

Visual Analysis of Documents with Semantic Graphs

Delia Rusu, Blaž Fortuna, Dunja Mladenič, Marko Grobelnik, Ruben Sipoš

Department of Knowledge Technologies

Jožef Stefan Institute, Ljubljana, Slovenia

{delia.rusu, blaz.fortuna, dunja.mladenic, marko.grobelnik, ruben.sipos}@ijs.si

ABSTRACT

In this paper, we present a technique for visual analysis of documents based on the semantic representation of text in the form of a directed graph, referred to as *semantic graph*. This approach can aid data mining tasks, such as exploratory data analysis, data description and summarization. In order to derive the semantic graph, we take advantage of natural language processing, and carry out a series of operations comprising a pipeline, as follows. Firstly, named entities are identified and co-reference resolution is performed; moreover, pronominal anaphors are resolved for a subset of pronouns. Secondly, subject – predicate – object triplets are automatically extracted from the Penn Treebank parse tree obtained for each sentence in the document. The triplets are further enhanced by linking them to their corresponding co-referenced named entity, as well as attaching the associated WordNet synset, where available. Thus we obtain a semantic directed graph composed of connected triplets. The document's semantic graph is a starting point for automatically generating the document summary. The model for summary generation is obtained by machine learning, where the features are extracted from the semantic graph structure and content. The summary also has an associated semantic representation. The size of the semantic graph, as well as the summary length can be manually adjusted for an enhanced visual analysis. We also show how to employ the proposed technique for the Visual Analytics challenge.

Categories and Subject Descriptors

I.2.7 [Natural Language Processing]: Text analysis.

General Terms

Algorithms, Design.

Keywords

Natural language processing, text mining, document visualization, semantic graph, triplet, summarization.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

VAKD'09, June 28, 2009, Paris, France.
Copyright 2009 ACM 978-1-60558-670-0...\$5.00.

1. INTRODUCTION

Visual Analytics incorporates, among others, knowledge discovery, data analysis, visualization and data management. The goal of this research field is to derive insight from dynamic, massive, ambiguous and often conflicting data [6]. Providing visual and interactive data analysis is a key topic in Visual Analytics. Data mining tasks, such as data description, exploratory data analysis and summarization can be aided with such visualizations. Visual exploration and analysis of documents enables users to get an overview of the data, without the need to entirely read it. The document overview offers a straightforward data visualization by listing the main facts, linking them in a way that is meaningful for the user, as well as providing a document summary.

In order to respond to this challenging task, we present a document visualization technique based on semantic graphs derived from subject – predicate – object triplets using natural language processing. This technique can be applied for providing documents and their associated summary with a graphical description that enables visual analysis at the document level. The triplets are automatically extracted from the Penn Treebank [10] parse tree which was generated for each sentence in the document. They are further processed by assigning their co-referenced named entity, by solving pronominal anaphors for a subset of pronouns and by attaching their corresponding WordNet [3] synset. Finally, the semantic graph is built by merging the enhanced triplets.

Moreover, this semantic representation is not only useful for visualizing the document, but it also plays an important part in deriving the document summary (as proposed in [7, 12]). This is obtained by classifying sentences from the initial text, where the features are extracted from the document and its semantic graph. The size of the semantic graph, as well as the summary length are not fixed, and this characteristic improves visual analysis. Furthermore, the document summary is also provided with a semantic graphical representation.

There are several tools dedicated to document corpus visualization, which are helpful in data analysis. Some are focused on a particular kind of data, such as news collections [5], other are developed for general text either based solely on the text of the documents in the corpus [4], or ontology-based, taking advantage of, for example, an ontology representing the users' knowledge or interests [14]. While these approaches explore and analyze a collection of documents as a whole, providing the overall picture of the text corpus, we perform a more in depth visual exploration and analysis of a single document.

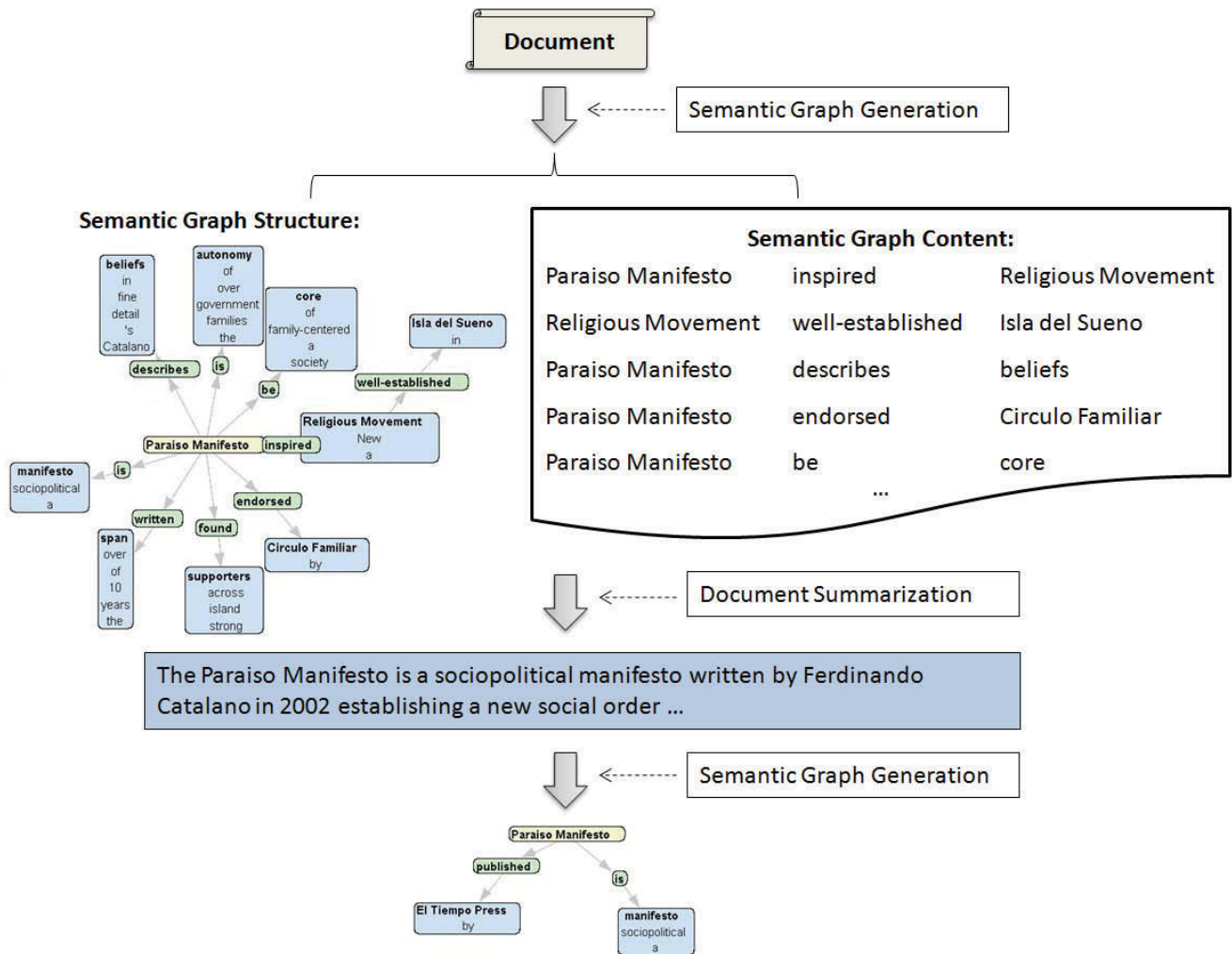


Figure 1. The document analysis process.

Moreover, we detail the main facts and connect them in a semantic structure, as well as provide a visual description for the document summary. Other visualization tools focus on tracking of story evolution: evolutionary theme patterns discovery, summary and exploration. The work described in [13] takes advantage of graphs to represent the development of a story; these graphs consist of elements of a co-occurrence network, disregarding synonymy relationships among the elements. In contrast to this approach, we construct a semantic graph where the building blocks are enhanced triplets linked to WordNet synsets. These triplets are much more than co-occurring entities, they are considered the core of the sentence, salient enough to carry the sentence message. They are connected taking into account the synonymy relationship among them.

Previous work related to visualizing a single document has focused on highlighting named entities, facts and events in the given text, or on using the human created structure in lexical databases for revealing concepts within a document. Our purpose is to further analyze the concepts in a text fragment, determine the

connections among them and visually represent this in the form of a semantic graph, either of the entire document or of its summary.

The Calais Document Viewer¹ creates semantic metadata for the user submitted text, in the form of named entities, facts and events, which are highlighted and navigable; the RDF output can also be viewed and captured for analysis in other tools. DocuBurst [1] presents the concepts within a document in a radial, space filling tree structure, using WordNet's IS-A hyponymy relationship. In the case of our system, named entities and facts or concepts represent the starting point; they are further refined in order to enable the construction of a semantic description of the document in the form of a semantic directed graph. The nodes are the subject and object triplet elements, and the link between them is determined by the predicate. The initial document, its associated facts and semantic graph are then employed to automatically generate a summary, which can also be visualized in the form of a graph.

¹ Calais url: <http://www.openalais.com/>

The paper is organized as follows. We start with an overview of the document visualization process in Section 2, continuing with a description of semantic graphs in Section 3, document summaries in Section 4 and an application of the described technique to the Visual Analytics challenge in Section 5. The paper concludes with several remarks.

2. DOCUMENT VISUAL ANALYSIS

The document visualization process is described in Figure 1, where an example document from the Visual Analytics Challenge² is used.

It starts with the original document, which is further processed and refined in order to obtain the set of subject – predicate – object triplets as well as its associated semantic graph. Next, the semantic graph structure and content serve as input for the document summarizer, which automatically generates a summary of sentences from the text. The approach considered for summarization is sentence extraction. This summary can also be visualized in the same way as the original document, by associating it with a semantic description.

3. SEMANTIC GRAPHS

The *semantic graph* corresponds to a visual representation of a document’s semantic structure. As proposed in [7] a document can be described by its associated semantic graph, thus providing an overview of its content. The graph is obtained after processing the input document and passing it through a series of sequential operations composing a pipeline (see Figure 2):

- *Text preprocessing*: splitting the original document into sentences;
- *Named entity extraction*, followed by named entity co-reference resolution and pronominal anaphora resolution;
- *Triplet extraction* based on a Penn Treebank parser;
- *Triplet enhancement* by linking triplets to named entities and semantic normalization via assigning each triplet its WordNet synset;
- *Triplet merger into a semantic graph* of the document.

In what follows, we are going to further detail the aforementioned pipeline components as proposed in [12].

3.1 Named Entity Extraction, Co-reference and Anaphora Resolution

The term *named entities* refers to names of people, locations or organizations, yielding semantic information from the input text. For named entity recognition we consider GATE (General Architecture for Text Engineering)³; it was used as a toolkit for natural language processing. For people we also store their gender, whereas for locations we differentiate between names of cities and of countries, respectively. This enables co-reference resolution, which implies identifying terms that refer to the same entity. It is achieved through consolidating named entities, using text analysis and matching methods.

² IEEE VAST 2008 Challenge url: <http://www.cs.umd.edu/hcil/VASTchallenge08/tasks.html>

³ GATE url: <http://gate.ac.uk/>

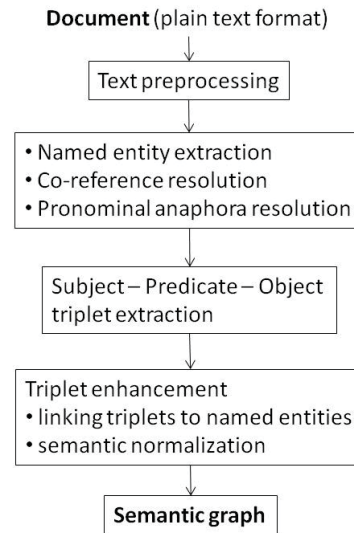


Figure 2. The semantic graph generation pipeline.

We match entities where one surface form is completely included in the other (for example “Ferdinando Catalano” and “Catalano”), one surface form is the abbreviation of the other (for example “ISWC” and “International Semantic Web Conference”), or there is a combination of the two situations described above (for example “F. Catalano” and “Ferdinando Catalano”).

Figure 3 represents an excerpt of a document with two annotated named entities and their corresponding co-reference (we eliminate stop words when resolving co-references – for example in the case of “EBay Inc”, “Inc” will be eliminated, as it is a stop word).

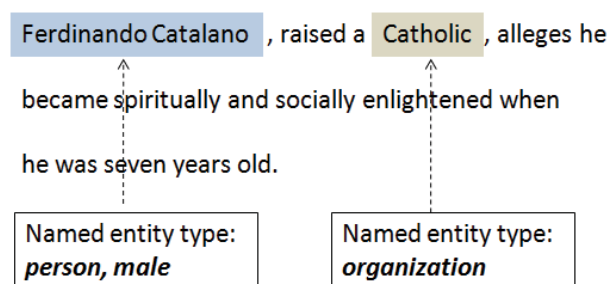


Figure 3. A document excerpt with two annotated named entities (a person and an organization).

We highlight the named entities found within the document, distinguishing between the three different entity types: people, locations and organizations, as illustrated in Figure 3.

Moreover, we resolve anaphors for a subset of pronouns: {*I, he, she, it, they*}, and their objective, reflexive and possessive forms, as well as the relative pronoun *who*. For solving this task, we take advantage of the co-referenced named entities and try to identify, for each pronoun belonging to the considered subset, its corresponding named entity. In the previous example (see Figure

3), the person named entity (“Ferdinando Catalano”) would be a good candidate to replace the pronoun “he”.

The pronominal anaphora resolution heuristic can be described as follows. We start by identifying pronouns in the given document, and search for each pronoun possible candidates that could replace it. The candidates receive scores, based on a series of antecedent indicators (or preferences) [12]: givenness, lexical reiteration, referential distance, indicating verbs and collocation pattern preference. The candidate with the highest score is selected as the pronoun replacement.

3.2 Triplet Extraction

We envisage the “core” of a sentence as a *triplet* consisting of the *subject*, *predicate* and *object* elements and assume that it contains enough information to express the message of a sentence. The usefulness of triplets resides in the fact that it is much easier to process them instead of dealing with very complex sentences as a whole.

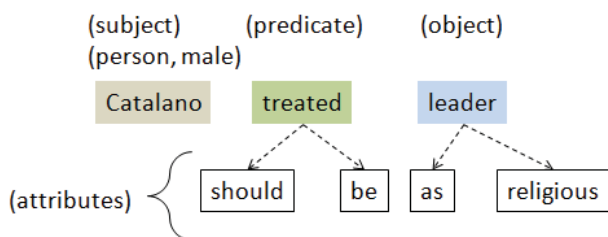


Figure 4. A triplet (Catalano – treated – leader) extracted from the sentence “*Followers claim the Paraiso Manifesto has inspired a New Religious Movement and Catalano should be treated as a religious leader.*”.

Triplets are extracted from each sentence independently, without taking text outside of the sentence into account. We apply the algorithm for obtaining triplets from a Penn Treebank parser output described in [11], and employ the statistical Stanford Parser⁴. The extraction is performed based on pure syntactic analysis of sentences. The rules are built by hand, and use the shape of the parse tree to decide which triplets to extract. Figure 4 shows a triplet (Catalano – treated – leader) extracted from the sentence “*Followers claim the Paraiso Manifesto has inspired a New Religious Movement and Catalano should be treated as a religious leader.*”. Aside from the main triplet elements (subject, predicate, object), the image also depicts the predicate and object attributes (*should*, *be* and *as religious*) – these are the words which are linked to the predicate and object in the parse tree.

As in the case of named entities, triplets are also highlighted differently, according to the triplet element type: subject, predicate or object. This convention is kept in the next phase of the pipeline, when building the semantic graph. Therefore, the triplet elements are much easier to identify within the graph structure.

⁴ Stanford Parser url: <http://nlp.stanford.edu/software/lex-parser.shtml>

3.3 Triplet Enhancement and Semantic Graph Generation

The semantic graph is utilized in order to represent the document’s semantic structure. Our approach is based on the research presented in [7] and further developed in [12]. While in [7] semantic graph generation was relying on the proprietary NLPWin linguistic tool [2] for deep syntactic analysis and pronominal reference resolution, we take advantage of the co-referenced named entities as well as the triplets extracted from the Penn Treebank parse tree and derive rules for pronominal anaphora resolution and graph generation. For generating the graph, triplets are first linked to their associated named entity (if appropriate). Furthermore, they are assigned their corresponding WordNet synset. This is a mandatory step, preceding the semantic graph generation, as it enables us to merge triplet elements which belong to the same WordNet synset, and thus share a similar meaning. Hence we augment the compactness of the graphical representation, and enable various triplets to be linked based on a synonymy relationship. We obtain a directed semantic graph, the direction being from the subject node to the object node, and the connecting link (or relation) is represented by the predicate.

Figure 5 presents a semantic sub-graph of a text excerpt. Semantic graph visualization was achieved through adapting the Prefuse visualization toolkit⁵ in a Java applet embedded in the web page. The graph layout is a dynamic force-directed one, yielding a spring graph, scalable to several hundred nodes.

The semantic graph generation system components were evaluated by comparing their output with the one of similar systems, as described in [12]. The evaluation was performed on a subset of the Reuters RCV1 [8] data set. For co-reference resolution, the comparison was made with GATE’s co-reference resolver; our co-reference module performed about 13% better than GATE. In the case of anaphora resolution, we compared the outcome of our system with two baselines that considered the closest named entity as a pronoun replacement: one baseline also took gender information into account, whereas the other did not. We obtained good results, particularly in the case of the masculine pronoun *he*.

4. DOCUMENT SUMMARY

The document summary is a means of retrieving a more synthetic text by extracting sentences from the original document. This is automatically obtained starting from the initial document and its corresponding semantic representation. The technique involves training a linear SVM classifier to determine those triplets that are useful for extracting sentences which will later compose the summary. The features employed for learning are associated with the triplet elements and obtained from the document content (linguistic and document attributes) and from the graph structure (graph attributes) [12]. Table 1 lists several examples of features that were used for learning. All in all, there are 69 features distributed among the triplet elements: 26 for the subject, 11 for the predicate, 26 for the object, and 6 sentence attributes (associated to the sentence that generated the triplet).

⁵ Prefuse url: <http://prefuse.org/>

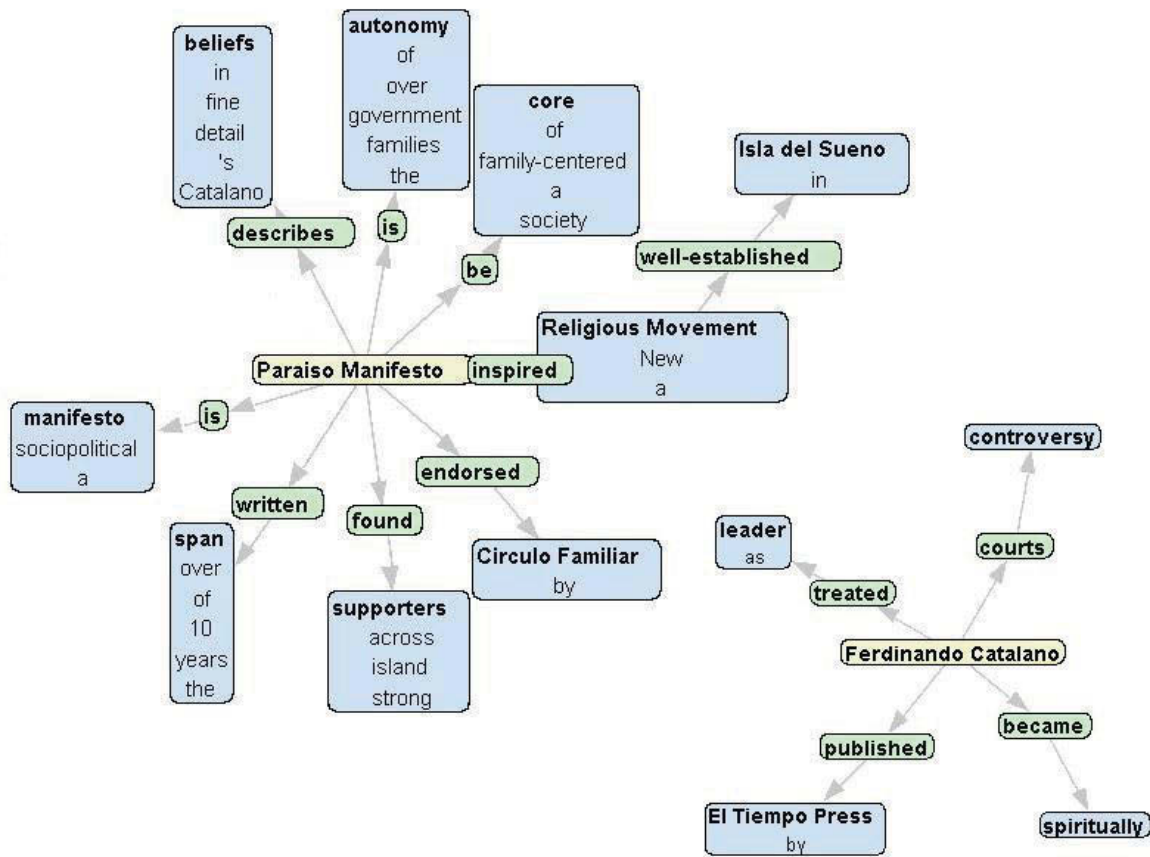


Figure 5. A semantic sub-graph of the “Paraiso Manifesto” Wikipedia article provided by the Visual Analytics Challenge.

Table 1. Examples of features used for learning.

Feature categories	Examples
<i>Linguistic</i>	<ul style="list-style-type: none"> the triplet type: subject, predicate or object, the Penn Treebank tag, the depth of the linguistic node extracted from the Penn Treebank parse tree, the part of speech tag;
<i>Document</i>	<ul style="list-style-type: none"> the location of the sentence within the document, the triplet location within the sentence, the frequency of the triplet element, the number of named entities in the sentence, the similarity of the sentence with the centroid (the central words of the document);
<i>Graph</i>	<ul style="list-style-type: none"> hub and authority weights, page rank, node degree, the size of the weakly connected component the triplet element belongs to;

For training the linear SVM model and for the evaluation of the document summary, we utilize the DUC (Document Understanding Conferences)⁶ datasets from 2002 and 2007, respectively, and compare the results with the ones obtained in the 2007 update task, as described in [12]. Thus we can compare the performance of our system with similar summarization applications that participated in the DUC 2007 challenge, for example, that generate compressed versions of source sentences as summary candidates and use weighted features of these candidates to construct summaries [9], or that learn a log-linear sentence ranking model by maximizing three metrics of sentence goodness [15].

The DUC datasets contain news articles from various sources like Financial Times, Wall Street Journal, Associated Press and Xinhua News Agency. The 2002 dataset comprises 300 newspaper articles on 30 different topics and for each article we have a 100 word human written abstract. The DUC 2007 dataset comprises 250 articles for the update task and 1125 articles for

⁶ DUC url: <http://duc.nist.gov/>

the main task, part of the AQUAINT dataset⁷; the articles are grouped in clusters and 4 NIST assessors manually create summaries (of 100 or 250 words) for the documents in the clusters. As training data we used the DUC 2002 articles, as well as the DUC 2007 main task articles, while the DUC 2007 update task articles were used for testing. We extracted triplets from the training and test data, and learned which triplets appear in the summaries. If we order the classified triplets by the confidence weights of their class we obtain a ranked list of triplets. In order to build the summary of a document, we trace back the sentences from which the triplets were extracted.

The summarization process, described in Figure 6, starts with the document semantic graph. The three types of features abovementioned are then retrieved. Further, the triplets are classified with the linear SVM, and then the sentences from which the triplets are extracted are identified, thus obtaining the document summary. The summary length is interactively determined by the user, for an enhanced visual analysis. As sentences have an associated SVM score (the one of the triplets extracted from the sentence), the summary will be composed of those sentences that received the highest score. In order to keep the summary readable, we maintain the same sentence ordering that appears in the original text. Because the training data was formed of newspaper articles, we expect the results to generalize to other news corpora, as well as Wikipedia articles.

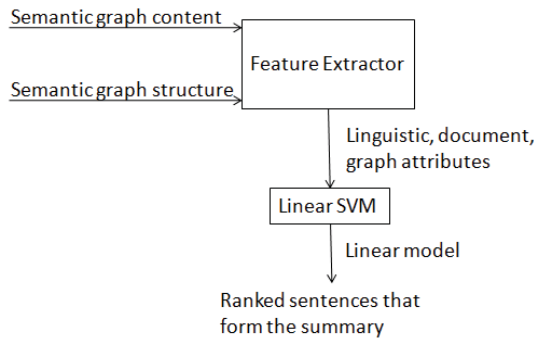


Figure 6. The summarization process.

5. VISUAL ANALYTICS CHALLENGE

The Visual Analytics challenge proposes a set of 4 mini-challenges, which combined form an overall challenge, concerning a fictitious, controversial socio-political movement. The datasets provided to solve the challenges are synthetic: a blend of computer- and hand-generated data. As a starting point, the organizers offer background material for both the overall challenge and the individual challenges in the form of a Wikipedia article page accompanied by discussions related to the article content.

⁷ AQUAINT url: <http://www ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2008T25>

The visual analysis techniques that we have described throughout the paper are very useful in exploring the documents provided as a starting point, that is, the Wikipedia article describing the “Paraiso Manifesto”, as well as the discussion page. By applying the pipeline components to these documents, we get an insight on the main issues mentioned in the text. The list of facts for the two documents, their associated semantic graphs and document summaries offer a good starting point for data analysis.

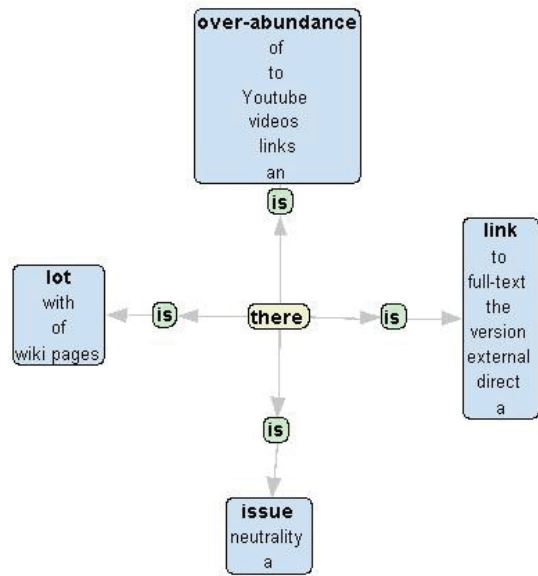


Figure 7. A semantic sub-graph generated from the Wikipedia Discussion page, under “POV Pushing”.

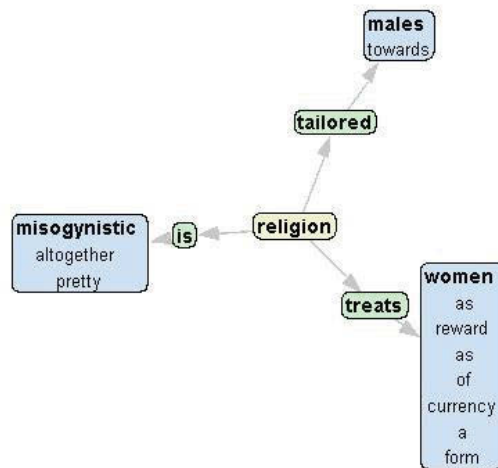


Figure 8. A semantic sub-graph generated from the Wikipedia Discussion page, under “Distinctive Doctrines”.

We have shown a sub-graph obtained from the Wikipedia article describing the “Paraiso Manifesto” movement (see Figure 5). The other sub-graph is centered on the “Ferdinando Catalano”

node (not shown entirely in the figure). We have also generated semantic graphs for the Wikipedia discussion page on the "Paraiso Manifesto". We also show two sub-graphs obtained by processing the text under "POV Pushing" and "Distinctive Doctrines" (Figure 7 and Figure 8 respectively). Figure 9 shows an example of two document summaries, generated based on the Wikipedia Discussion page. The first two-sentence long summary was obtained from the "POV Pushing" sub-section, whereas the latter summary corresponds to the "Distinctive Doctrines" sub-section.

Thus we can use our system as a first step in solving either the overall challenge or the sub-challenges, by analyzing the documents provided as background material.

Wikipedia Discussions Page, under "POV Pushing"
2 sentence-long summary:

There also is an over-abundance of links to Youtube videos produced by Catalano's Pirate Radio programs, this is unsavoury.
Disciples of any religious group will lie and twist the truth to keep up appearances and that is what is happening on this Wiki page dedicated to Catalano.

Wikipedia Discussions Page, under "Distinctive Doctrines"
2 sentence-long summary:

The religion treats women as a form of currency or a reward, not giving them any say in who they will marry, where they will live, and how many children they will have.
However, the women believe they are given to the men but what the men do is none of their business.

Figure 9. Two document summaries obtained by processing parts of the Wikipedia discussion page.

6. CONCLUSIONS

In this paper we presented a document visualization technique based on semantic graphs. We showed that this technique can be applied not only to the original document, but also to its automatically generated summary. Each of the system components were detailed, starting with the semantic graph generation pipeline composed of named entity recognition, triplet extraction and enhancement, semantic graph construction, and concluding with the document summarization process. The runtime of our approach mainly depends on the document size and sentence complexity. The main bottleneck is represented by sentence parsing, which we intend to overcome by using a faster parser.

Regarding future improvements, we aim at extending the system by adding several components such as a more sophisticated named entity recognizer module, and a new triplet extraction module. To further refine the document overview through semantic graphs and summaries, we intend to integrate external resources that would enhance the semantic representation, as well as the document summary.

Based on the feedback obtained from several users we can conclude that the presented document visualization using

semantic graphs is promising and can be helpful for the user. However, more experiments evaluating the usefulness of the proposed visualization approach are needed for firm conclusions.

7. ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency and the IST Programme of the EC SMART (IST-033917) and PASCAL2 (IST-NoE-216886).

8. REFERENCES

- [1] Collins, C. 2006. DocuBurst: Document Content Visualization Using Language Structure. In Proceedings of IEEE Symposium on Information Visualization, Poster Session. Baltimore.
- [2] Corston-Oliver, S.H. and Dolan, B. 1999. Less is more: eliminating index terms from subordinate clauses. In Proceedings of the 37th Conference on Association for Computational Linguistics, College Park, Maryland.
- [3] Fellbaum, Ch. 1998. WordNet: An Electronic Lexical Database. MIT Press.
- [4] Fortuna, B., Grobelnik, M and Mladenić, D. 2005. Visualization of Text Document Corpus. Informatica Journal 29, pp. 270-277.
- [5] Grobelnik, M. and Mladenić, D. 2004. Visualization of news articles. Informatica Journal 28, pp. 375-380.
- [6] Keim, D. A., Mansmann, F., Schneidewind J., and Ziegler, H. 2006. Challenges in visual data analysis. In Proceedings of IEEE International Conference on Information Visualization, pages 9 -16.
- [7] Leskovec, J., Grobelnik, M. and Milic-Frayling, N. 2004. Learning Sub-structures of Document Semantic Graphs for Document Summarization. Workshop on Link Analysis and Group Detection (LinkKDD) at KDD 2004 (Seattle, USA, August 22 – 25, 2004).
- [8] Lewis, D. D., Yang, Y., Rose, T. G., Li, F. 2004. RCV1: A New Benchmark Collection for Text Categorization Research. Journal of Machine Learning Research, Vol. 5.
- [9] Madnani, N., Zajic, D., Dorr, B., Ayan, N. F. and Lin, J. 2007. Multiple Alternative Sentence Compressions for Automatic Text Summarization. In Proceedings of the Document Understanding Conference (DUC).
- [10] Marcus, M. P., Santorini, B. and Marcinkiewicz, M. A. 1993. Building a large annotated corpus of English: the Penn Treebank. Computational Linguistics, Volume 19.
- [11] Rusu, D., Dali, L., Fortuna, B., Grobelnik, M. and Mladenić, D. 2007. Triplet Extraction from Sentences. In Proceedings of the 10th International Multiconference "Information Society - IS 2007" (Ljubljana, Slovenia, October 8 – 12, 2007). 218 – 222.
- [12] Rusu, D., Fortuna, B., Grobelnik, M. and Mladenić, D. 2009. Semantic Graphs Derived From Triplets With Application In Document Summarization. Informatica Journal.

- [13] Subašić, I. and Berendt, B. 2008. Web Mining for Understanding Stories through Graph Visualisation. In Proceedings of the International Conference on Data Mining (ICDM), pp. 570-579.
- [14] Thai, V, Handschuh, S. and Decker, S. 2008. IVEA: An information visualization tool for personalized exploratory document collection analysis. In Proceedings of the European Semantic Web Conference (ESWC), pp. 139-153.
- [15] Toutanova, K., Brockett, C., Gamon, M., Jagarlamudi, J., Suzuki, H. and Vanderwende, L. 2007. The PYPHY Summarization System: Microsoft Research at DUC2007. In Proceedings of the Document Understanding Conference (DUC).