Empiric Experiments with Text Representing Centroids

Mario Kubek[1], Thomas Böhme[2], and Herwig Unger[1]
[1] FernUniversität in Hagen, Lehrgebiet Kommunikationsnetze, Universitätsstr. 27, Hagen, Germany
[2] Technische Universität Ilmenau, Institut für Mathematik, Weimarer Straße 25, Ilmenau, Germany
Email: {mario.kubek, herwig.unger}@fernuni-hagen.de, thomas.boehme@tu-ilmenau.de

***Abstract*—Centroid terms are comfortable instruments to represent texts, compare them semantically and to even (hierarchically) cluster sets of documents using them. Their determination depends on their topical and conceptual context, i.e. the dynamically changing knowledge of a user represented by the co-occurrence graph. Herein, important properties of centroids as well as their applicability for tasks in natural language processing and text mining shall be discussed and their use justified by a set of experiments. Based on the obtained results, a new approach to detect fine-grained similarities between text documents is derived.**

***Index Terms*—centroid term, co-occurrence graph, document similarity, text processing**

I. Introduction

Text centroids –inspired from the centre of mass in physics– have been introduced in [1] to represent sentences, paragraphs or whole texts by a single representing term. In addition, it could be shown that a distance measure among centroids may be defined which can be used to determine sematic text similarities and distances as well as to derive a hierarchical clustering algorithm [2] based on them. The introduction of centroids changes the methods of comparing texts in a significant manner. Two major approaches with practical relevance might be distinguished:

- The pairwise processing of two documents typically using the block- or cosine distance. These methods are based on the vector space model [3] following the bag-of-words principle and work with any kind of term vectors of the two documents and consider –by nature– only words contained in one or the other document. These methods are quite simple but do not work well if the texts are written by authors using different sets of words to describe similar topics.
- The consideration of texts in the context of a corpus as used for instance in the technique Latent Semantic Indexing (LSI) [4] requires the calculation of the term-document matrix of a whole corpus and the computational expensive determination of a lower-dimensional approximation of its original semantic space. Document vectors in this lower-dimensional space can then be compared in the same manner.

The new, centroid-based method presented in [1] represents texts by a single centroid term (which is not necessarily contained in the document), which must be usually calculated, only once. Then, every comparison operation is just a single distance measurement on the respective co-occurrence graph, which can be considered a condensed, compact and easily extendible representation of the knowledge of an entity (e.g. a user) at a given moment and can be used to determine the centroid terms. Since the entity's knowledge may change/update/extend and texts may be subject to different modifications (merge, split, edit), properties of centroids must be investigated in a more detailed manner than it has been done so far, what is the goal of the presented work. After a short introduction on centroids, the influence of an entity's knowledge (i.e. the co-occurrence graph) for their calculation will be discussed followed by a detailed consideration of centroid properties obtained from a set of empiric experiments.

II. Fundamentals

To understand our subsequent discussions, some basic notations need be introduced. Any two words w_i and w_j are called co-occurrents, if they appear together in one sentence (or any other well-defined environment or context). This co-occurrence relation may be used to define a graph $G=(W, E)$. Therefore, the set of words of a document corresponds to the set of nodes $w_a \in W$ and two nodes are connected by an edge $(w_a, w_b) \in E$, iff w_a and w_b are co-occurrents. A weight function $g((w_a, w_b))$ can be introduced to represent the frequency of a co-occurrence in a document, while usually only co-occurrences of a high significance $g((w_a, w_b)) > 0.5$ are taken into account. For this filtering, the used weight function must yield values between 0 and 1 (both inclusive) as it is the case with e.g. the Dice coefficient [5].

In a next step, a distance must be defined in G. Two words are close, if $g((w_a, w_b))$ is high. If $(w_a, w_b) \in E$ (i.e. the words involved are co-occurrents) their distance $d(w_a, w_b)$ is easily to be defined as

$$d(w_a, w_b) = \frac{1}{g(w_a, w_b)}. \quad (1)$$

Manuscript received January 2, 2017; revised June 28, 2017.

doi: 10.18178/lnit.5.1.23-28

Otherwise, let us consider the shortest path $p=\{(w_1, w_2), (w_2, w_3),..,(w_k, w_{\{k+1\}})\}$ with $w_1=w_a$, $w_{\{k+1\}}=w_b$ and $(w_i,w_{\{i+1\}}) \in E$ and define

$$d(w_a, w_b) = \sum_{i=1}^{k} d(w_i, w_{i+1}). \quad (2)$$

If there is no path between any two words w_a and w_b, $d(w_a, w_b)=\infty$ shall be set. The definition of the centroid term $\chi(D)$ of a document D uses all N words $\{w_1, w_2,.., w_N\} \in D$, which can be reached from any term t in G. Therefore, the average distance $d(D, t)$ of all words in D to the term t can be obtained by

$$d(D, t) = \frac{\sum_{i=1}^{N} d(w_i,t)}{N}. \quad (3)$$

The centroid term $\chi(D)$ is defined to be the term with

$$d(D, \chi(D)) = MINIMAL.$$

Note, that $\chi(D)$ does not necessarily occur in D. Let χ_1 be the centroid term of D_1, and χ_2 the centroid term of D_2, then $d(\chi_1, \chi_2)$ can be understood as the distance of the two documents D_1 and D_2.

III. Properties of Centroids

A. Background

The authors have always argued that the centroid-based computation is close to thinking mechanisms in the human brain. Mostly the feeling about similarities of words and documents and their sorting within ontological categories is learnt in a long process. Hereby, every new text source can not only be classified but is also used again to refine the knowledge of the individual. First rough working approximations are learnt fast and seem then to be stable for long times. It is observed that a few known keywords are enough to classify even completely unknown sources in the right manner. From the authors' point of view, semantic relations among words and their associated senses are the reason for these effects. Starting with WordNet [6], those relations have been put into graph-based models. Later, different forms of co-occurrence graphs have been found to be a good approximation for the human's intuition for word and term associations, confirmed by stimulus-response experiments [7]. The following experiments shall give some more justification for those thoughts. For all of the exemplary experiments (many more have been conducted) discussed herein, linguistic preprocessing has been applied on the documents to be analysed whereby stop words have been removed and only nouns (in their base form), proper nouns and names have been extracted. In order to build the undirected co-occurrence graph G (as the reference for the centroid distance measure), co-occurrences on sentence level have been extracted. Their significance values have been determined using the Dice coefficient [5]. The particularly used sets of documents to create G and to calculate the centroid terms will be described in the respective subsections. Interested researchers may download these sets from: http://www.docanalyser.de/cd-properties-corpora.zip.

B. Stability of the Co-occurrence Graph

The first experiment shall confirm the fast convergence and stability of the co-occurrence graph which is a prerequisite for its use as a dynamic knowledge base of the individual or local computing node. A co-occurrence graph may be constructed from a text corpus in an iterative manner by successively adding co-occurrences from one document after another and finally removing all non-significant co-occurrence relations. During this learning process, the number of nodes and edges added to the co-occurrence graph for each new incoming document is converging. As especially high-weight edges representing significant co-occurrences are of interest for the centroid determination, Fig. 1 shows that the number of these edges converges quickly when a well-balanced corpus is used to construct G.

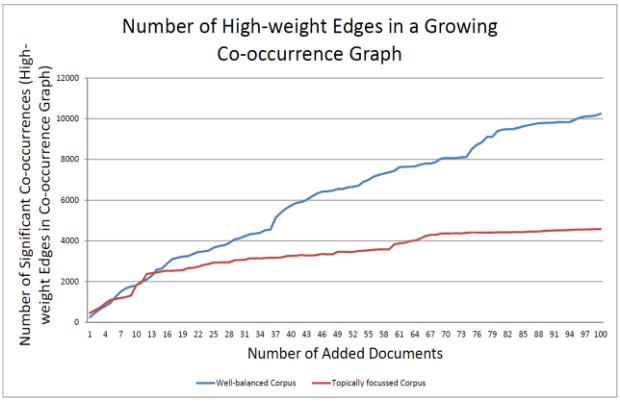

Figure 1. Convergence of the number of high-weight edges in a growing co-occurrence graph.

The effect is even stronger for a topically focussed corpus as the terminology used in it does not vary greatly. The topically well-balanced corpus from dataset 3.2.1 used in this experiment contains 100 randomly chosen online news articles from the German newspaper 'Süddeutsche Zeitung' from the months September, October and November of 2015 and covers 19 topics. The topically focussed corpus from the same dataset contains 100 articles on the European migrant crisis (a hotly discussed topic in late 2015) from the same newspaper and the same period.

During the learning process, the probability that new words (nodes) are included is drastically decreasing, also the nodes' ranks (according to their outdegrees) only seldom changes. The node with rank 1 has the highest connectivity. Fig. 2 shows that most of the rank changes occur at the nodes with higher ranks (in this case, rank changes have been determined after 100 documents have been added to the collection) only. These are usually nodes with low connectivity and often have been added to the co-occurrence graph just recently. For this experiment, the topically focused corpus from dataset 3.2.1 has been used again.

Centroid terms –as defined in the previous section– do not change frequently, even when the co-occurrence

graph is growing. For this experiment, only one document has been added to the co-occurrence graph in each time step.

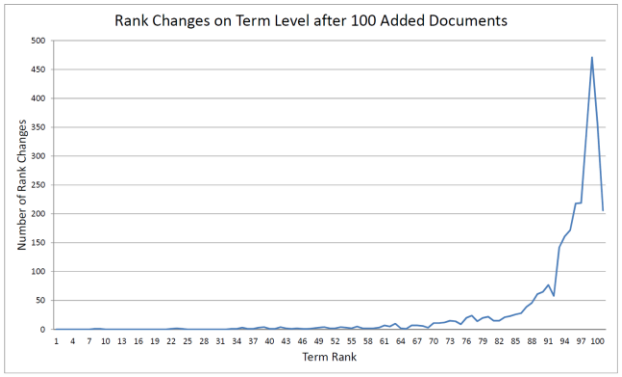

Figure 2. Stability of term ranks in a grown co-occurrence graph.

As it can be seen in Fig. 3, the 'movement' of a reference document's centroid term (calculated after each time step) stabilises quickly. However, the convergence time depends on the order in which documents are added to the co-occurrence graph. Here, the topical orientation of and similarity between them plays an important role.

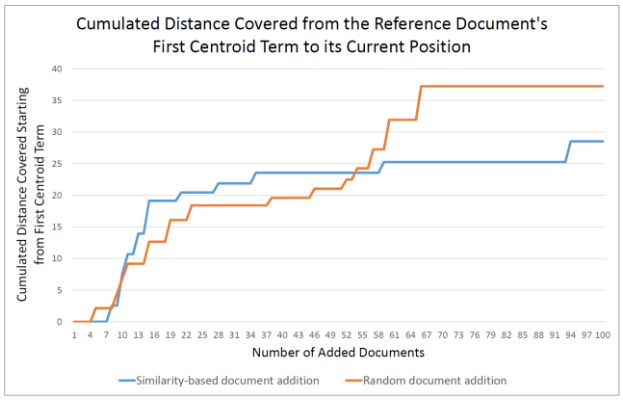

Figure 3. Changes of document centroids in a growing co-occurrence graph.

If similar documents to the reference document are added first (blue curve) and other, topically dissimilar documents afterwards, then the centroid term changes rarely (almost never) during their addition. If, however, the documents are added randomly (orange curve), then the convergence time increases. The reason for this observation is that topically similar documents (which mostly influence the centroid term's position) can be added at any time. Thus, the probability that the centroid term changes at any time is increased, too.

The corpus from dataset 3.2.3 used for this experiment contains 100 newspaper articles from 'Süddeutsche Zeitung' which cover three topical categories 'car' (34 articles), 'finance' (33 articles) and 'sports' (33 articles). The reference document used was 'Schmutzige Tricks' (an article on the car emissions scandal). Therefore, it is not surprising that mainly in the first 34 time steps during the similarity-based document addition (in which the 34 car-related documents are added) the centroid term's position of this article is changed. Even so, in both cases, the centroid term's jumping distance between two consecutive 'positions' in the co-occurrence graph is low. However, it is still possible that the centroid term (usually just slightly) changes when the co-occurrence graph significantly grows as this process changes the distances between all nodes as well.

C. *Uniqueness of Centroid Terms*

Unfortunately, it is not easy to generate the centroid term of two parts of text from knowing their separate centroids. It costs once an effort of O(W3) to construct the distance matrix of the co-occurrence graph G and then an additional O(W) to determine for every document any possible centroid candidate; an effort which must be definitely reduced in the future (although the calculation must be carried out only once or –at least– not very often for every document). However, since the co-occurrence graph's stability is quickly reached, centroids usually need to be calculated only once when documents first appear in the corpus or after larger time periods, in which significant changes of the underlying knowledge have occurred due to incoming sets of new documents.

In fact, it might happen in a given co-occurrence graph that one or more terms have the same, minimal average distance to all terms of the text or document. This would mean that the centroid term is not uniquely defined and more than one term could represent the document. In particular, this complies with reality –some documents, especially interdisciplinary ones– may not by clearly assigned to the one or another category. However, the subsequently explained practical experiences justify that this case in fact might only appear extremely rarely.

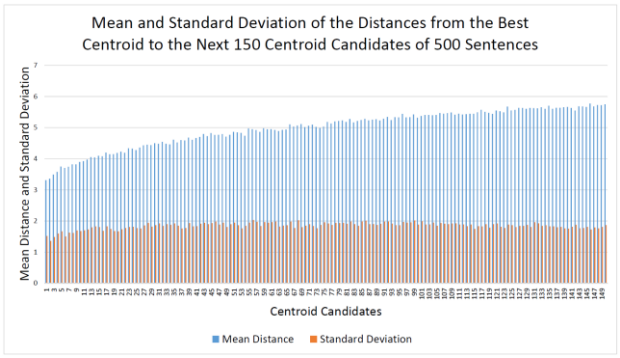

Figure 4. Distances between centroid candidates.

In Fig. 4, it is shown that in general there is a significant distance between the best (the actually chosen one) centroid term and the next 150 potential centroid candidates closest to it.

This experiment has been conducted using 500 randomly selected sentences from the mentioned Wikipedia corpus for which their respective centroid terms have been determined while avoiding a topical bias. The results show that the mean distance from the best centroid to the potential centroid candidates gradually increases, too.

Although the standard deviation is relatively large, it stays constant. However, even when taking this value into account as well, the mean distance in the co-occurrence

graph between the best centroid and its e.g. 10 closest centroid candidates is still large enough to come to the conclusion that the determined centroid (its position) is in general the best choice to represent a given textual entity. Further research needs to be carried out to find out, if and how much the centroid candidates' topical orientation or focus generally differs from the centroid term's one.

At this point, it must also be mentioned that the well-known superposition principle may not be applied to text centroids, i.e., if $\chi(D_1)$ and $\chi(D_2)$ are the centroid terms of two pieces of text (or documents) D_1 and D_2, the following usually holds:

$$\chi(D_1 \cup D_2) \neq \chi(\chi(D_1) \cup \chi(D_2)). \quad (4)$$

Fig. 5 illustrates one respective counterexample. Given the presented graph, the nodes Z_1 and Z_2 are the centroids for the sets of nodes {a, b, c} and {u, v, w,} respectively. These sets could represent the terms contained in two sentences. The following distances between the nodes can be extracted: $d(Z_1, x)=1$ for $x \in \{a, b, c\}$; $d(Z_2, x)=1$ for $x \in \{u, v, w\}$; $d(Z_4, Z_1)=d(Z_4, Z_2)=2$; $d(Z_3, Z_1)=d(Z_3, Z_2)=3$; $d(Z_3, x)=2$ for $x \in \{a, b, c, u, v, w\}$.

The centroid of Z_1 and Z_2 is node Z_4 and not Z_3 which is, however, the centroid of all single nodes contained in these sets. In particular, this fact contradicts with the hope to reduce the needed effort to calculate the centroid of a set of documents in an easy manner. It shows once more that there are significant differences between the text centroids and their physical analogon, the centre of mass, due to the discrete character of the co-occurrence graph.

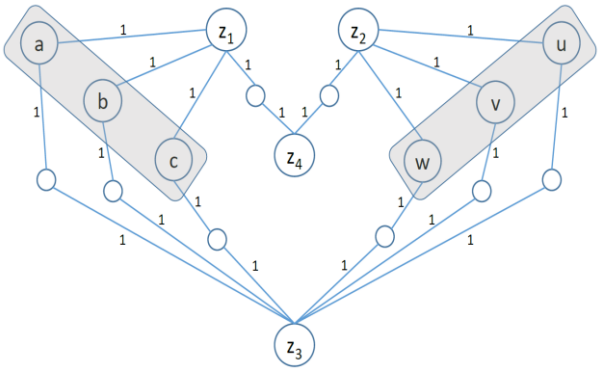

Figure 5. Counterexample for the calculation of centroids.

However, practical experiments have shown, that $\chi(D_1 \cup D_2)$ and $\chi(\chi(D_1) \cup \chi(D_2))$ are not too far from each other. As it can be seen in Table I for an example using the Wikipedia-article 'Measles', those centroids have an average low distance between 2 and 3 while the maximum distance of two terms in the co-occurrence graph used was 18. The distance between the centroids of all sentence centroids in a fixed section of the article and the direct centroid of this section (in this case, its sentence boundaries have not been considered and its terms have been directly used to determine the centroid) is shown for all 11 sections.

TABLE I. DISTANCES OF SPECIFIC CENTROIDS OF SECTIONS IN THE WIKIPEDIA-ARTICLE 'MEASLES'

Number of section	Centroid of all sentence centroids in section	Centroid of section	Distance of both centroids
1	measles	treatment	3,63
2	virus	measles	2,31
3	HIV	HIV	0
4	infection	diagnosis	3,40
5	aid	measles	3,24
6	infection	risk	2,52
7	infection	symptom	3,44
8	net	prevention	1,53
9	malaria	prevention	2,23
10	aid	interaction	1,51
11	research	health	2,23

Summarising, it can be mentioned that

- The centroid of one term is the term itself, the centroid of two terms is usually a node close to the middle of the shortest path between them,
- The centroid is usually not the most frequent or most central term of a document,
- Usually, the centroid is uniquely defined although two or more terms may satisfy the condition to be the centroid,
- The centroid of a text or document can be a term which is not contained in its set of words and
- The centroid of two or more documents is usually not a node on any shortest path among their centroid terms or a star point with the shortest distance to them.

Nevertheless, finding a representing term to pieces of text also brings with it significant advantages, which shall be discussed in the following section.

D. Hierarchies of Centroids

Although –differing from semantic approaches– the assigned centroid terms may not represent any semantic meaning of the given text, they are in each case a formally calculable, well-balanced extract of the words used in the text and their content relations. This approach has been used to define the distance of documents [1] and to determine cluster hierarchies [2], too. Additionally, centroids may be used to detect topical shifts, i.e. subsequent changes in sections, paragraphs or sections of texts may be analysed, where usually classic methods offer only a pairwise comparison of the similarity of text fragments and do not take additional structural information of the given texts into account.

Fig. 6-7 show for the two structurally similar Wikipedia-articles 'Measles' and 'Chickenpox' the obtained dendrograms of centroid terms if the centroids of sentences are set in relation with those of paragraphs, sections and the whole documents. The centroids of those text fragments have been calculated by directly taking all terms contained in them into account and not by computing the centroids of the centroids of the respective next lower structural level.

As an example for the article ‘Measles’, 11 sections are contained in it while the first section’s centroid is ‘treatment’, the second section’s centroid is ‘infection’ and so on. Furthermore, the sixth section (on the treatment of measles with the centroid ‘risk’) contains five paragraphs with up to 5 sentences in them. For each of those paragraphs and each of the sentences contained in them, the computed centroids are presented as well.

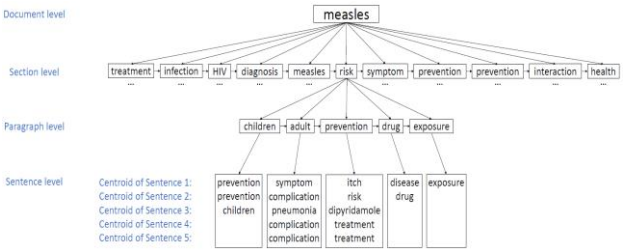

Figure 6. Hierarchy of centroids obtained from sentences, paragraphs, sections and the entire Wikipedia-article ‘Measles’.

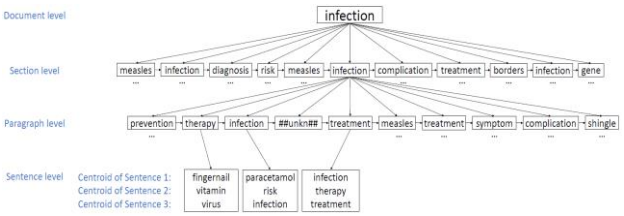

Figure 7. Hierarchy of centroids from sentences, paragraphs, sections and the entire Wikipedia-article ‘Chickenpox’.

The section on the treatment of chickenpox (also the sixth section) contains in contrast to the article ‘Measles’ 10 paragraphs. As it can be seen in the depicted lists of sentence centroids for both articles and even on paragraph level, the terms such as ‘paracetamol’, ‘vitamin’, ‘drug’ and ‘dipyridamole’ are more specific than on section level. The centroid term of the fourth paragraph in the treatment section of the article ‘Chickenpox’ could not be properly determined (‘##unkn##’) as the one sentence in this paragraph contains only one term (the noun ‘antiviral’) which is, in addition to it, not existing in the given co-occurrence graph as it does not co-occur with any other term in the used Wikipedia corpus. Alternatively, as stated in the previous section, the term ‘antiviral’ could have been chosen as the centroid term instead.

However, it can be noticed from this example and similar cases that the determination of centroid terms is partly difficult and their quality is reduced as well when the textual context used for this purpose is small, the co-occurrence graph does not contain required terms or isolated and small clusters in it are addressed. In practice, the handling of exceptional cases like this must be specified.

When comparing both dendrograms (and the centroid terms in them), it is also possible to come to the conclusion that both articles exhibit a similar topical structure. Even on section level, it is recognisable that the articles first deal with the general description of the diseases followed by usual diagnostic methods applied. Then they deal with the treatment of the diseases and discuss their epidemiology.

The interesting aspect is not only the decomposition of segments into sub-topics but the traces obtained from the left-to-right sequence of centroid terms and topics on the same level of the tree. From the examples above, it can be concluded that

- The diseases covered in the selected articles are similar,
- The articles exhibit in their structural composition one and the same style,
- The centroid terms in the lower structural levels are more specific than in the upper levels (the number of terms in the sentences and paragraphs used to calculate them is of course lower and their centroids are determined by a smaller, more topically specific context),
- A distance calculation of equally-ranked centroids on the same structural level will result in an estimation of how semantically close the respective descriptions (in this case those diseases) are to each other (due to these considerations, the authors detected the similarity of two diseases, whose English names are similar but differ from e.g. German ones, i.e. ‘Measles' and ‘German Measles' (Rubella)) as well as
- A continuous distance check of paths or sequences of section- or paragraph-based centroid terms of very similar documents can show where exactly their semantic or topical differences lie.

Therefore, it is sensible that future cluster building solutions take into account these findings. Also, they present a new direction to compute the centrality [8], [9] of words in a co-occurrence graph in order to find proper generalising terms for contents and to perform further semantic derivations from the position of centroids and their traces in the co-occurrence graph.

IV. Conclusion

As the behaviour and changes of centroids as well as their determining context, the co-occurrence graph, are hard to derive in a theoretic manner, a set of experiments in well-defined environments have been conducted.

The results justify the practicability and usability of centroid terms as text representatives. It could also be demonstrated that the used context behaves in a stable manner and especially its extension in a knowledge learning process does not influence the situation. Further publications will investigate mechanisms to reduce the effort needed to determine centroids by utilising neighbourhood effects in co-occurrence graphs.

References

[1] M. Kubek and H. Unger, “Centroid terms as text representatives,” in *Proc. ACM Symposium on Document Engineering*, New York, 2016, pp. 99–102.

[2] M. Kubek and H. Unger, “Towards a librarian of the web,” in *Proc. 2nd International Conference on Communication and Information Processing*, New York, 2016, pp. 70–78.

[3] G. Salton, A. Wong, and C. S, Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613-620, 1975.
[4] S. Deerwester, *et al.*, "Indexing by latent semantic analysis," *Journal of the American Society of Information Science*, vol. 41, no. 6, pp. 391– 407, 1990.
[5] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297-302, 1945.
[6] G. A. Miller, "WordNet: A lexical database for English," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.
[7] G. Heyer, U. Quasthoff, and T. Wittig, *Text Mining - Wissensrohstoff Text*, Bochum: W3L-Verlag, 2006.
[8] G. Erkan and D. R. Radev, "LexRank: Graph-based lexical centrality as salience in text summarization," *Journal of Artificial Intelligence Research*, vol. 22, no. 1, pp. 457-479, 2004.
[9] I. Cantador, D. Vallet, and J. M. Jose, "Measuring vertex centrality in co-occurrence graphs for online social tag recommendation," in *Proc. ECML/PKDD Discovery Challenge*, Aachen, 2009, pp. 17-33.

Mario Kubek is a researcher at the Chair of Communication Networks of the FernUniversität in Hagen. He received his PhD in 2012 with a thesis on locally working agents to improve the search for web documents. His research focus is on natural language processing, text mining and semantic information retrieval in large distributed systems. His further research interests include topic and trend detection in diachronic text corpora and contextual information processing in mobile computing environments.

Thomas Böhme is a professor for mathematics at the Institute for Mathematics of the Technische Universität Ilmenau (TU Ilmenau). He obtained his PhD with a work on spatial representations of finite graphs in 1988 from the Technische Hochschule Ilmenau and his habilitation with a work on cycles in embedded graphs from TU Ilmenau in 1999. His main research focus is on game theory, especially learning in repeated games. Second, he is conducting research in the field of graph theory. Also, he is interested in distributed algorithms.

Herwig Unger received his PhD with a work on Petri Net transformation in 1994 from the Technische Universität Ilmenau and his habilitation with a work on large distributed systems from the University of Rostock in 2000. Since 2006, he is a full professor at the FernUniversität in Hagen and the head of the Chair of Communication Networks. His research interests lie in the areas of self-organization, adaptive and learning systems, Internet algorithms, simulation systems as well as information retrieval in distributed systems.