Chapter 7
Social Semantic Web Mining and Big Data Analytics

7.1 Introduction

In the context of Big Data Analytics and Social Networking, Semantic Web Mining is an amalgamation of three scientific areas of research: (1) Social Networking (2) Semantic Web and (3) Big Data Analytics, including Web Mining and Data Mining.

Semantic Web is defined and used to provide semantics or meaning to the data on the web, while Big Data Analytics and data mining are aiming to identify and extract interesting patterns from homogeneous and less complex data in or outside the web.

In view of the huge growth of the sizes of data on the web and social networks, we have the Big Data scenario coming up and in view of the growth of the complexity of meaning of data in the web we have the Semantic Web coming up. Due to the rapid increase in the amount of Semantic Data and Knowledge in various areas such as biomedical, genetic, video, audio and also social networking explosion, there could be a transformation of the entire data scenario into a perfect target for Big Data Analytics leading to both in the terms Semantic Web Mining and Social Semantic Web Mining. The Petabyte scale of huge data sizes which are required to be mined on the web which can be either a normal (syntactic) web- or knowledge-based Semantic Web; we have research possibilities of Social Semantic Web Mining Techniques. Semantic Web is changing the way in which data is collected, deposited and analyzed [1] on the web.

7.2 What Is Semantic Web?

We define Semantic Web as a Web which provides meaning to its contents as against the present 'Syntactic Web' which carries no meaning but only display. It is about providing meaning to the data from different kinds of web resources so as to enable machine to interpret or 'understand' these enriched data to precisely answer and satisfy the requests from the users of the Web [2–4]. Semantic Web is new generation

C. S. R. Prabhu et al., *Big Data Analytics: Systems, Algorithms, Applications*,
https://doi.org/10.1007/978-981-15-0094-7_7

Web 2.0 or beyond. Semantic Web is an extension of the present (Syntactic) Web (WWW) to enable users to represent meaning of their data and also share it with others.

Why Semantic Web?

Semantic Web was initiated for working on specific problems [5]: (1) to overcome the limitation of data access in the web as: instead of retrieving all kinds of documents from the web for a given query but focus better with more knowledge-based best fit searches and (2) the delegation of task problems (such as integrating information) by supporting access to data at web scale and enabling the delegation of certain classes of tasks.

Finally, it is the objective of Semantic Web to represent knowledge on the web instead of simple data where meaning is not machine understandable. This calls for techniques of knowledge representation (KR) being taken from artificial intelligence (AI).

Ontology in Semantic Web represents the knowledge repositories.

7.3 Knowledge Representation Techniques and Platforms in Semantic Web

XML

Extended markup language (XML) is a mechanism to represent all types of data. XML can be read by computers for interoperability between applications on the web. XML pages generated by applications based on schema can be read by humans also and XML can be interpreted by computers. XML provided data interoperability but not meaning—the data in XML could not be described for its meaning—meaning/semantics could not be described. It could be used only for defining syntax. Semantics integration and interoperability (e.g., British units and Metric units) calls for explicitly semantic descriptions to facilitate integration. Prior to XML, one had to code by hand the modules to retrieve the data from the data sources and also construct a message to send to other applications. XML facilitated building systems that integrate distributed heterogeneous data. XML allows flexible coding and display of data by using metadata to describe the structure of the data (using DTD on XSD). The first step in achieving data integration is to take raw data such as text, tables or spreadsheets and convert them to well-formed XML documents. The next step is to create DTD on XSD for each data source to analyze/document its structure. XML is incorporated in web services which can interoperate among themselves by using SOAP for invocation.

RDF

Resource description framework (RDF) can be read by machines, i.e., computers. Resources (web sites) are linked to form the web—to give meaning to resources and links a few new standards and languages are being developed—to allow precisely describe the resources and their links. RDF and OWL are standards that enable the web to be able to share documents and data RDF is semantic definition of the details of the description of a web site.

7.4 Web Ontology Language (OWL)

Web ontology language (OWL) is more complex language with better machine interpretability than RDF.

OWL primarily defines the nature of resources and their inter-relationships [6]. To represent the information in Semantic Web, the OWL uses ontology.

7.5 Object Knowledge Model (OKM) [7]

The RDF and RDFS provide the mechanism for knowledge representation using triplets (subject-predicate-object). This is only to the sentence level of Semantics which is being captured. More deeper semantics can be captured if the Object Knowledge Model (OKM) is deployed which gives greater or word level semantics in terms of root, suffix indicating the case, gender, person, etc., depending on whether the word is a noun or verb, etc. (as identified after part of speed (POS) tagging). OKM represents complete natural language semantics of a sentence in a web page. Thus, instead of user identifying the metadata in terms of RDF entities, etc., it will be jointly to analyze the sentences in the web page themselves to identify not only the metadata in terms of entities and their properties and relationships with other entities, but also details of such relationships including case, number, gender and person.

Ontology

Ontologies are similar to taxonomies but use richer semantic relationships among terms and attributes and also strict rules on how to specify terms and relationships (Ontologies were originally from artificial intelligence (AI) used for inferencing but are also recently being applied to semantic web area).

An ontology is a shared conceptualization of the world. Ontologies contain definitional aspects such as high-level schemas, associated aspects such as entities, attributes, inter-relationships between entities, domain vocabulary and factual knowledge—all connected in a semantic manner [8–11].

Ontologies provide a common understanding of a particular domain.

The domain can be communicated between people, organizations and application systems using ontologies. Ontologies provide specific tools to organize and provide useful descriptions of heterogeneous content.

The major uses of ontologies

1. To assist in communication between human beings.
2. To achieve interoperability among software systems.
3. To improve the design and quality of software systems.

Technically, an ontology model is likely an ordinary object model in object-oriented programming. It contains classes, inheritance and properties.

In many cases, ontologies are considered as knowledge repositories.

7.6 Architecture of Semantic Web and the Semantic Web Road Map

Figure 7.1 indicates the architecture and the road map of the Semantic Web as indicated by W3C (Copyright).

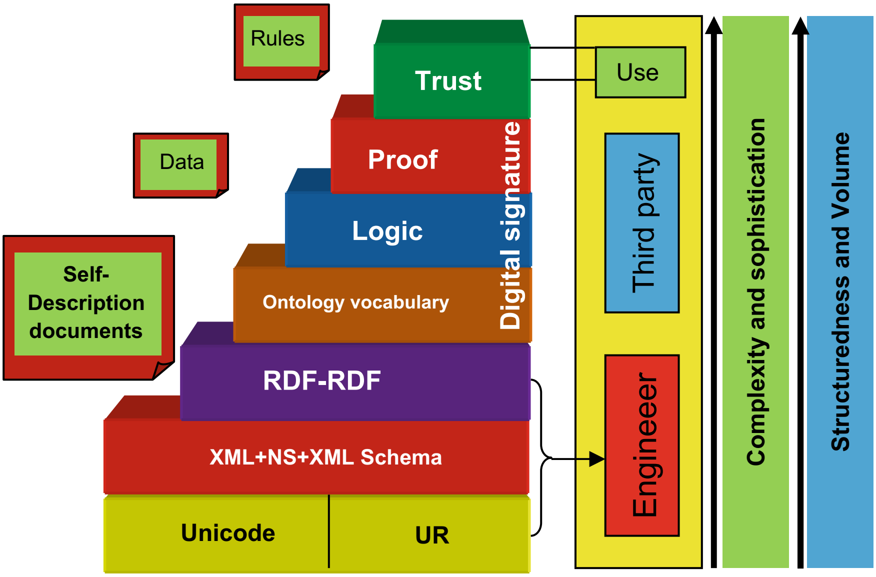

Fig. 7.1 Semantic web layer architecture [4]

7.7 Social Semantic Web Mining

The World Wide Web includes the social web or social networking-based web in addition to the normal conventional web. It represents the vast repository where millions of people on thousands of sites across hundreds of diverse cultures and languages and leave trail or footprint about various aspects of their lives, both professional and personal. Social systems are understood to be purely systems of communication and only communication, but therefore leave trails by analyzing what we can analyze in the social life itself: the political trends, cultural trends, media responses, trends of business and financial transactions and also Science.

There are several automated methods and techniques dealing with various aspects of information organization and retrieval that can be applied and used in the social science research based on Social Networks. Especially in the context of elections in the USA, India and other countries, political attitudes of voters become very important and it becomes very critical to analyze the trends and predict the results of the elections. The social web is also known to have influenced political discourse and a lot of trails about political standpoints and political views/sentiments of the citizens can be found on all over the web the use of Big Data Analytics.

Design and development of repositories of heterogeneous content with central themes is possible. Studies, surveys, databases and data warehouses on central themes can be prepared. Such themes could be specific problems such as attitudes toward government policies and projects or social trust among people, biases, prejudices, etc. Automated search engines are used to form certain hypothesis or model can be prepared by researchers and such hypothesis can be verified from the existing data repositories.

A separate section outlines the possibilities of automated multimedia processing by describing Speech Recognition and Machine (Computer) Vision techniques.

A separate section gives natural language processing (NLP) techniques overview and a separate section provides a survey of sentiment analysis techniques, and in a separate section, we describe Recommender systems and user profiling which allows us to automatically generate profiles of users based on their behaviors in a given domain and accordingly suggest what could be interesting to them. A separate section provides the concept Semantic Wiki Systems which extend usual wiki systems like Wikipedia with semantic description (e.g., metadata) and thus allow the creation of vast text and multimedia repositories that can be analyzed automatically based on AI techniques like automated reasoning. A separate section identifies the challenges and a separate section identifies the possible issues and research directions.

Data mining or knowledge discovery in large input data is a broad discipline under which web mining is a branch or subdomain. Web mining deploys techniques of data mining on web information so as to extract useful information from the web. The outcome is valuable knowledge extracted from the web.

Web mining can be classified into (a) Web content mining, (b) Web structure mining and (c) Web usage mining. In (a) Web content mining, the text and images in web pages are required to be mined, requiring to adopt natural language processing

(NLP) techniques. Web structure Mining is a technique to analyze and explain the link and structure of web. Web usage mining can be used to analyze how websites have been used much as determining the navigational behavior of users in terms of user behavior and therefore linked to social and psychological sciences.

Web is a source of invaluable social data by virtue of the ability the users to create, search and retrieve web pages containing valuable information.

Social Networking activities are a part of this mass data creation trend. They can be considered as special instance of Web Mining, i.e., social Web mining. Social Networking is understood to be a continuous networking activity to connect individual users to other individuals or groups of users on a daily, online or even real-time basis. This will produce many social relationships, including closeness and intimacy between individuals or groups. Experiences in life, ideas and thoughts are data which including video and photograph are shared among people. A variety of Social Networking Services is available today as:

Full-fledged personal Social Networking as Facebook and Twitter.

Professional Social Networking as LinkedIn
Blogging services as Word press
Video-sharing services as Vimeo
Photograph-sharing services as Flickr
e-commerce services as Quickr and many others as Amazon, Snapdeal, Flipkart, etc.

All these different Social Networking are continuously generating very large quantities of information on a nonstop, continuous basis, reflecting the instantaneous behavior of their users, reflecting their true life, relationships, interpersonal communication, thereby creating ready to use datasets, fit for analysis.

Although data is available, analyzing it is not easy. Several and different types of data exist and they are required to be analyzed accordingly, using appropriate tools technology and methodologies. Structured data in databases can be analyzed using SQL type of queries in DBMS such as Oracle. Data in data warehouses can be analyzed using OLAP tools and also data mining techniques and tools. Machine learning techniques such as Classification, Clustering and Regression can be deployed on the data, if it is adequately structured to find out trends, patterns and even forecast the possible future happenings on the data. This type of data could be from online sources as exchange rates, weather, traffic, etc.

Text data in web pages, emails or in blogs can be analyzed using NLP techniques of artificial intelligence (AI).

Semantic Web techniques can be deployed on creating metadata and Semantic Web Documents on web pages. They will be in RDF, RDFS and OWL kind of environments.

Image data, both static or still photographs and video clips, can be analyzed by image processing techniques (e.g., flash applications).

Since there is a time dimension also in the WWW, the temporal analysis of data is possible. This enables the trend analysis as in OLAP or regression. Adding geographical dimension to temporal data is of great value. Analyzing Social Networking data in time and using location information produces valuable results. For example, in a Cycle Race by the following the spectator's comments, it is possible keep track of the Race in real time. Analyzing data of social networks with all components as time and location (temporal and geographical) can lead to information which can make it easier to follow and also to keep check of the voters in an election. By gathering such information, it is possible to analyze it and find out why and how voters are thinking about the Party and the Candidate. Based on that data, it is possible to influence the voters by knowing what is important wherein a locality of population. In 2012 election in the USA and in 2014 election in India, a detailed analysis of social networking activity of the voters was extensively used to reach the voters and conveying them information which influenced their voting behavior and decide voting patterns to win the election finally. Those candidates who used such techniques benefited immensely.

Context in data is critical. Unless the context is understood, the data cannot make any sense or any meaning for the purpose of analysis. Sometimes, statements in Social Networks can be misleading, especially when they are sarcastic and therefore they mean the opposite of their true intention. Suddenly, they may lose their sarcasm. For example, in a satirical portal or online magazine, if the true context of the background is not understood, then the opposite meaning will be interpreted by the software or for analysis purpose. Therefore, it becomes extremely important to know the context for data creating, publication and consumption.

Social and Conceptual Analysis and Tag Clouds (II)

The new Science of Networks or Network Theory allows and enables us to study networks of any origin or subject domain. In the context of social Web analytics, we have two types of networks as relevant and of interest: social and conceptual. While social network provides the linkages (communication, friendship, interactions, trade, cooperation, etc.) between social entities (people, organization, social groups, political parties, countries, etc.), the conceptual networks provide insights into structure (homonymy, mutual context, synonymy, hyperlink, etc.) and dynamics (evolution of context) of concepts (words, ideas, phrases, symbols, web pages, etc.).

On the Social Semantic Web, both Social Networks and Conceptual Networks are ubiquitous. If two people can be connected on the social web, it can be any of the following: connection, friendship, co-opinion, cooperation, classmates, co-voter ship or they may have liked the same audio or same video on a news casting site or two people have similar political interest or attitude. When we multiply this situation for millions of individuals on the Social Semantic Networks, we can deploy social network analysis techniques as finding connected components or clustering or classification to identify networks of people who form groups of similar features such as similar political ideology or attitudes or belonging to the constituency of same political attitude.

If we observe such Social Networks over time, we might also reason about the dynamics of evolution of groups how and when, where groups or subgroups were formed. By having such data, we can develop agent-based models (A&M) that might predict future behavior based on the past behavior.

In addition, Social Networks can be analyzed so as to provide visualization of appealing, yet informative insights into a social system and find its peculiarities, if any, calling for appropriate action.

On the other hand, conceptual networks are not so directly observable or obvious: two web pages with one linked to the other or two keywords used in the same article, two tags on the same online video, two concepts extracted from the same text, etc.

Tag Clouds, Tag Maps, folksonomies, visualization and many other usable methods exist for analyzing Conceptual Networks.

Tag Clouds

Content created by humans can be tagged. Such tags are metadata on the context. What are Tags? Tags are keywords that can provide insights to readers of the contents of a given web page. Multiple tags can be created for better accuracy.

In Social Tagging, the so-called Tag Clouds are implemented which are visualization of tags provided by users. The most used tags are shown in large-sized letters and less used tags in small-sized letters. Tools such as 'Wardle' even have basic NLP capabilities and are capable of auto-summarizing the text in form of beautiful keyword clouds.

Tag or keyword clouds are not following network analysis method, but they do create and present visual representations of text and the context that provides an insight into the matter with a simple look on it.

Topic Maps

Topic Maps are a methodology for Text Visualization. Let us take a context of text, for example, in newspaper articles, we want to analyze. If we wish to see the development of certain topics keywords in time in the given context of a textual discourse, then we can do such visualization by deploying Topic Maps. If a topic is used more often in a given text, then it covers greater surface on the Topic Map. Even though such mechanisms are used in online social blogging or forms, they can be used in any series of text with spatio-temporal frame.

7.8 Conceptual Networks and Folksonomies or Folk Taxonomies of Concepts/Subconcepts

Conceptual networks comprise of clusters to give well-connected components. Complex conceptual network visualization may have hundreds of thousands of nodes. The visualization of complex network is an art and a science. Visualization has to be appealing, informative and concise. Numerous algorithms for complex network visualization have been proposed and developed, each for different types of networks.

Processes on Networks

There are techniques that allow in the study the dynamics and evolution of networks. Examples such as virus spreading or rumors spreading can be represented as Network

Processes. It is possible to develop agent-based models that allow us to understand the spreading of certain themes, whose behavior is to be studied. For example, in order to study the process in which the particular political attitude can spread through a social network, we might be required first to harvest the communication between people on the network (forum or blog), then we can conclude text (messages, articles, blogging, past, etc.) that deal with the name subject and also temporal data (time of publication). Then by deploying advanced analytical techniques to (a) construct the social network of people involved in the discourse, (b) identify different attitudes toward the actor (politician) concerned by using NLP techniques or sentiment analysis and finally (c) define the actual process of information spreading in the form of an agent-based model.

7.9 SNA and ABM

Another distinct method that enables analysis of a distinct set of primarily social entities and their interaction is by using the combined power of Social Network Analysis (SNA) and Agent Based Modeling (ABM) so as to enable huge amounts of data to be embedded into user friendly models.

The SNA and ABM go together. ABMs are modeling agents. They interact on the basis of a mathematical model which is also area-specific. The SNA plays a role in unveiling the structure of interaction of the modeled agents. Both models complement each other. As an example, 'Recipe World' (by Fontana and Torna) stimulates the emergence of a network out of decentralized autonomous reaction. Agents are provided with recipes or steps to be taken to achieve a goal. Agents are activated and the network emerges as an effect.

This model is based on four distinct sets: (1) The actual work populated by entities and their actual network (2) in an ABM (agent-based model) with agent which base their behavior on the orders and recipes derived from (1) above (3) represents the network generated by (2). The events are: the possibility of populating (3) and (4), respectively, by knowing (4) representing them as data on the network and using influence and a kind of reverse engineering.

Agent-based model (ABM) can be created as an outcome of reverts engineering effort by combing the previously mentioned web mining and network analysis. Data (D) can be gathered from social web mining to model a political network (C), comparing data about political parties and their interactions of their voters, leaders, NGO and also Government organizations.

In turn, this network (C) can serve as the basis for a ABM (B) have allowing the simulation of real-world entities and their actual network (A).

7.10 e-Social Science

In the context of more and more digitization of life, the term e-Social Science refers to the use of electronic data in social studies, especially social networking, where much of interaction and social life is increasingly now only in online mode and online world only. The online records of social activity offer a great opportunity for the social science researchers to observe communicative and social patterns in much larger scale than ever before. Having removed the human element, it becomes easy to sample data at arbitrary frequencies and observe the dynamics in a much more reliable manner. A large diversity of topics can be studied in e-Social Science. There are many diverse electronic data sources in the Internet for e-Social Science research. These include electronic discussion networks, private, corporate or even public groups such as Usenet and Google groups, blogs and online communities, web-based networks, including web pages and their links to other WebPages.

The expansion of the online universe also means a growing variety of social settings that can be studied online. Extraction of information from web pages also called information retrieval or information extraction can extract networks of a variety of social units such as individuals or institution. The most important technical issue or hurdle to overcome is the problem of person name disambiguation, i.e., matching of references to actual persons they represent. Person names are neither unique nor single: a person may have different ways and methods of writing or even different spellings of his own name and also several persons can have the same name. Even if Father's Name is associated with a few duplicates and ambiguities may appear. When Father Name and Address combination is attached to a single name then ambiguity or duplication may be reduced, if not eliminated. However, a unique identity number (as Social Security Number or PAN Number or Aadhar Number) can disambiguate the name. Semantic Web can aim at identifying a named object identity formation for all entities. Semantic Web sites such as 'Live Journal' that already import data in a Semantic Format do not suffer from this problem. In their case, a machine-processable description is attached to every web page, containing the same information that appear on the web page but in a way that individuals are uniquely identified. This method can also be used for merging multiple electronic data sources for analysis.

Speech Recognition and Computer Vision

The Social Semantic Web is not plain text but has substantial audio and video content which needs to be processed and understood for its knowledge content by using appropriate advanced techniques of Speech Recognition and Computer Vision.

Speech recognition comprises of human speech-to-text conversion. Various recorded speeches need to be processed to directly identify the speaker and then given the speech so it convert to text. Speech recognition technology and speech-to-text conversion techniques are very much advanced but have a serious limitation of language dependency. It requires substantial text base in that specific language and speech to text system for that specific language only. If such language base or speech to text system is not available for a new language, one has to develop from the beginning such a system and that is a long-drawn process which also a difficult task.

Computer vision is a field of computer science that aims at allowing the computer systems to analyze images and/or video data in order to recognize the objects that are present on a given image/video. Unlike speech, there is no language or depending on a language in computer vision, but there is even a worse problem that for every type of object that needs to be recognized there needs a system that recognizes that object. This is more true for person recognition or biometrics where one needs to trace the system to recognize every single person of interest (this process is called enrollment).

Even then, there do exist certain types of objects which can more or less be recognized by even single system like text or any image or in a video. Such text can be used along with the recognized speech.

Natural Language Processing (NLP)

Natural language processing, or NLP, considered a part of artificial intelligence, aims at acquiring an understanding of the contents of a given text in any particular natural language such as English or French or Hindi. This is applicable to spoken languages also, in addition to the recorded text. To quote Cohen, natural language processing (NLP) comprises hardware and/or software components which have a capability to analyze or synthesize spoken or written language, similar to the humans. The main problem for achieving this objective is the complexity of the natural language as characterized by two well-known constraints: ambiguity and context sensitivity. Further, languages always keep evolving in time.

Every natural language has a grammar that controls such evolution and restricts changes by formulating logical rules (grammar) for the formation of words, phrases and sentences which are compulsorily required to be adhered to by the user of the language if it has to be a valid language construct. In addition to logical rules of grammar, every language requires a Lexicon or Dictionary. These two (the grammar and the Lexicon) are the characteristic requirements for the language itself. Then, there are specific linguistic tools that make it easier for the algorithms to access, decompose and make sense of sentences already available for use in processing of NLP data. Such tools include:

- Tokenizers or sentence delimiters which detect sentences based on delimiters (as a fullstop).
- Stemmers and POS taggers—morphological analyzers that link a word with its root and therefore identify the part of speech (POS) as noun, verb, adjective, etc.
- Noun phrase and name recognizers—labeling works with noun phrases (e.g. adjective + noun) and recognizing names.
- Parsers and grammar.
- Recognizing well-formed phrases and sentence structures in a sentence.

Steps involved in NLP can be summarized as:

Firstly, the entire text is required to be broken down into sentences. Secondly, the sentences are required to be broken down into words and words to be tagged for part of speech (POS). Then, each word so identified is to the broken down for its root and its suffix to indicate the meaning of the word by looking into the Lexicon by using Stemmers.

Stemmers can be ineffectual (identify the tense, gender, etc.) or derivational which link up several words to the same roots (say the same verb root for several words).

History of NLP can be traced back to ancient times to Sanskrit grammar associated with Panini and Patanjali as the grammarians who successfully processed the language of Sanskrit using the human brain alone without any computer systems.

In modern times, NLP developed since the 1950s with (a) symbolic approach and (b) statistical approach of language processing.

Symbolic approach is outlined above with the grammar as the basis for processing the language.

Statistical natural language processing is based on machine learning algorithms. Machine learning algorithms learn from training data and use that learning during testing phase. This approach requires corpora of the language with handmade annotations of labels such as nouns and verbs. Using such corpora, learning is performed by learning algorithms. Such learning is utilized in testing phase. Decision trees can be used.

NLP has many applications in processing some corpora of text, auto-summarization, automated classification of documents and other similar applications.

7.11 Opinion Mining and Sentiment Analysis

This amazing technology enables us to identify the sentiment or feelings of the user (a person who had these feelings while writing some text or message). This has many applications including: (1) election time attitude and sentiment analysis, (2) product or services research for marketing purposes, (3) recommendation systems, (4) campaign monitoring, (5) detection of ‘flames’, (6) public health monitoring (e.g., emotional/mental status, detecting potential suicidal victims), (7) government intelligence (8) trend monitoring (social, cultural and political trends etc.).

Sentiment analysis toward a political personality is very critical to an election. This needs to be integrated with newspaper feedback also. For performing sentiment analysis using NLP, we require to maintain a database of sentiments analyzed already. Normally, the sentiments are classified as positive or negative only (with various degrees of polarity between these two extremes) and therefore complex qualitative sentiments are hard to detect and hence not implemented. Sentiment analysis systems have a huge commercial value and therefore not freely available for usage.

Recommender Systems and User Profiling

Recommender Systems recommend items available through the same systems to its user. Online shopping sites as Amazon, eBay, news portals and other portals offer the Recommendation Services in their own contexts. The factors that are to be considered are: (1) the context of each information node with descriptors generated by the use of NLP tools, (2) user profiles which are generated based on browsing data and (3) user preferences. The Sources of data for recommender systems include: data from web server access logs, proxy server logs, browser logs, registration data, user sessions or transactions, cookies, user queries, bookmark data, mouse clicks and clickstream data (or scrolls) and any other data that gets generated during the interaction of the users with system. In one approach by Seva (in 2014), user profiles were generated from user-visited documents, and newly generated content was recommended based on the underlying taxonomy-based classification. The objective is to provide individual personalized recommendations (usually user groups only are provided recommendations) clusters of users are given recommendations by the system.

7.12 Semantic Wikis

Wikis allow users to collaboratively work on a repository of content.

Semantic Wikis include ideas from Semantic Web (whose objective is to bring in machine-readable metadata to web pages in order to allow computer systems to understand the content they are processing).

Therefore, Semantic Wikis can be used to manage large repositories of textual and multimedia (audio or video) data on which Semantic Web features such as automated reasoning, complex querying mechanism and automated categorizations of content are possible.

When collaboratively edited, metadata as keywords, categories, classification, etc., is added to Wikipedia's a social taxonomy can be generated (as happened in the case of TaOPIs). This taxonomy allows for automated summarizing of text using built-in queries. Intelligent Agents can then be built and deployed to final conceptually similar terms and descriptors of each concept described in Semantic Wikis. They provide an excellent tool to organize research data and yield the needed structural metadata for data mining. A Collaborative System can be thus built.

7.13 Research Issues and Challenges for Future

As was clear from the above sections, Social Semantic Web Mining and Big Data Analytics can provide a substantial addition to Social Science research methodology.

On any specific topic if Social Science research is required to be conducted, then it is possible to create a large Semantic Wiki, added with Semantic Social Network

Mining, audio, video, text inputs from surveys, studies and TV programs. Such a topic could be just political, for example (say on poverty and immigration).

One possible goal will be to identify the main actors (politicians), leaders, their key statements, the position of government, etc. All these can be achieved through NLP techniques and also through Social Semantic Conceptual Network Analysis and also expert opinions. Specialized user groups can be identified through user profiling. The discussions between interested groups can be analyzed using Topic Maps. Sentiments can be measured and analyzed to track the spread of attitudes and concepts and Social Concept Networks can be identified. The above was just an example. The real challenges which can be addressed to be solved in future could be outlined as:

(1) The huge 'Big' data is heterogeneous, structured, semi-structured, unstructured, text, audio and video. This has to be really integrated. They are from different file and data formats. They need to be cleaned up, integrated into a single Big Data Repository for Analysis purposes.
(2) NLP adoption is a challenge given the multiple natural languages generating huge text, audio to analyzed. Do we have multilingual NLP tools and techniques to meet these challenges? (OKM offers on approach for solving this problem).
(3) Sources of data for a specific purpose or topic or objective are required to be identified.
(4) Unique Named Object Identity to provide 'single version of truth' in the Semantic Web is still elusive with multiple dispersed, distributed sources of data which exist severally today across the web and social networks. This is required to be developed.

7.14 Review Questions

1. What is Semantic Web? What is Semantic Web Road Map?
2. What are the various knowledge representation (KR) techniques in Semantic Web? Explain each one of them.
3. What is Object Knowledge Model (OKM)? Explain its application.
4. What is Social Semantic Web Mining?
5. What are conceptual networks? Explain.
6. What are SNA and ABM? Explain.
7. Explain the evolution of natural language processing (NLP).
8. Explain various stages involved in NLP.

References

1. N. Lavrac, A. Vavpetic, L. Soldatova, I. Trajkovski, P.K. Novak, Using ontologies in semantic data mining with SEGS and G-SEGS, in *Proceedings of the 14th International Conference on Discovery Science*, Espoo, 5–7 Oct 2011, pp. 165–178
2. O. Mustapasa, A. Karahoca, D. Karahoca, H. Uzunboylu, Hello world, web mining for e-learning. Procedia Comput. Sci. **3**(2), 1381–1387 (2011). https://doi.org/10.1016/j.procs.2011.01.019
3. D. Jeon, W. Kim, Development of semantic decision tree, in *Proceedings of the 3rd International Conference on Data Mining and Intelligent Information Technology Applications*, Macau, 24–26 Oct 2011, pp. 28–34
4. V. Sugumaran, J.A. Gulla, *Applied Semantic Web Technologies* (Taylor & Francis Group, Boca Raton, 2012)
5. J. Domingue, D. Fensel, J.A. Hendler, *Handbook of Semantic Web Technologies* (Springer, Heidelberg, 2011)
6. A. Jain, I. Khan, B. Verma, Secure and intelligent decision making in semantic web mining. Int. J. Comput. Appl. **15**(7), 14–18 (2011). https://doi.org/10.5120/1962-2625
7. C.S.R. Prabhu, Extending a semantic e-governance grid with OKM, a frame based knowledge representation framework for enabling semantic—knowledge search on the contents of the web pages, in *7th International Conference on E-Government (ICEG-2010)*Bangalore, India, 22–24 Apr 2010, Paper No. 6
8. H. Liu, Towards semantic data mining, in *Proceedings of the 9th International Semantic Web Conference*, Shanghai, 7–11 Nov 2010, pp. 1–8
9. V. Nebot, R. Berlanga, Finding association rules in semantic web data. Knowl. Based Syst. **25**(1), 51–61 (2012). https://doi.org/10.1016/j.knosys.2011.05.009
10. V. Nebot, R. Berlanga, Mining association rules from semantic web data, in *Proceedings of the 23rd International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems*, Cordoba, 1–4 June 2010, pp. 504–413
11. M. Schatten, J. Seva, B. Okusa Duric, An introduction to social semantic web mining and big data analytics for political attitudes and mentalities research, AI Lab, Faculty of Organization and Informatics, University of Zagreb, Croatia