Semantic Web Mining of Un-structured Data: Challenges and Opportunities

Manoj Manuja manoj_manuja@infosys.com
Principal, Education & Research Dept,
Infosys Technologies Ltd.,
Chandigarh, India

Deepak Garg dgarg@thapar.edu
Thapar University,
Patiala, India

Abstract

The management of unstructured data is acknowledged as one of the most critical unsolved problems in data management and business intelligence fields in current times. The major reason for this unresolved problem is primarily because of the actuality that the methods, systems and related tools that have established themselves so successfully converting structured information into business intelligence, simply are ineffective when we try to implement the same on unstructured information. New methods and approaches are very much necessary. It is a known realism that huge amount of information is shared by the organizations across the world over the web. It is, however, significant to observe that this information explosion across the globe has resulted in opening a lot of new avenues to create tools for data management and business intelligence primarily focusing on unstructured data. In this paper, we explore the challenges being faced by information system developers during mining of unstructured data in the context of semantic web and web mining. Opportunities in the wake of these challenges are discussed towards the end of the paper.

Keywords: Semantic Web, Web Mining, Unstructured Data.

1. INTRODUCTION

The last few years have seen growing recognition of information as a key business tool for the success of the organizations across the world. The organizations which effectively identify, accumulate, study, scrutinize and thereafter act upon the information are definite winners in this new “information age”. Further to this, the realization of “web” has critically changed the perspective of how the organizations extract information from the available data in today’s world of dynamic business.

Therefore, the most important differentiator between a successful and an unsuccessful business is how an organization manages its data. The critical aspect in today’s business scenario is how data is converted into Information and subsequently how information is converted into knowledge.

2. BUSINESS DILEMMA IN A LARGE ORGANIZATION

It is very crucial to extract knowledge from un-structured data which is available in various formats and generated by heterogeneous sources across a big organization.

According to projections from Gartner, professionals will spend anywhere between 30 to 40 % of their office time in managing various documents, which is 20% more than what they used to spend on similar activities 10-15 years ago. Similarly, Merrill Lynch has anticipated that data which is unstructured will amount to more than 85 percent of all information available in a company.

It is very easy to extract useful knowledge from structured data using proven algorithms and patterns. But the problem comes when we have unstructured data to work with. It becomes very difficult to extract knowledge from the un-structured data because of non-availability of proven algorithms, schemas, patterns and information systems. Through this paper, we shall explore various challenges in the field of unstructured data mining using semantic web techniques and also available opportunities in the wake of these challenges.

3. DATA, INFORMATION AND KNOWLEDGE

In a practical real life scenario, data is available in three forms:

- **Unprocessed data** which is gathered in real time,
- **Extracted data** which gives us information and
- **Processed data** which provides us useful knowledge [1].

Knowledge is being used to determine and analyze the specifics of a given situation. Knowledge is a credence that is factual, vindicated, and relies on no false theories.

3.1 Un-Structured Data

It refers to computerized information that is either not having any data model or cannot be directly used by a computer program [1]. In other words, data with some form of structure may be characterized as unstructured if its structure does not reflect a useful schema to get a desired processing task.

Most of the business information exists as unstructured data – commonly appearing in e-mails, blogs, discussion forums, wikis, official memos, news, user groups, chatting scripts on social networking sites, project reports, business proposals, public surveys, research and white papers, marketing material, official and business presentations and most of the web pages on WWW.

3.2 Different forms of un-structured data

We have different types of un-structured data generated by different users across the globe:

- **Business Data**: This type of data is primarily generated in a business organization. Although a small part of it is structured data like employee information, salary details, company's balance sheet etc, but a large part of it is simply unstructured like customer communication and feedback, client presentations, minutes of project / team meeting, official memos and many more.
- **Social networking data**: This type of data is purely un-structured. Users basically use SMS type language which is not easily understandable by even human beings. Product reviews, and feedbacks are another important part of this database. Chat scripts are also constituents of such data.
- **General communication data**: This type of data mainly constitutes emails, blogs, wikis, news, discussion forums etc. Although templates to capture this data are structured but the contents inside those text blocks are mainly un-structured.
- **Audio-Visual data**: The data in the form of audio and video files is available in huge quantity across the world. There is no defined pattern available while we browse these files.

4. CURRENT SCENARIO OF DATA MINING OVER WEB

Keeping in view of the potential opportunity to extract business focused knowledge from the colossal amount of data available on www, a structured approach is being followed by data administrators and managers across the world when we talk about data mining over web. Below are the major fields which are being explored in terms of finding knowledge from un-structured data:

4.1 Data mining with a focus towards mining unstructured data

A lot of unstructured data is noisy text [2]. Spontaneous communication (such as e-mails, discussion forums, SMS, blogs, and collaborative web portals) contains noisy text and processing noise. We can define "noise" in text as the difference of any type found between the original and received text.

In the context of unstructured content [3], there is no conceptual and data type definitions available in textual documents, and we find it very tricky to extract information from the content [4]. Therefore, proficient algorithms duly supported by human intercessions are necessary to make the unstructured data smoothly readable and understandable by a computer machine [5]. A vast proportion of this unstructured data contains informal and semi-formal, internal and external communications of a given organization [6]. Usually humans can understand such text straightaway. However, with enormous quantity of such data content being available nowadays, both online and inside the enterprise, it becomes critical to mine such text using computers as it becomes very difficult and complex for a human being to mine huge data manually.
We can think of using available data mining generalized models to represent unstructured data also but with very less efficiency and proper outcome. There are a few algorithms available to extract useful information from unstructured data including Opinion mining [7] from noisy text data, but a generalized, rugged approach is still missing.

4.2 Semantic Web Mining

The semantic web is based on the visualization of Tim Berners-Lee [8], the inventor of the World-Wide-Web (WWW). According to him, "The semantic web is not at all visualized as a separate web but it is an expansion of the existing one, in which information is given well-defined sense and significance, better enabling PCs and people to work in cooperation."

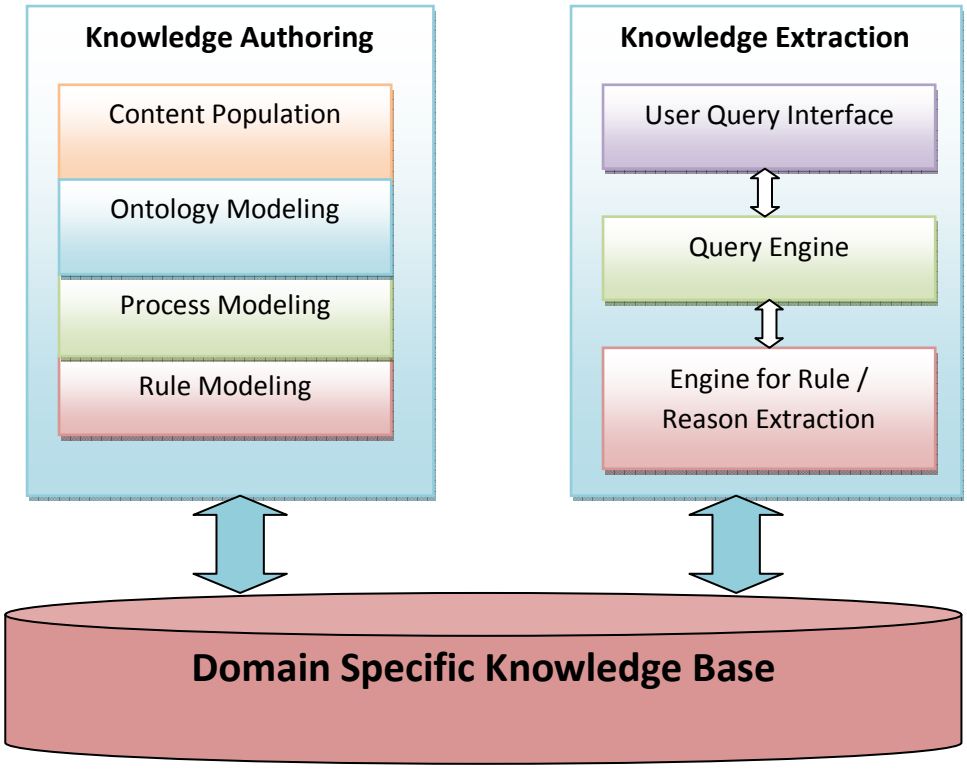

FIGURE 1: Semantic Web Solution Architecture

Semantic web mining intends at two emergent research areas of semantic web and web mining [9,10]. The idea is to improvise the results of web mining by taking advantage of the new semantic structures on the Web; and also, making use of web mining, for building up the semantic web by extracting similar meanings, useful patterns, structures, and semantic relations from existing web resources.
Figure 1 shows proposed solution architecture for semantic web mining. The architecture is primarily divided into three logical modules, namely knowledge extraction block, knowledge

authoring block and domain specific knowledge base. The availability of prevailing search engines has to a great extent improved our ability to carry-out a meaningful data search on the web. But, such search option is still primarily restricted to structured data. In semantic technology, the focus is generally to formulate flexible data model (called Triples) from the user friendly domain query. Semantic search engines are yet to prove themselves in the huge periphery of web search.

4.3 Ontologies Extraction

Extracting ontology from the web is a challenging task [10]. Ontology extraction and modeling use a lot of existing resources, like text, thesauri, dictionaries, databases and similar resources. Techniques from several related research areas e. g., machine learning, information retrieval [11], etc, are combined, and are applied together to discover the `semantics' in the data and to make them plain and clear.

A few systems have already been developed by research community across the world to extract ontology [12]. Several standards have been developed to implement the layered structure of the Semantic Web, such as the Resource Description Framework (RDF) [13] and Web Ontology Language (OWL) [14]. Resource Description Framework (RDF) is being used by people to represent metadata of web pages which can be processed by a machine. It describes a data model to represent all relations between different resources. Still, this “similar meaning and relation extraction” work is yet to mature on the global information retrieval platform simply because of the non-availability of proven processes, standards and systems.

5. CHALLENGES IN SEMANTIC WEB MINING OF UNSTRUCTURED DATA

There are many challenges when it comes to mining of the unstructured data in the context of semantic web:

i. **Structured-data mining focused search engines:** The emergence of some great search engines has significantly improved our skillful capability to search for data on the web; however, such search tool is vastly restricted to structured data only. Not many search engines are available in public domain which specifically addresses the requirement of mining / searching unstructured data flavored with semantic search. This is the biggest challenge being faced by industry and academia alike.

ii. **No standardized web form structure:** We can search for and extract information available as HTML, but till date, we are not proficient to gain easy access to the hidden web. It is very difficult to get to the accurate web form, and even harder to find a suitable truthful web application and related service. When we find the accurate web form or web service, then there is a supplementary step to understand its schema and reformulate the user's query to fit that schema. While human beings do this on a regular basis, one form at a time, it is very complex and cumbersome to automate the process of query reformulation, and therefore we cannot leverage the wealth of information residing behind web forms and services for the masses.

iii. **Non-availability of standard Semantics:** We cannot apply the techniques for exploiting corpora of documents directly for searching unstructured data. The main reason is that searching unstructured data requires an understanding of its underlying semantics. This structure is normally specified by the schema. However, in specifying these semantics, the actual words used and the information clustering purely depends more on the developer's whim, and little variations may result in a very different semantics altogether. Thus, it is a big challenge to have standard semantics available for general usage.

iv. **Lack of global Standards:** Very less international standards are available on Semantic Web Mining. Some big organizations and universities are working on it with little success. Non-availability of broad, rugged and internationally recognized set of standards addressing amalgamated mix of semantic web, web mining and unstructured data mining is the crucial challenge being faced by researchers across the world.

v. **Lack of proven frameworks:** There are some challenges involved in large scale integration on the web [15] namely the realization of the mining framework, the robustness of mining techniques, and the exploration of holistic insight.

vi. **Non-standard implementation:** There is a huge implementation challenge to develop a database management system for administering the entire process of information extraction in an efficient and effective manner [16]. Some industries and academic establishments have come up with their own KDIS (knowledge driven information systems), but all these systems are proprietary to individual organizations or universities. Because of this, the research in this area is not getting opened up for wider coverage and business implementation.

vii. **Non-standard Information Systems:** If we are to design such a system which can help the users to mine unstructured data, how should it look for? What will be the system capabilities? The key challenges include data model and representational issues; need for newer index structures; standardization for Information Extraction (IE); data cleaning and its fusion; accurate relationships in the context of IE and probabilistic databases; and lastly the role of knowledge which user possess and the iterative nature of user interaction. These are a few challenges researchers face during information system development.

viii. **No support for audio / video data mining:** There is any standard support available which can help the end-users to extract valuable and handy information from audio and video files available in huge quantity across the world.

ix. **Less-explored Ontology framework:** Developing the ontology for the large scale databases is a great challenge in itself. There are so many industry verticals and domains available across the organizations. Developing Ontologies for these verticals and domains is a major challenge.

x. **Lack of availability of best practices:** Data vocabulary complemented by content relevance is the operational challenges during the development of semantic web mining tools. There are no best practices available for this development as the technology area is not yet matured and lot of new developments are happening across the world in this field.

6. OPPORTUNITIES FOR SEMANTIC WEB DATA MINERS IN THE FIELD OF UN-STRUCTURED DATA

The unprecedented success of www has unfolded the true potential of two fast-emerging research areas of semantic web and web mining. Both of these areas complement each other and open up new opportunities for the researchers across the world. As discussed in preceding sections, the majority of the available data on web is totally un-structured which can be understood by human-beings only. But the amount of data suggests that the same can be processed by machines efficiently. Hence, there is a good opportunity for semantic and web miners to explore this situation to provide next level of mining paradigm to the world.

The intact opportunity in the field of semantic web mining can be elaborated and split into two unique parts as “semantic” – “web mining” or “semantic web” – “mining”. In the past few years, there have been many attempts at “breaking the syntax barrier” on the web [17].

Analysis of challenges mentioned in previous section gives us an ample opportunity to explore semantic web mining of un-structured data and extract huge amount of knowledge available un-tapped at www. A few opportunities are suggested below:

- To develop web mining techniques that will enable the power of www to be realized. These constitute development of web metrics and measurements, process mining, temporal evolution of the Web, web services optimization, fraud and threat analysis, and web mining and privacy.
- To design and develop search engines specifically focused towards mining un-structured databases. This is the need of the hour as the success and failure of semantic web mining of unstructured data will primarily depend on the availability of suitable and relevant search engines.
- To design and develop information systems for exploring unstructured data available in bulk on web primarily extracting content, structure and usage mining. An enterprise system

integrating these three spheres of web mining is very critical to the success of this field of research [15].

- To design and develop knowledge extraction and un-structured text processing algorithms which are either available as proprietary algorithms with some big organizations or not available at all for general usage and exploration.
- To design and develop models for concept and ontology extraction from unstructured data. This extraction is very important after the enormous explosion of social networking.
- To design and develop an ontology modeling algorithm which also addresses rule and process modeling in a particular industry vertical or domain area.
- There are not many KDIS available across the organizations which can help them feel the pulse of their customers, employees and vendors. Mining the opinion from a huge unstructured data available on www is one of the hottest research areas in current time. Extracting sentiment and opinion of customers' feedback is an exciting problem to work on.
- A few enterprises and research groups are working to make standards for semantic web, web mining and semantic web mining. This is one of the most critical fields in today's world which will provide directions to the research community on semantic web mining.
- To develop a user-friendly database management system to manage the entire process of information extraction [16]. To develop a data model which addresses representational issues of un-structured data with newer index structures is an important unexplored field. An end-to-end solution which may provide knowledge extraction and retrieval will be a big opportunity for the developers to develop.
- To standardize process for Information Extraction (IE); design efficient algorithms for data cleaning and fusion; design mechanism to find out relationships between uncertainty management in the context of IE and probabilistic databases.

7. CRITICAL ANALYSIS

Research in the field of semantic search engines is focused on various approaches and classification theories. Miller et al. [18] talked about Navigational Searches which points to the classification of documents based upon the intention of the user. Mangold [19] focused on architecture, coupling, user context, query modification, transparency, structure of ontology and relevant technology as parameters to realize semantic search. In another critical research on semantic search engine [20], it is pointed out that augmenting traditional keyword search with semantic techniques is considered as the important parameters to implement the semantic search engine.

Hildebrand et al. [21] suggested a search system based upon query construction in section with custom search algorithms. Dietze and Schroeder [22] suggest a new classification approach based on 9 criteria which include structured/unstructured file, text mining type, type of documents, number of documents, Ontologies, clustering, result type, highlighting, scientifically evaluated. Dong et al. [23] present a extended classification with semantic search algorithm based on the Graph, methodology on distributed hash tables and logics-based Information retrieval.

8. COMPARISON OF SEMANTIC SEARCH ENGINES

In the current search scenario, there's no denial about the super power and unquestionable popularity of the Google search engine, where results are based on page rankings and proprietary algorithms. But there are some very innovative ways available to search the web, using semantic search engines. A semantic search engine will definitely ensure more closely suggested relevant results based on the ability to understand the definition and user-specific meaning of the word or term that is being searched for. Semantic search engines are able to better understand the context in which the words are being used, resulting in smart, relevant results with more user satisfaction.

A comparison of semantic search engines is shown in Table – 1.

Semantic Search Engine	URL	Main Approach	Features
Kngine	http://kngine.com/	Based upon "Concepts"	It contains more than 8 million concepts
Yebol	http://www.yebol.com/	Based upon patented algorithms paired with human knowledge	Yebol automatically clusters and categorizes search terms, Web sites, pages and contents, instead of the common "listing" of Web search queries.
Hakia	http://hakia.com/	Based upon "Credible"	It divides the results into Web, News, Blogs, Twitter, Image and Video, and can be re-listed according to relevance.
Duckduckgo	http://duckduckgo.com/	Based upon classic search, information search	If we search for a term that has more than one meaning, it will give us the chance to choose what you were originally looking for, with its disambiguation results.
EVRI	http://www.evri.com/	Based upon Information search	Search results can be filtered into *Articles, Quotes, Images* and *Tweets*.
Truevert	http://www.truevert.com/	Based upon "Green Search Engine"	All results are filtered and organized from one specific perspective – with the topic of environmental awareness in mind.

TABLE 1: Comparison of Semantic Search Engines

9. SOME SEMANTIC WEB BASED WEB SITES

Although, there are many web pages available on www which are using semantic web as the base technology, we are sharing a few popular web sites as shown in Table 2.

Web Pages supported by semantic web	URLs
Brickipedia	http://lego.wikia.com/wiki/LEGO_Wiki
Familypedia	http://familypedia.wikia.com/wiki/Family_History_and_Genealogy_Wiki
Semantic MediaWiki	http://smwtest.wikia.com/wiki/Semantic_MediaWiki_Test_Wiki
Books Wiki	http://bookswiki.wikia.com/wiki/Books_Wiki
SuperWikia	http://super.wikia.com/wiki/Main_Page
Governance Wiki	http://government.wikia.com/wiki/Giki
Yellowikis	http://yellowikis.wikia.com/wiki/Main_Page
MyWikiBiz	http://mywikibiz.com/Main_Page
Common Sense Wiki	http://commonsense.wikia.com/wiki/Common_Sense_Wiki
Animepedia	http://anime.wikia.com/wiki/Animepedia
Creative Commons Wiki	http://wiki.creativecommons.org/Main_Page
semanticweb.org	http://semanticweb.org/wiki/Main_Page

TABLE 2: Semantic web based web sites

10. CONCLUSION

Semantic web mining is relatively new sub-field of data mining. It has a vast scope for investigation keeping in view of the availability of tons of unstructured data on WWW. Lack of available global standards on this subject opens up a enormous prospect for the research community to focus on this area in a big way. Non-availability of a rugged database management system to manage semantic web mining opens up new avenues for the researchers to develop KIMS (Knowledge extraction management system) for unstructured data available on the web. A user-oriented semantic search engine is the need of the day. These fields if explored in a right manner will provide unlimited opportunities to extract knowledge from the goldmine of unstructured data available across the globe.

11. REFERENCES

[1] J. Han, M. Kamber. (2001) "Data Mining Concepts and Techniques". Academic Press, Morgan Kaufmarm Publishers. ISBN 1-55860-489-8.

[2] W. Fan, L. Wallace, S. Rich, Z. Zhang. (2006, September). "Tapping the power of text mining". Communications of the ACM. Volume 49, Issue 9. pp. 76 – 82

[3] D. Bitton, F. Faerber, L. Haas, J. Shanmugasundaram. (2006). "One platform for mining structured and unstructured data: dream or reality?". Proceedings of the 32nd international conference on Very large data bases. pp. 1261 – 1262

[4] R. J. Mooney, R. Bunescu. (2005). "Mining knowledge from text using information extraction"; ACM SIGKDD Explorations Newsletter. pp. 3 – 10

[5] R. Ghani and Carlos. (2006, December). "Data mining for business applications". KDD-2006 workshop. Volume 8, Issue 2. pp. 79 – 81

[6] M. Rajman, R. Besancon. (1997). "Text Mining - Knowledge extraction from unstructured textual data". In Proceedings of the 7th IFIP Working Conference on Database Semantics (DS-7). pp. 7-10

[7] L. Dey , S. K. M. Haque. (2009, July). "Studying the effects of noisy text on text mining applications". Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data. Barcelona, Spain

[8] Tim Berners-Lee. "Semantic Web Roadmap". http://www.W3.org/

[9] B. Berendt, A. Hotho, and G. Stumme. (2002). "Semantic Web Mining and the Representation, Analysis, and Evolution of Web Space". Proceedings of the First International Semantic Web Conference on The Semantic Web. pp. 264 – 278

[10] B. Berendt, A. Hotho, G. Stumme. (2002). "Towards Semantic Web Mining"; ISWC '02: Proceedings of the First International Semantic Web Conference on The Semantic Web; Publisher: Springer-Verlag

[11] A Maedche. (2002). "Ontology Learning for the Semantic Web"; Kluwer. ISBN: 0792376560

[12] M. Niepert, C. Buckner, J. Murdock, C. Allen. (2008). "InPhO: a system for collaboratively populating and extending a dynamic ontology". Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries, Pittsburgh PA, PA, USA. pp. 429-429

[13] Resource Description Framework (RDF) Schema Specification. (2000) In W3C Recommendation.

[14] Web Ontology Language (OWL), http://www.w3.org/2004/OWL/.

[15] K. Chang, B. He, Z. Zhang (2004, December). "Mining semantics for large scale integration on the web: evidences, insights, and challenges". ACM SIGKDD Explorations Newsletter, Volume 6 , Issue 2. pp. 67-74.

[16] A. Doan, R. Ramakrishnan, S. Vaithyanathan. (2006). "Managing information extraction: state of the art and research directions". Proceedings of the ACM SIGMOD international conference on Management of data. pp. 799 – 800

[17] G. Stummea, A. Hotho, B. Berendt. (2006). "Semantic Web Mining State of the art and future directions". Journal of Web Semantic. Web Semantics: Science, Services and Agents on the World Wide Web 4. pp. 124–143.

[18] Miller, Guha, R., McCool, R. (2003). "E. Semantic Search". Proceedings of the WWW'03, Budapest.

[19] C Mangold (2007). "A survey and classification of semantic search approaches". International Journal of Metadata, Semantics and Ontologies, pp. 23–34.

[20] D. Buscaldi, P. Rosso, E. S. Arnal (2005). "A wordnet-based query expansion method for geo-graphical information retrieval". Working Notes for the CLEF Workshop.

[21] M. Hildebrand, J. Ossenbruggen, and L. Van Hardman (2007). "An analysis of search-based user interaction on the semantic web". Report, CWI, Amsterdam, Holland.

[22] F. Figueira, J. Porto de Albuquerque, A. Resende, Geus, P. Lício de Geus, G. Olso (2009). "A visualization interface for interactive search refinement". 3rd Annual Workshop on Human-Computer Interaction and IR, Washington DC. pp. 46-49.

[23] H. Dong, FK Hussain, and E. Chang (2008). "A survey in semantic search technologies". 2nd IEEE International Conference on Digital Ecosystems and Technologies.