Volume 3, Issue 4, April 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Exploiting Semantic Structures for Building the Semantic Web

R.Preethi
Dept. of CSE
Bharath University, India.

C. Anuradha
Dept. of CSE
Bharath University, India.

Abstract- Semantic Web Mining aims at combining the two fast-mounting research areas Semantic Web and Web Mining. This project analyzes the junction of trends from both areas. The consequences of web mining can be enhanced by exploiting semantic constitutions in the web and the web mining techniques are used to construct the semantic web. These web mining techniques can be used for mining the Semantic Web itself. The Semantic Web is enriched by machine-processable information which supports the user in his tasks. It aims at converting a web dominated by unstructured and semi structured data into a web of data. Given the enormous size even of today's Web, it is unfeasible to physically enrich all of these resources. Therefore, computerized schemes for eruditing the relevant information are increasingly being used. Apache jena framework is anticipated for the automatic interpretation in semantic web mining. It provides an automatic interpretation and a common framework that allows data to be shared and reused across application, venture and community limitations. Duplication of data can be avoided and it provides more accurate and meaningful data for an user query.

Keywords- Web Mining, Semantic Web, Apache jena, Knowledge Discovery, World Wide Web.

1. Introduction

The two fast-developing research areas Semantic Web and Web Mining build both on the success of the World Wide Web. They complement each other well because they each address one part of a new challenge posed by the great success of the current world wide web. The temperament of most data on the Web is so unstructured that they can only be interpreted by humans, but the amount of data is so huge that they can only be processed efficiently by machines. The Semantic Web addresses the first part of this challenge by trying to make the data machine explicable, while Web Mining addresses the second part by semi-automatically extracting the useful knowledge hidden in these data. Semantic Web Mining aims at combining the two areas Semantic Web and Web Mining. This revelation improves the results of Web Mining by exploiting semantic structures in the Web, and make use of Web Mining techniques for building the Semantic Web. These techniques can be used for mining the Semantic Web itself. The wording Semantic web Mining emphasizes this spectrum of possible interaction between both research areas. It can be read both as Semantic Web Mining and as Semantic Web Mining. Users rely on the semantic information in transcript corpora that is implicitly exploited by geometric methods. Some methods also examine the structural characteristics of data; they profit from standardized syntax like XML. Here, we concentrate on markup and mining approaches that refer to an explicit conceptualization of entities in the respective domain. These relate the syntactic tokens to background knowledge represented in a model with formal semantics. When we use the term "semantic", we thus have in mind a formal logical model to represent knowledge. The aim of this paper is to give an overview of where the two areas of Semantic Web and Web Mining meet today. We will first describe the current state of the two areas and then confer, using an example, their grouping, thereby outlining future enhancement. We will provide references to typical approaches. Most of them have not been developed unambiguously to close the gap between the Semantic Web and Web Mining, but they robust naturally into this scheme. Web mining techniques can be applied to help creating the Semantic Web. This is not a scalable solution for a wide-range application of Semantic Web technologies. Recent developments include the mining of sites that become more and more Semantic Web sites and the development of mining techniques that can tap the expressive power of Semantic Web knowledge representation.

2. Related Work

The current research proposes a apache jena framework that exploits knowledge-based semantic information to integrate text-summarization and web page-segmentation technologies, thus improving the overall approach effectiveness. The following sections overview the state of apache jena framework, text summarization and web mining.

2.1 Apache Jena

Apache Jena is a Java framework for building Semantic Web applications. Jena provides a compilation of tools and Java libraries to help you to develop semantic web and connected data apps, tools and servers. The Jena structure includes an API

for understanding, processing and writing resource description framework data in XML, N-triples and Turtle formats, an ontology API for handling web ontology language and resource description framework ontologies, a rule-based inference engine for reasoning with resource description framework and web ontology language data sources, stores to admit large numbers of RDF triples to be efficiently stored on disk, a query engine acquiescent with the newest SPARQL specification and servers to allow RDF data to be published to other applications using a variety of protocols, including SPARQL.

2.2 Components of Apache Jena

The components of the apache jena are RDF, SPARQL, inference, SQL database.

RDF:

RDF is a Resource Description Framework, which is used to transfer data underlying any platform. RDF has an XML syntax and many who are familiar with XML will think of RDF in terms of that syntax. But RDF should be interpreted in terms of its data model.

SPARQL:

SPARQL is used to querying and updating RDF models. It's a query language used to process data to and from RDF model. As a query language, SPARQL is "data-oriented" in that it only queries the information held in the models; there is no inference in the query language itself. Of course, the Jena model may be 'smart' in that it provides the impression that certain triples exist by creating them on-demand, together with OWL analysis. SPARQL does not do anything other than take the description of what the application wants, in the form of a query, and proceeds that data, in the form of a set of bindings or an RDF graph.

Inference:

Inference is heart of the system which contains rule engine and inference algorithms to derive consequences from RDF models. The rule engine has set of rules which must be compromised by the query. The Jena inference subsystem is designed to allow a range of inference engines or reasoners to be inserted into Jena. Such engines are used to develop additional RDF assertions which are entailed from some base RDF together with any optional ontology information and the axioms and rules associated with the reasoner. The major use of this method is to support the use of languages such as RDFS and OWL which allow additional facts to be inferred from instance data and class descriptions.

SQL Database:

It is a query language used to construct the jena model. SDB is a component of jena for RDF storage and query specifically to support SPARQL. The storage is provided by an SQL database and many databases are supported, both Open Source and ownership. An SDB store can be accessed and managed with the provided command line scripts and via the Jena API.

Text Summarization:

A summary is a text produced by one or more other texts, expressing significant information of unique texts, and no longer than half of the original texts. Actually, text summarization techniques intend to diminish the reading exertion by maximizing the information density that is prompted to the reader. Summarization techniques can be categorized into two approaches: in extractive methods, summaries stem from the accurate extraction of words or sentences, whereas abstractive methods create unique summaries by using natural language generators.

Web Mining:

Web mining is the use of data mining techniques to automatically discover and extract information from web documents and services. The application areas include resource finding, information selection, generality and data analysis. Web mining includes three main sub-areas: web content mining, web structure mining, and web usage mining. The former area covers the analysis of the contents of web resources, which in general encompass dissimilar data sources such as texts, images, videos and audio, metadata and hyperlinks are often classified as text content.

3. Experimental Result

The framework can effectively tackle the web mining task when the input was a news-text, which mainly dealt with a single event. A web page, however, often collects diverse textual resources, each describing a specific, homogenous set of topics. Hence, the apache jena framework was designed to evaluate the ability of the proposed framework to identify the most informative subsections of a web page. A set of new documents were generated by assembling the news originally provided. Each new document eventually included four news articles and covered four different topics. Then, the list of credentials was processed by the proposed framework, which was expected for each manuscript to select as the most relevant topics those that were chosen in the set up. As a result, one can conclude that the performances attained by the framework in terms of ability to identify the relevant topics in an heterogeneous document are very promising.

4. Conclusion

This project introduces a framework that can effectively support advanced Web mining tools. The proposed system addresses the analysis of the textual data provided by a web page and exploits semantic networks to achieve multiple goals:
1) the recognition of the most significant topics;
2) the assortment of the sentences that better correlates with a given topic;
3) the automatic summarization of a textual resource.
4) the automatic interpretation of a textual content.
The eventual framework exploits those functionalities to tackle text summarization. The semantic characterization of text is indeed a core aspect of the proposed methodology, which takes advantage of an theoretical representation that expresses the informative content of the basic textual resource on a cognitive basis. The present approach, though, cannot be categorized under the Semantic Web area, as it does not depend on semantic information already embedded into the Web resources. In the proposed methodology, semantic networks are used to characterize the content of a textual resource according to semantic domains, as opposed to a predictable bag of words. Experimental results proved that such an approach can yield a coarse-grained level of sense distinctions, which in turn favors the identification of the topics actually addressed in the Web page. In this regard, investigational results also showed that the system can emulate human assessors in evaluating the relevance of the single sentences that compose a text. A future direction of this research can be the integration of the content-driven segmentation approach with conventional segmentation engines, which are more sloping toward the analysis of the inherent structure of the Web page.

5. Future Enhancement

Future works may indeed be focused on the integration of semantic orientation approaches into the proposed framework. These techniques are fetching more and more important in the web, where one may need the automatic analysis of fast-changing web elements like customer reviews and web status data. In this regard, the present structure may provide content-filtering features that support the selection of the data to be analyzed.

References

[1] S. Acharyya and J. Ghosh. Context-sensitive modeling of web-surfing behaviour using concept trees.
[2] C.C. Aggarwal. Collaborative crawling: Mining user experiences for topical resource discovery.
[3] C.C. Aggarwal, F. Al-Garawi, and P.S. Yu. Intelligent crawling on the world wide web with arbitrary predicates.
[4] C.C. Aggarwal, S.C. Gates, and P.S. Yu. On the merits of building categorization systems by supervised clustering.
[5] J. Allan, editor. Topic Detection and Tracking: Event-based Information Organization.
[6] S.S. Anand, M. Mulvenna, and K. Chevalier. On the deployment of web usage mining.
[7] C.R. Anderson, P. Domingos, and D.S. Weld. Relational Markov models and their application to adaptive web navigation.
[8] Pascal Auillans, Patrice Ossona de Mendez, Pierre Rosenstiehl, and Bernard Vatant. A formal model for topic maps.
[9] P. Baldi, P. Frasconi, and P. Smyth, editors. Modeling the Internet and the Web. Probabilistic Methods and Algorithms.

