Centroid Terms and their Use in Natural Language Processing

Mario M. Kubek and Herwig Unger

Chair of Communication Networks
FernUniversität in Hagen, Germany

Abstract: The calculation of semantic similarities between text documents plays an important role in automatic text processing. For example, algorithms to topically cluster and classify texts heavily rely on this information. Standard methods for doing so are usually based on the bag-of-words model and thus return only rough estimations regarding the relatedness of texts. Moreover, they are unable to find generalising terms or abstractions describing the textual contents. Therefore, a new graph-based method to determine centroid terms as text representatives will be introduced. It is shown, that – among further application scenarios – this method is able to compute the similarity of texts even if they have no terms in common. In first experiments, its results and advantages will be discussed in detail.

1 Motivation

After only a few lines of reading, a human reader is able to determine which category of texts and which abstract topic category a given document belongs to. This is a strong demonstration of how well and fast the human brain, especially the human cortex, can process and interpret data. It is able to not only understand the meaning of single words – as representations of real-world entities – but a certain composition of them [1], too.

In many text mining applications, the topical grouping of texts and terms/words contained in them are common tasks. In order to group semantically related terms, unsupervised topic modeling techniques such as LDA [2] have been successfully applied. This technique tries to infer word clusters from a set of documents based on the assumption that words from the same topic are likely to appear next to each other and therefore share a related meaning. Here, deep and computationally expensive (hyper)parameter estimations are carried

out and for each word, the probability to belong to specific topic is computed in order to create those constructions. The graph-based Chinese Whispers algorithm [3] is another interesting clustering technique that can be used in the field of natural language problems, especially to semantically group terms. It is usually applied on undirected semantic graphs that contain statistically significant term relations found in texts.

Also, it is usual to apply the k-means algorithm [4] to group terms. For this purpose, it necessary to determine their semantic distance. Here, several methods can be applied. The frequency of the co-occurrence of two terms in close proximity (in a window of n words or on sentence level) is a first indication for their semantic distance. Terms that frequently co-occur together are usually semantically related. Several graph-based distance measures [7, 8] consult manually created semantic networks such as WordNet [5], a large lexical database containing semantic relationships for the English language that covers relations like polysemy, synonymy, antonymy, hypernymy and hyponymy (i.e. more general and more specific concepts), as well as part-of-relationships. These measures apply shortest path algorithms or take into account the depth of the least common subsumer concept (LCS) to determine the closest semantic relationship between two given input terms or concepts. It is also common to measure the similarity of term contexts [6] that contain terms that often co-occur with the ones in question. Technically, these contexts are realised as term vectors following the bag-of-words model.

The same approach is applied when the semantic similarity or distance of any two documents should be determined. Here, the term vectors to be compared contain the texts' characterising terms and their score (typically, a TF-IDF-based statistic [9] is used) as a measure for their importance. The similarity of two term vectors can be determined using the cosine similarity measure or by calculating the overlap of term vectors, e.g. using the Dice coefficient [10]. The commonly used Euclidean distance and the Manhattan distance are further examples to measure the closeness of term vectors at low computational costs.

However, in some cases, these measures do not work correctly (with respect to human judgement), mostly if different people write about the same topic but are using a completely different vocabulary for doing so. The reason for this circumstance can be seen in the isolated view of the words found in documents to be compared without including any relation to the vocabulary of other, context-related documents. Moreover, short texts as often found in posts in online social networks or short (web) search queries with a low number of de-

criptive terms can therefore often not be correctly classified or disambiguated. Another disadvantage is that these measures cannot find abstractions or generalising terms by just analysing the textual data provided. For this purpose, static lexical databases such as WordNet [5] must be consulted as a reference. Despite their usefulness, these resources are – in contrast to the human brain – not able to learn about new concepts and their relationships.

In order to address these problems, this article presents a new graph-based approach to determine centroid terms of text documents. It is shown that those terms can actually represent text documents in automatic text processing, e.g. to determine their semantic distances. In the next section, the fundamentals of this method are presented. Afterwards, section 3 describes its mathematical and technical details. Section 4 proves the validity of this approach by explaining the results of first experiments. In section 5, the method's working principles and advantages are discussed. Section 6 presents numerous application scenarios for it in the fields of text mining and information retrieval while also elaborating on technological aspects of its practical implementation. Section 7 summarises the article and suggests further application fields of the introduced method.

2 Fundamentals

For the approach presented herein, co-occurrences and co-occurrence graphs are the basic means to obtain more detailed information about text documents than term frequency vectors etc. could ever offer. The reason for this decision is that co-occurrence graphs are able to accumulate a certain knowledge obtained from a few selected or all documents of a text corpus while (at least to some extent) maintaining the semantic connection of terms found in them.

Two words w_i and w_j are called *co-occurrents*, if they appear together in close proximity in a document D. The most prominent kinds of such co-occurrences are word pairs that appear as immediate neighbours or together in a sentence. A *co-occurrence graph* $G = (W, E)$ may be obtained, if all words of a document or set of documents W are used to build its set of nodes which are then connected by an edge $(w_a, w_b) \in E$ if $w_a \in W$ and $w_b \in W$ are co-occurrents. A weight function $g((w_a, w_b))$ indicates, how significant the respective co-occurrence is in a document. If the significance value is greater than a pre-set threshold, the co-occurrence can be regarded as significant and a semantic relation between the words involved can often be derived from it. Commonly used significance measures are the Dice coefficient [10], the mutual information measure [11], the Poisson collocation measure [12] and the log-likelihood ratio [13].

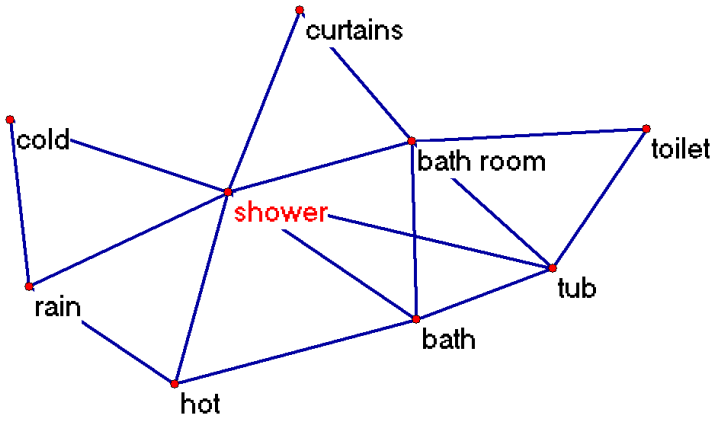

Fig. 1: A co-occurrence graph for the word "shower"

A co-occurrence graph – similarly to the knowledge in the human brain – may be built step by step over a long time taking one document after another into consideration. From the literature [6] and own experiments (see Figure 2) it is known that the out-degrees of nodes in co-occurrence graphs follow a power-law distribution and the whole graph exhibit small-world properties with a high clustering coefficient as well as a short average path length between any two nodes. This way, a co-occurence graph's structure also reflects the organisation of human lexical knowledge.

The use of the immediate neighbourhood of nodes in a co-occurrence graph has been widely considered in literature, e.g. to cluster terms [3] and to determine the global context (vector) of terms in order to evaluate their similarity [6] or to derive paradigmatic relations between them [14]. In the authors' view, indirect neighbourhoods of terms in co-occurrence graphs (nodes that can be reached only using two or more edges from a node of interest) and the respective paths with a length ≥ 2 should be considered as well as indirectly reachable nodes may still be of topical relevance, especially when the co-occurence graph is large. The benefit of using such nodes/terms in co-occurrence graphs has already been shown by the authors for the expansion of web search queries using a spreading activation technique applied on local and user-defined corpora [15]. The precision of web search results can be noticably improved when taking those terms into account, too.

The field of application of indirect term neighbourhoods in co-occurrence graphs shall be extended in the next section by introducing an approach to determine centroid terms of text documents that can act as their representatives in further text processing tasks. These centroid terms can be regarded as the texts'

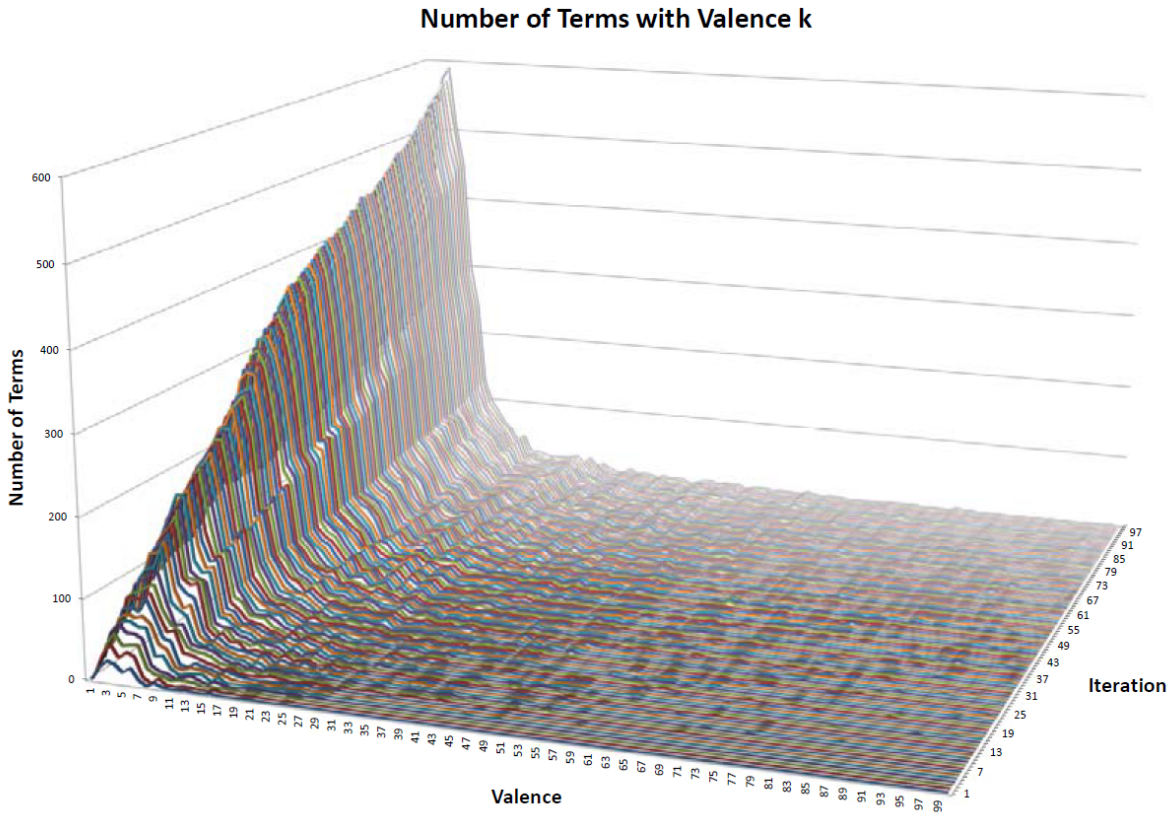

Fig. 2: Distribution of out-degrees in a co-occurrence graph over time

topical centers of interest (a notion normally used to describe the part of a picture that attracts the eye and mind) that the authors' thoughts revolve around.

3 Finding Centroid Terms

In physics, complex bodies consisting of several single mass points are usually represented and considered by their so-called center of mass, as seen in Figure 3. The distribution of mass is balanced around this center and the average of the weighted coordinates of the distributed mass defines its coordinates and therefore its position.

For discrete systems, i.e. systems consisting of n single mass points $m_1, m_2, .., m_i$ in a $3D-$space at positions $\vec{r}_1, \vec{r}_2, .., \vec{r}_i$, the center of mass $\vec{r}_s$ can be found by

$$\vec{r}_s = \frac{1}{M} \sum_{i=1}^{n} m_i \vec{r}_i, \tag{1}$$

whereby

$$M = \sum_{i=1}^{n} m_i. \tag{2}$$

Usually, this model simplifies calculations with complex bodies in mechanics by representing the whole system by a single mass at the position of the center of mass. Exactly the same problem exists in automatic text processing: a whole text shall be represented or classified by one or a few single, descriptive terms which must be found.

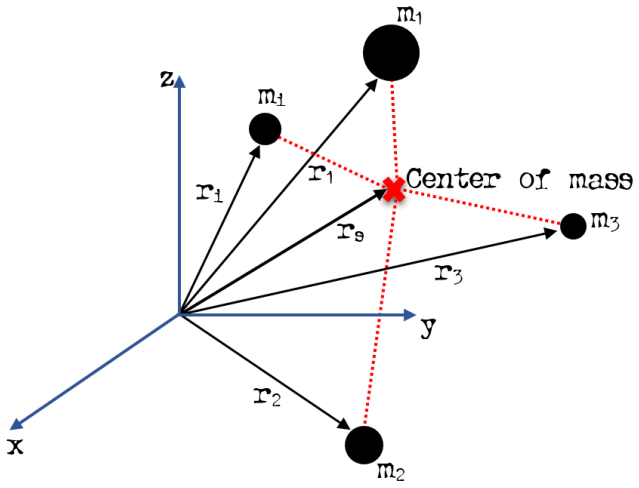

Fig. 3: The physical center of mass

To adapt the situation for this application field, first of all, a *distance d* shall be introduced in a co-occurrence graph G. From literature it is known that two words are semantically close, if $g((w_a, w_b))$ is high, i.e. they often appear together in a sentence or in another predefined window of n words. Consequently, a distance $d(w_a, w_b)$ of two words in G can be defined by

$$d(w_a, w_b) = \frac{1}{g((w_a, w_b))}, \tag{3}$$

if w_a and w_b are co-occurrents. In all other cases (assuming that the co-occurrence graph is connected[1]) there is a 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 $w_i, w_{i+1} \in E$ for all $i = 1(1)k$ such that

$$d(w_a, w_b) = \sum_{i=1}^{k} d((w_i, w_{i+1})) = MIN, \tag{4}$$

[1]This can be achieved by adding a sufficiently high number of documents to it during its building process.

whereby in case of a partially connected co-occurrence graph $d(w_a, w_b) = \infty$ must be set. Note, that differing from the physical model, there is a distance between any two words but no direction vector, since there is no embedding of the co-occurrence graph in the $2-$ or $3-$dimensional space. Consequently, the impact of a word depends only on its scalar distance.

In continuation of the previous idea, the distance between a given term t and a document D containing N words $w_1, w_2, .., w_N \in D$ that are reachable from t in G can be defined by

$$d(D, t) = \frac{\sum_{i=1}^{N} d(w_i, t)}{N}, \tag{5}$$

i.e. the average sum of the lengths of the shortest paths between t and all words $w_i \in D$ that can be reached from it. Note that -differing from many methods found in literature- it is not assumed that $t \in D$ holds! Also, it might happen in some cases that the minimal distance is not uniquely defined, consequently a text may have more than one centroid term (as long as no other methods decide which one is to use). In order to define the centroid-based distance ζ between any two documents D_1 and D_2, let t_1 be the center term or *centroid term of* D_1 with $d(D_1, t_1) = MIN$. If at the same time t_2 is the centroid term of D_2,

$$\zeta(D_1, D_2) = d(t_1, t_2) \tag{6}$$

can be understood as the semantic distance ζ of the two documents D_1 and D_2. In order to obtain a similarity value instead,

$$\zeta_{sim}(D_1, D_2) = \frac{1}{1 + \zeta(D_1, D_2)} \tag{7}$$

can be applied.

It is another important property of the described distance calculation that documents regardless of their length as well as single words can be assigned a centroid term by one and the same method in a unique manner. The presented approach relies on the preferably large co-occurrence graph G as its reference. It may be constructed from any text corpus in any language available or directly from the sets of documents whose semantic distance shall be determined. The usage of external resources such as lexical databases or reference corpora is common in text mining: as an example, the so-called difference analysis [6, 16] which measures the deviation of word frequencies in single texts from their frequencies in general usage (a large topically well-balanced reference corpus is needed for this purpose) is an example for it. The larger the deviation is, the more likely

it is that a term or keyword of a single text has been found. Furthermore, the presented distance measure is not only based on a physical analogon and bears (at least to a certain extent) resemblance to the well-known difference analysis as discussed, the measure's approach is brain-inspired, too. Further considerations in this respect will be discussed in section 5.

In the following section, the quality and properties of the centroid terms and the new centroid-based distance measure shall be investigated and discussed.

4 First Experiments

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 [10]. The particularly used sets of documents will be described in the respective subsections[2].

4.1 Centroids of Wikipedia Articles

As the centroid terms are the basic components for the centroid-based distance measure, it is useful to get a first impression of their quality in terms of whether they are actual useful representatives of documents. Table 1 therefore presents the centroid terms of 30 English Wikipedia articles. The corpus used to create the reference co-occurrence graph G consisted of 100 randomly selected articles (including the mentioned 30 ones) from an offline English Wikipedia corpus from `http://www.kiwix.org`. It can be seen that almost all centroids properly represent their respective articles.

4.2 Comparing Similarity Measures

In order to evaluate the effectiveness of the new centroid-based distance measure, its results will be presented and compared to those of the cosine similarity measure while the same 100 online news articles from the German newspaper "Süddeutsche Zeitung" from the months September, October and November of 2015 have been selected (25 articles from each of the four topical categories 'car',

[2]Interested researchers can download these sets (1,3 MB) from `http://www.docanalyser.de/cd-corpora.zip`

Table 1: Centroids of 30 Wikipedia articles

Title of Wikipedia Article	Centroid Term
Art competitions at the Olympic Games	sculpture
Tay-Sachs disease	mutation
Pythagoras	Pythagoras
Canberra	Canberra
Eye (cyclone)	storm
Blade Runner	Ridley Scott
CPU cache	cache miss
Rembrandt	Louvre
Common Unix Printing System	filter
Psychology	psychology
Religion	religion
Universe	shape
Mass media	database
Rio de Janeiro	sport
Stroke	blood
Mark Twain	tale
Ludwig van Beethoven	violin
Oxyrhynchus	papyrus
Fermi paradox	civilization
Milk	dairy
Corinthian War	Sparta
Health	fitness
Tourette syndrome	tic
Agriculture	crop
Finland	tourism
Malaria	disease
Fiberglass	fiber
Continent	continent
United States Congress	Senate
Turquoise	turquoise

'travel', 'finance' and 'sports' have been randomly chosen) for this purpose. As the cosine similarity measure operates on term vectors, the articles' most important terms along with their scores have been determined using the extended PageRank [17] algorithm which has been applied on their own separate (local)

co-occurrence graphs (here, another term weighting scheme such as a TF-IDF variant [9] could have been used as well). The cosine similarity measure has then been applied on all pairs of the term vectors. For each article A, a list of the names of the remaining 99 articles has been generated and arranged in descending order according to their cosine similarity to A. A most similar article can therefore be found at the top of this list.

In order to apply the new centroid distance measure to determine the articles' semantic distance, for each article, its centroid term has been determined with the help of the co-occurrence graph G using formula 5. The pairwise distance between all centroid terms of all articles in G has then been calculated. Additionally, to make the results of the cosine similarity measure and the centroid distance measure comparable, the centroid distance values have been converted into similarity values using formula 7.

The exemplary diagram in Figure 4 shows for the reference article (”Abgas-Skandal – Schummel-Motor steckt auch in Audi A4 und A6”) its similarity to the 50 most similar articles. The cosine similarity measure was used as the reference measure. Therefore, the most similar article received rank 1 using this measure (lower bars). Although the similarity values of the two measures seem uncorrelated, it is recognisable that especially the articles with a low rank (high similarity) according to the cosine similarity measure are generally regarded as similar by the centroid distance measure, too. In case of Figure 4, the reference article dealt with the car emissions scandal (a heavily dicussed topic in late 2015). The articles at the ranks 3 (”Abgas-Affäre – Volkswagen holt fünf Millionen VWs in die Werkstätten”), 7 (”Diesel von Volkswagen – Was VW-Kunden jetzt wissen müssen”) and 12 (”Abgas-Skandal – Was auf VW- und Audi-Kunden zukommt”) according to the cosine similarity measure have been considered most similar by the centroid distance measure, all of which were indeed related to the reference article. The strongly related articles at the ranks 1, 4, 6 and 9 have been regarded as similar by the centroid distance measure, too. In many experiments, however, the centroid distance measure considered articles as similar although the cosine similarity measure did not.

Here, another implicit yet important advantage of the new centroid distance measure becomes obvious: two documents can be regarded as similar although their wording differs (the overlap of their term vectors would be small or even empty and the cosine similarity value would be very low or 0). The article at rank 49 (”Jaguar XF im Fahrbericht – Krallen statt Samtpfoten”) is an example

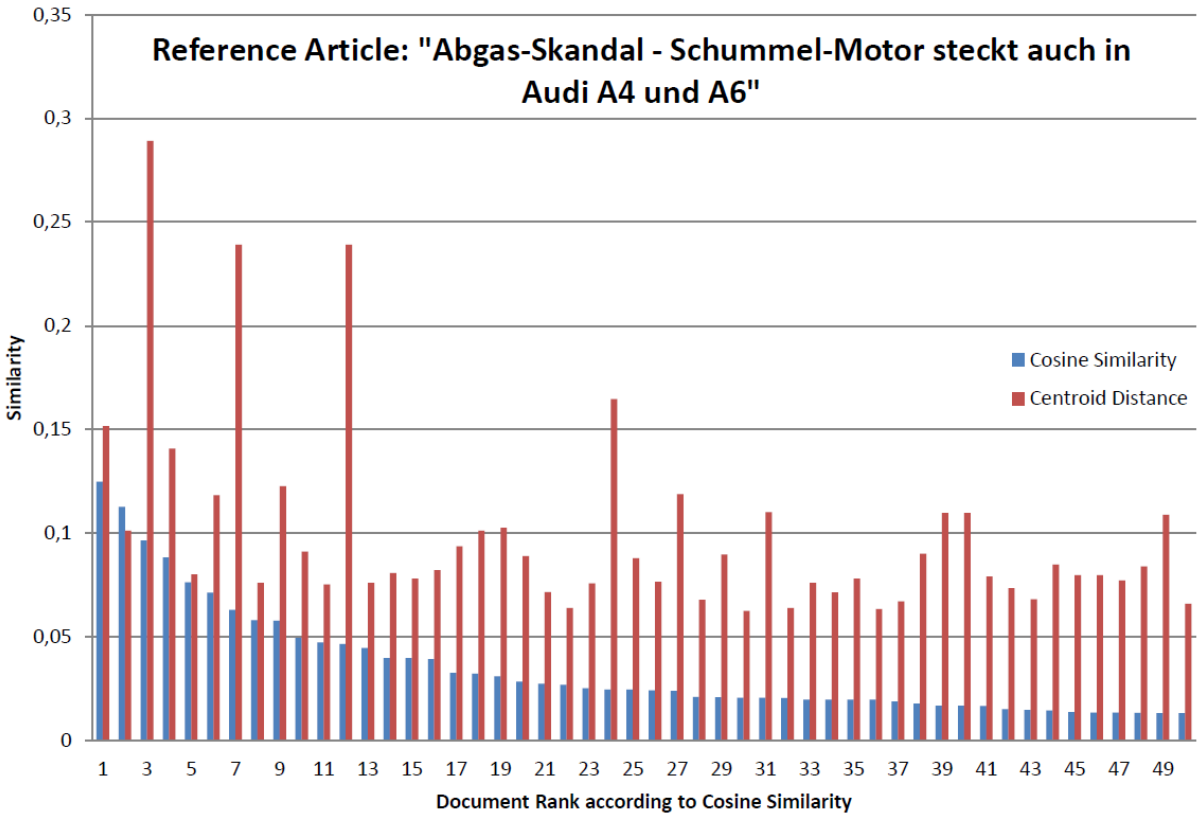

Fig. 4: Cosine similarity vs. centroid distance (topic: car emissions scandal)

for such a case. The centroid distance measure uncovered a topical relationship to the reference article, as both texts are car-related and deal with engine types.

Figure 5 depicts another case of this kind: the article with rank 29 received the highest similarity score from the centroid distance measure. A close examination of this case revealed that the centroids of the reference article (”Deutschland – Ausgebucht”) and the article in question (”Briefporto – Post lässt schon mal 70-Cent-Marken drucken”) are located close to each other in the reference co-occurrence graph. The reference article’s main topic was on financial investments in the German hotel business and the article at rank 29 dealt with postage prices of Deutsche Post AG. Both articles also provided short reports on business-related statistics and strategies.

4.3 Searching for Text Documents

The previous experiments suggest that the centroid distance measure might be applicable to search for text documents, too. In this sense, one might consider a query as a short text document whose centroid term is determined as described

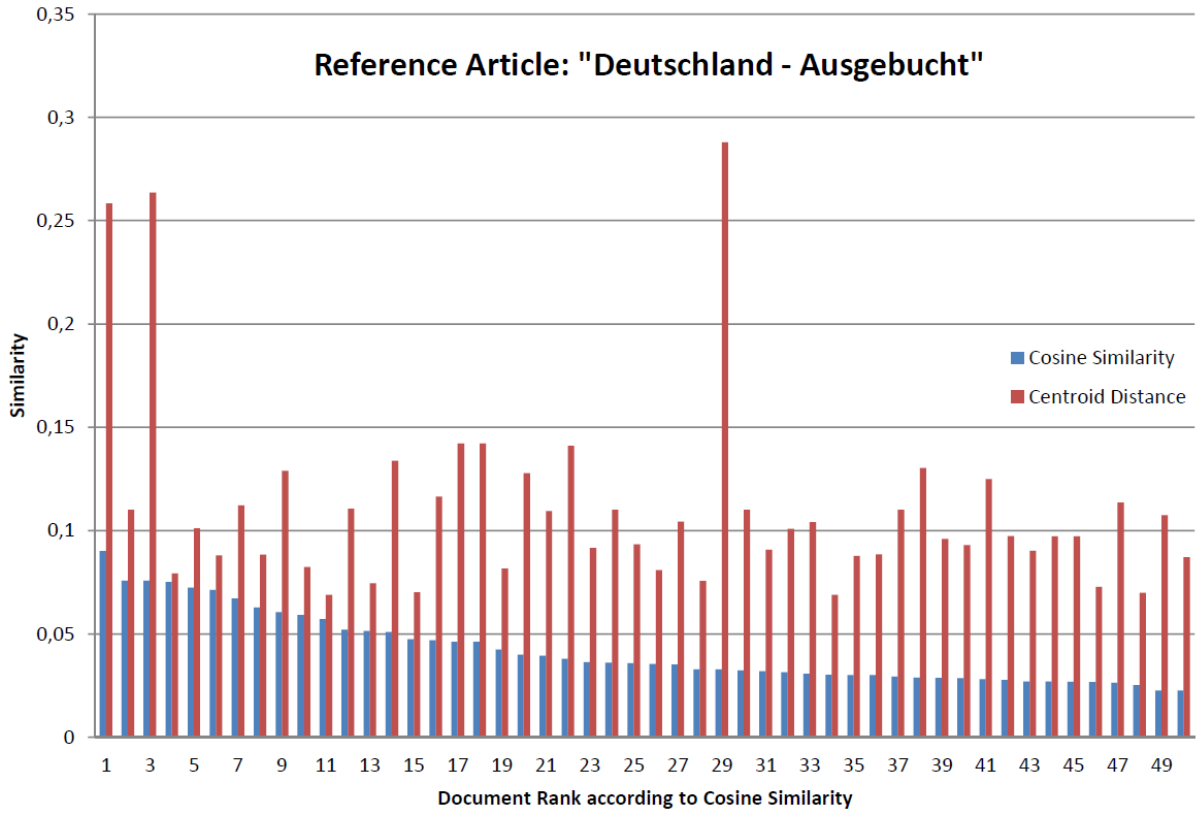

Fig. 5: Cosine similarity vs. centroid distance measure (topic: business-related statistics and strategies)

before and the k documents whose centroid terms are closest to the query's centroid term are returned as matches. These k nearest neighbours are implicitly ranked by the centroid distance measure, too. The best matching document's centroid term has the lowest distance to the query's centroid term.

The following two tables show for two exemplary queries "VW Audi Abgas" (centroid term: "Seat") and "Fußball Geld Fifa" (centroid term: "Affäre") their respective top 10 articles from the German newspaper "Süddeutsche Zeitung" along with their own centroid terms whereby the distances from the queries' centroid terms to all 100 mentioned articles' centroid terms in the co-occurrence graph G have been calculated.

It can be seen that most of the documents can actually satisfy the information need expressed by the queries. This kind of search will, however, not return exact matches as known from the popular keyword-based web search. Instead, documents will be returned that are in general topically related to the query. As the query and the documents to be searched for are both represented by just one centroid term, an exact match is not possible when applying this approach.

Table 2: Top 10 documents for the query "VW Audi Abgas" (Seat)

Filename of News Article	Centroid Term
auto_abgas-skandal-vw-richtet...	Audi
geld_aktien-oeko-fonds-schmeissen-volkswagen-raus	Ethik
auto_bmw-siebener-im-fahrbericht-luxus-laeuft	S-Klasse
auto_abgas-affaere-volkswagen-ruft...	Schadstoffausstoß
auto_abgas-skandal-schummel-motor...	Schadstoffausstoß
geld_briefporto-post-laesst-schon...	Marktanteil
auto_abgas-skandal-was-auf-vw-und-audi...	EA189
auto_abgas-skandal-acht-millionen-vw-autos...	Software
auto_diesel-von-volkswagen-was-vw-kunden...	Motor
auto_abgas-affaere-schmutzige-tricks	Motor

Table 3: Top 10 documents for the query "Fußball Geld Fifa" (Affäre)

Filename of News Article	Centroid Term
sport_affaere-um-wm-mehr-als-nur-ein-fehler	Fifa
sport_angreifer-von-real-madrid-karim-benzema...	Videoaufnahme
sport_affaere-um-wm-vergabe-zwanziger-schiesst...	Zwanziger
sport_affaere-um-fussball-wm-wie-beckenbauers...	Organisationskomitee
sport_affaere-um-wm-zwanziger-es-gab-eine...	Organisationskomitee
sport_affaere-um-wm-vergabe-zwanziger-legt...	Gerichtsverfahren
sport_affaeren-um-wm-vergaben-die-fifa...	Zahlung
sport_affaere-um-wm-netzer-wirft-zwanziger...	Fifa-Funktionär
geld_ehrenamt-fluechtlingshilfe-die-sich...	Sonderausgabe
sport_affaere-um-wm-wie-zwanziger-niersbach...	Präsident

However, this method can still be of use when a preferably large set of topically matching documents is needed. This kind of recall-oriented search is of interest e.g. for people that want to get an overview of a topic or during patent searches when exact query matches might lower the chance of finding possibly relevant documents that nevertheless do not contain all query terms but related terms instead. A typical precision-oriented search would then be harmful. In these cases, a search system would first determine documents that contain the input query terms using its inverted index and then rank these documents by computing e.g. their term vectors' cosine similarity with the query. That means that a highly relevant document will contain (almost) all query terms.

In order to optimise both recall using the centroid distance measure and also the precision for the k top documents (precision@k) using a variant of the aforementioned procedure, it might be sensible to calculate a combined rank that factors in the rankings of both approaches. Also, it is imaginable to use the centroid distance measure (as a substitute for the Boolean model) to pre-select those documents that are in a second step ranked according to the cosine similarity measure. Still, other well-known techniques such as expanding queries using highly related and synonymous terms [15] are suitable options to increase recall as well. More experiments in this regard taking all these approaches into account will be conducted.

Also, in the experiments presented herein, mostly topically homogeneous texts (except for the book analyses) have been used in order to demonstrate the validity of the centroid distance measure and the role of centroid terms as text representatives. In future experiments, it will be interesting to evaluate the effectiveness of this approach when it is applied on more topically heterogeneous documents.

4.4 Analysing Full Books

Additionally, full books (not in combination with other texts) have been analysed to determine their centroid terms. In these cases, the books' own co-occurrence graphs G have been used to determine their important terms and to find their respective centroid terms (one for each book). In case of the English King James version of the Holy Bible, the centroid term determined that has the shortest average distance in the book's (almost fully connected) graph G to all other 7211 reachable terms is 'Horeb'. This experiment has been repeated while only using the k ($k = 25, 50, 75 \ldots$) most frequent terms for this purpose. Here, besides 'Horeb' and others, the terms 'God' and 'gladness' have been determined as the centroid terms. It is to be pointed out that all of these terms have a low distance to each other in the co-occurrence graph G, meaning they are all good representations of the text no matter what actual centroid term is used for further considerations and applications. This also shows, that it is sufficient to take into account only a few prominent terms of a text in order to determine its centroid term in the co-occurrence graph G while at the same time the algorithm's execution time is drastically lowered.

5 Discussion

The presented approach of using a reference co-occurrence graph to determine the semantic distance of texts is brain-inspired, too. Humans naturally, unconsciously and constantly learn about the entities/objects and their relationships surrounding them and build representations of these perceptions in form of concept maps as well as their terminologies in their minds. New experiences are automatically and in a fraction of a second matched with those previously learned. The same principle is applied when using the centroid distance measure. An incoming text A – regardless of whether it was previously used to construct the co-occurrence graph G or not – whose centroid term shall be found, must at least partially be matched against G. In this sense, G takes on the role of the brain and acts as a global and semantic knowledge base. The only prerequisite is that the graph G must contain enough terms that the incoming text's terms can be matched with. However, it is not necessary to find all of A's terms in G for its at least rough topical classification. The human brain does the same. A non-expert reading an online article about biotechnology may not fully understand its terminology, but can at least roughly grasp its content. However, in doing so, this person will gradually learn about the new concepts, a process that is not yet carried out in the herein presented approach. In later publications, the inclusion of this process will be examined.

In order to find proper centroid terms for documents whose topical orientation is unknown, it is important to construct the co-occurrence graph G from a preferably large amount of texts covering a wide range of topics. That is why, in the previous section, the 100 documents to build the respective corpora have been randomly chosen to create G as a topically well-balanced reference. However, the authors assume that topically oriented corpora can be used as a reference when dealing with documents whose terminology and topical orientation is known in advance, too. This way, the quality of the determined centroid terms should increase as they are expected to be better representations for the individual texts' special topical characteristics. Therefore, a more fine-grained automatic classification of a text should be possible. Further experiments are planned to investigate this assumption.

The bag-of-words model that e.g. the cosine similarity measure solely relies on is used by the centroid-based measure as well, but only to the extend that the entries in the term vectors of documents are used as anchor points in the reference co-occurrence graph G (to 'position' the documents in G) in order to determine their centroid terms. Also, it needs to be pointed out once again that a

document's centroid term does not have to occur even once in it. In other words, a centroid term can represent a document, even when it is not mentioned in it.

However, as seen in the experiments, while the cosine similarity measure and the centroid distance measure both often regard especially those documents as similar that actually contain the same terms (their term vectors have a significantly large overlap), one still might argue that both measures can complement each other. The reason for this can be seen in their totally different working principles. While the cosine similarity measure will return a high similarity value for those documents that contain the same terms, the centroid distance measure can uncover a topical relationship between documents even if their wording differs. This is why it might be sensible to combine both approaches in a new measure that factors in the results of both methods. Additional experiments in this regard will be conducted.

Additionally, the herein presented experiments have shown another advantage of the centroid distance measure: its language-independence. It relies on the term relations and term distances in the reference co-occurrence graph G that has been naturally created using text documents of any language.

6 Application Scenarios

The presented centroid distance measure can naturally be applied by text mining algorithms that topically cluster or classify documents. These algorithms make heavy use of similarity and distance measures in order to group semantically similar documents or terms. Here, the new measure can be perfectly applied as an alternative to the well-known measures mentioned above. It will be especially useful, when it comes to grouping topically documents that – despite their topical relatedness – have only a limited amount of terms in common.

However, as shown in the experiments, search applications can make use of this measure, too. Also in this case, documents can be found that do not even share a single query term, yet are highly relevant to the query. Even so, as users are often interested in documents that actually contain the entered query terms but make mistakes in finding the right terms for their information needs, it might be sensible to expand the original query terms with the determined centroid term along with some of its neighbouring terms in the co-occurrence graph G. Matching documents containing these terms could be ranked in reverse order of their similarity to the expanded query. By using this approach, the search results' recall and precision are both expected to increase as common terms in

a topical field (the included centroid term and/or its immediate neighbours) as well as the original query terms are used to find matching documents. This approach will be examined and discussed in further publications.

Interactive search applications such as "DocAnalyser" [18] that aim at helping users to find topically similar and related documents in the World Wide Web could benefit from employing the centroid distance measure, too. Starting with a document of the user's interest, the application could determine the document's centroid term as described before and send this term (to increase the search results' recall) as well as some characteristic terms of the document as an automatically formulated query to a web search engine which will (hopefully) return relevant documents.

From the technological point of view, it becomes obvious that it is necessary to be able to manage large graph structures efficiently and effectively. Graph databases such as Neo4j [19] are specifically designed for this purpose. They are also well-suited to support graph-based text mining algorithms [20]. This kind of databases is not only useful to solely store and query the herein discussed co-occurrence graphs, with the help of the property graph model of these databases, nodes (terms) in co-occurrence graphs can be enriched with additional attributes such as the names of the documents they occur in as well as the number of their occurrences in them, too. Also, the co-occurrence significances can be persistently saved as edge attributes. Graph databases are therefore an urgently necessary tool as a basis for future and scalable text mining solutions.

7 Conclusion

A new physics-inspired method has been introduced to determine centroid terms of particular text documents which are strongly related to them and yet do not need to occur in them. As text representatives, these terms are useful to determine the semantic distance and similarity of text documents. Especially, texts with similar topics yet different descriptive terms, may be classified more precisely than by commonly used measures. As the text length's influence does not play a role in doing so, even short texts or (search) queries may be matched with other texts using the same approach. It may therefore be applied in future (decentralised) search engines and text clustering solutions.

References

[1] Hawkins, J., Blakeslee, S.: *On Intelligence*, Times Books, New York, NY, USA, 2004

[2] Blei, D. M., Ng, A. Y., Jordan, M. I.: Latent Dirichlet Allocation, In: *The Journal of Machine Learning Research*, Vol. 3, pp. 993–1022, 2003

[3] Biemann, C.: Chinese Whispers: An Efficient Graph Clustering Algorithm and its Application to Natural Language Processing Problems, In: *Proceedings of the HLT-NAACL-06 Workshop on Textgraphs-06*, pp. 73–80, ACL, New York City, 2006

[4] MacQueen, J. B.: Some Methods for Classification and Analysis of Multivariate Observations, In: *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, pp. 281–297, University of California Press, 1967

[5] Miller, G. A.: WordNet: A Lexical Database for English, In: *Communications of the ACM*, Vol. 38, Issue 11, pp. 39–41, Nov. 1995

[6] Heyer, G., Quasthoff, U., Wittig, T.: *Text Mining - Wissensrohstoff Text*, W3L Verlag Bochum, 2006

[7] Budanitsky, A., Hirst, G.: Evaluating WordNet-based measures of semantic distance, In: *Computational Linguistics*, Vol. 32, Issue 1, pp. 13–47, 2006

[8] Resnik, P.: Using information content to evaluate semantic similarity, In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pp. 448–453, Montreal, Canada, 1995

[9] Baeza-Yates, R. A., Ribeiro-Neto, B.: *Modern Information Retrieval*, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999

[10] Dice, L. R.: Measures of the amount of ecologic association between species, In: *Ecology*, Vol. 26, No. 3, pp. 297–302, 1945

[11] Church, K. W., Hanks, P.: Word association norms, mutual information, and lexicography, In: *Computational Linguistics*, Vol. 16, Issue 1, pp. 22–29, Mar. 1990

[12] Quasthoff, U., Wolff, C.: The poisson collocations measure and its application, In: *Workshop on Computational Approaches to Collocations*, Wien, Austria, 2002

[13] Dunning, T.: Accurate methods for the statistics of surprise and coincidence, In: *Computational Linguistics*, Vol. 19, Issue 1, pp. 61–74, MIT Press, Cambridge, 1993

[14] Biemann, C., Bordag, S., Quasthoff, U.: Automatic acquisition of paradigmatic relations using iterated co-occurrences, In: *Proceedings of LREC2004*, Lisboa, Portugal, 2004

[15] Kubek, M., Witschel, H. F.: Searching the Web by Using the Knowledge in Local Text Documents, In: *Proceedings of Mallorca Workshop 2010 Autonomous Systems*, Shaker Verlag Aachen, 2010

[16] Witschel, H. F.: *Terminologie-Extraktion: Möglichkeiten der Kombination statistischer und musterbasierter Verfahren*, Ergon-Verlag, Würzburg, 2004

[17] Kubek, M., Unger, H.: Search Word Extraction Using Extended PageRank Calculations, In: *Autonomous Systems: Developments and Trends*, Studies in Computational Intelligence, Vol. 391, pp. 325–337, Springer Berlin Heidelberg, 2011

[18] Kubek, M.: DocAnalyser - Searching with Web Documents. In: *Autonomous Systems 2014, Fortschritt-Berichte VDI*, Vol. 10, Nr. 835, pp. 221–234, VDI-Verlag Düsseldorf, 2014

[19] Website of Neo4j, `https://neo4j.com/`, 2016, Last retrieved on 07/22/2016

[20] Efer, T.: Text Mining with Graph Databases: Traversal of Persisted Token-level Representations for Flexible On-demand Processing, In: *Autonomous Systems 2015, Fortschritt-Berichte VDI*, Vol. 10, Nr. 842, pp. 157–167, VDI-Verlag Düsseldorf, 2015