2010 14th International Conference Information Visualisation

Visualization of Web Page Content Using Semantic Technologies

Lorand Dali
Department of Knowledge Technologies
Jožef Stefan Institute
Ljubljana, Slovenia
lorand.dali@ijs.si

Dunja Mladenić
Department of Knowledge Technologies
Jožef Stefan Institute
Ljubljana, Slovenia
dunja.mladenic@ijs.si

Abstract— **This paper presents a system for visualizing the information contained in the text of a web page. The goal of the visualization is to help the users better and faster understand the text on a web page and/or find related content on the internet. These visualizations are possible due to the use of text mining, natural language processing and semantic web technologies. Our system tries to make these technologies instantly accessible to a wide variety of users reading a wide variety of web pages. This high coverage of both users and content can be achieved because the system is implemented as an extension to Firefox, one of the most popular browsers, and because the visualizations are computed on the fly for any page the user happens to be reading at a given moment.**

Visual content enhancement, semantic technologies, text mining

I. INTRODUCTION

Understanding information provided at Web pages is often non-trivial, especially if the pages are not carefully prepared or are taken out of the context. For instance, long pages with poor structure or, pages that assume familiarity with some related content not directly provided. In this paper we propose a system that integrates state-of-the-art technology of text mining, natural language processing and semantic web providing intuitive and easy to use extension of internet browser.
The functionality of the system is the following:

- Presenting the keywords extracted from the text as a list ordered by importance
- Finding web pages which are related to the current one
- Finding related news articles
- Finding pictures which are related to the text of the current web page
- Classification of the current web page into the open directory project[1]
- Summarization
- Showing a word cloud for the current page
- Extraction of triplets from text
- Visualization as a semantic graph
- Annotation with semantic entities form the open data cloud

[1] http://www.dmoz.org/

Figure 1 shows a screenshot of the system which contains the slidebar with the visualization options on the left, and the page being analyzed on the right. Currently the view which shows the summary of the web page, is selected. The visualization options in the slidebar on the left are represented by icons each corresponding to one of the functionality from the list above.

The paper is structured as follows. Section II briefly describes related work., Section III describes the features of the system, while the implementation details are given in Section IV. Section V describes the experiments, and the last section draws the conclusions.

II. RELATED WORK

In this section we mention three systems similar to ours: Glydo[2], Headup[3] and SimilarWeb[4]; all three of which are available as Firefox extensions.

Glydo is a toolbar which provides information related to the current web page. Its features consist of: recommendation of similar web pages, similar videos on youtube, related news stories, information about important named entities.

Headup annotates semantic entities on a page by underlining them. The system finds additional information related to the annotated entities, and presents it in a silverlight widget. It is also possible for the user to select a piece of text which is not annotated and get information related to the selection. The related information provided for a semantic entity is the following: summary of the wikipedia article about the selected entity, related pages, related news and blog posts, videos from youtube, pictures from flickr, search results for the selected entity.

SimilarWeb is a system recommending related pages, related news and related tweets for any web page the user is visiting. Similarly to our system, the information is displayed in the sidebar of the browser.

In our opinion all three of the related systems are very good and useful, but none of them provides as much functionality as our system, which in addition provides a summary[1]of the current page, triplets[2], semantic graphs[3][6], word cloud, and classification[5]. According to our experiments these features are the most helpful for

[2] http://www.glydo.com/
[3] http://www.headup.com/
[4] http://www.similarweb.com/

1550-6037/10 $26.00 © 2010 IEEE
DOI 10.1109/IV.2010.49

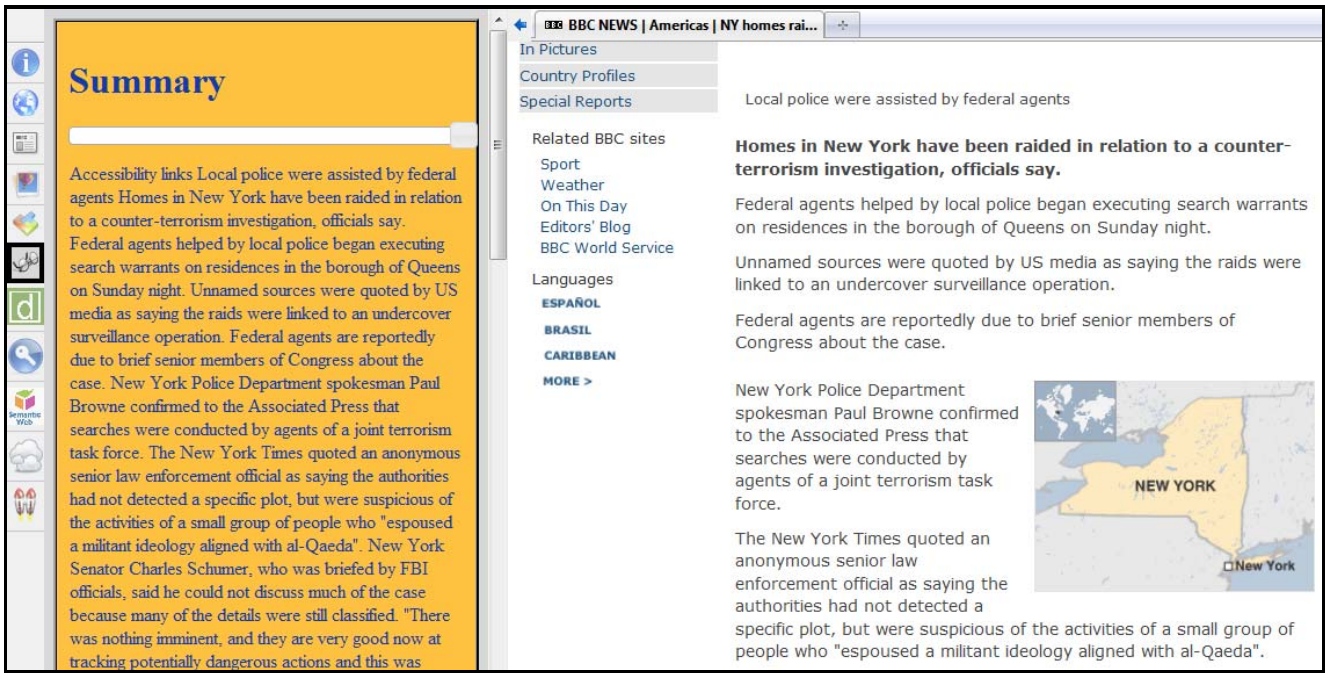

Figure 1 Screenshot. On the right is the page being analyzed, on the left is the slidebar with the visualization options. In this picture the summary view is selected

better understanding a given page. What our system doesn't have is related videos and related tweets.

III. FUNCTIONALITY DESCRIPTION

The system is a Firefox extension and offers in the slidebar[5] several options for visualizing the text of the web page in the currently focused tab. The following subsections explain what these options are and how they work.

A. *Keyword extraction*

One of the most common and straightforward methods of analyzing text is the extraction of the most important words as keywords. Our system implements the TextRank[4] keyword extraction algorithm which is an unsupervised graph based method. The text is represented as a graph having where the nodes are the words from the text. The weights of the edges are computed based on how many times two words co-occur and how close they are to each other. Finally the PageRank[9] algorithm is used to rank the words and extract the higher ranking ones as keywords.

B. *Related web pages, news articles and pictures*

This feature tries to suggest to the user content which is relevant to the current Web page. Related web pages, news and articles are found by using the BOSS[6] search services provided by Yahoo. First, the important keywords are extracted from the current page using functionality described in Section III.A, then, with the top ranked keywords, Yahoo web search, news search and image search is performed.

C. *Classification*

When doing classification we try to find the relevant topic categories into which the given page fits best. We are using DMoz (Open directory) taxonomy of topic categories with associated Web pages, as a large, well established taxonomy that is commonly used in research experiments of document classification[5]. Figure 2 shows the results. Because dmoz is a hierarchy of categories, the classification is also hierarchical, e.g. arts → television → programs → talk shows. For classifying we have used the TextGarden[7]library, and the hierarchical algorithm is described in detail in[5]

D. *Summarization*

From summarization in general we expect to reduce the amount of text to be read; so we eliminate much of the textual content while still retaining the most important parts. The summarization approach adopted is extractive; so the main task is to of the algorithm is to select the most important sentences form the original text.Hence we have a problem of sentence ranking which we solve by the TextRank[4]method, which is unsupervised and graph based. Once the ranking of the sentences is obtained, the user can adjust with a slider the number of sentences he would like to have in the summary. A summary example can be seen in Figure 1.

[5] The slidebar is a piece of user interface in Firefox which can show information at the side of the browser and can be hidden or shown at the will of the user

[6] http://developer.yahoo.com/search/boss/

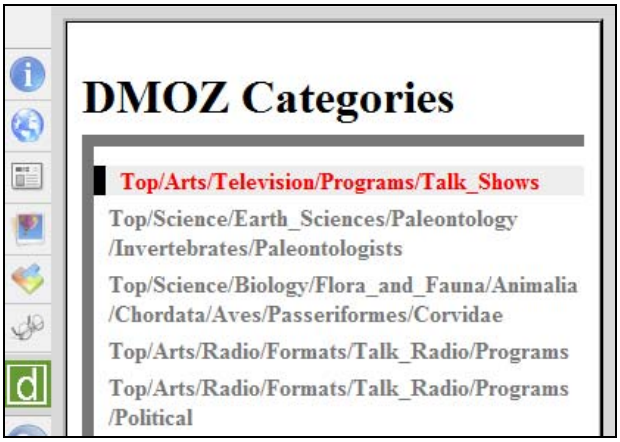

Figure 2 Classification into the open directory project

E. *Word Cloud*

A word cloud is a type of text visualization where the words of the text are displayed, and their sizes are directly proportional with their frequency of occurrence. We have used the wordle[7] word cloud software to generate the word cloud images which are displayed.

F. *Triplet extraction*

Triplets[2]are like short statements or facts. They are extracted from each sentence and are usually made of the subject the predicate and an object form the sentence. Examples of triplets would be (Jay Leno - has - tonight show) or (week guest list - includes - Sarah Palin). We believe that it is useful to show such triplets to the user because they are sentences reduced to their essence and the user can see at a glance who did what to whom. We have used a Web service provided as a part of content enrichment systems[8]. The algorithms for extracting triplets from a sentence are described in detail in[2]

G. *Semantic Graph*

This type of visualization allows the user to see the text as a directed graph as shown in Figure 3. The nodes are important words and phrases occurring in the text. In our visualization, the yellow nodes represent verbs while the green nodes are nouns. The edges show subject → predicate (from green to yellow) and predicate → object (from yellow to green) relations. The orientation of the edges is determined by the fact that the edges are wider at the start than at the end. Three consecutive nodes where both the first and the last are nouns (green) make up a subject → verb → object triplet. These are the same triplets described previously.

Apart from the visual role, the semantic graph also has summarization purposes. The user can adjust the size of the graph, and when the size decreases, the less important nodes disappear. To achieve this, a node ranking model has been learned using supervised learning. The features are graph features and text features. More details about node ranking and how to build and use semantic graphs can be found in[6]

The semantic graph is displayed as a Java applet. For computing the layout and displaying the semantic graph, we have used the TouchGraph[8] library.

H. *Semantic annotation*

By annotation of semantic entities we mean extracting named entities (people, locations and organization) and linking them to the open data cloud[9]. Figure 4 shows how we display the semantic entities and their links in the slidebar. The annotations are obtained by using the Enrycher[10][8] web services.

IV. Implementation Details

We have implemented the system as a Firefox extension on top of the jetpack[11] platform. Jetpack is a Firefox extension which facilitates the development of other extensions by exposing much of the Firefox user interface through JavaScript APIs. Moreover, jetpack supports the jQuery[12] library which is a very useful tool for web development because it enables the usage of the AJAX technology. Hence the jetpack extension is a prerequisite to run our extension. On the browser side, the document ready, tab focus and slide select events are handled. In other words, when a page is (re)loaded the plain text is extracted from the html and HTTP web services are called to do the necessary processing on the server side. The necessary processing i.e. the web services which must be called is determined by the slide which is currently selected in the slide bar. If the user selects another slide (i.e. another visualization option), then new web services, which generate the required visualization, are called. If the user focuses another tab, then the information displayed in the slidebar is updated in order to show the visualization of the newly focused page. All the HTTP web services are called from the JavaScript event handlers using the AJAX technology.

On the server side, the processing algorithms (classification, summarization, keyword extraction, semantic graph etc.) are implemented as HTTP web services which when called from AJAX get the text as input and send the processing results in JSON or XML format as a response. This service oriented architecture increases the modularity of the system and makes it easy to replace or improve any of the preprocessing algorithms.

The processing is done in a lazy manner, i.e. the text is not processed completely at the beginning, instead the services are called only when needed. For instance if the user does not request to see a summary then the text is never summarized.

[7] http://www.wordle.net/

[8] http://touchgraph.sourceforge.net
[9] http://linkeddata.org/
[10] http://enrycher.ijs.si
[11] https://jetpack.mozillalabs.com
[12] http://jquery.com/

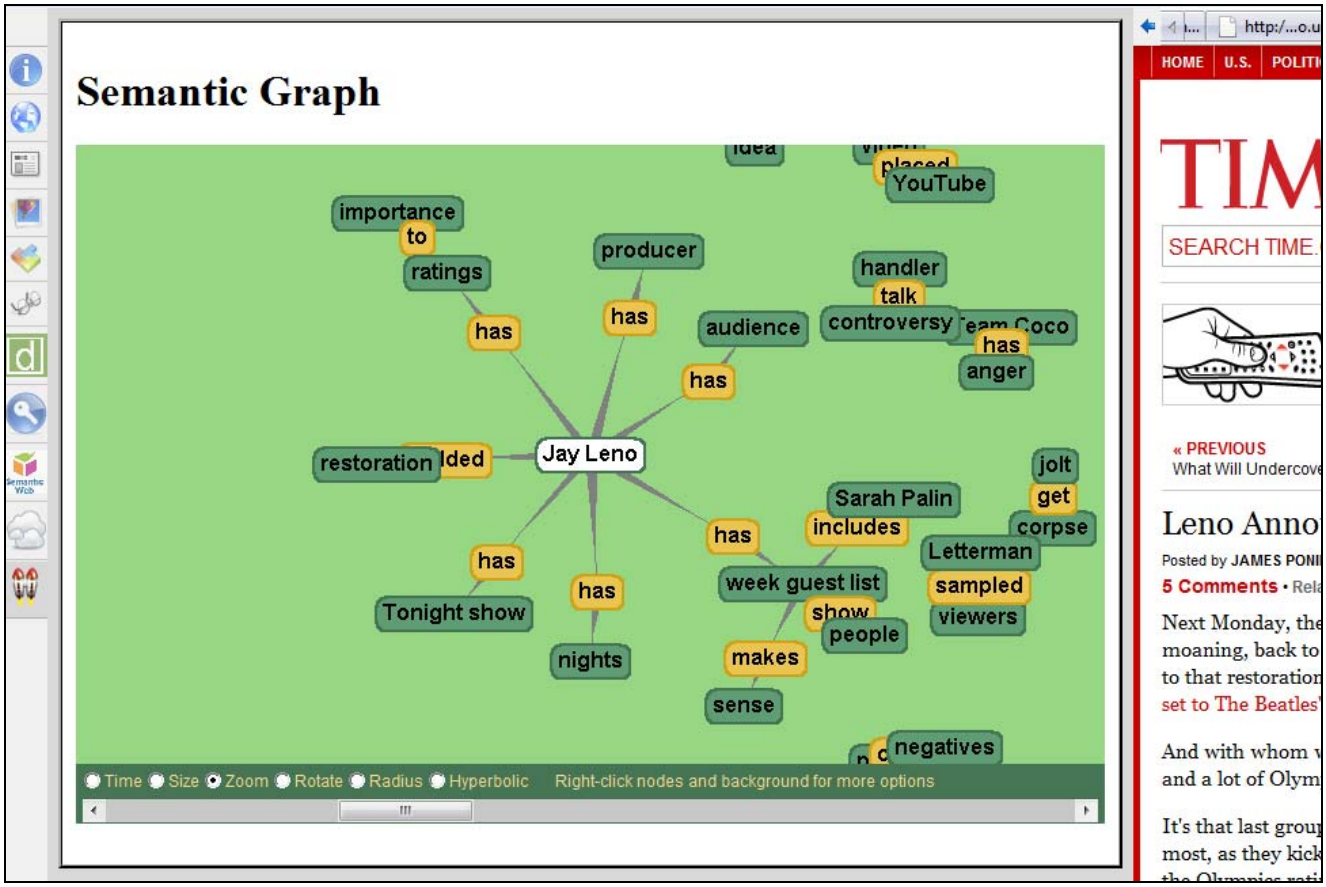

Figure 3 Semantic graph visualization of an article about Jay Leno

V. EXPERIMENTS

In the experiments we want to see how much information about the text on the web page our visualization methods are giving. We have designed the experiments to estimate how well the users can understand the main information of a Web page without seeing the original content and relying on using our software to obtain visualizations for web pages whose content was removed. In other words they are trying to find out what a web page is about by only looking at the visualizations and not seeing the page itself. After the users formulated their assumptions and were told what the web page is actually about, they were asked to rate each of the visualization functionality of the system.

We have used three pages for evaluation, one Wikipedia article about a French nobleman, a news article about the effect of social network sites on television and a blog post which makes a parallel between playing blackjack and trading on the stock market. Seven users were involved in the experiments., After trying to guess the content of these articles from their visualizations, rated each visualization functionality according to how helpful it was from 1(not helpful) to 5(very helpful). The users were asked to spend about three minutes for each article. We cannot claim to have come to some solid conclusions from such a small experiment, but nevertheless here is what we have observed. The users were able to guess what the web pages were about, but they often missed the finer points and details related to the story. The most difficult page was the news article describing the effects on social networks on television. Based on the feedback from the users we think this is due to the diversity of topics covered in the article.

Table 1 Rating of the visualizations

Visualization	Points
Summary	35
Semantic Graph	28
Triplets	25
Word Cloud	22
Classification	21
Keywords	19
Related News	17
Related Pages	16
Semantic Entities	15
Related Pictures	8

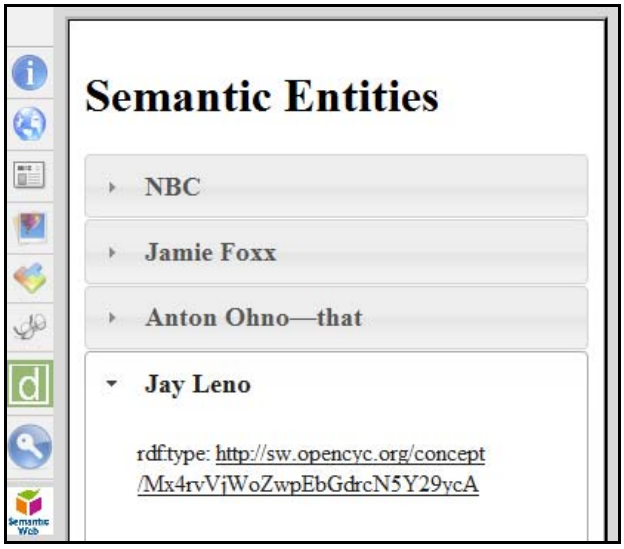

Figure 4 Semantic Entities

The Wikipedia article about the French nobleman was the one about which the users got the closest idea. This article is a descriptive one focusing on this French nobleman.

For the blog post making a parallel between blackjack and stock trading the results were quite opposing. Three of the users observed this parallel and totally guessed what the blog post is about while the others were rather confused and skeptical about the appearance of the two apparently unrelated topics.

Table 1 shows each visualization type and how many rating points it got, ordered from best to worst (maximum is 35 – getting 5 points from each of the seven users, while the minimum is 7). Not only was the summary the clear winner but it also was favored by all the users. Triplets and semantic graphs, the runners-up, show similar information but the users preferred either one or the other. Some liked the interactive nature of the graph while others found the simplicity of the triplets more readable, An observation is that the users found the semantic graph very useful for the page which focuses on a single topic; for the pages with diverse content, like the second one, the graph was not much help. Another important lesson which we learned is that most of the users, although they had one or two favorite visualization methods, checked all the available visualizations in turn from top to bottom. Only two users went directly to what they found the most helpful. This tells us that we should reorder the tabs in the slidebar and put summary as the most useful functionality at the top.

Related pages, related news and related pictures were of quite poor value to the users, maybe also because they are more suited to suggest related content than to explain the given one. The list of semantic entities didn't give the users much insight either.

VI. Conclusions

The paper presented a system aiming to help the users better and faster understand the text on a web page and/or find related content on the internet. In order to evaluate how helpful each visualization method is, we have performed experiments with several users who tried to guess the textual content of pages only looking at the visualizations provided. It turned out that the summary, the graphs and the triplets were of real help, the other visualizations being of questionable value. We also found out that the visualizations our software provides (especially the semantic graph) are most effective if applied to a text with a single focus in its topic. Additionally, it turned out that the order of the visualization options in the user interface matters, and that the more popular methods should come first.

References

[1] Mihalcea, R., Language independent extractive summarization. Proceedings of the ACL 2005 on Interactive poster and demonstration sessions. p. 52, 2005

[2] Rusu, D., Dali, L., Fortuna, B., Grobelnik, M. and Mladenic, D. Triplet Extraction from Sentences. In Proceedings of the 10th International Multiconference "Information Society - IS 2007". Ljubljana, Slovenia. pp. 218 -- 222. Ljubljana, Slovenia, October 2007

[3] Leskovec, J., Grobelnik, M., Milic-Frayling, N.. Learning Sub-structures of Document Semantic Graphs for Document Summarization. In Proceedings of the 7th International Multi-Conference Information Society IS 2004, Volume B. pp. 18-25, 2004.

[4] Mihalcea, R. and Tarau, P. TextRank: Bringing order into texts. In Proceedings of EMNLP, vol. 4, pp. 404 -- 411, Barcelona 2004.

[5] Grobelnik, M. and Mladenic, D. Simple classification into large topic ontology of Web documents. Journal of Computing and Information Technology. nr. 4, vol. 13, p. 279, 2005

[6] Rusu, D., Fortuna, B., Mladenic, D., Grobelnik, M. and Sipos, R. Document Visualization Based on Semantic Graphs. In Proceedings of the 13th International Conference Information Visualisation (IV09). Barcelona, Spain. pp. 292 -- 297. July 2009.

[7] Grobelnik, M., Mladenic, D. Text Mining Recipes,Springer-Verlag, Berlin; Heidelberg; New York, 2006

[8] T. Stajner, D. Rusu, L. Dali, B. Fortuna, D.Mladenic, M. Grpbelnik, Enrycher - service oriented text enrychment, SiKDD2009, October 16th, 2009, Ljubljana, Slovenia

[9] A. Arasu, J. Novak, A. Tomkins and J. Tomlin, "PageRank Computation and the Structure of the Web: Experiments andAlgorithms", Technical Report, IBM Almaden Research Center, Nov. 2001.