European Quarterly of Political Attitudes and Mentalities EQPAM
Volume 4, No.1, January 2015
ISSN 2285 – 4916
ISSN-L 2285 - 4916

An Introduction to Social Semantic Web Mining & Big Data Analytics for Political Attitudes and Mentalities Research

Markus Schatten, Jurica Ševa, and **Bogdan Okreša Đurić**
Artificial Intelligence Laboratory
Faculty of Organization and Informatics
University of Zagreb
Croatia

Date of submission: January 20th, 2015 *Date of acceptance: January 24rd, 2015*

Abstract

The social web has become a major repository of social and behavioral data that is of exceptional interest to the social science and humanities research community. Computer science has only recently developed various technologies and techniques that allow for harvesting, organizing and analyzing such data and provide knowledge and insights into the structure and behavior or people on-line. Some of these techniques include social web mining, conceptual and social network analysis and modeling, tag clouds, topic maps, folksonomies, complex network visualizations, modeling of processes on networks, agent based models of social network emergence, speech recognition, computer vision, natural language processing, opinion mining and sentiment analysis, recommender systems, user profiling and semantic wikis. All of these techniques are briefly introduced, example studies are given and ideas as well as possible directions in the field of political attitudes and mentalities are given. In the end challenges for future studies are discussed.

Keywords: social semantic web, web mining, big data analytics, network theory, network visualization, agent based modeling, natural language processing, sentiment analysis

1. Introduction

The world wide web and especially the social web represents a vast repository where '*millions of people, on thousands of sites, across hundreds of diverse cultures, leave "trails" about various aspects of*

Corresponding Author: Dr. Markus Schatten, Assistant Professor
Affiliation: Head of the Artificial Intelligence Laboratory, Faculty of Organization and Informatics, University of Zagreb, Croatia
Address: Pavlinska 2, 42000 Varaždin
e-mail: markus.schatten@foi.hr

European Quarterly of Political Attitudes and Mentalities - EQPAM, Volume4, No.1, January 2015, pp. 40-62.
ISSN 2285 – 4916
ISSN–L 2285 – 4916

EQPAM Volume 4 No.1 January 2015
ISSN 2285 – 4916
ISSN L 2285 – 4916

their lives' (Schatten, 2011, p. 13). If we, for a moment, adopt the interpretation of Niklas Luhmann that social systems are systems of communication and only communication (Luhman, 1984) we might find a justification that by analyzing the "trails"[1] we can analyze the social system itself: politics, culture, media, business, technology, science etc.

As it has already been outlined in our previous studies (Neumann et al. 2014, Voinea & Schatten 2014)[2] there are numerous more or less automated methods and techniques dealing with various aspects of information organization and retrieval that can be applied and used in social science research. Especially for the field of political attitudes and mentalities research, due to the fact that the social web has tremendously influenced the political discourse and lots of "trails" about political standpoints can be found on all over the world wide web, the use of advanced big data, web mining as well as semantic web and agent based techniques seems to be a fruitful direction for future studies.

Example problems that might be addressed using advanced computing techniques could be design and development of heterogeneous content repositories with central themes (studies, surveys, texts, multimedia and other more structured data about a certain idea or problem like attitudes towards government or social trust among people in a provided geographical location) as well as techniques for automated search and analytics of such repositories to assist researchers in forming hypotheses, evaluating them or developing agent based models based on empirical data as predictive mechanisms.

In order to develop such a repository as well as mechanisms for analyzing such vast collections of (often unstructured) data including tools for the development of agent-based simulation models is a major challenge and numerous advanced computing techniques would have to be employed in order to implement it. In the following a number of such techniques will be reviewed and examples for possible studies will be provided. In general, these techniques are related to the social semantic web (sometimes referred to as Web 3.0) and big data analytics.

The rest of this paper is organized as follows: in *Section 2* we provide an overview of web mining in general with a special accent on social web mining where most (social) data could be harvested from. In *Section 3* we give an overview of network theory (especially social and conceptual network analysis techniques) that might be useful and provide examples of interesting studies some of which we have conducted. In the ending subsection we present a connection between social network analysis (SNA) and agent based simulation models as outlined by (Fontana & Terna 2014) and give guidelines on how this approach can be enriched by web mining techniques.

In *Section 4* we give an overview of the possibilities of (automated) multimedia processing and analysis by describing speech recognition and machine vision techniques. *Section 5* gives a short overview of natural language processing (NLP) that allows computers to actually understand and analyze human language. In *Section 6* we present opinion mining and sentiment analysis techniques, which, as opposed to NLP which analyzes semantics (meaning), allow computers to analyze feelings a person had when writing some text. In *Section 7* we deal with so called *recommander* systems and user profiling, which allow us to automatically generate profiles of users based on their behavior on some system (for example a news portal) and use this profile to suggest new content that might be interesting for the user (e.g. a new article is proposed based on previous interests of the user).

[1] Which represent the results of social systems structural coupling to ICT, see for example (Schatten et al., 2009) or more recently (Schatten 2014) for an in-depth discussion.

[2] This paper represents a comprehensive extension of the ideas outlined in (Neumann et al. 2014, Voinea & Schatten 2014)

ISSN 2285 – 4916
ISSN-L 2285 - 4916

Section 8 presents the concept of semantic wiki systems, which extend the usual wiki systems like Wikipedia with semantic descriptors (e.g. meta-data) and thus allow the creation of vast text and multimedia repositories that can be analyzed automatically based on artificial intelligence techniques like automated reasoning. In *Section 9* we discuss the challenges that have to be addressed and draw our conclusions in *Section 10*.

2. (Social) Web Mining

In order to get to know the discipline of web mining, it is advisable to grasp the basic concept of data mining, a more general approach to data. Data mining, in its broadest term, can be defined as a process of extracting knowledge, especially informative knowledge, from a large collection of data. Web mining is considered a more specialized version of data mining, i.e. "*web mining is the means of utilizing data mining methods to induce and extract useful information from web data information*" (Xu et al., 2010). It is evident that domain of web mining lies heavily in the world wide web area and that the final product is information, or, better yet, informative knowledge based on and extracted from, the provided data.

Typically, web mining is divided into three types describing different aspects of web data and different techniques they use: web content mining, web structure mining and web usage mining. According to Ting, "*web content mining analyzes content on the web, such as text, graphs, graphics, etc*" (Ting, 2008) wherefore the main technology used in this type of web mining is natural language processing, described in further detail below. "*Web structure mining is a technique to analyze and explain the links and structure of websites*" (Ting, 2008). The most suitable concept for use in this technique, and the most convenient theory, is network theory, also described later. The main interest of this type of web mining is in creation of computer programs, called crawlers and harvesters, with the main goal of automated extraction and construction of websites' structure. Lastly, "*web usage mining can be used to analyze how websites have been used, such as determining the navigation behavior of users*" (Ting, 2008). This type of web mining is the one which is connected the most with social and psychological sciences, since it is mainly concerned with users and their behavior while using services of the Web.

It is self-evident that the web survived great development since its start, let alone in the last couple of years. Nowadays, with the maturity of web 2.0 technologies, which allow users free and uninhibited creation and publication of data on the web, where this same data becomes instantly available to a multitude of users, web is becoming an invaluable source of social data.

Social networking applications are a part of this mass data creation trend, successfully evolving a special instance of web mining, i.e. social web mining. Ting defines social networking as a concept "*usually formed and constructed by daily and continuous communication between people*" (Ting, 2008), therefore including many different social relationships, e.g. closeness among individuals or groups. Experience, thought and feelings sharing has always been incorporated in the life of humans, hence it is no wonder that the age of web 2.0 uncovered on-line social networking as arguably the most massively used feature of the web. Rapid growth of different types of communication which provide good platforms for users to communicate and share data, incorporating on-line social networking into everyday lives, include photo sharing services (e.g. Flickr), video sharing services (e.g. Vimeo), professional networking services (e.g. LinedIn), full-fledged social networking services (e.g. Facebook), blogging services (e.g. Wordpress), micro-blogging services (e.g. Twitter), instant messaging services (e.g. Skype) and many others. All these services made interpersonal communication, relationships and human behavior readily available on-line, creating easy-to-use datasets for analysis. According to Ting,

EQPAM Volume 4 No.1 January 2015
ISSN 2285 – 4916
ISSN L 2285 – 4916

"*history of social networks analysis is […] dating back more than a hundred years to around the 1900s*" (Ting, 2008)

and mostly in the field of sociology. One might argue that scientists conducting experiments around the 1900s could not predict the amount of data available today. Computing power used nowadays is capable of analyzing and reasoning on not only data provided by services mentioned above, but content in comment sections of news portals and similar services, popular wiki pages and huge collections of data, all available on the Web.

Although data is available, analyzing it is not a job easily done. Several different types of data exist, demanding corresponding method of analysis to be used: structured data, usually formatted in tables and easily read can be analyzed in a manner similar to an ordinary database (e.g. exchange rates); semi-structured data, including text data widely available on the web and containing a lot of information, demands improved methods of analysis including NLP (e.g. blogging services); unstructured data, mostly consists of images, video or audio files, and web applications (e.g. flash applications).

The world wide web allows for adding a temporal dimension to the data and, by extend, to the analysis. Collecting data through time creates a dataset which contains indications on how a certain trend developed through time. For example, following a certain Twitter hashtag during the time of presidential elections in Croatia in 2014, one could have conducted an elaborate analysis to depict the state of the people in the country and abroad, thus following popularity of one candidate over the other, grading and predicting their actions through time. Adding geographical dimension to the temporal data, it might be possible to follow an event through time and space. For example, analyzing the *Giro d'Italia* bicycling event tweets, it might be possible to follow the race in real time, following comments spectators make. Similarly, following presidential candidates on tour before elections might result in interesting insight in their actions, and the effect those actions have on the people.

Analyzing data with all its components (e.g. temporal and geographical) can lead to information which can make easier following and keeping check of your voters in elections. Furthermore, the gathered information can make you more successful in influencing them and knowing *what*, and *where* is a topic of great importance or more interest to the local population. 2012 U.S. presidential election candidates worked hard on their social networking propaganda, influencing youth, but gaining information on other parts of population as well.

There is a potential problem though, in the context. Data is indeed beneficial if one is sure of the subject the data is about. Should the data lack context, analysis becomes a tough problem. Statements might become sarcastic, or lose their sarcasm. Simple sentences could have their meaning inverted. For example, should one not know the nature of a satirical portal, one might draw conclusions based on the content of an article, misinterpret and completely miss the general idea. It is, therefore, very important to know the context of data creation, publication and consumption.

3. Social & Conceptual Network Analysis

The "*new science of network*" as Barabasi likes to call network theory, allows us to study networks regardless of their origin. In the context of social web analytics, two types of networks are of particular interest: social and conceptual. While the former represent linkages (communication, friendship, interaction, trade, cooperation etc.) between social entities (people, organizations, social groups, countries, political

European Quarterly of Political Attitudes and Mentalities EQPAM
Volume 4, No.1, January 2015
ISSN 2285 – 4916
ISSN-L 2285 - 4916

parties etc.), the latter provides insights into the structure (synonymy, mutual context, homonymy, hyperlink etc.) and dynamics (evolution of context) of concepts (words, ideas, phrases, symbols, web pages etc.).

Both social and conceptual networks are ubiquitous on the social web. For example, apart from the obvious that two people are connected if they are friends on a social networking site like Facebook, two people might be connected if they have commented on the same topic on some forum, have liked the same article or video on some news portal or podcasting site, have bookmarked the same web page or are subscribed to the same news feed. As one can see, if this topic, article, video and/or news feed has a political context we might argue that these two people have a similar political interest or attitude. If we now multiply this situation on hundreds or thousands of users, we can use social network analysis techniques (like finding connected components or clustering) to identify sub-networks of people that form groups of similar political mentalities. Depending on the particular criteria of network formation (the context of the observed network) we might form hypotheses about the actual criteria. For example, if the criterion is a video about a political speech of some political actor, we might find that some speeches of different actors actually have the same or a similar group of interested followers. Additionally, if we observe such social networks over time we might also reason about the dynamics an evolution of groups: how and when did they form? Having such data available, would allow us for example, to develop agent based models that might predict this observed behavior.

Apart from being analyzed using SNA methods, social network can be visualized to provide appealing, yet informative insights into a social system. For example, in a study we have visualized a social network of co-authors on a ICT related conference over the years (see figure 1). While such visualizations can be beautiful, they provide us with important information: in this case we can see that a certain core of authors participates with a paper every year with more or less the same co-authors, while the majority of authors participate only once and never comes back. This allows us to draw certain conclusions about this conference and might indicate to the conference organizers that some things should be changed.

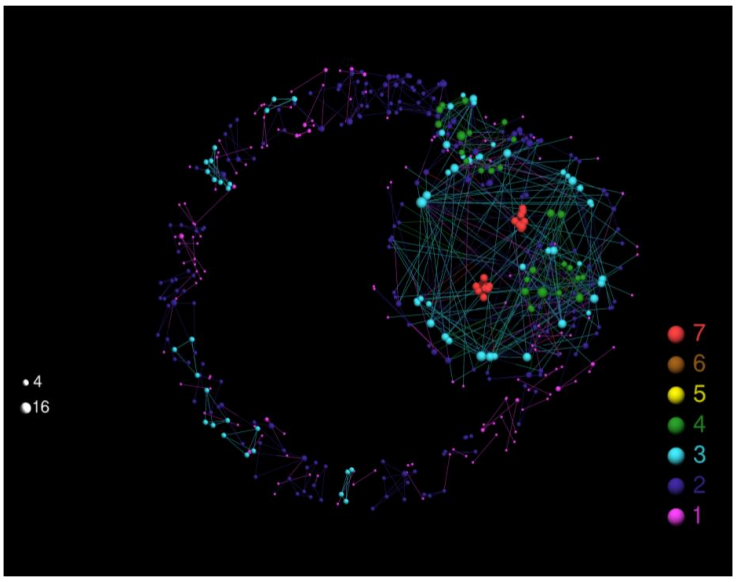

Figure 1.
Social network of co-authors from an ICT conference (Schatten et al. 2011)

EQPAM Volume 4 No.1 January 2015
ISSN 2285 – 4916
ISSN L 2285 – 4916

Conceptual networks, on the other hand, aren't maybe that obvious: two web pages where one links to the other, two keywords used on the same article, to tags used on the same on-line video, two extracted[3] concepts which appear in the same text etc. Such conceptual networks can be studied in a number of ways: in the following we will show four such techniques (tag clouds, topic maps, folksonomies and visualizations), but there are multitudes of other usable methods.

3.1. Tag Clouds

Various tagging applications are pervasive across social media. On most blogs users are allowed to tag content with keywords they believe accurately describes the content. These tags can be considered to be certain meta-data about the actual content: users who have read / reviewed the content have in a way preprocessed it and provided us with keywords that can give us important insights about it without even reading it. Especially, when this process is multiplied, e.g. when multitudes of users tag the same content, due to a certain collective intelligence these tags provide an accurate descriptor of the content.

On a lot of systems that use social tagging so called tag clouds (Smith 2008) are implemented which are simple visualizations of the tags provided by the users. Usually, the most used tags are shown in a bigger font, while less used tags are smaller. Some applications like *Wordle*[4] even have basic NLP capabilities and are able to auto-summarize text in form of beautiful keyword clouds.

While keyword clouds are not in the common sense a conceptual network analysis method, they still allow us to create a visual representation of text and context that provides an insight into the matter with a simple look on it. We have used this technique in a number of studies to visualize important keywords on research papers. For example, **Figure 2** shows a keyword cloud which was assembled by analyzing research papers dealing with data mining applications in tourism. From this visualization, one can easily comprehend which were the most important approaches, methods and technologies used in a great number of studies.

Figure 2.
A keyword cloud visualization of data mining in tourism related papers (Pejić Bach et al. 2013)

[3] By using NLP techniques for example, see below.

[4] Available at http://www.wordle.net

ISSN 2285 – 4916
ISSN-L 2285 - 4916

Similar keyword clouds can be constructed for arbitrary sets of articles with keywords / tags which might also be generated through NLP techniques. As an example we have implemented a platform which has been used for web mining, keyword cloud visualization as well as conceptual network analysis of the Croatian Scientific Bibliography (CROSBI).[5] The implemented system, which is still work in progress, used a number of technologies including web scraping (Scrapy),[6] an object-relational database (PostgreSQL)[7] and an advanced scripting language (Python).[8] The selection of technologies wasn't arbitrary, Scrapy allowed for easy implementation of a number of harvester agents that collected and extracted data from semi-structured documents that were stored in a specially designed PostgreSQL database. PostgreSQL was selected due to its unique text mining and NLP capabilities that allowed for automated dictionary based text stemming. PostgreSQL uses dictionaries to eliminate words that should not be considered in a search (so called stop words), and to normalize words so that different derived forms of the same word will match (lexemes). Python was used to glue these technologies together and provide analysis related features.

CROSBI is a social application in which Croatian scientists provide bibliographic data about their publications (Schatten, 2013). A usual entry includes authors, title, type of publication, abstract, keywords, link to document (if available), language, databases the publication is abstracted in, scientific field, category, as well as a number of additional fields depending of document type like journal name or publisher.

In order to test a number of simple and advanced text search techniques all currently available[9] bibliographic entries were harvested with a total of 385,236 distinct publications. Six searching techniques were tested and analyzed including:

- simple (naive) keyword search – keywords were searched regardless of syntax and grammar;
- NLP[10] enhanced keyword search – keywords were normalized and then searched;
- NLP enhanced title search – title was normalized and vectorized before searching;
- NLP enhanced abstract search – abstract was normalized and vectorized before searching;
- simple graph based keyword search – keyword graphs in form of folksonomy based conceptual networks (see Mika, 2007; Schatten et al., 2011; Schatten 2013 for more detailed descriptions) regardless of syntax and grammar were searched;
- NLP enhanced graph based keyword search – as in the previous method, but this time keywords were normalized.

In order to identify publications dealing with political attitudes the following keywords were used for searching: 'political influence', 'political persuasion', 'political attitude', 'contextual theory', 'social context', 'primary group', 'reference group'; as well as their Croatian translations: 'politički utjecaj', 'političko uvjeravanje', 'politički stav', 'kontekstualna teorija', 'društveni kontekst', 'primarna grupa', 'referentna grupa' respectively.

5 Available at http://bib.irb.hr
6 Available at http://scrapy.org
7 Available at http://www.postgresql.org
8 Available at http://www.python.org
9 Harvesting was conducted on October 10th 2014
10 Due to a lack for support for the Croatian language by PostgreSQL, only English words were normalized in all cases.

EQPAM Volume 4 No.1 January 2015
ISSN 2285 – 4916
ISSN L 2285 – 4916

The outlined methods gave different collections of documents in the results and likewise levels of accuracy: 24 (accuracy 87.5 %), 35 (82.9 %), 48 (87.5 %), 933 (52,5 %), 5211 (65 %), and 6245 (57.5%)[11] respectively. As one can see, results from the first three search methods yield much more accurate results, but on the other hand provide only a small number of documents. On the contrary, the last three methods provide a much larger collection of documents, but their relevancy is also much lower.

In order to further analyze these results conceptual networks of the used keywords were constructed: two keywords were considered to be related if they appear on the same publication.

The conceptual networks were visualized using keyword clouds with Wordle (**Figure 3**). As one can see from these visualizations the Croatian keywords provided a much better descriptor of the field (most important keywords translate to: social, context, political, influence, politics, adults, research, identity, ethics, public, community, education, values, sociology, planning, literacy, journal, trust etc.) since in English publications were more oriented towards history related research. This is also the reason why methods 5. and 6. gave lower accuracy.

(a) (b)

Figure 3.
Keyword Clouds of (a) English and (b) Croatian word

The current system can easily be adapted to use any other web based system as input (only the harvesting agents would need to be changed) or change the keywords to be used.

3.2. Topic Maps

Topic maps are a more advanced conceptual network visualization technique. Topic maps show the development of topics in a certain textual discourse through time. The more often a certain concept was used in some text, the greater surface on the topic map. Peaks show when a certain concept was most often used.

[11] First three methods were evaluated in full, last three based on a random sample N = 40 (confidence level ~ 20 %).

European Quarterly of Political Attitudes and Mentalities EQPAM
Volume 4, No.1, January 2015
ISSN 2285 – 4916
ISSN-L 2285 - 4916

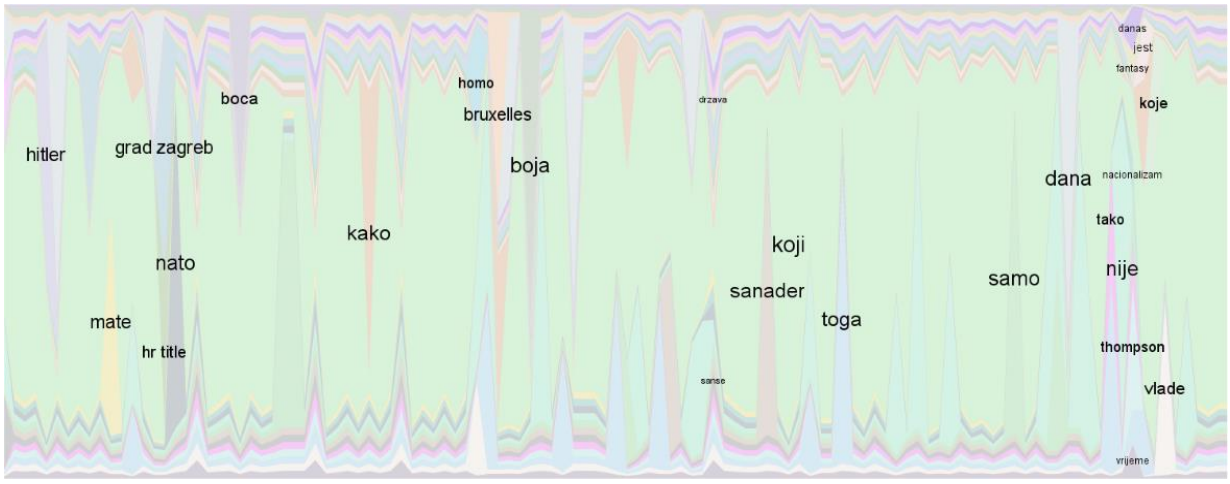

Figure 4.
A topic map from a major Croatian political blogging site (Ponjavić et al. 2010)

For example, **Figure 4** shows a topic map generated from data harvested from a major Croatian political blogging site.[12] It was used to identify most important concepts in the given time-frame, as well as the time-line of the topics that were analyzed.

While in this context topic maps were used on on-line (social blogging or forum etc.) text, they are applicable to any series of text that have a common spatio-temporal frame. For example, one might take a number of political newspaper articles from a given time frame (e.g. the 80's) from a given space (e.g. Yugoslavia), automatically extract keywords (using NLP techniques for example) and then visualize the development of topics.

3.3. Folksonomies

The term folksonomy coming from *folk* and *taxonomy* was coined by Mika (2007) to label a certain mathematical structure in a special context. He was analyzing the social bookmarking site *Delicious* on which users were able to bookmark web pages they encountered and add keywords to describe each page. These keywords were then merged from all users and used to facilitate a "socially powered" search engine. What Mika has observed was that each bookmark has three components: (1) a user who made it, (2) a web page that is being tagged, and (3) a keyword that describes the page in the mind of the user.

This observation gave rise to the idea to model the set of data obtained by bookmarking as a tripartite hypergraph (which represents the folksonomy) in which every (hyper)node consists of the three outlined components. By using a procedure called graph folding it is possible to construct bipartite graphs and moreover to connect keywords based on various criteria and thus form conceptual networks. For example, two keywords can be considered to be connected if they have been used on the same web page, and likewise if they have been used by the same user. The former criteria has been shown to yield an intuitively well-structured conceptual network of connected concepts.

[12] The data was previously filtered based on another analysis SNA techniques in order to identify authors with certain (network based) characteristics.

EQPAM Volume 4 No.1 January 2015
ISSN 2285 – 4916
ISSN L 2285 – 4916

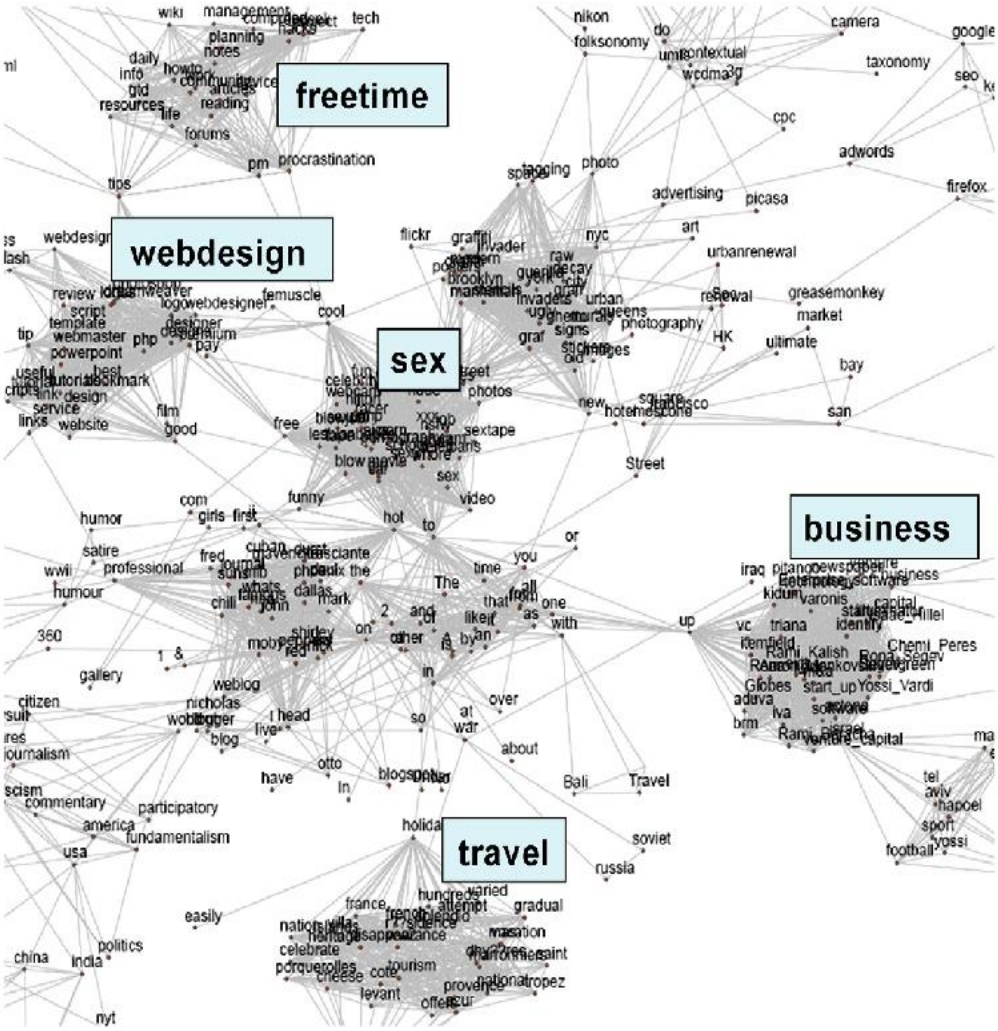

Figure 5.
Folksonomy visualization of a social tagging system called Delicious (Mika 2007)

This conceptual network can further be clustered to yield well connected components as shown on figure 5. As one might see, these components show what where the major themes of web pages users of *Delicious* bookmarked: free time, web design, sex, business and travel.

While this folksonomy model seems to be specific for bookmarking sites, we have shown in (Schatten et al. 2011) and (Schatten 2013) that it is applicable to almost any social content related data like a bibliography for example. Also, the model is not constrained on only three dimensions, but any number of dimensions can be used depending on the actual dataset. For example in (Schatten 2013) we have used 4 dimensions to analyze the Croatian scientific bibliography - each hypernode consists of (1) an author who

European Quarterly of Political Attitudes and Mentalities EQPAM
Volume 4, No.1, January 2015
ISSN 2285 – 4916
ISSN-L 2285 - 4916

has participated in writing some scientific article, (2) a keyword provided by the authors, (3) the actual article, and (4) the scientific field the author(s) have categorized the paper.[13]

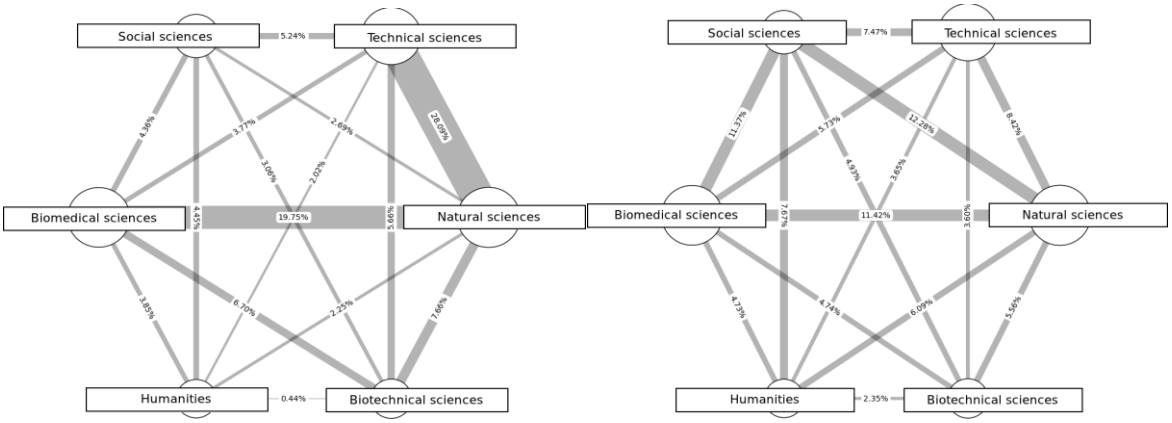

Figure 6.
Connections among scientific fields in Croatia (left - social criteria, right - conceptual criteria)

The last dimension allowed us to compare the connection between scientific fields on a global level according to two criteria: (1) social (scientific collaboration) and (2) conceptual (same keywords are used in both fields). **Figure 6** shows both such obtained networks and indicated discrepancies: e.g. some fields are mutually conceptually well connected but there was only little collaboration between them.

As one can see, the folksonomy model can be used to analyze any repository of tagged text: bibliographies, news articles with keywords, tagged blog entries, categorized wiki entries etc. The "tags" can be keywords provided by authors or other people, categories, labels, geographical or other origin, field of study, titles, topics etc.

3.4. Complex Conceptual Network Visualizations

To come back to the study outlined in the previous subsection, the constructed conceptual networks of the different scientific fields were quite complex, with hundreds of thousands of nodes. Since such data sets are hard to represent analytically or in form of tables, the field of complex network visualization has emerged.

[13] In Croatia according to the national classification there are six scientific fields (social sciences, humanities, natural sciences, technical sciences and biomedical sciences).

EQPAM Volume 4 No.1 January 2015
ISSN 2285 – 4916
ISSN L 2285 – 4916

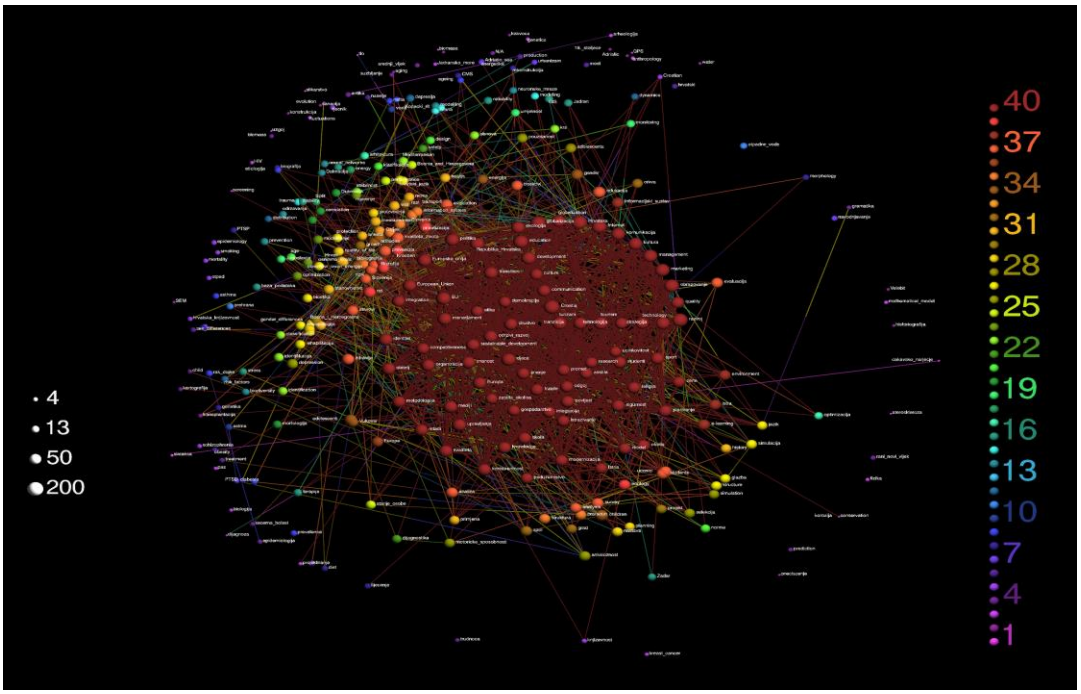

Figure 7.
Complex conceptual network visualization of most important concepts in the social sciences according to an analysis of the Croatian Scientific Bibliography (Schatten 2013)

The visualization of complex networks is an art and a science. Visualizations have to be appealing, informative and concise. Numerous algorithms for complex network visualization have been proposed and developed, each of which adjusted to different kinds of networks. For example, on Figure 7 we have used the k-core decomposition algorithm to visualize the conceptual network of keywords used in the social sciences on the Croatian scientific bibliography. What emerged in the red inner core, are the most important concepts the social sciences have dealt with in the target time frame.

3.5. Processes on Networks

While the previous examples more or less exclusively dealt with the (static) structure, one needs to mention that there are also methods that allow us to study the dynamics and evolution of networks. One particular type of methods that we want to point out here are virus spreading models (Pastor-Satorras & Vespignani 2001) which describe the mathematics of how viruses spread through a networks of people. While there seems to be no obvious connection to social science, we need to point out that a very similar mechanism is behind the spreading of rumors (Moreno et al. 2004) and information (Yang et al. 2010).

By using empirical data acquired through web mining one could develop agent based models that allow us to understand the spreading of certain themes that we want to study. For example, if we want to model a process in which some political attitude towards a certain political actor has spread through a social network, we might first harvest the communication data between people on a particular network (for example a forum, a blogging site or some news portals which allow user commenting). Then we extract only those texts (e.g. messages, articles, blog posts etc.) that deal with the political actor in question as well as temporal data (time of publication). We might then employ advanced analysis techniques to (1) construct

European Quarterly of Political Attitudes and Mentalities EQPAM
Volume 4, No.1, January 2015
ISSN 2285 – 4916
ISSN-L 2285 - 4916

the social network of people involved in the discourse, (2) identify the different attitudes towards the actor by using NLP techniques or sentiment analysis, and (3) define the actual process of information spreading in form of an agent based model for example.

3.6. SNA & ABM

A method different from those mentioned above enables analysis of a distinct set of, primarily social, entities and their interaction, is reachable by using combined power of social network analysis (SNA) and agent-based modeling (ABM), since this coalescence allows "embedding a huge amount of data in user-friendly models" (Fontana & Terna, 2014). These models ease information retrieval from the provided data through techniques of both SNA and ABM. The role of ABM is modeling agents and their interaction based on mathematical and area-specific models, whilst the role of SNA lies in unveiling the structure of interaction of the modeled agents and efficiency and stability of the network. The benefit of using SNA and ABM together, according to Fontana & Terna, is that one takes care of the problems of the other, e.g. SNA creates serious problems with exploring possible sets of nodes' configurations, since it can only be achieved by the means of combinatorics, producing an exponential number of possible sets, yet ABM makes this problem evanish, on account of the number of agents in a model being limited only by computational power. Fontana & Terna stress that the number of possible configuration remains enormous, but it is possible to mitigate this problem.

In order to successfully represent the idea of combining SNA and ABM, Fontana and Terna devised *recipeWorld* - "*an agent-based model that simulates the emergence of networks out of a decentralized autonomous interaction*" (Fontana & Terna 2014). The idea is that modeled agents are given recipes (variable number of steps to be taken in order to achieve a given end), and "*they are activated, following their internal rules and capabilities, by the events, and the network emerges as a side effect*" (Fontana & Terna 2014).

This proposed model is based on four distinct sets: (A) is the actual world populated by entities and their actual network; (B) is an ABM with agents which base their behavior on the orders and recipes derived from (A); (C) represents the network generated by (B), as represented in **Figure 8.**

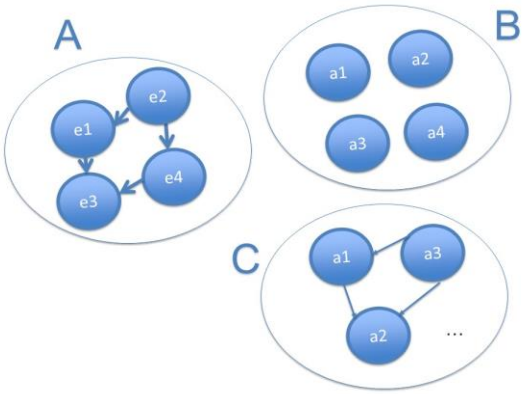

Figure 8.
Sets A, B and C (Fontana & Terna 2014)

EQPAM Volume 4 No.1 January 2015
ISSN 2285 – 4916
ISSN L 2285 – 4916

Fontana and Terna propose the possibility of populating the sets (C), (B) and (A), respectively, by knowing (D), representing known data on the network, and using inference and a sort of *reverse engineering*, as shown in **Figure 9.**

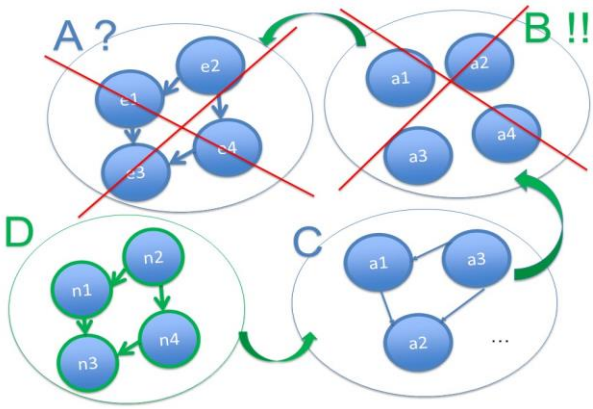

Figure 9.
Process of obtaining A, B, and C by knowing D.

Combination of the above approach with the aforementioned methods, such as web mining and network analysis, presents one with the possibility of using this *reverse engineering* approach and creating agent-based models based on gathered social web data. For example, social web mining might provide one with enough data (D) to model a political network (C) showing e.g. a number of political parties and their interaction, or some other social entities, including voters, political figures, numerous governmental and non-governmental organizations, societies, or even countries. In turn, this network (C) can serve as the basis for an ABM, (B), hence allowing for simulation of real-world entities and their actual network (A).

4. Speech Recognition & Computer Vision

Since the social web does not only include textual data, but also multitudes of audio, video and image data, in order to analyze it in an automated or semi-automated fashion, one needs to employ advanced computing techniques.

Speech recognition is a technique that transforms recorded audio of human speech to text. This can be used to transcript various audio tapes and videos that include human speech, identify different speakers and then use the above mentioned techniques to analyze it. While speech recognition is a well-developed technology, there is a serious drawback of using it to study specific materials: it is in most cases language dependent, which means that for each language one wants to analyze, one needs and adequate speech to text system for that particular language. If there is no such system available, and there often isn't for some languages, one needs to develop such a system which is a long and difficult undertaking.

Computer vision is a field of computer science that aims on allowing computers to analyze image and/or video data in order to recognize objects that are on the given image/video. As opposed to the previous one, these methods aren't language dependent, but are even worse object dependent, which means that for every type of object one wants to recognize, there needs to be an adequate system which recognizes it. This is especially true for person recognition or biometrics (see Bača et al. 2006) where one needs to train the system to recognize every single person of interest (this process is called enrollment).

European Quarterly of Political Attitudes and Mentalities EQPAM
Volume 4, No.1, January 2015
ISSN 2285 – 4916
ISSN-L 2285 - 4916

Still, there are some types of objects which can more or less be recognized by even simple system like text on an image or in a video. Such text can then be used in conjunction with the recognized speech.

5. Natural Language Processing

Natural Language Processing (NLP) can be defined as a process of making a computer system understand the meaning behind written and/or audio data that contains language based information. To give a more comprehensive definition one can quote (Cohen, 2004) which states that NLP "*is normally used to describe the function of software or hardware components in a computer system which can analyze or synthesize spoken or written language*". The goal of systems that analyze and process textual or audio data in a given natural language is to achieve a level of language understanding that is characteristic for humans and their language processing abilities. The main problem in achieving this goal is the complexity of the phenomenon that is the target of NLP: natural language. One of the examples is the word '*bank*': it can be a financial institution, a river shore, relying on something etc. The other problem is that language itself is a living entity that evolves during time. The language itself consists of grammar (the logical rules of combining words of the language to make comprehensive sentences that have meaning) and the lexicon (the words and the situations of their usage). There are specific linguistic tools that make it easier for the algorithms to access, decompose and make sense of a sentence already available for use in processing NLP data. Those tools are as follows (Cohen, 2004):

- Sentence delimiters and tokenizers – detecting sentence boundaries and determining parts of sentences (based on punctuation characters)
- Stemmers and taggers – morphological analysis that links the word with a root form and word labeling giving information if a word is a noun, verb, adjective etc.
- Noun phrase and name recognizers – labeling the words with the noun phrase (e.g. adjective + noun) and recognizing names
- Parsers and grammars – recognizing well-formed phrase and sentence structures in a sentence.

The entire process can be broken down in to several steps. In general, the entire text is first broken down into paragraphs, paragraphs into sentences and sentences into individual words that are then tagged (to recognize parts of speech among other) before the parsing begins. The full suite of tools available are *sentence delimiters, tokenizers, stemmers* and *parts of speech taggers*, but they are not used in full in all situations. The role of sentence delimiters and tokenizers is the determination of the scope of the sentences and identifying the members of a sentence. Sentence delimiters try to find the boundaries of a sentence, which can be a hard task since the usual sentence endings (e.g. period) can represent other meaning. They are usually created by using expression rules. Tokenizers segment a list of characters into meaningful units that are called tokens. Creating tokens is language dependent as there are differences in how different languages mark word breaks (Latin based languages use white space as a word delimiter). They are usually created using rules, finite state machines, statistical models and lexicons. Stemmers are used to find out the root form of a word. There are two types of stemmers:

a) inflectional, which express the syntactic relations between words of the same part of speech (focus is on grammatical features such as present/past or singular/plural)
b) derivational, which try to bind different words to the same root (e.g. kind/unkind share the same root). They are supported by the use of rules and lexicons (they relate any form of a word to its root form).

EQPAM Volume 4 No.1 January 2015
ISSN 2285 – 4916
ISSN L 2285 – 4916

Parts of speech (POS) taggers label the grammatical type of a word, assigning labels and deciding if a word is a noun, a verb, an adjective, etc. Since the same sentence can have more than one meaning, often POS taggers tag the same word with more than one label. POS taggers can be

- rule based, that rely on linguistic knowledge to rule out tags that are syntactically incorrect, or
- stochastic, that rely on training data and assign tags based on frequency probabilities (they are computed from the supplied training set that was matched by hand).

Although POS taggers are very useful in determining the structure and type of words in a sentence, some tasks require their more specific use. At that point, noun phrase parsers and name recognizers are used. Their goal is to identify major constituents of a sentence (e.g. a name).

The history of NLP and its development has seen several main approaches in implementing previously mentioned tools. The early 1950's, which are regarded as the decade where NLP started,

- Symbolic approach
- Statistical approach

Symbolic Approach is based on the existence of a predefined grammar that is then used in data analysis. The grammars consist of known grammatical patterns and their meaning which are then looked upon in the analyzed data. Due to the limitations of grammar based approach, researchers started to look at the possibilities of other approaches. The statistical approach is based on machine learning and data mining and this approach tries to use available data and allow for the algorithms to learn based on them. The basis of this approach is the existence of a corpora which is, in fact, a hand annotated with the appropriate linguistic labels (e.g. noun, first name, person, etc.) or concepts (e.g. person, building, organization etc.) and an learning algorithm which then uses the corpora and tries to "learn" new concepts from an unlabeled data set based on the rules extracted from the corpora. This approach itself has also seen a change in its approach. Primarily, the approach was based on decision trees (as similar to the grammars used in symbolic approach) which are derived from a number of if else statements in order to analyze the data to modern, statistical approaches where each recognized concept is defined through a weighting scheme/system. This approach is the basis for all modern NLP algorithms, although their combination is also a frequent approach.

As previously stated, the main current approach to NLP is the use of statistical through machine learning and data mining algorithms. Although this approach has many advantages over symbolic approach one has to have a corpus, a previously annotated data set used as the basis for the algorithms used in the implemented algorithms' learning phase. As it is to be expected, the languages with large linguistic support are languages that have many users (e.g. English). Central and Eastern European languages are, unfortunately, still largely unsupported as the necessary tools for a good computational language understanding are still either missing or highly undeveloped. The overall support for Central and Eastern European languages is given in **Figure 10** and **Figure 11**. A good starting point for these languages is the META-NET repository (Kapetanović, Tadić, & Brozović-Rončević, 2012).

European Quarterly of Political Attitudes and Mentalities EQPAM
Volume 4, No.1, January 2015
ISSN 2285 – 4916
ISSN-L 2285 - 4916

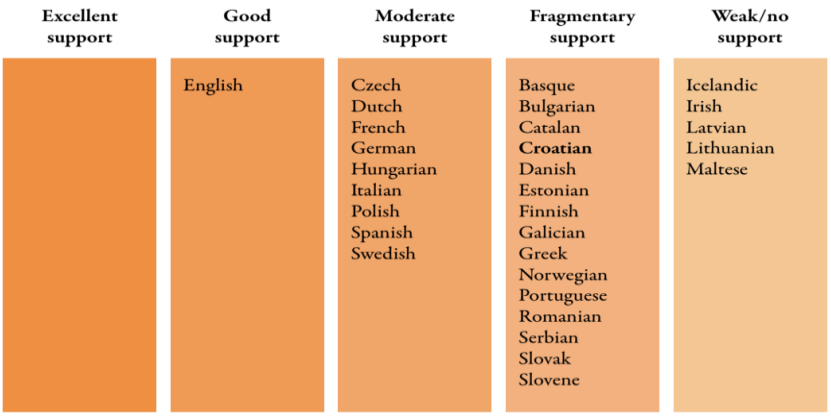

Figure 10.
Speech and text resources: state of support for 30 European languages (Kapetanović et al., 2012)

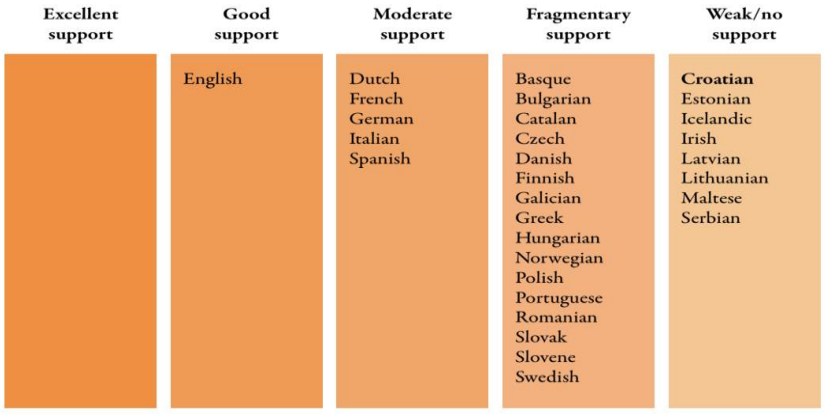

Figure 11.
Text analysis: state of language technology support for 30 European languages (Kapetanović et al.., 2012)

NLP can find its application in numerous analyses including but not limited to finding most important concepts in some corpora of text, auto-summarization, automated classification of documents and similar.

6. Opinion Mining & Sentiment Analysis

Sentiment analysis or opinion mining refers to the application of NLP, computational linguistics, and text analytics to identify and extract subjective information in source materials (see Pang and Lee 2008). This amazing technology allows us to identify the sentiment (e.g. the feelings) a person had while writing some text or message.

There are numerous application areas of sentiment analysis including but not limited to: (1) product/services research for marketing purposes, (2) better search engines, (3) campaign monitoring (e.g. political, business, showbiz, etc.), (4) recommendation systems (see below), (5) detection of "flames", (6)

EQPAM Volume 4 No.1 January 2015
ISSN 2285 – 4916
ISSN L 2285 – 4916

public health (e.g. emotional/mental status, detecting potential suicide victims etc.), (7) government intelligence, (8) trend monitoring (e.g. social, cultural, political etc.) and similar.

The possible use of sentiment analysis in the study of political attitudes seems to be obvious: for example one might want to study the sentiments a social group has had towards a given political actor over some period of time. What one needs to do is: (1) harvest all communication data from one or more social media applications or other sources like newspapers and similar in the given time frame, (2) filter the data and find only those that contain comments on the actual actor that we want to analyze, and (3) feed this data to the sentiment analysis tool (or tools) for the given language to obtain sentiments over time.

Again, similarly to speech recognition and NLP, sentiment analysis is language dependent. In order to use sentiment analysis on a given text, there has to be a sentiment analysis database or system for the given language. Sadly, there are no sentiment analysis systems for some languages, for example only recently there were initiatives to implement a Croatian sentiment database (Schatten 2012; Glavaš et al. 2012). What also needs to be mentioned here is that types of sentiments in various systems are limited: often only positive or negative (with various degrees between these two polarities) are recognized. Complex sentiments are hard to detect, and thus often not implemented or an active field of study. Additionally, since opinion mining applications have a potentially big commercial value, they are often not freely available.

7. Recommender Systems & User Profiling

Recommendation systems are computer systems whose goal is to recommend items, available through the system itself, to its users. Recommendation systems are employed in numerous domains: on-line shopping sites (e.g. Amazon, eBay etc.), news portals and other. One of the major application areas of NLP tools, discussed previously, are contextual recommendation systems. Contextual recommendation systems are based on the idea that it is possible to recommend items of interest to users of the computers systems based on the meaning derived from the recommended items context. In that process it takes to consideration three factors:

- the context of each information node, with descriptors generated by the use of NLP tools
- user profiles which are generated based on generated browsing data and
- users preferences.

The available data sources that are used in obtaining the data needed for the analysis include data from web server access logs, proxy server logs, browser logs, registration data, user session or transactions, cookies, user queries, bookmark data, mouse clicks and scrolls (clickstream data) or any other data as the result of interaction between the user and the recommendation system (Singh & Singh, 2010). Data collected in this manner is the base of user profiles. User profile can be defined as the set of "*all the available personal information about a user*" (Carmagnola, Cena, & Gena, 2007). There are numerous approaches in building a contextual recommendation system. Ševa (2014) presents one of the developed approaches in recommendation systems. This research was based on the Open Directory Project data and the classification taxonomy developed based on it. In this research effort user profiles were generated based on visited documents and newly generated content was recommended based on the underlying taxonomy-based classification. The focus of this work was to achieve an individual recommendation system. In majority of recommendation systems, due to the number of users, recommendations are not generated on an individual level but based on user groups/clusters. User data is still collected for each individual user and the data itself is processed in the same manner. The only difference is the use of analyzed data where similar users are grouped into interest groups or clusters and

ISSN 2285 – 4916
ISSN-L 2285 - 4916

newly generated content is then recommended (or not recommended) to the entire cluster and all cluster users. This approach is not only applicable in the field on news portals or on-line retailers but can also be employed in a number of other cases. One of the fields where this approach is very useful is the possibility of mining for political sentiment in a country/region. Users can be connected to various political agendas and one can model social entities and create a interest-based social network based on the created user profiles and the context of their interaction with the analysed computer system.

8. (Semantic) Wikis

The semantic web (Berners-Lee et al. 2001) is the idea of adding machine readable meta-data to web pages in order to allow computers to understand the content they are processing. The world wide web, in its current form, is made for people: people can search for data, find various sources and services, integrate and use them. When trying to do that by using a computer program, it becomes a non-trivial, hard to solve problem. By adding semantics in form of so called ontologies[14] computer can use various advanced artificial intelligence techniques to reason about the data and use it in various contexts.

Wikis, as mentioned previously, are an interesting social web application which allows users to collaboratively work on a repository of content. Wikipedia is the obvious example, where millions of users contribute to create a huge on-line encyclopedia. Semantic wikis include ideas from the semantic web in order to facilitate automated reasoning, complex querying mechanisms and automated categorization of content (see Schaffert 2006; Schatten 2010 for example systems, or Auer et al. 2007 for a very interesting project called DBPedia).

Semantic wikis can be used to manage big repositories of textual and multimedia data. For example in a recent study (Schatten et al. 2014) we have analyzed a large corpus of organization theory literature[15] by using a semantic wiki system called TaOPis. By collaboratively editing content and just adding meta-data about the content (e.g. keywords, categories and similar) a social taxonomy emerged shown on figure 8. This taxonomy allowed for automated summarizing of text using built-in queries. Later we implemented two intelligent agents to find conceptually similar terms and descriptors of each concept described on the semantic wiki.

[14] One needs to mention here that the term ontology is not used in the context of philosophy, but computer sciences where it represents a formal description of a domain.
[15] More than 230 various references have been considered including books, articles and Websites.

EQPAM Volume 4 No.1 January 2015
ISSN 2285 – 4916
ISSN L 2285 – 4916

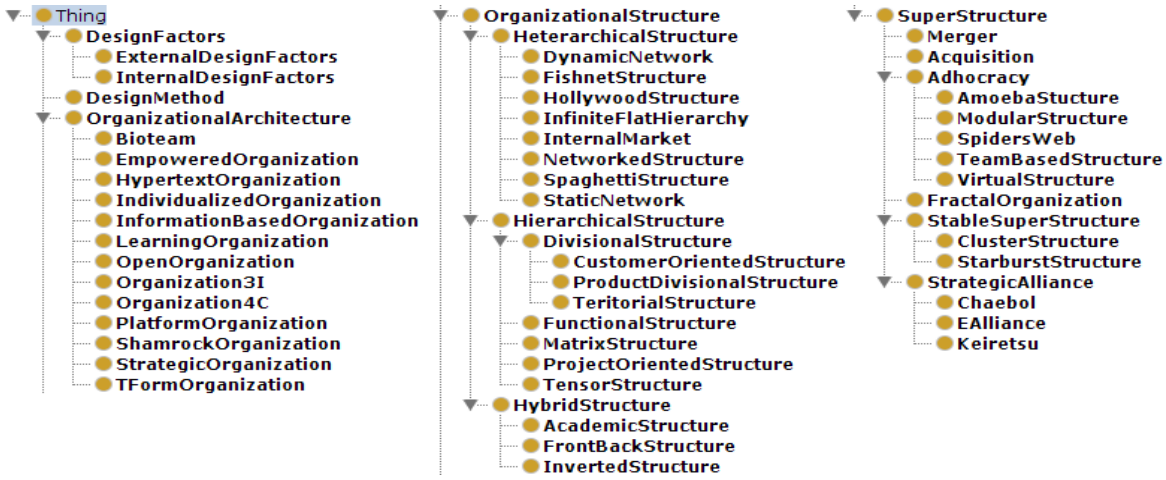

Figure 12.
Taxonomy of organizational forms exported from an organization theory semantic wiki (Schatten et al. 2014)

When working with large *corpora* of textual or multimedia content, on one hand, and in a collaborative environment, on the other, semantic wikis provide an excellent tool to organize research data and yield needed structural meta-data for data mining. With only few additions to contemporary semantic wiki systems, a scientific collaboration platform can be easily established.

9. Discussions & Challenges

As one can see from the previous contemplation, social semantic web mining and big data analytics can provide a substantial addition to social science research methodology. To come back to the example problems outlined in the introduction, assume that we are a group of scientists distributed across Europe, and want to assess the impact of the communist legacy for the development of democracy and Europeanization in Eastern Europe. We might start by establishing a heterogeneous content repository to gather all relevant studies, surveys and quantitative data from reliable scientific and professional sources. Afterwards, we could enrich this repository with (possibly to be digitized) data gathered from news, TV and radio archives as well as data available on the social semantic web. Adapted semantic wiki systems with cloud computing facilities would be suitable outlet to start such an undertaking. Such a repository, could allow for the creation of so called shadow configurations as outlined by Voinea (Voinea & Schatten 2014) by connecting various sources through adding meta-data (conceptual descriptors).

After the repository has been built, we might work on preparing and preprocessing data of interest, e.g. extracting text and labels from audio and video material, transforming data from various sources into a common format, creating structured databases with most relevant data for faster access etc. Now we have everything setup to analyze the data in various ways.

One possible analysis might be to identify most important political actors in the dataset and map them to relevant speeches, comments and news items as well as other actors. This might be achieved through mining and NLP techniques as well as social and conceptual network analysis but also expert opinions. A next step might be to identify social groups of interest which could be achieved through user profiling. Then we might analyze the content of discussion inside or between these social groups by

ISSN 2285 – 4916
ISSN-L 2285 - 4916

visualizing discourse with topic maps, measuring the sentiment over time or modeling the spread of various attitudes and concepts through their networks. Each social group and/or actor might be represented with a conceptual network (similar to perceived fields of competence in the sense of Neumann et al. 2014) which might be represented as a simple tag cloud or as complex network visualization. One of the important features of network theory is that we can model different systems, represented with heterogeneous data using the same formalism: this is important since it allows us to compare such systems. The various results of analyses could be cross-compared to find relations between processes that were going on and thus lead us to hypotheses about how one conceptual network dealing for example with communist terminology has influenced another dealing for example with Europeanization. Such hypotheses might be modeled using agent based techniques and then simulated and evaluated.

Still, there are important challenges in such research projects that need to be outlined here: (1) there are vast amounts of data available (this is why this field is called big data) and often such mining and analysis can be like looking for the needle in the haystack; (2) heterogeneous data is not simply integrated, especially when having the first challenge in mind - various file formats, different web sites, different types of data (e.g. audio, video, text etc.), have to be cleansed, reorganized and stored in a way that makes analysis possible; (3) different languages (which is a special issue of the previous challenge) can yield major problems - non-existing tools for NLP and/or sentiment analysis, heterogeneous tools which have to be integrated etc. Thus, before starting such a project all the above mentioned challenges have to be considered in detail, e.g.: (1) sources of data have to be identified and limited to a reasonable level, (2) time and effort for data digitalization, cleansing and reorganization have to be taken into account including adequate storage facilities (hardware or cloud services), (3) languages to be analyzed have to be known in advance and depend in general on the available tools - some tools might be implemented, but the implementation of others might not be feasible.

10. Conclusion

In this paper we have given an introduction to social semantic web and big data analytics for possible use in social science research projects. Methods like social web mining, social and conceptual network analysis, speech recognition, computer vision, natural language processing, sentiment analysis, recommender systems, user profiling as well as semantic wikis have been identified as possibly fruitful additions to social science methodology. The methods have been illustrated with example studies where possible as well as ideas of possible applications in concrete research scenarios. Afterwards a more detailed research problem dealing with political attitudes in post-communist countries of Eastern Europe as well as a possible solution using the outlined techniques has been described. In the end important challenges have been identified and recommendations on how to approach these challenges have been given.

References

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z. (2007). DBpedia: A nucleus for a web of open data, Springer Berlin Heidelberg, pp. 722-735.

Bača, M., Schatten, M., & Rabuzin, K. (2006). Framework for systematization and categorization of biometrics methods. Information and Intelligent Systems: IIS 2006.

Berners-Lee, T., Hendler, J., Lassila, O. (2001). The semantic web. Scientific american, 284(5), pp. 28-37.

Carmagnola, F., Cena, F., Gena, C. (2007). User modeling in the social web. In Knowledgebased intelligent information and engineering systems 4694, p. 745.

EQPAM Volume 4 No.1 January 2015
ISSN 2285 – 4916
ISSN L 2285 – 4916

Cohen, K. B. (2004). Natural Language Processing for On-line Applications: Text Retrieval, Extraction and Categorization (review). Language, 80 (1), pp. 178-178.

Fontana, M., Terna, P. (2014) From Agent-based models to network analysis (and return): the policy-making perspective, SwarmFest 2014, University of Notre Dame.

Glavaš, G., Šnajder, J., Bašić, B. D. (2012). Semi-supervised acquisition of Croatian sentiment lexicon. In Text, Speech and Dialogue, Springer Berlin Heidelberg, pp. 166-173.

Luhmann, N. (1984) Soziale Systeme: Grundriß einer allgemeinen Theorie Suhrkamp.

Kapetanović, A., Tadić, M., Brozović-Rončević, D. (2012). The Croatian language in the digital age / Hrvatski jezik u digitalnom dobu. In White paper series (p. 101). Springer.

Mika, P. (2007) Ontologies are us: A unified model of social networks and semantics. Web Semantics: Science, Services and Agents on the World Wide Web. 5(1), pp. 5–15.

Neumann, M., Srbljinović, A., Schatten, M. (2014) "Trust me, I know what I'm doing!" Competence Fields as a Means of Establishing Political Leadership. European Quarterly of Political Attitudes and Mentalities. 3(2), pp. 18-33.

Pang, B., Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and trends in information retrieval, 2(1-2), 1-135.

Pastor-Satorras, R., Vespignani, A. (2001). Epidemic spreading in scale-free networks. Physical review letters, 86(14), pp. 3200.

Pejić Bach, M., Schatten, M., Marušić, Z. (2013) Data Mining Applications in Tourism: A Keyword Analysis, In Hunjak, T., Lovrenčić, S., Tomičić, I. (Eds.) Proceedings of the 24th Central European Conference on Information and Intelligent Systems, Varaždin : Faculty of Organization and Informatics, pp. 26-32.

Moreno, Y., Nekovee, M., & Pacheco, A. F. (2004). Dynamics of rumor spreading in complex networks. Physical Review E, 69(6), 066130.

Schaffert, S. (2006). IkeWiki: A semantic wiki for collaborative knowledge management. In Enabling Technologies: Infrastructure for Collaborative Enterprises, 2006. WETICE'06. 15th IEEE International Workshops on. IEEE, pp. 388-396.

Schatten, M. (2010). Programming Languages for Autopoiesis Facilitating Semantic Wiki Systems, PhD thesis. University of Zagreb.

Schatten, M. (2011) Mining Social and Semantic Network Data on the Web, Seminar za metodologijo in informatiko. Novo Mesto, Faculty of Information Studies.

Schatten, M. (2012). Opinion Mining on News Portal Comments - Towards a Croatian Sentiment Database. Seminar za metodologijo in informatiko. Novo Mesto, Faculty of Information Studies.

Schatten, M. (2013) What do Croatian Scientist Write about? A Social and Conceptual Network Analysis of the Croatian Scientific Bibliography. Interdisciplinary description of complex systems. 11(2), pp. 190-208.

Schatten, M. (2014) Structural Couplings of Organizational Design and Organizational Engineering, In Magalhães, Rodrigo (Ed.) Organization Design and Engineering - Coexistence, Cooperation or Integration, United Kingdom : Palgrave Macmilllan, pp. 184-201.

Schatten, M., Bača, M., Ivanković, (2009) Public Interfaces as the Result of Social Systems Structural Coupling, In M. Mertik, M. & Damij, N. (Eds.) Proceedings of the 1st International Conference on Information Society and Information Technologies ISIT 2009, Faculty of information studies in Novo mesto.

Schatten, M., Grd, P., Konecki, M., & Kudelić, R. (2014). Towards a Formal Conceptualization of Organizational Design Techniques for Large Scale Multi Agent Systems. Procedia Technology, 15, pp. 577-586.

Schatten, M., Rasonja, J., Halusek, P., Jakelić, F. (2011) An Analysis of the Social and Conceptual Networks of CECIIS 2005 – 2010, In Hunjak, T., Lovrenčić, S., Tomičić, I. (Eds.) Proceedings of the 22nd Central European Conference on Information and Intelligent Systems, Varaždin : Faculty of Organization and Informatics, 2011. pp. 259-264.

Singh, B., Singh, H. K. (2010). Web Data Mining research: A survey. In 2010 ieee international conference on computational intelligence and computing research. IEEE, pp. 1-10.

Smith, G. (2008) Tagging: People-Powered Metadata for the Social Web, Berkeley, New Riders.

Ševa, J. (2014). Web News Portal Content Personalization using Information Extraction Techniques and Weighted Voronoi Diagrams (PhD thesis). University of Zagreb.

Ting, I-H. (2008). Web mining techniques for on-line social networks analysis. In Service Systems and Service Management, 2008 International Conference on, 2008, pp. 1-5.

ISSN 2285 – 4916
ISSN-L 2285 - 4916

Voinea, C. F., Schatten, M. (2014) Recovering the Past. Eastern European Web Mining Platforms for Reconstructing Political Attitudes, In Voinea, C. F. (Ed.) European Conference on Political Attitudes and Mentalities ECPAM, Bucharest, Romania.

Xu, G., Zhang Y., Li, L. (2010). Web Mining and Social Networking: Techniques and Applications, Springer Science & Business Media.

Yang, J., Yao, C., Ma, W., Chen, G. (2010). A study of the spreading scheme for viral marketing based on a complex network model. Physica A: Statistical Mechanics and its Applications, 389(4), pp. 859-870.