Centroid Terms as Text Representatives

Mario M. Kubek
Chair of Communication Networks
University of Hagen
Universitätsstr. 27, Hagen, Germany
mario.kubek@fernuni-hagen.de

Herwig Unger
Chair of Communication Networks
University of Hagen
Universitätsstr. 27, Hagen, Germany
herwig.unger@fernuni-hagen.de

ABSTRACT

Algorithms to topically cluster and classify texts rely on information about their semantic distances and similarities. Standard methods based on the bag-of-words model to determine this information return only rough estimations regarding the relatedness of texts. Moreover, they are per se unable to find generalising terms or abstractions describing the textual contents. A new method to determine centroid terms in texts and to evaluate their similarity using those representing terms will be introduced. In first experiments, its results and advantages will be discussed.

Keywords

text processing, centroid term, co-occurrence graph, document similarity

1. INTRODUCTION

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 order to be able to topically classify unseen content, the brain acts as a knowledge base. It tries to match the terms (words that carry a meaning) in such a document with previously learned terminology and can, in doing so, instantly and unconsciously perform at least a rough classification. At the same time, it gradually and constantly learns about new concepts. Also, it automatically abstracts from topical details and can associate a highly specific document with its more general topics (and its representing terms). As an example, given an article about steering wheels, the more general topics/terms 'car parts' or 'car' could be found.

In many text processing applications, the topical classification and grouping of texts are common tasks. For this purpose, it is necessary to measure the semantic distance or similarity of the documents to be clustered or classified.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

DocEng '16, September 12 - 16, 2016, Vienna, Austria

© 2016 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ISBN 978-1-4503-4438-8/16/09...$15.00

DOI: http://dx.doi.org/10.1145/2960811.2967150

The usual approach is to represent the documents by their term vectors -following the -bag-of-words model- which contain the texts' characterising terms and their score (typically, a TF-IDF-based statistic [2] is used) as a measure for their importance. The similarity of two term vectors can be determined e.g. using the cosine similarity measure or by calculating their overlap, e.g. using the Dice coefficient [3].

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 decriptive 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 [4] 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 text processing tasks, 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 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. 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, too.

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 [3], the mutual information measure [5], the Poisson collocation measure [6] and the log-likelihood ratio [7].

The use of the immediate neighbourhood of nodes in a co-occurrence graph has been widely considered in literature, e.g. to cluster terms [8] and to determine the global context (vector) of terms in order to evaluate their similarity [9] or to derive paradigmatic relations between them [10]. 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 on local and user-defined corpora [11]. The precision of web search results could be noticably improved when taking them 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' 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 1. 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 text processing: a whole text shall be represented or classified by one or a few single, descriptive terms which must be found.

To adapt the situation for text processing, 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 k words. Consequently, a distance $d(w_a, w_b)$ of two words in G can be defined by

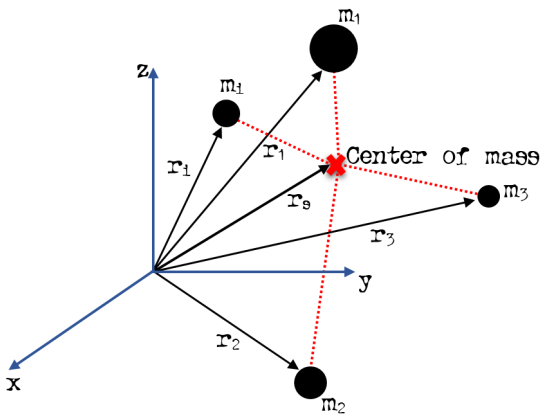

Figure 1: The physical center of mass

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

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

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

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). The centroid terms of documents can now be used to define the centroid-based distance ζ between any two documents D_1 and D_2. Therefore, let t_1 be the term with $d(D_1, t_1) = MIN$, then we call this t_1 the center term or *centroid term of* D_1. 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

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

D_1 and D_2. In order to obtain a similarity value instead,

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

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 processing: as an example, the so-called difference analysis [9] 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.

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

4. FIRST EXPERIMENTS

For all of the examplary 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 [3]. 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 25 English Wikipedia articles. The corpus used to create the reference co-occurrence graph G consisted of 100 randomly selected articles (including the mentioned 25 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.

Table 1: Centroids of 25 Wikipedia articles

Title of Wikipedia Article	Centroid Term
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
Universe	shape
Mass media	database
Stroke	blood
Mark Twain	tale
Ludwig van Beethoven	violin
Oxyrhynchus	papyrus
Fermi paradox	civilization
Milk	dairy
Health	fitness
Tourette syndrome	tic
Agriculture	crop
Malaria	disease
Fiberglass	fiber
Continent	continent
United States Congress	Senate
Turquoise	turquoise

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', '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 [12] algorithm which has been applied on their own separate (local) co-occurrence graphs (here, another term weighting scheme such as a TF-IDF variant [2] 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 other 99 articles has been generated and arranged in descending order according to their cosine similarity to A. An article's 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 examplary diagram in Figure 2 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 (blue 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 2, 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

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

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

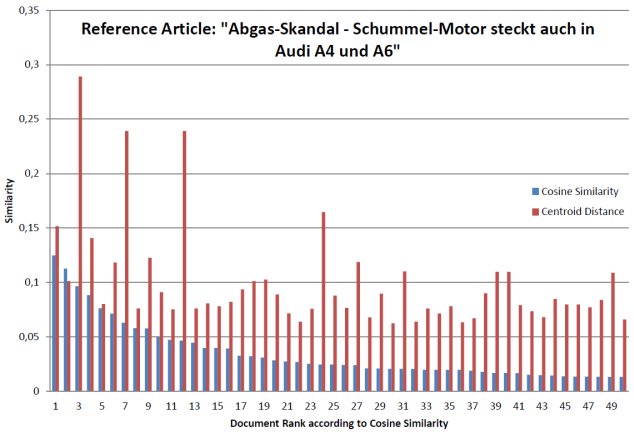

Figure 2: Cosine similarity vs. centroid distance (topic: car emissions scandal)

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 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.

5. DISCUSSION

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 never mentioned in it.

While the cosine similarity measure considers especially those documents as similar that actually use the same words (their term vectors have a significantly large overlap), the centroid diatance measure can uncover a topical relationship between documents even if their wording differs. Due to their completely different working principles, it might be sensible to combine both approaches in a new measure that factors in the results of both methods. Further 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. 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 with commonly known 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.

7. REFERENCES

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

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

[3] L. R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, July 1945.

[4] G. A. Miller. WordNet: A Lexical Database for English. *Commun. ACM*, 38(11):39–41, Nov. 1995.

[5] K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, Mar. 1990.

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

[7] T. Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, Mar. 1993.

[8] C. Biemann. Chinese whispers: An efficient graph clustering algorithm and its application to natural language processing problems. In *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*, TextGraphs-1, pages 73–80, Association for Computational Linguistics, Stroudsburg, PA, USA, 2006.

[9] G. Heyer, U. Quasthoff, and T. Wittig. *Text Mining: Wissensrohstoff Text – Konzepte, Algorithmen, Ergebnisse*. IT lernen. W3L-Verlag, Herdecke, Germany, 2008.

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

[11] M. Kubek and H. F. Witschel. Searching the web by using the knowledge in local text documents. In *Proceedings of Mallorca Workshop 2010 Autonomous Systems*, pages 75–79. Shaker Verlag, Aachen, Germany, 2010.

[12] M. Kubek and H. Unger. Search word extraction using extended pagerank calculations. In *Autonomous Systems: Developments and Trends*, pages 325–337. Springer Berlin Heidelberg, 2012.