A Survey of Semantic based Solutions to Web Mining

K. Sridevi¹ and Dr. R. Umaram²

¹Assistant Professor,
Department of Computer Science, Nehru Memorial College,
Puthanampatti 621005, Trichy District, Tamilnadu, India.

²Associate Professor in CS,
Sri Sarada College for Women, Salem, Tamilnadu, India.

Abstract: An effective retrieval of the most relevant documents from the Web is difficult due to the large amount of information in all types of formats. Researches have been conducted on ways to improve the efficiency of Information Retrieval (IR) systems. To arrive to suitable solutions in IR systems, machines need additional semantic information that helps in understanding Web documents. This is made true by an intelligent web called the Semantic Web, which offers users the ability to work on shared meaningful knowledge representations on the web. Semantic Web makes the Web content meaningful to computers and it intends to support machine-processing capabilities. Using Semantic Web is a way to increase the precision of IR systems. This paper focuses on the various Semantic-based approaches in Web mining research.

Keywords: Information Retrieval, Semantic Web, Ontology, Search Results Clustering, WordNet, Personalized Search.

1. INTRODUCTION

Information Retrieval [1] is the technology for providing the required content based on the request from the user. Current information retrieval techniques are unable to exploit the semantic knowledge within documents and hence cannot give precise answers to precise questions. Using Semantic Web [2] aims at enhancing the ability of both people and software agents to find documents, information and answers to queries on the Web. Semantic Web which is meaningful web proposed by Sir Tim Berner’s Lee, which can play an important role towards the achievement a new Web architecture. The objective of this paper is to present the basic ideas on Semantic Web, Semantic Web architecture, various Semantic-based approaches to Web mining and some of the Semantic Web tools and languages.

1.1 Semantic Web

In general, semantics is the study of meaning. If a computer understands the semantics of a document, it doesn’t just interpret the series of characters that make up that document, but help to separate meanings from data, document content, or application code, using technologies based on open standards. The current WWW has a huge amount of data that is often unstructured and usually only human understandable. Semantic Web is an extension of current Web which offers to add structure to the present Web. The Semantic Web [3] aims to address this problem by providing machine interpretable semantics to provide greater machine support for the user. The effort behind the Semantic Web is to add semantic annotation to Web documents in order to access knowledge instead of unstructured material, allowing knowledge to be managed in an automatic way.

The goal of the Semantic Web is to develop enabling standards and technologies designed to help machines understand more information on the Web so that they can support richer discovery, data integration, navigation, and automation of tasks. The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries.

1.2 Semantic Web Architecture

The development of the Semantic Web proceeds in layers, one above another allowing for a more standardized way of developing. As it is being built on existing technology it allows developers to roll out parts of technology and implementing them without realizing the full capabilities of the Semantic Web. The functionality of each layer with reference to the above layered architecture is represented below with Semantic Web Layered Architecture [4].

![Figure 1: A Semantic Web Layered Architecture](image-url)
Uniform Resource Identifiers (URIs):
The development of the Semantic Web is heavily influenced by the fact that anyone can name or describe anything. To be able to describe things there needs to be a way to reference or identify them, both the current web and the Semantic Web use URIs for this task. The purpose of an URI is to unambiguously specify an identifier to represent a resource in a uniform way, identifying information representation constructs including classes, properties and individuals. As there is no ambiguity, it becomes possible to aggregate all data that refers to a given resource. It is the use of URIs that gives the Semantic Web a fundamental benefit over other technologies. URIs provides users and software to know exactly what it is they are being referred to, they are globally unique and each occurrence of the same identifier means the same thing. By having resources labeled in this way, makes it easier to integrate data sources that have been created independently.

XML (Extensible Mark-up Language):
The HTML program is not extensible. That is, it has specifically designated tags that require universal agreement before changes can be made. Web site developers had no way of adding their own tags, the solution was XML. It offered developers a way to identify and manipulate their own structured data. XML is a data model and uses a schema language (XML schema), to constrain the format, not the meaning of the data. The schema expresses shared vocabularies, define structure, content and semantics and will allow machines to carry out rules made by developers. Using metadata to describe what the data type is and the format it is in. The term metadata means data about data, the concept is to provide structured information that describes, locates and explains information resources making it easier for resources to be retrieved. This layer aims to be a baseline for structuring data on the web but without semantics. It is a mechanism used to describe data in a way that can be understood by the upper layers and can be interoperable.

Resource Description Framework (RDF):
The purpose of RDF is to provide a standard framework for making statements about resources and their attributes, making assertions about resources. RDF provides away to model information but does not provide a way of specifying semantics, or what the information means. There are several features to RDF:
• Statements are generic and can describe any domain
• RDF can be distributed (like HTML), allowing for growth of a knowledge base
• Statements can be exchanged by heterogeneous applications and interpreted without loss of meaning
• RDF enables the use of inference allowing queries to be answered
RDF and XML form the basic relational language layer of the Semantic Web architecture. To express a common meaning, a vocabulary needs to be developed and the Resource Description Framework Schema provides the platform for such a vocabulary.

Resource Description Framework Schema (RDFS):
RDFS is a schema language that provides basic structure by using classes and properties, the structures are formally defined and builds on the RDF foundation. The schema provides additional descriptive features and a language for describing the expanded vocabulary. It is a universal language that lets developers describe resources using their own vocabulary. By using RDFS the classes and properties can be arranged in generalization /specialisation hierarchies.
The function of RDF and RDFS is to provide metadata to upper technologies placed on the layers on the top, in which that metadata can be exchanged and reused between these technologies or between these technologies and other applications. The weakness in the expressive power of RDFS is what led to the development of more expressive languages for the Semantic Web.

Web Ontology Language (OWL)-The Ontology Layer:
With the influence of reasoning systems, Description Logics and web languages, the Web Ontology Language (OWL) was developed and can be used for defining ontologies. OWL has been built upon RDF and RDFS and has the same XML based syntax. OWL satisfies the Semantic Web’s requirements of providing minimal input from humans and supporting software requirements for a language with explicit meaning. OWL adds additional vocabulary to ontologies, extending RDFS with ontological constructs for describing object-oriented classes, properties and individuals. The ontology language uses RDF and RDFS, XML Schema data types and OWL namespaces.
The main function of layer is the provision of semantics which produces a web of meaning. Ontologies are helpful to clearly represent objects and also the relation ship between them it may be direct or inverse relationship. Using ontologies helps machines process meaning and facilitate sharing of information.

Rules, Proof & Trust Layer:
Rules Layer is supposed to be used as a framework for making new inferences how these inferences should be expressed for the implementation of the Semantic Web. Proof layer is incorporated to verify why the results generated by the agents should be believed or in other words the authenticity of the agent behavior is corroborated. Trust layer is to provide a mechanism for trust and confidence between Information sources and information users (man or machine).

Communicating Agent Layer:
This layer needs to perform the interoperability functions between various horizontal layers (Unicode to Proof) and the vertical layer crypto.
2. SEMANTIC-BASED APPROACHES TO WEB MINING

Semantic Web Mining aims at combining the two areas Semantic Web and Web Mining. The process of building the Semantic Web is currently an area of high activity. There are plenty of semantic solutions to mine the current Web. An overview of some of the Semantic-based solutions made by researchers has been represented below.

2.1 The Ontology Approach

2.1.1 An Ontology

Due to the unstructured and semi-structured nature of Web pages, it is a challenging task in categorizing and extracting content from the Web. In this ontology [5][6] plays a major role. Ontology being represented as a set of concepts and their inter-relationships relevant to some knowledge domain. The knowledge provided by ontology is extremely useful in defining the structure and scope for mining Web content. Ontology is defined as an explicit specification of a set of objects, concepts, and other entities that are presumed to exist in some area of interest and the relationships that hold them.

As implied by the above general definition, an ontology is domain dependent and it is designed to be shared and reusable. Usually, ontologies are defined to consist of abstract concepts and relationships (or properties) only [7].

The Semantic Web is an efficient way to represent data on the World Wide Web, or as a database that is globally linked, in a manner understandable by machines, to the content of documents on the Web. Semantic technologies represent meaning using ontologies and provide reasoning through the relationships, rules, logic, and conditions represented in those ontologies.

2.1.2 Ontology-based Web Mining

Overview of Web Mining:

Web Mining [8] refers to the discovery of knowledge from Web data that include Web pages, media objects on the Web, Web links, Web log data, and other data generated by the usage of Web data. Web Mining can be classified into (a) Web content mining, (b) Web structure mining and (c) Web usage mining.

Web content mining refers to mining knowledge from Web pages and other Web objects. Web structure mining refers to mining knowledge about link structure connecting Web pages and other Web objects. Web usage mining refers to the mining of usage patterns of Web pages found among users accessing a Website. Among the three, Web content mining is perhaps studied most extensively due to the prior work in text mining. The traditional topics covered by Web content mining include:

**Web page classification:** This involves the classification of Web pages under some pre-defined categories that may be organized in a tree or other structures.

**Web clustering:** This involves the grouping of Web pages based on the similarities among them. Each resultant group should have similar Web pages while Web pages from different resultant groups should be dissimilar.

**Web extraction:** This involves extracting HTML elements, term phrases, or tuples from Web pages that represent some required concept instances, e.g., person names, location names, book records, etc.

**Web Mining and Ontologies:** Semantic Web provides a very flexible framework for content based retrieval. Semantic web would serve as a good integration platform for content based retrieval.

**Ontology-based Web page classification:** Ontologies can be used as background semantic structures for Web mining. For example, instead of categorizing Web pages into categories, ontology-based Web page classification may classify Web pages as concept instances and Web page pairs as relationship instances. This allows Web pages to be searched using more expressive search queries involving search conditions on concepts and/or relationships.

**Ontology-based Web clustering:** ontology-based Web clustering can use HTML elements corresponding to concept instances as features to derive more accurate clusters.

**Ontology-based Web extraction:** In ontology-based Web extraction, one may address the problem of extracting both HTML elements as concept instances and finding related pairs of HTML elements.

**Ontology-based Web site structure mining:** ontology-based Web site structure mining can derive linkage pattern among concepts from Web pages for Website design improvements.

Apart from creation of ontology, the following operations can be done on ontology, on which researches have been made.

**Merge** of ontologies means creation of a new ontology by linking up the existing ones. Conventional requirement is that the new ontology contains all the knowledge from the original ontologies, however, this requirement does not have to be fully satisfied, since the original ontologies may not be together totally consistent. In that case the new ontology imports selected knowledge from the original ontologies so that the result is consistent. The merged ontology may introduce new concepts and relations that serve as a bridge between terms from the original ontologies.

**Mapping** from one ontology to another one is expressing of the way how to translate statements from ontology to
the other one. Often it means translation between concepts and relations. In the simplest case it is mapping from one concept of the first ontology to one concept of the second ontology. It is not always possible to do such one to one mapping. Some information can be lost in the mapping. This is permissible, however mapping may not introduce any inconsistencies.

Alignment is a process of mapping between ontologies in both directions whereas it is possible to modify original ontologies so that suitable translation exists (i.e., without losing information during mapping). Thus it is possible to add new concepts and relations to ontologies that would form suitable equivalents for mapping.

Refinement is mapping from ontology A to another ontology B so that every concept of ontology A has equivalent in ontology B, however primitive concepts from ontology A may correspond to non-primitive (defined) concepts of ontology B. Refinement defines partial ordering of ontologies.

Unification is aligning all of the concepts and relations in ontologies so that inference in one ontology can be mapped to inference in other ontology and vice versa. Unification is usually made as refinement of ontologies in both directions.

Integration is a process of looking for the same parts of two different ontologies A and B while developing new ontology C that allows to translate between ontologies A and B and so allows interoperability between two systems where one uses ontology A and the other uses ontology B. The new ontology C can replace ontologies A and B or can be used as an interlingua for translation between these two ontologies. Depending on the differences between A and B, new ontology C may not be needed and only translation between A and B is the result of integration. In other words, depending on the number of changes between ontologies A and B during development of ontology C the level of integration can range from alignment to unification.

Inheritance means that ontology A inherits everything from ontology B. It inherits all concepts, relations or axioms and there is no inconsistency introduced by additional knowledge contained in ontology A. This term is important for modular design of ontologies where an upper ontology describes general knowledge and a lower application ontology adds knowledge needed only for the particular application. Inheritance defines partial ordering between ontologies.

Applications of Ontology-based Web Mining:
Ontology-based Web mining, like traditional Web mining, is useful to many different applications. These applications are grouped under the following two classes.

Improved search to Web data: With additional ontological semantics, Web data can be indexed by their concepts and relationships to support expressive search queries. A more expressive query model can support very precise information search and reduce the amount of irrelevant Web information in the results [9].

Better browsing capabilities: Similar to searching, Web pages can be browsed based on their ontology concepts and relationships instead of following Web links only. If Web pages are the concept instances, relationship instances can be created as some virtual links between Web pages. Other than selecting Web pages belonging to concepts of interest, one can thus navigate the virtual links between Web pages enriching the browsing experience [10].

2.2 Semantic based Approaches to Web Search Results
Traditional Web (Web 1.0) is a web of documents. Finding documents is the main goal of information retrieval. There were some improvements in IR (Information Retrieval) on the Web since tf-idf (term frequency inverse document frequency) concerning using other information than just documents themselves. One of those approaches is analyzing link structure used in HITS and Google PageRank. Another approach may be using time metadata to enable filtering based on document publishing date as used e.g. in Google Blog Search. For example, To redefine the above, in [11] a semantic based approach has been used to discover semantically similar terms in documents and query terms in WordNet.

2.3 Web Search Results Clustering
Giving user a simple and uncomplicated web search result representation is an active area of Information Retrieval research. Traditional search engines use the hyperlink structure of the web to retrieve documents or pages and give them in a ranked fashion to the user. Retrieving relevant information from web, containing enormous amount of data, is a highly complicated research area. A landmark research that contributes to this area is web clustering which efficiently organizes a large amount of web documents into a small number of meaningful and coherent groups. Various techniques aim at accurately categorizing the web pages into clusters automatically. Various new techniques and algorithms have been proposed for grouping web search results into clusters to make them refined, meaningful, and relevant to the query [12][13].

2.4 Semantic based Personalized Search
Personalization aims to find a subset of Web data that matches the interest profile of a user or a group of users. This can be achieved by recommending Web pages or Websites to the users, or by filtering Web pages that are of interest to the users [14]. For example, this can done
by analyzing the historical data recording user accesses to Web data, and mining the topics relevant to a user by clustering previously accessed Web pages based on content similarities. When a new Web page is found to be similar to one of the clusters, it can be routed to the user. As Web pages are annotated with ontology entity labels, the grouping of Web pages accessed by a user can be more effectively done leading to more effective content recommendation.

Personalized search takes advantage of Semantic Web standards (RDF and OWL) to represent the content and the user profiles. Personalization of Web data access can be effectively used for improving the precision and recall in search, particularly by re-ranking the search results based on the learner's past activities.

3. SEMANTIC WEB & ONTOLOGY TOOLS & LANGUAGES

3.1 Markup Ontology Languages

Ontology Languages are formal languages used to construct ontologies. They allow the encoding of knowledge about specific domains and often include reasoning rules that support the processing of that knowledge. These languages use a markup scheme to encode knowledge, most commonly with XML.

Languages for representing ontology:

- DAML+OIL
- Ontology Inference Layer (OIL)
- Resource Description Framework (RDF)
- RDF Schema (RDFS)
- Web Ontology Language (OWL)

DAML+OIL is a successor language to DAML and OIL that combines features of both. In turn, it was superseded by Web Ontology Language (OWL). DAML may refer to DARPA Agent Markup Language, a markup language for the Semantic Web. DAML+OIL is a semantic markup language for Web resources. It builds on earlier W3C standards such as RDF and RDF Schema, and extends these languages with richer modelling primitives. DAML+OIL provides modelling primitives commonly found in frame-based languages.

RDF is a framework for representing information about resources in a graph form. It was primarily intended for representing metadata about WWW resources, such as the title, author, and modification date of a Web page, but it can be used for storing any other data. RDF Schema is a set of classes with certain properties using the RDF extensible knowledge representation language, providing basic elements for the description of ontologies, otherwise called RDF vocabularies, intended to structure RDF resources. These resources can be saved in a triplestore to reach them with the query language SPARQL, a protocol for accessing RDF data. SPARQL can be used for querying ontologies and knowledge bases directly as well. The Web Ontology Language (OWL) is a family of knowledge representation languages for authoring ontologies. The languages are characterised by formal semantics and RDF/XML-based serializations for the Semantic Web.

3.2 Ontology Development and Editing/Semantic Web Tools

The construction of an ontology itself is an ongoing research topic. The construction process can be manual with the help of some ontology editing tools [15] such as OntoEdit, OilEd, SWOOP (Semantic Web Ontology Overview and Perusal), Protégé, etc.

OntoEdit: OntoEdit [16] is based on IBM Eclipse framework. OntoEdit is a development environment for ontology design and maintenance. It supports multilingual development, and the knowledge model is related to frame-based languages. OntoEdit is based on an open plug-in structure. Data about classes, properties, and individuals may be imported or exported via different formats, such as RDF/RDFS, OWL and other formats.

OilEd: OilEd [17] is an ontology editor allowing the user to build ontologies using DAML+OIL, the language that inspire the actual OWL standard. The current versions of OilEd do not offer a full ontology development environment, but provides enough functionality to allow users to build ontologies. Data can be imported from DAML+OIL, OWL RDF/XML, and OIL text formats. OilEd can save ontologies as DAML+OIL documents only.

SWOOP: SWOOP (Semantic Web Ontology Overview and Perusal) [18] is a simple, scalable, hypermedia inspired OWL ontology browser and editor written in Java. Other familiar web-browser look and feel features include an address bar to load ontological entities, history buttons and bookmarks. SWOOP has been designed in keeping with the W3C OWL recommendations and has reasoning support (Pellet, an OWL inference engine). Another facility is the multiple ontology environment whereby entities and relationships across various ontologies can be seamlessly compared, edited and merged. All ontology editing in SWOOP is done inline with the HTML renderer, using different color codes and font styles to emphasize ontology changes, e.g. diverse representations for added, deleted or inferred axioms. Undo/redo options are provided with an ontology change log and a rollback option. SWOOP could import ontologies from OWL, XML, RDF and text formats. These formats could be used to save the edited ontologies. The overall tool architecture is based on MVC (Model-View-Controller) design pattern.
Protégé: Protégé [19] is a free, open-source Java-based platform that provides a growing user community with a suite of tools to build domain models and knowledge-based applications with ontologies. Protégé implements a rich set of knowledge-modeling structures and actions that support the creation, visualization, and manipulation of ontologies in various representation formats. Further, Protégé can be extended by way of a plug-in architecture and a Java-based Application Programming Interface (API) for building knowledge-based tools and applications. This application is written in Java and heavily uses Swing to create the rather complex user interface.

3.3 Ontology Repositories

Although the Web improves the visibility of centralized ontology development, it is hard to achieve a universal ontology for everything due to huge space complexity. Hence, distributed ontology development is preferred in the Semantic Web, i.e., small ontologies are authored by different sources in an incremental fashion. To reuse existing ontologies, effective web-based tools are in great need to browse, search and navigate distributed ontologies. The technical highlights of some of the popular repositories for publishing and searching ontologies on the Web are detailed below.

DAML Ontology Library: DAML Ontology Library[20] indexes user submitted ontologies and provides browse/search services. It organizes ontologies by their URI, users’ annotations supplied during ontology submission (e.g. submission date, keyword, open directory category, funding source, submission organization), the defined class/property, or the used namespace. Users can run sub-string queries over a defined class/property.

SchemaWeb: SchemaWeb [21] provides services similar to DAML ontology library with better human/machine user interface (i.e. both HTML and web service interface). It adds more services: (i) for human user, it provides fulltext search service for indexed ontologies, and a customizable resource search interface by letting users specify triple pattern; (ii) for machine agents, it searches the “official” ontology of a given namespace or the resource with user specified triple pattern; it also navigates RDF graph through RDFS properties (i.e. subClassOf, subPropertyOf, domain, range), and publishes RSS feeds about new ontology submissions.

W3C’s Ontaria: W3C’s Ontaria [22] stores RDF documents (including ontologies) and provides search/navigation services in the repository. It allows a user to (i) browse a RDF file as a list of triples, a list of used properties, or a list of populated classes, and (ii) browse relations between RDF files.

Semantic Web Search: Semantic Web Search [23] provides an object-oriented view of the Semantic Web, i.e. it indexes instances of well-known classes including rdfs:Class, rdf:Property, foaf:Person, and rss:Item. It partially supports ontology search by finding instances of rdfs:Class and rdf:Property; however, its search results are biased to terms from the namespace of WordNet.

Swoogle: Swoogle [24] indexes millions of Semantic Web documents (including tens of thousand of ontologies). It enables users to search ontologies by specifying constraints on document metadata such as document URLs, defined classes/properties, used namespaces, and RDF encoding. Moreover, it provides detailed metadata about ontologies and classes/properties in an object-oriented fashion. It has an ontology dictionary that enables users to browse the vocabulary (i.e. over 150KB URIrefs of defined/used classes and properties) used by SemanticWeb documents, and to navigate the SemanticWeb by following links among classes/properties, namespace and RDF documents. In addition, it is powered by automatic and incremental Semantic Web document discovery mechanisms and updates statistics about the use of ontologies in the Semantic Web on a daily basis.

3.4 Ontology Language Processors/Frameworks

An ontology construct conveys descriptive semantics, and its actionable semantics is enforced by inference. Hence, effective tools, such as parsers, validators, and inference engines, are needed to fulfill the inferencability objective. The following are some of the popular tools for processing semantic web ontology processors which support the actionable semantics.

Jena: Jena [25] is a popular open-source. It provides sound and almost complete inference support for OWL. Current version of Jena also partially supports OWL inference and allows users to create customized inference engines. Apache Jena is a Java framework for building Semantic Web applications. Jena provides a collection of tools and Java libraries to help to develop semantic web, tools and servers.

The Jena Framework includes:

- an API for reading, processing and writing RDF data in XML, N-triples and Turtle formats
- an ontology API for handling OWL and RDFS ontologies
- a rule-based inference engine for reasoning with RDF and OWL data sources
- stores to allow large numbers of RDF triples to be efficiently stored on disk
- a query engine compliant with the latest SPARQL specification
servers to allow RDF data to be published to other applications using a variety of protocols, including SPARQL

**Racer**: Racer [26] is a description logic based reasoner. It supports inference over RDFS/DAML/OWL ontologies through rules explicitly specified by the user.

**Pellet**: Pellet [27] is a ‘hybrid’ reasoner that can deal both TBox reasoning as well as non-empty ABox reasoning. It is used as the underlying OWL reasoner for SWOOP ontology editor and provides in-depth ontology consistency analysis.

4. CONCLUSION

Semantic-based Web data mining is a combination of the Semantic Web and Web mining. Web links documents to documents whereas Semantic Web links data to data. Web mining results help to build the Semantic Web. Semantic Web supports Universal data representation (using RDF), Reusable data models (using RDF, RDFS, and OWL), W3C Standard query language (SPARQL). Information validation and classification (using Reasoners). In Semantic Web, Ontology plays a major role. Ontologies offer an efficient way to reduce the amount of information overload by encoding the structure of a specific domain and offering easier access to the information for the users. The Semantic Web research extends to improve Ontology modeling, reuse methodologies and methods, Ontology extraction, comparison, mapping, merging, evaluation and reliability measurement. Semantic Web knowledge management deals in improving classification, clustering, searching, content creation and annotation of Semantic Web data. The knowledge of Semantic Web makes Web mining easier to achieve and also can improve the effectiveness of Web mining.

REFERENCES


[17] OilEd: http://oiled.man.ac.uk/


[22] http://www.w3.org/2004/ontario/
   r.f.moeller/racer/