A Network Based Approach for the Visualization and Analysis of Collaboratively Edited Texts

Tobias Hecking
University of Duisburg-Essen
Lotharstraße 63/65
47048 Duisburg, Germany
hecking@collide.info

H. Ulrich Hoppe
University of Duisburg-Essen
Lotharstraße 63/65
47048 Duisburg, Germany
hoppe@collide.info

ABSTRACT
This paper describes an approach for network text analysis and visualization for collaboratively edited documents. It incorporates network extraction from texts where nodes represent concepts identified from the words in the text and the edges represent relations between the concepts. The visualization of the concept networks depicts the general structure of the underlying text in a compact way. In addition to that, latent relations between concepts become visible, which are not explicit in the text. This work concentrates on evolving texts such as wiki articles. This introduces additional complexity since dynamic texts lead to dynamic concept networks. The presented method retains the user information of each revision of a text and makes them visible in the network visualization. In a case study it is demonstrated how the proposed method can be used to characterize the contributors in collaborative writing scenarios regarding the nature of concept relations they introduce to the text.

General Terms
Algorithms, Visualization, Experimentation

Keywords
Network Visualization, Network Analysis, Natural Language Processing, Collaborative Writing, Learning Analytics

1. INTRODUCTION
Network text analysis is the task of extraction and analysis of networks from text corpora. In those networks the nodes are concepts identified from the words in the text and the edges between the nodes represent relations between the concepts. The visualization of concept networks can help to depict the general structure of the underlying text in a compact way. In addition to that, latent relations between concepts become visible, which are not explicit in the text. Thus, approaches for visualizing texts as networks allow analysts to concentrate on important aspects without reading large amounts of the texts. Several network analysis techniques can be applied to identify important concepts, perform concept clustering, as well as comparative analysis of different texts [11].

Existing applications for network text analysis include the identification of key phrases [10], mining of relations between real world entities [6], as well as the extraction of complete concept ontologies and concept maps with labelled edges [18].

This work concentrates on the relations between concepts that can be found in evolving and collaboratively edited texts such as wiki articles. This introduces additional complexity since dynamic texts lead to dynamic concept networks. The presented method retains the user information of each revision of a text which allows for characterizing the contributors in collaborative writing scenarios regarding the nature of concept relations they introduce to the text. The resulting visualization is a concept network with colored edges where each edge color is allocated uniquely to a specific contributor. In further analysis steps, network centrality measures are calculated that give additional information about the contribution of each editor.

The outline of this paper is as follows: Section 2 gives the theoretical background of this work and highlights significant research work in the area of network text analysis. The general idea of our visualization and analysis approach is presented in section 3. Section 4 focuses on the concrete implementation. This incorporates the applied natural language processing chain, as well as the description of network analysis methods.

2. Background

2.1 Collaborative Writing Activities in Education
Collaborative writing activities are a common task in educational scenarios [3, 13]. Users can learn actively by creating artefacts but can also learn passively by consuming artefacts created by others [14].
It could be shown that user generated content is relevant to learners in addition to tutor provided content [13]. With the emergence of online communities such as Wikipedia collaborative knowledge building takes place with open scale in terms of the number of contributors. There is some evidence that individual and collective knowledge co-evolves through collaborative editing of epistemic artefacts in open online environments [9]. In general collaborative writing requires different rhetorical and organizational skills of the editors [8], and thus, the learner generated artefacts are a valuable data source for analysis.

This motivates the development of methods that makes collaborative writing processes visible in order to understand and improve the application of collaborative text writing in educational settings.

2.2 Visualization Approaches for Collaborative Writing
Several methods have been developed to represent evolving texts with multiple editors in a visual way. One of the first approaches

for the visualization of evolving wiki articles is the History Flow method [17]. In this approach each contributor has assigned a unique color. Each revision of the evolving text is then represented as a sequence of blocks that represent the sections of the document. The blocks are colored according to the author who has edited the section and the size of the block corresponds to the amount of text. This does not only depict the insertion and removal of text sections by the users but additionally allow for the identification of edit wars between authors. In contrast to this page centric view, the iChase method [12] visualizes activities of a set of authors across multiple wiki articles as heatmaps. Southavilay et al. [16] extend the pure depiction of the amount and location of text edits done by a user by incorporating topic modeling. Therefore, they apply latent dirichlet allocation [4] in order to identify the contributions of users to the particular topics covered in a document. Based on the identified topics the evolution of topics as well as collaboration networks of users on particular topics can be analyzed.

2.3 Representing Mental Models as Graphs

Networks are a common representation for relations between entities of various kinds. Schvaneveldt et al. [15] argue that networks between entities based on proximities induced by people have a psychological interpretation. They assume that cognitive concepts such as memory organization and mental categories are reflected in the network structure. The pathfinder algorithm [15] derives a network of concepts from proximity data. Such proximities could be induced, for example, by associations made by a person. In general, it is also possible to derive such proximity data between concepts described in natural language texts [20].

One of the first approaches that utilize computational tools to extract mental models from text has been described by Carley [5]. After the identification of relevant words in a text, the words are linked based on syntactical analysis of the sentences of a text.

This approach has been further developed by Diesner et al. [6] and implemented in the software tool Automap where an analyst can specify a metamatrix of concepts and concept classes. This enables the identification of relations between entities of different types from text corpora, for example, people and organizations.

3. Visualization Approach

This paper extends network extraction from texts to dynamically evolving and collaboratively edited documents. When networks extracted from texts are considered as the author's mental model of the domain, as described in section 2.3, the aggregation of the networks extracted from several revisions of a collaboratively edited text can be interpreted as the joint representation of the individual mental models of all authors.

The basic assumption is that different authors introduce different concepts and relations to the text. In order to make these differences visible the author information is additionally incorporated into the network representation.

Each connection between concepts that can be extracted from the text can be labeled with the author who established it. In the small example in Figure 1 the little piece of text was produced by two different authors. Each author has assigned a unique color - in this case blue and red. The edges of the resulting network can then be colored according to the author who was the first who introduced the concept relation in the text.

This not only allows for a characterization of the underlying document in terms of concept relations but also a characterization of the contributors. Central concepts that are used by different authors but linked to different other concepts indicate different associations or views of the authors. Furthermore, the visualization approach additionally depicts which authors concentrate on thematic areas and which authors tend to relate concepts from different sub topics, for example, by writing a summary.

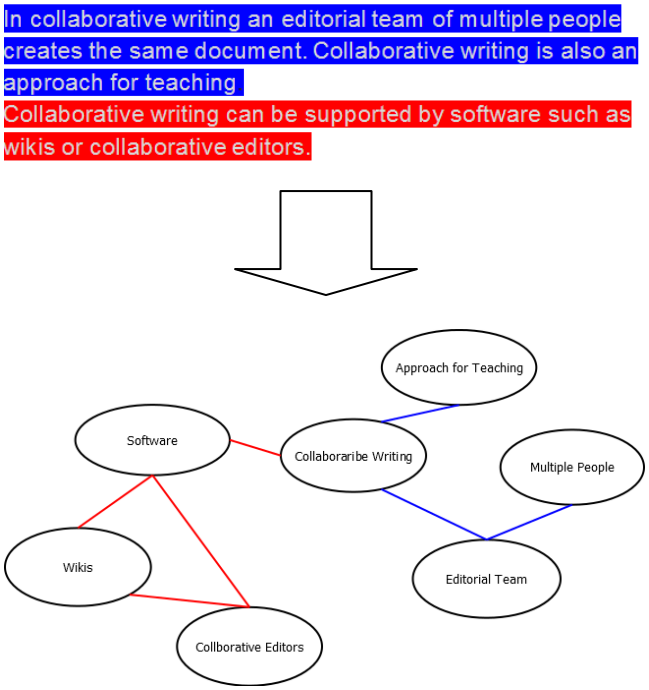

Figure 1 A concept network extracted from a text edited by two different authors. The authors are represented by color.

By calculating network measures on the concept network a further quantitative characterization of the authors is possible as described in section 4.3.

4. Implementation

This section outlines details of the implementation in two perspectives. In particular, these are word network extraction using natural language processing, and network analysis.

4.1 Extracting Concept Networks from Texts

The extraction of networks from text requires several natural language processing components. In this work the DKPro toolkit [7] was used. It is based on the Apache UIMA[1] framework and provides a large variety of natural language processing algorithms that can be combined in a flexible way. The process of the extraction of word networks from a single document is depicted in Figure 2. First, a preprocessing step is often required for text gathered from the web in order to remove wiki or HTML markup. Further, in this step irrelevant content can be filtered from the document. For example, Wikipedia pages often contain a large reference section and a list of related web resources. These parts are important for the wiki article itself but are a source of noise when the actual content of the article should be analyzed. In the

[1] https://uima.apache.org/

second step, the phrases representing concepts in the text have to be identified, and after that, connected to a network by using a proximity measure in step 3. Since the result might contain phrases with slightly different spelling which actually refer to the same semantic concept the entity resolution step merges those candidate phrases to a single concept. Concepts and relations can then be encoded as a network that is used for further processing. In the following the steps 2 to 4 are described in more detail.

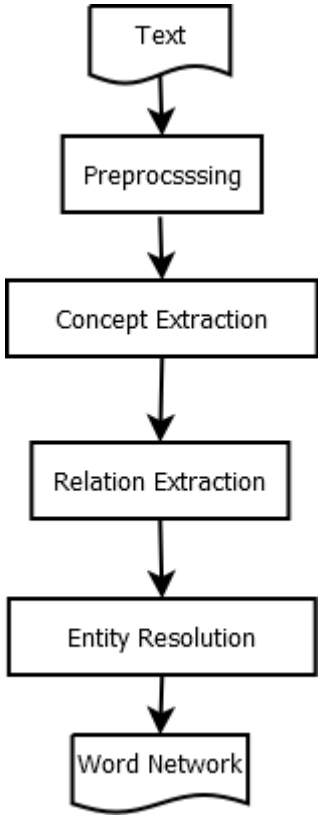

Figure 2 Process chain for the extraction of work networks from texts.

4.1.1 Concept Extraction

For the identification of the concepts in the input text noun phrase chunking was applied. First, the text is segmented into its sentences. Then part-of-speech (POS) tagging (using the Stanford PSO tagger[2]) is applied to label each word according to its function in its sentence. A naive solution for the extraction of concepts from the text would be to take each noun identified by the POS tagging as one concept. However, often one concept is described by more than one word. For example the phrase "Approach [NN] for [for] teaching [NN]" would result in two concepts, namely "Approach" and "Teaching", which does not really reflect the meaning of the phrase. Thus, noun phrase chunking is applied where the POS labeled words are chunked to meaningful noun phrases. This is done with the OpenNLP chunker[3], which identifies noun phrases according to certain rules. For example, the words "Approach [NN] for [for] teaching [NN]" are then identified as one single noun phrase.

4.1.2 Relation Extraction

After all concepts in the text are identified they have to be connected to a concept network according to a certain proximity measure. In this work, an edge between two concepts becomes established if the concepts co-occur in a sliding window of *n* words in at least one of the sentences in the text. This approach is straight forward but works well in practice [6, 10].

4.1.3 Entity Resolution

As already mentioned entity resolution is necessary in order to identify nodes in the network that represent the same concept and to merge them into single nodes. For example the noun phrases "Wiki" and "The Wikis" can be merged to the same concept "Wiki". In order to solve this problem, first all noun phrases have to be normalized using lemmatization. After that the concepts are compared pairwise by substring similarity [1]. If the similarity exceeds a value of 0.7 the concepts are merged and labeled with the shorter label of the two concepts.

4.2 Networks from Different Revisions

In order to extract an aggregated network from different revisions of a collaboratively edited text, the process chain described in section 4.1 is applied to each revision of the text in temporal order from the oldest to the latest revision. Each revision of the text was done by a single author. The edges in the network of the first revision are labeled with the author of this initial revision. Then in the first aggregation step all edges that are part of the network extracted from the second revision but do not exist in the network of the first revision are labeled with the author of the second revision and added to the previously extracted network. This proceeds until each revision has been processed. As described in section 3 the author information attached to the edges can then be visualized by using different colors for each author.

Since the aggregated network contains every noun phrase that has been used by the authors as a concept node, the network can be very large and likely contains concepts that are not relevant for the domain. Those concepts are often not well connected. Thus, in a preprocessing step the *k*-core [2] of the network is computed such that the resulting network contains only concepts with at least *k* connections to other concepts of the core. The resulting network has a reduced number of nodes, and the visualization concentrates on the most important concepts according to the connectedness to other core concepts in the network.

4.3 Quantitative Characterization of Contributors

For quantitative analysis the nodes (concepts) and edges can be ranked according to network centrality measures [19]. In this work concepts are ranked according to eigenvector centrality and betweenness centrality. The eigenvector centrality is a recursive measure and assigns a weight to each node according to the number its neighbors while the connections are weighted according to the centrality of the neighbors. This gives high weight to concepts that have many connections to other important concepts.

Edges are ranked according to the edge-betweenness centrality. The edge-betweenness centrality assigns high weights to edges that often occur on shortest paths between any pair of nodes.

[2] http://nlp.stanford.edu/software/tagger.shtml

[3] https://opennlp.apache.org/documentation/1.5.2-incubating/manual/opennlp.html

In order to use the network measures for a characterization of the authors of the document an aggregation is necessary. For the node centric centralities, namely node-betweenness and eigenvector centrality the centrality contribution of an author *A* can be calculated by equation 1:

$$nc\_contrib(A) = \sum_{(c_i,c_j)\in E:lab(e)=A} \frac{cent(c_i,c_j)}{2|(c_i,c_j) \in E: lab(e) = A|} \quad (1)$$

This result is the average centrality of nodes that are incident to edges labeled with author *A*.

The edge-betweenness contribution of author *A* is the average of all edges labeled with author *A* (equation 2):

$$eb_{contrib(A)} = \sum_{e\in E:lab(e)=A} \frac{eb\_cent(e)}{|e \in E: lab(e) = A|} \quad (2)$$

An author with a high contribution in terms of edge-betweenness centralilty could be interpreted as someone who relates different parts of the text and introduces relations between concepts of different sections. This could, for example, be someone who creates a comprehensive summary of a longer wiki article. Authors with high contribution to the eigenvector centrality of the concepts can be those who work on important sections of the text and establish many relations between important domain concepts.

5. Case Study

As a case study the described method was applied to a wiki article on media economy created during a master level university course in a study program on Applied Cognitive Science and Media Science. The relations between the concepts are based on a sliding window with the size of 4 words. Figure 3 depicts the 5-core of the resulting aggregated concept network. The size of the nodes corresponds to the number of connections in order to support the visual discovery of important concepts. It can be directly seen from the visualization that the concept "media combination" is most central. Four of the six authors relate this concept to other concepts as it can be seen by counting the different colors of the incident edges. The highest coverage of the edges has the author who has pink as assigned color. Other contributors relate concepts more according to certain sub topics like communication (see blue edges).

The results for the quantitative characterization of the contributors are presented in Table 1. It is important to mention that reducing the network to its 5-core has mainly presentation purposes. Thus, for more reliable results the calculations were performed on the 2-core of the network in which more concept are present.

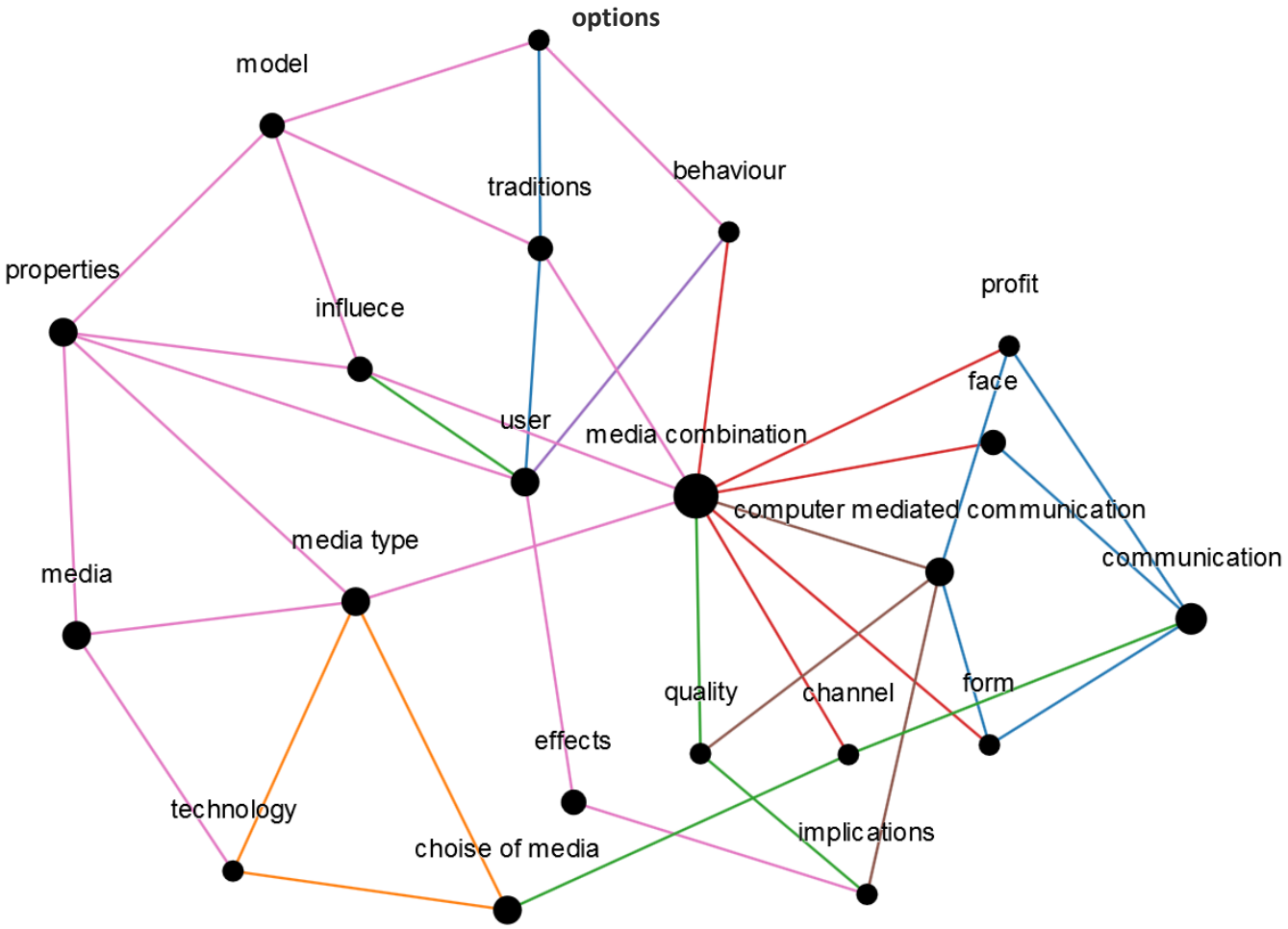

Figure 3 5-core of the aggregated concept networks extracted from a wiki article on media economics.

Table 1 Centrality contributions of the authors. EVC: Eigenvector centrality, NBC: Node-betweenness centrality (normalized), EBC: Edge betweenness centrality.

Author	Color	EVC	NBC	EBC
Student 1	Pink	0.20	0.07	**161.02**
Student 2	Red	**0.71**	**0.16**	95.85
Student 3	Green	0.35	0.07	80.17
Student 4	Blue	0.16	0.05	73.19
Student 5	Orange	0.1	0.04	111.45
Student 6	Brown	0.35	0.08	81.47

Student 1 has by far the highest contribution to the edge betweenness centrality. This is reasonable because this student did a reworking of large parts of the article and was highly involved in the shaping of the particular sections of the text. Student 2 has the highest scores regarding the node based centrality measures. However, the average edge-betweenness centrality is only moderate. This indicates that this student concentrated on the core topic of the article. This can also be seen in Figure 3 where the red edges of student 2 are all incident to the central concept.

6. CONCLUSION AND FURTHER WORK

The research presented in this paper describes an approach for the extraction of concept networks from text that incorporates author information in the visualization. In contrast to other existing visualizations of evolving texts our approach focuses rather on the relations between concepts than on the amount of text that is produced by individual authors. The case study has shown that the method is promising and can contribute to the analysis of collaborative text writing. In educational scenarios the proposed method enables tutors to investigate how students relate important domain concepts, and therefore, gain insights into their (possibly different) mental conceptualization. Thus, different views and focuses of students become visible. In future work the visualization will be integrated in an interactive application that supports the visual exploration of the resulting network through improved node and edge highlighting as well as facilities for data gathering and network reduction using *k*-core analysis. Regarding the interpretation and the analysis of the extracted networks the concept extraction can be adapted in such a way that the concepts and relations can be weighted by an expert according to their importance for the domain. This would result in more compact networks. In further evaluation the student characterizations derived from the colored word network can be related to self-assessment and characterizations made by a tutor.

7. REFERENCES

[1] Bär, D., Zesch, T. and Gurevych, I. DKPro Similarity: An Open Source Framework for Text Similarity. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations.* (Sofia, Bulgaria). Association for Computational Linguistics, 2013, 121-126.

[2] Bader, G. D. and Hogue, C. W. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics, 4, 1 (2003), 2.

[3] Belanger, Y. and Thornton, J. Bioelectricity: A Quantitative Approach Duke University's First MOOC. (2013) Technical Report, Duke University.

[4] Blei, D. M., Ng, A. Y. and Jordan, M. I. Latent dirichlet allocation. J.Mach.Learn.Res., 3(mar 2003), 993-1022.

[5] Carley, K. and Palmquist, M. Extracting, representing, and analyzing mental models. Social forces, 70, 3 (1992), 601-636.

[6] Diesner, J. and Carley, K. M. Revealing social structure from texts: meta-matrix text analysis as a novel method for network text analysis. *Causal mapping for information systems and technology research: Approaches, advances, and illustrations*, 2005, pp. 81-108.

[7] Eckart de Castilho, R. and Gurevych, I.In *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT*. Association for Computational Linguistics, 2014, 1-11.

[8] Flower, L. and Hayes, J. R. A Cognitive Process Theory of Writing. College Composition and Communication, 32, 4 (1981), pp. 365-387.

[9] Harrer, A., Moskaliuk, J., Kimmerle, J. and Cress, U. Visualizing wiki-supported knowledge building: co-evolution of individual and collective knowledge. In Anonymous *International. Symposium on Wikis*. 2008 19:1-19:9.

[10] Mihalcea, R. and Tarau, P. TextRank: Bringing order into texts. In Proceedings of the EMNLP. Association for Computational Linguistics, (Barcelona, Spain), 2004, 404-411.

[11] Paranyushkin, D. Identifying the pathways for meaning circulation using text network analysis. Technical Report Nodus Labs, Berlin, (2011).

[12] Riche, N. H., Lee, B. and Chevalier, F. iChase: Supporting Exploration and Awareness of Editing Activities on Wikipedia. In *Proceedings of the International Conference on Advanced Visual Interfaces.* (Roma, Italy). ACM, New York, NY, USA, 2010, 59-66.

[13] Sabrina Ziebarth and Hoppe, H. U. Moodle4SPOC: A Resource-Intensive Blended Learning Course. In *Proceedings of the European Conference on Technology Enhanced Learning*. (Graz, Austria), 2014, 359-372.

[14] Scardamalia, M. and Bereiter, C. Computer Support for Knowledge-Building Communities. The Journal of the Learning Sciences, 3, 3 (1993), pp. 265-283.

[15] Schvaneveldt, R. W., Durso, F. T. and Dearholt, D. W. Network structures in proximity data. Psychol. Learn. Motiv., 24 (1989), 249-284.

[16] Southavilay, V., Yacef, K., Reimann, P. and Calvo, R. A. Analysis of collaborative writing processes using revision maps and probabilistic topic models. In *Proccedings of the Learning Analytics and Knowledge Conference*. (Leuven, Belgium), 2013, 38-47.

[17] Viegas, F. B., Wattenberg, M. and Dave, K. Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the SIGCHI conference on Human factors in computing systems.* ACM, 2004, 575-582.

[18] Villalon, J. J. and Calvo, R. A. Concept Map Mining: A definition and a framework for its evaluation. In *Proceedings of the International Vonference on Web Intelligence and Intelligent Agent Technology, 2008,* IEEE, 2008, 357-360.

[19] Wasserman, S. and Faust, K. *Social Network Analysis: Methods and Applications.* Cambridge University Press, 1994.

[20] Wild, F., Haley, D. and Bulow, K. Monitoring conceptual development with text mining technologies: CONSPECT. In *Proceedings of eChallenge*, 2010, 1-8.