storage goes, the Internet is unrivaled. It has information on almost every topic conceivable, including papers and multimedia resources, and anybody with an Internet connection may access this material instantly. The decentralized nature of the Web is a major factor in its popularity. Many computers host web sites, and any document on the web may link to other papers on the same or other systems. This means that everyone, anywhere in the globe may contribute to the Web, which means that it can expand at an exponential rate as its usage becomes more widespread.

One downside of the Internet is that it is so massive. Finding relevant information is become more and more of a challenge as the amount of data accessible grows exponentially. Search engines like Google and Alta Vista, as well as directories like Yahoo!, may be helpful, but they are not foolproof. Still, finding the "right" document might seem like searching for a needle in a haystack for many people.
More than that, consumers want to do something when they use the Web; it's not enough to just find a content. A user may, for instance, be interested in locating the most affordable desktop computer, organizing and purchasing a trip to Goa, or reserving a reasonably cost Goa restaurant to accompany their planned evening movie. In order to do these activities, you will often need to visit many sites, integrate their material, and reason about them. Modern directories and search engines simply cannot handle this kind of work, but will they be able to in the future? The fact that computers weren't originally intended to handle data on the Web is the biggest problem. Web pages do not include any information that aids a machine in understanding the meaning of the content, however they do contain specific instructions that tell a computer how to display text or where to go when a link is clicked.

Therefore, for a computer to effectively interpret a web page, it has to comprehend the content. However, comprehending natural language is a notoriously difficult and unresolved subject.Researchers have also started to look at the possibility of creating a Semantic online by linking explicit meaning with online content further in that direction. This method asks writers to characterize their papers in a knowledge representation language instead of depending on NLP to glean this meaning from preexisting texts.

2. INCONVENIENT LOCATIONS THAT RESULT IN INTELLIGENT DATA RETRIEVALS

At the moment, all online searches rely on keywords, meaning that

data is accessed by searching text. that are accessible and connect to relevant websites. As a consequence, the user can see data that isn't relevant to their needs. Hyperlinks to online material all across the globe make resources available on the modern web. The machines do not comprehend these relationships, but they establish the physical connections. As a result, robots can't comprehend the linkages since there aren't enough associations that describe them. Data on the web must be connected in a manner that computers can understand. By adding computer-readable semantic annotations to online pages, the goal is to build a smarter web. Computers would provide better results if they could decipher the content of web pages.

An intelligent search engine needs to fetch the few pages from the billions available. The problem issues three topics:

1 Accuracy of keywords entered.

2 The way the search engines map the Net.

Ensuring that web pages are appropriately tagged to help search engines understand their content. We need smarter, faster, more meaningful online searches to replace the current, inefficient ones [15]. Annotating online resources using semantic markup that captures information about their meaning is the core principle that it is founded on. In order to facilitate advanced search capabilities for software agents that navigate between pages, we aim to establish a standard framework that enables data sharing and reuse across applications, enterprise, and community boundaries while also providing structure to the meaningful content of Web Pages.

I. EVOLUTION OF SEMANTIC WEB

The first article on the semantic web appeared in Scientific American in 2001 and traced its development from a web mostly consisting of material for human reading to one that computers could manipulate. The Semantic Web is an interconnected system of data that can be accessed and used in practical ways by applying a semantic theory to the interpretation of symbols. According to semantic theory, systems may communicate with one another via a "meaning" system that relies on the logical relationships between words.

Despite the article's assumptions to the contrary, it is still challenging to recruit data to a specific use context on the modern web. The 2001 proposal for the semantic web's architecture came from Tim Berners-Lee.

II. SEMANTIC WEB VS CLASSICAL WEB

Unlike traditional webs that rely on keyword connections, the Semantic Web is an advanced network of interconnected ideas. Typically seen as highlighted text on browsers, an HTML anchor tag (link) is a keyword reference to another document on the classical web. It provides a word or phrase that links to another page. However, it is unclear from the link precisely which paper is which. HTML hyperlinks aren't very helpful for figuring out how files relate to one another, and the link text isn't often clear either. The Resource Description Framework (RDF) is a new standard that allows for much more precise descriptions of relationships between entities. Because RDF allows for the linking of any things or ideas, it defines much more than just documents. That ideas, not documents, may be connected is the core principle of the Semantic Web.

III. PROBLEMS WITH CURRENT WEB

Despite the fact that the World Wide Web is fantastic and has brought many advantages that have revolutionized our society [3]. Unfortunately, the web as it now stands cannot meet the demands of modern computing in terms of robustness, scalability, and dynamic distribution. Information structure, an improved search mechanism, and the exposure of information semantics are the primary goals of new web technologies [4].

We need a new vision and architecture for the web because of the following constraints of our present web [3, 4].

Part A: Searching a Single Document [8]

The fact that information can typically be retrieved from a single web page and a single document is a major issue with the current state of information representation and retrieval on the Web [4]. It is also very difficult to retrieve information collectively that is spread out over multiple documents and web pages.

B. Search Is Constrained to Keywords Excluding Semantics

There are many instances where the search returns no results because the keywords we entered did not match any of the papers that were searched. This occurs, however, even when the sought-after material is included in the indexed papers; the problem is that the indexed articles may use a different language and terminology.

Section C: Superfluous and Unneeded Details

Furthermore, a keyword-based query often returns an overwhelming quantity of results, which is a major drawback of keyword-based searches. Sorting through the mountain of received data to find what you need is a tedious and error-prone process, and the vast majority of it is useless.

D. Representing Information in a Semi-structured Form

Our present-day web is not yet capable of supporting complex information representation due to its document-centric nature [3]. The majority of the data accessible on the internet nowadays is in an unstructured or semi-structured state. While the information is well-suited for direct human usage and information sharing on HTML-based free format web pages, they are completely unsuitable for automated information interchange, retrieval, and processing by software agents (machines).

IV. CONCEPTS USED TO MAKE SEMANTIC WEB A REALITY

A. Metadata

It is data about data, and it is used for searches all the time in the actual world. Consider the following scenario: you want to borrow some electronic books from the library. It is common practice for libraries to include a search function that lets patrons organize books

by various criteria such as author, title, topic, and more. Author, title, ISBN, publication date, and, most importantly, book location are just a few of the many helpful pieces of information included in this list. Metadata, which consists of information about information such as author, title, and topic, may be used to get the desired information, in this example, the book's location. Metadata, however, is superfluous since you may search the library's collection for a specific book by name. It is clear that this is a foolish approach. Metadata is also useful for purposes other than searching, but that is by far the most prevalent use case. Some other critical business-related data is stored behind the scenes as well.

Section B: RDF (Resource Description Framework)

A information processing system, Resource Description system (RDF) details the connections between resources, their attributes, and their values. The following principles form its foundation:

a. Resource: Anything may be specified using RDF expressions. Any kind of resource, including whole or partial web pages, has a unique Uniform Resource Identifier (URI).

b. A resource's property is what the World Wide Web Consortium calls "a specific aspect, characteristic, attribute, or relation used to describe a resource" (RDF Model and Syntax Specification, W3C, 2001). A property may also have other properties, which makes it a resource in and of itself.

c. Statements: Assets, attributes, and values are all part of a statement. The "subject," "predicate," and "object" are the three distinct components.

Example:

Statements can be represented as a graph in RDF.Consider a simple example:

B. TRIPLES

C. Each RDF encoded triple represents a simple sentence's subject, predicate, and object. The information is shown as subject-predicate-object triples, which may be written as . This means that subject s has a predicate, or connection, with object o. The component that designates the object of the statement is called the subject. In a declarative statement, the predicate is the component that describes the subject's attributes or properties. The component that specifies the worth of that attribute is called an object.

Section C. XMLS, RDF/XML, and XML (Extended Markup Language)

For the purpose of documenting and transferring relationships, RDF offers an XML-based syntax known as RDF/XML. XML allows users to construct their own tags with clear definitions; the language also facilitates the addition of arbitrary structure to documents without really conveying any content. A syntactic framework for data description is provided by the XML [7] standards.

D. Lexicons

When it comes to representing knowledge, ontologies are king because they let us logically link massive volumes of data [5]. Formal theories that facilitate the exchange and reuse of knowledge are known as ontologies. The semantics of semi-structured data may be more clearly represented using these. These provide for high-level automated assistance with data acquisition, maintenance, and access. In the context of information exchange, "ontology" refers to a description (similar to a formal program specification) of the ideas and connections that may exist for an individual or group of individuals [3]. Definitionally, ontology is a domain-specific, purpose-built model. The formal definition of ontology is a philosophy of logic. The terminology that agents use to communicate assertions and queries is defined by a shared ontology, which is important from a pragmatic standpoint.

As a consequence, OWL (Web Ontology Language) came into being. One way to describe concepts and the connections between them is using OWL, which is a collection of standardized XML components and properties. The main goal in creating it was to provide computers a standard method to read and interpret online material. In comparison to RDF, OWL is a more robust language that offers better machine compatibility and provides a schema.

REFERENCES

"The Semantic Web Revisited," by Tim Berners-Lee, Nigel Shadbolt, and Wendy Hall, published in May/June 2006 by IEEE.

"The future of search," by Ram Mohan Rao, published in DIGIT in August 2005.

"The Semantic Web: prospects and challenges," Michael Wilson and Brian, 2007 IEEE, http://ieeexplore.ieee.org/iel5/11087/35308/01678469.pdf?isnumber=35308&arnumber=1678469

"The Semantic Web: What, Why, How, and When," by Frank van Harelen and VrijeCS.VU.NL/~frankh/postscript/IEEE-DS04.pdf, published in March 2004 by IEEE Distributed Systems Online.

"A Survey on the Use of Ontologies in the Semantic Web" Avancha Sasikanth, Ding Li, Kolari Pranam, Zhongli Ding, Finin Tim, and Joshi Anupam "Semantic Web Services," by Sheila A. McIlraith, Tran Cao Son, and Honglei Zeng, published in 2001 by IEEE Intelligent Systems, may be found at http://ieeexplore.ieee.org/iel5/5254/19905/00920599.pdf. "Semantic Web Technology—the next generation of the Internet," C. Anantaram [7] CSI September 2006

In a semantic web context, Yoo, Myaeng, and Yun, Lee developed a "universal information retrieval system." 2005 at the IEEE

In "Semantic Web: Challenges and new requirements,"

A. Min, Andjomshoaa, Ferial, and Roland Paper presented at the 16th International Workshop on Database & Expert Systems Applications in 2005 by the IEEE Computer Society The paper "Characterizing the Semantic Web on the Web" was presented by Li Ding and Tim Finin in November 2006 at the 5th International Semantic Web Conference (ISWC'06). "Dependable Semantic

Web" by Bhavani, Eric, and Dock

CS 2002, IEEE Search on the Semantic Web, in IEEE Computer, 10(38):62-69, 2005, by Li Ding, Tim Finin, Anupam Joshi, Yun Peng, Rong Pan, and Pavan Reddivari.

"Semantic Web and Semantic Web Services" by Martin Hepp, published in March/April 2006 by the IEEE Internet Computer Society.

In "The Semantic Web—A guide to the future of XML, Web Services, and Knowledge Management," Michael C. Daconta, Leo J. Obrst, and Kevin T. Smith publish their work with Wiley Publishing, Inc.

"Swoogle: A Search and Metadata Engine for the Semantic Web" in Proceedings of the 13th ACM Conference on Information and Knowledge Management, 2004 by Li Ding, Tim Finin, Anupam Joshi, Rong Pan, R. Scott Cost, Yun Peng, Pavan Reddivari, Joel Sachs, and Vishal C. Doshi.