Interactive, Topic-oriented Search Support by a Centroid-based Text Categorisation

Mario Kubek, Herwig Unger

Abstract—Centroid terms are single words that semantically and topically characterise text documents and so may serve as their very compact representation in automatic text processing. In the present paper, centroids are used to measure the relevance of text documents with respect to a given search query. Thus, a new graph-based paradigm for searching texts in large corpora is proposed and evaluated against keyword-based methods. The first, promising experimental results demonstrate the usefulness of the centroid-based search procedure. It is shown that especially the routing of search queries in interactive and decentralised search systems can be greatly improved by applying this approach. A detailed discussion on further fields of its application completes this contribution.

Keywords—Search algorithm, centroid, query, keyword, co-occurrence, categorisation.

I. Motivation

GOOGLE, Bing and other well-known centralised search engines are still the ultimate tools to locate any wanted content in the Internet. With a plenty of (seemingly fitting) results usually returned within an extremely short time, most users do not doubt that this technology works well and is a seminal approach to manage their information needs. The search works even so well, that researchers observe a 'Google effect' [1], i.e. that it becomes easier to search for some information than to keep it in mind.

However, it becomes hard to evaluate the real performance and service quality of those giant information collectors. At first, recall (i.e. the fraction of the relevant documents that are successfully retrieved) and precision (i.e. the fraction of retrieved documents that are relevant to the query) have been defined for this purpose [2]. However, both parameters are only to be determined if the whole search space and an entire result list are known: in most cases this is simply impossible to realise today for a common user. Differing from recall, precision is a quite subjective parameter and since mostly the first 10 to 30 query results are considered, it has been refined to 'precision@10, 20, 30' or included in other parameters like the 'F1 score' with a questionable success [3].

Mostly, the decision on whether documents are considered relevant or non-relevant directly depends on the membership of the keywords of a query in the considered document sets; only in some cases, synonyms, cross-language translations and alternative writings are considered. However, the biggest obstacle for an objective evaluation is the problem of coverage: mostly, users assume that any available document in the WWW (World Wide Web) is considered to generate the result set presented. However, while according to estimates, the size of the WWW amounts to over 100 billion pages [4], search engines have indexed less than 10 percent of it and even have very different document sets indexed [5].

Dr.-Ing. Mario Kubek and Prof. Dr.-Ing. habil. Herwig Unger are with the Department of Communication Networks at the FernUniversität in Hagen, 58084 Hagen, Germany,(e-mail: kn.wissenschaftler@fernuni-hagen.de, phone +49 2331 9871155).

Some authors and publications like [6] argue from the point of view of databases that any information shall be kept only once and any copy (like the search engine's repository) will cause consistency problems and may result in an tremendous effort to solve them. In this paper, the out-of-date keyword-based description for the information wanted shall be decried and a way out will be given.

Any keyword-based description of content works well, as long as the user can precisely describe her or his information need or has memorised a set of promising, constructive terms from former search sessions. Anyhow, the use of frequently used queries or personal topics of interest may catch users easily in so-called 'filter bubbles' [7], which hinder them to find new items as documents of well-known topics will be presented primarily. Keyword-based searches usually fail, if:

- several authors use a different wording for one and the same topic,
- exact keywords (e.g. for market innovations and trends) are not yet known such that these items must be referred to with quite general or even inappropriate terms and
- the keywords shall describe a topical area (but do not need to be contained in the documents to be found).

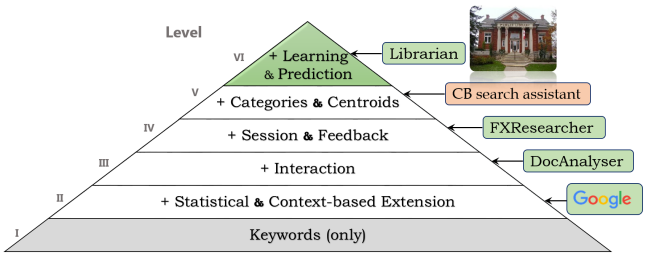

Fig. 1. Different levels of search

By recalling the slow but efficient service in an old-fashioned library and combining it with the advantages of the centroid-based (CB) text analysis, a way to significant improvements in today's Internet (Fig. 1) search may be paved.

II. Principles of Centroid-based Search Procedures

A. Fundamentals

The definition of centroids in [9] is based on co-occurrence graphs [8], which have been transformed into distance graphs (by interpreting the reciprocal value of the significance as distance in a new graph which is isomorphic to the co-occurrence graph) and –therefore– a metric space. Using the physical analogue of the *centre of mass*, *text centroids* $\chi_G(D)$ have been defined to be the node (term) in the reference co-occurrence graph G with the minimal average distance to all words of the respectively analysed document D. Such a centroid can be defined for (longer) texts in the same manner as for short or even one-word search queries. In already published articles, it was shown that

1) centroids represent the topics of a document [9] well,
2) can be used to determine the similarity of texts and/or queries [10],
3) represent formal calculable categories for a group of texts (which do not necessarily have to represent semantic classes in human thinking) [11],
4) are therefore directly useable to build hierarchical, tree-like document structures [12] and finally that
5) the respective structures can be established in a fully decentralised (P2P) system by random walkers [13].

These results may now be directly used to improve the (decentralised) search in the WWW, if the respective, hierarchical structures have been built and are maintained as shown in [13].

B. Diversity and Speciality

As a first step, a quantitative evaluation of texts or queries by their **diversity** and **speciality** may be defined. Therefore, let us consider any two words w_i, w_j of a query or text Q, furthermore let $d(w_i, w_j)$ be their distance in the distance graph obtained from the respective reference co-occurrence graph G and let us postulate that all words $w_i, w_j \in Q$ are also nodes of that co-occurrence graph. Then, we can define that

- the supremum

$$\mu = sup(d(w_i, w_j)|(w_i, w_j) \in (Q \times Q)) \tag{1}$$

  obtained from all pairs of words w_i, w_j in a text/query Q is called its **diversity** μ.
  The lower this value is, the more a query targets a designated, focussed topic area, while high values of the diversity indicate a more general, common request.
- the term

$$\sigma = \frac{1}{1 + sup(d(\chi_G(Q), w_i)|_{w_i \in Q})} \tag{2}$$

  obtained from the distance of the centroid $\chi_G(Q)$ to the query Q with all contained words w_i is called its **speciality** σ.
  Of course, diversity and speciality are related to each other, whereby the former mostly expresses the most extreme (semantic) differences in a text/query.

An interpretation of these measures and in particular the definition of the diversity may directly result in an easy and practicable evaluation of search queries as realised within a search tool shown in Fig. 2.

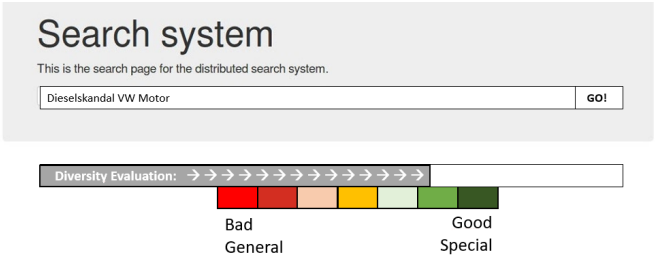

Fig. 2. Evaluation of search queries by their diversity

High values for the diversity (possibly normalised using the average distance in the distance graph) stand for unspecialised or comprehensive requests with a huge number of possible, mostly non-relevant results. Small diversity values represent highly specialised requests, for which usually fewer, very well matching search results may be offered.

C. Search in Hierarchical Structures

In [13], it has been shown, how hierarchical tree-like structures can be built by random walkers, which

- are managed by a peer within a P2P-system that is created as an extension of a standard Apache Tomcat server,
- include and represent every single web document,
- react automatically to faults as well as to the addition or removal of documents and
- carry out the query and document classification strictly using dichotomies on the basis of the centroid method.

The obtained, distributed tree-like structures (the grey-shaded parts in Fig. 3) can now be used for routing queries arriving at any nodes. Starting from the root, search queries need to be forwarded to the son node whose centroid has the minimal distance to the centroid of the query or whose area of topical responsibility contains the respective query centroid.

As long as the speciality of the query is high (Fig. 3), the decision on where to forward it to is unique and this process can be carried out easily. Problems occur in case of low speciality values of the search query as it may be positively answered by documents offered by the leaves of different subtrees, i.e. the routing is not uniquely defined.

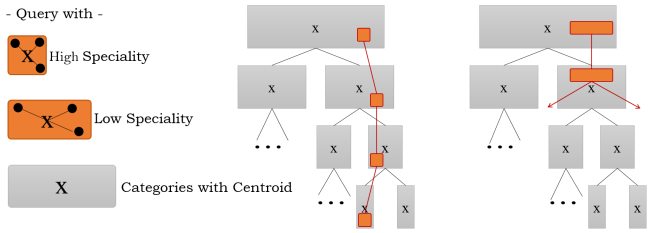

Fig. 3. Routing of search queries depending on their speciality

In this case, several strategies can be applied:

1) the most relevant alternative while exclusively depending on the distance or membership of the centroid term is chosen, what may, however, reduce recall,
2) the query is duplicated and forwarded to both son nodes of the recent position, what may significantly increase traffic volumes especially in case of multiple application.
3) an interaction (i.e. a new quality of search, see Fig. 1) is initiated, whereby
   - additional keywords are presented to the user,
   - context- or session-related information is used to suggest further keywords,
   - centroids and further keywords of the son nodes in the hierarchical structure are suggested as additional criteria and keywords to the user and
   - neighbours of centroids or second and third centroid candidates of the request or document are used to solve ambiguity problems

   in order to finally make a more precise decision on where the request must be routed to.

Experimental results of the suggested, holistic methodology will be presented in the following section for the first time.

III. Experiments and Discussion

The following experiments have been conducted in order to show that centroid terms can successfully be applied in search processes. The two datasets "Corpus-100" and "Corpus-1000" used to construct the co-occurrence graphs for the determination of centroid terms consist of either 100 or 1000 topically classified (topical tags assigned by their authors) online news articles from the German newspapers "Süddeutsche Zeitung" and "Die Welt". Each dataset consists of 25 respectively 250 randomly chosen articles from the four topical categories 'car' (German tags: 'Auto', 'Motor'), 'money' (German tags: 'Geld', 'Finanzen'), 'politics' (German tag: 'Politik') and 'sports' (German tag: 'Sport'). Interested readers may download the used datasets and analysis results (10.1 MB) from [14].

In order to build the (undirected) respective co-occurrence graphs, linguistic preprocessing has been applied on these documents whereby sentences have been extracted, stop words have been removed and only nouns (in their base form), proper nouns and names have been considered. Based on these preparatory works, co-occurrences on sentence level have been extracted. Their significance values have been determined using the Dice coefficient [15] and its reciprocal has been used to represent the distance between the terms involved. These values and the extracted terms are persistently saved in an embedded Neo4j [16] graph database using its property-value store provided for all nodes (represent the terms) and relationships (represent the co-occurrences and their significances).

The analysed articles stem from the same datasets and their representing centroid terms have been determined using these co-occurrence graphs and persistently stored afterwards. Furthermore, each article's nouns, proper nouns and names have been stored in an inverted index in order to be able to answer Boolean queries as well. All results are provided along with the mentioned datasets.

A. Exp. 1a: Centroid-based Search

In the first set of experiments presented here, the goal is to show that for automatically generated and topically homogeneous queries consisting of three terms with a low diversity (neighbouring terms in the co-occurrence graph of dataset "Corpus-1000") from the four mentioned topics, the first 10 (Top-10) returned documents are actually relevant. In order to test the applicability of centroid terms in this process, the centroids of these queries have been determined along with their distances to the centroid terms of the articles in the co-occurrence graph of dataset "Corpus-1000". Then, for each query, the articles have been listed in ascending order by this distance. Furthermore, these queries have been sent to Google as well. It has been made sure that the generated queries make sense, i.e. that humans would actually formulate them.

While it is not directly possibly to compare the results of the much smaller dataset "Corpus-1000" with the huge result set of Google, a general evaluation of the retrieval quality –especially in an interactive setting– is possible. Therefore and because of the shortcomings of the well-known evaluation metrics recall and precision, for each query, the set of the Top-10 results of both approaches (centroid-based approach and Google) have been assigned marks by five test persons while applying the German school grading scale. The marks 1 (very good), 2 (good), 3 (satisfactory), 4 (sufficient) and 5 (insufficient) were possible.

The evaluation results of 20 randomly generated queries (five of each topic) are presented in Table I. In it, the average and rounded marks of all five test persons are presented for each query. As it is possible to see in these results, generally, the Top-10 results obtained from the centroid-based approach are mostly relevant. This also implies that mostly useful documents from the correct topical orientation have been returned first from the set of all possible 1000 articles. Although, Google's results were generally more relevant, its much larger knowledge base is certainly one reason for this outcome. The overall, average mark for the centroid-based approach was 2,0 (good) and 1,4 for Google (very good). However, a closer look at the results reveals that Google was not able to interpret all queries correctly. As an example, for the car-related query 5, Google's results were largely related to heating technology and industrial engines and only two relevant results had a relation to car emissions.

Another observation is that the sports-related terms seem to have a higher discriminative power as the centroid-based approach returned in almost all cases relevant results for this topic. The other topics contain more terms that can be found in many documents of different topical orientations. Especially money-related terms can be found in the documents of all topics. For instance, in the car-related documents that mostly deal with the car emissions scandal in 2015, financial penalties are often discussed. Therefore, these topics' queries and their respective centroids have a lower discriminative power leading to a slightly higher number of irrelevant search results.

TABLE I
EVALUATION OF CENTROID-BASED (CB) AND GOOGLE'S SEARCH RESULTS

Query No.	Query (topic: car)	Centroid	Mark CB	Mark Google
1	Bauzeit, VW-Konzern, Millionenseller	Bauzeit	3	2
2	Luftreinhalteplan, Software-Nachrüstung, Stuttgart	Luftreinhalteplan	2	1
3	Roller, Hersteller, Gefühl	Roller	3	2
4	Abgasrückführung, Grenze, Spritverbrauch	Abgasrückführung	2	1
5	Ruß, Abgasreinigung, Rohrleitung	Ruß	2	4
Query No.	**Query (topic: money)**	**Centroid**	**Mark CB**	**Mark Google**
6	Loft, Monatsmiete, Quadratmeter	Loft	2	1
7	mTan, Auftrag, Verfahren	mTan	3	1
8	Mischfond, IMC, Investmentfond	Mischfond	2	1
9	Stoppkurs, Prozent, Zertifikat	Aktie	2	1
10	Verbraucher, Sicherheit, Sparer	Sparer	1	1
Query No.	**Query (topic: politics)**	**Centroid**	**Mark CB**	**Mark Google**
11	Flüchtlingsheim, Gesundheit, Albaner	Flüchtlingsheim	2	1
12	Asylpolitik, Region, Bundesjustizminister	Asylpolitik	2	2
13	Migration, Monat, Untersuchung	Migration	3	2
14	BAMF, BA, Mitarbeiter	BAMF	3	2
15	Flüchtlingslager, Damaskus, Europaparlament	Flüchtlingslager	1	1
Query No.	**Query (topic: sports)**	**Centroid**	**Mark CB**	**Mark Google**
16	Stadion, Fan, Lust	Stadion	2	1
17	Halbzeit, Spiel, Abpfiff	Halbzeit	1	1
18	Schiedsrichter, DFB, Hinweis	DFB	2	1
19	Spieltag, Weise, Bundesliga-Live	Spieltag	1	1
20	Stadion, Verhaltenshinweise, Mannschaft	Stadion	2	1
		Average mark:	**2,0**	**1,4**

TABLE II
RESULTS OF BOOLEAN RETRIEVAL AND PARTIAL MATCHING

Query No.	Exact matches	Partial matches
1	0	15
2	1	59
3	1	167
4	1	159
5	1	9
6	1	43
7	1	86
8	1	1
9	1	316
10	1	91
11	0	45
12	0	75
13	0	276
14	0	72
15	0	30
16	0	116
17	0	47
18	1	69
19	0	93
20	0	145

B. Exp. 1b: Centroid-based Search

Also, Boolean retrieval and partial matching has been applied on these 20 queries while using the inverted index of the dataset "Corpus-1000". These queries have been interpreted as conjunctive ones, which means that the goal was to find documents that contain all query terms. The results obtained from these tests are presented in Table II which shows the number of exact matches and the number of partial matches for each query. In only 10 cases, a document matching all terms could be found. As another noteworthy example, for query 8, only one exactly matching article and only one partially matching article could be found although 250 money-related (among them many relevant stock market news) articles were available. Also, queries 1 and 5 returned only a minor fraction of the relevant documents to be found. However, in most of the cases, partial matching returned at least an acceptable number of documents. However, the general finding for these retrieval approaches is that they are unable to identify a large quantity of relevant results. A low recall is to expected when applying them. Even so, the documents they return are mostly relevant as the query terms actually appear in them (at least partially).

C. Exp. 2: Classification of Query Topics

The second set of experiments investigate whether the queries' topical orientation can be correctly identified by

the centroid-based approach. In order to evaluate this automatic classification, for each query, the number of the Top-10 returned documents that match its topical orientation are counted and the average number as well as the median of topical matches have been determined based on these numbers afterwards. At first, this measurement has been carried out for the 20 low-diversity queries listed above. Table III presents the results of these measurements. As it is to be seen, for all topics, the topical matches outweigh the mismatches.

In accordance with the previous experiment, the sports-related queries generally returned more topically matching articles than the queries from the other topics. Generally, it can be deducted that the centroid-based search approach is for the most part able to correctly identify a query's topical orientation, especially when it is a low-diversity one.

TABLE III
DETECTION OF TOPICS OF LOW-DIVERSITY QUERIES

Topic	Average no. of topical matches	Median of topical matches
car	6,4	6
money	7,4	8
politics	6,6	6
sports	8,8	8

In order to evaluate the influence of the query diversity, the same experiment has been carried out for topically oriented queries with unspecified diversity as well. That means that the queries's terms –in contract to the previous experiments– do not necessarily have to be direct neighbours in the used reference co-occurrence graph generated from the dataset "Corpus-1000". Instead, the queries' terms have been automatically and randomly selected from a particular topic of interest. As a consequence, those queries might not always make sense for humans, but a mixture of low- and high-diversity queries has been obtained this way. In this experiment, for each of the four mentioned topics, 25 queries (altogether 200 queries) have been automatically generated.

TABLE IV
DETECTION OF TOPICS OF QUERIES WITH UNSPECIFIED DIVERSITY

Topic	Average no. of topical matches	Median of topical matches
car	4,3	4
money	4,26	5
politics	6,06	7
sports	6,42	8

As Table IV shows, the average number and the median of topical matches expectedly decreased. The reason for this result is that due to their generally higher diversity, the queries' centroid terms are more distant to the query terms as well and therefore often can not be clearly assigned to one or the other topic. Usually, those centroid terms are general ones (e.g. the German word 'Jahr', Engl. year) that have many connections to terms of all other topics in the co-occurrence graph. If such kind of terms are determined as centroid terms, the Top-10 documents returned are mostly from diverse topics as well. However, even in this experiment, the sports-related queries returned the best results. Their median of topical matches even did not decrease at all.

D. Exp. 3: Centroid Candidates

While the previous results indicate that centroid terms can in fact be useful in search or classification processes, there is still room for improvement. As an example, the finally chosen centroid term C which has by definition the shortest average distance to the terms of a query –these terms might stem from a user query or a text document alike– in the used co-occurrence graph, might not be the best selection to represent it. Instead, the usage of the second or third best centroid candidates might be an even better choice. The following experiments investigate this hypothesis.

Furthermore, the direct neighbours of C on the shortest paths to these candidates might be better descriptors than the sole centroid term C. Last but not least, the experiments' aim is to show that by using larger (and specially analysed) corpora to create the co-occurrence graph, the query interpretation is improved, too. For this purpose, the previously introduced datasets "Corpus-100" and "Corpus-1000" have been used. In order to obtain even more term relationships that the standard sentence-level co-occurrence analysis yields, second-order co-occurrences [17] have been extracted from the dataset "Corpus-1000" as well. Using this technique, relations between terms can be extracted even when they do not co-occur in sentences originally. Differing from the previous experiments, the queries consist of the 25 most frequent terms of the articles to be analysed from which the respective centroid term C and the next best centroid candidates CC_i with $i \geq 2$ are computed.

The following Tables V, VI, VII and VIII present for four randomly chosen articles from the German newspapers "Süddeutsche Zeitung" and "Die Welt" their respectively obtained analysis results. In the tables, the co-occurrence graph numbers refer to the datasets used (1 means "Corpus-100", 2 means "Corpus-1000" and 3 means "Corpus-1000" with second-order co-occurrence extraction). In the following elaborations, the second best centroid candidate is referred to CC_2. The next three centroid candidates are referred to in an analogous manner. The direct neighbours of centroid term C on the shortest path to CC_i are respectively referred to as (neighbouring terms) NT_i with $i \geq 2$.

The results consistently show that by using the larger co-occurrence graphs 2 and 3 the finally obtained centroid terms C better reflect the individual documents' topics. As an example, in Table V, a car-related document has been analysed. When using the small dataset "Corpus-100", the centroid term 'Modell' (Engl. model) is chosen, which is a rather general term. However, for the co-occurrence graphs 2 and 3, the centroid term 'VW' is returned. This term better reflects the article's content as it deals with the 2015 car emissions scandal which mainly revolves around the German car brand Volkswagen (VW). Likewise, these results confirm that the method to calculate centroid terms is inherently corpus-dependent.

Furthermore, in all test cases, the next best centroid candidates CC_i provide more useful insight into the document's topical focus and have for the largest part a more discriminative power than the actually chosen centroid term. For instance, while in case of the money-related article (Table VI), the actual

TABLE V
CENTROID CANDIDATES OF THE CAR-RELATED DOCUMENT AUTO_VW-ABGAS-SKANDAL-MILLIONEN-AUDIS-VON-ABGASAFFAERE-BETROFFEN.TXT

Basic Parameters		Centroid Candidates CC_i and Direct Neighbours NT_i of C on the shortest path to CC_i							
Co-occ. graph	Centroid C	CC_2	NT_2	CC_3	NT_3	CC_4	NT_4	CC_5	NT_5
1	Modell	Fahrzeug	Hersteller	VW	Passat	Million	Hersteller	Konzern	TDI
2	VW	Volkswagen	Marke	Dieselmotor	Dieselmotor	Fahrzeug	Fahrzeug	USA	USA
3	VW	Volkswagen	Volkswagen	Software	Software	Fahrzeug	Fahrzeug	Modell	Modell

TABLE VI
CENTROID CANDIDATES OF THE MONEY-RELATED DOCUMENT GELD_GELDWERKSTATT-VERLUST-VERHINDERN.TXT

Basic Parameters		Centroid Candidates CC_i and Direct Neighbours NT_i of C on the shortest path to CC_i							
Co-occ. graph	Centroid C	CC_2	NT_2	CC_3	NT_3	CC_4	NT_4	CC_5	NT_5
1	Prozent	Monat	Euro	Frage	Laufzeit	Euro	Euro	Million	Euro
2	Prozent	Euro	Euro	Jahr	Jahr	Bank	Bank	Geld	Geld
3	Aktie	Prozent	Prozent	Euro	Milliarde	Jahr	Milliarde	Milliarde	Milliarde

TABLE VII
CENTROID CANDIDATES OF THE POLITICS-RELATED DOCUMENT
POLITIK_FLUECHTLINGE-IN-EUROPA-SO-WURDE-BUDAPEST-WIEN-MUENCHEN-ZUR-HAUPTROUTE-FUER-FLUECHTLINGE.TXT

Basic Parameters		Centroid Candidates CC_i and Direct Neighbours NT_i of C on the shortest path to CC_i							
Co-occ. graph	Centroid C	CC_2	NT_2	CC_3	NT_3	CC_4	NT_4	CC_5	NT_5
1	Ungarn	Deutschland	Deutschland	Grenze	Grenze	Flüchtling	Flüchtling	Million	Flüchtling
2	Flüchtling	Mensch	Mensch	Deutschland	Deutschland	Land	Land	Euro	Million
3	Flüchtling	Ungarn	Ungarn	Grenze	Grenze	Mensch	Mensch	Deutschland	Deutschland

TABLE VIII
CENTROID CANDIDATES OF THE SPORTS-RELATED DOCUMENT SPORT_BUNDESLIGA-THOMAS-TUCHELS-ERFOLGREICHE-RUECKKEHR.TXT

Basic Parameters		Centroid Candidates CC_i and Direct Neighbours NT_i of C on the shortest path to CC_i							
Co-occ. graph	Centroid C	CC_2	NT_2	CC_3	NT_3	CC_4	NT_4	CC_5	NT_5
1	Platz	Europa	Dreier	Sonntag	Sonntag	Bayern	Punkt	Zuschauer	Halbzeit
2	Bayern	Trainer	Dortmund	Punkt	Punkt	Mannschaft	Mannschaft	Partie	Punkt
3	Dortmund	League	League	Punkt	Punkt	Bayern	Bayern	Spiel	Spiel

centroid term 'Prozent' (Engl. percent) obtained when using the co-occurrence graphs 1 and 2 has a rather low discriminative power, the centroid candidates 'Euro', 'Bank' (Engl. bank) and 'Geld' (Engl. money) better reflect the article's topical orientation. Therefore, these terms should not be neglected in classifying queries as well. In this example, the comparatively best centroid term (and candidate) is 'Aktie' (Engl. stock) which is returned when using the co-occurrence graph 3 as the article mainly deals with stock market investments.

Similarly, the direct neighbours NT_i of the centroid term C on the shortest path to the centroid candidates CC_i often make the analysed document's topic more explicit than the centroid term itself. In case of the politics-related article (it deals with the European migrant crisis from 2015) used in Table VII, topically fitting neighbouring terms of C are for instance Flüchtling (Engl. refugee) or Grenze (Engl. border) which describe the article's focus very well.

Also, it is noteworthy that in many cases (and especially when applying the much denser co-occurrence graph 3) the terms CC_i and NT_i are the same. This observation confirms that the nodes of the next best centroid candidates CC_i are in fact very close or even adjacent to the node of the centroid term C. Usually, only one hop from it is needed to reach them. This even suggests that it is possible to approximate further centroid candidates by simply visiting C's neighbouring nodes (preferably the ones that are closest to the analysed document's terms). By using this approach, there would be no need for computing them explicitly.

E. Discussion and Fields of Application

The previous experiments showed that centroid terms can be of benefit in both search and classification processes. While naturally centroid-based document retrieval cannot be used for finding exactly matching documents, more potential relevant documents can be found by this approach as the vocabulary mismatch problem is inherently solved. As an example, a search system applying it could return documents on cars, automobiles and vehicles when the query's centroid is 'car'. Due to its working principle, this solution is particularly useful in the fields of patent search and classification where besides exact document matches rough ones are of interest, too. A search system like DocAnalyser [18] whose task it is to find more similar and related web documents could benefit from this approach as well.

However, as indicated in section 2, the most benefit will be gained when applying it in decentralised search systems, particularly when it comes to forwarding incoming queries of different diversity to peers with matching contents. The peers can make a routing decision based on the distance of those queries' centroids to the centroids of other peers (their contents is analysed for this purpose) while referring to the own (dynamically growing) reference co-occurrence graph

which acts as a local knowledge base. Based on the results of the third set of experiments, this routing decision can be positively influenced by taking into account the neighbouring terms of the chosen centroid term in this graph as well the next best centroid candidates. These terms can be regarded as topical pointers as they provide a more precise semantic context and interpretation for the finally chosen centroid term, which might by itself not be the best representative for a query or document. By this means, an implicit word-sense disambiguation of the centroid terms is provided.

Furthermore, the peers can even use the technique of centroid-based document retrieval to answer incoming queries by themselves as the first set of experiments showed. However, for this purpose, other retrieval techniques based on the vector space model or probabilistic approaches might return results with a higher precision. Also, the application of highly performant and space-efficient Bloom filters [19] is a viable option at this search step. In order to gain benefit from all these retrieval approaches, their combined use might be sensible. In further experiments, this hypothesis will be investigated.

Last but not least, the herein presented as well as previously conducted experiments using English and German corpora have shown that the centroid-based text analysis and comparison is language-independent. A search system applying this approach only needs to select the correct language-specific co-occurrence graph before the actual, language-independent query or document analysis is carried out.

IV. Conclusion

This contribution investigated and evaluated several uses of text-representing centroid terms in search processes. It was found that the new method of centroid-based document retrieval is able to return both topically matching and relevant results. Also, an approach has been discussed to determine the correct word sense of a potentially ambigous centroid term by enriching it with more specific terms. In addition to these solutions, two new graph-based measures to quantitatively evaluate the diversity and speciality of queries in interactive search sessions have been presented. These values immediately indicate whether a user should reformulate a query. Also, they are of value in subsequent processing steps such as the routing of queries in decentralised search systems. Future research will investigate if the technique of centroid-based document retrieval can be successfully used in conjunction with traditional information retrieval methods in order to get benefit from all their approaches.

References

[1] B. Sparrow, J. Liu and D. M. Wegner, *Google effects on memory: Cognitive consequences of having information at our fingertips*, In Science, Vol. 333, pp. 776–778, 2011.

[2] C. Cleverdon, *The Cranfield Tests on Index Language Devices*, In Readings in Information Retrieval, pp. 47–59, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997.

[3] C. D. Manning, P. Raghavan and H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press, New York, NY, USA, 2008.

[4] J. B. Miller, *Internet Technologies and Information Services, 2nd Edition*, Libraries Unlimited, Santa Barbara, California, USA, 2014.

[5] A. van den Bosch, T. Bogers and M. de Kunder, *Estimating search engine index size variability: a 9-year longitudinal study*, In Scientometrics, Volume 107, Issue 2, pp. 839-856, 2016.

[6] M. Kleppmann, *Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems*, O'Reilly Media, 2017.

[7] E. Pariser, *The Filter Bubble: What the Internet Is Hiding from You*, Penguin Group, 2011.

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

[9] M. M. Kubek and H. Unger, *Centroid Terms as Text Representatives*, In Proceedings of the 2016 ACM Symposium on Document Engineering, DocEng '16, pp. 99–102, ACM, New York, NY, USA, 2016.

[10] M. M. Kubek and H. Unger, *Centroid Terms and their Use in Natural Language Processing*, In Autonomous Systems 2016, Fortschritt-Berichte VDI, Reihe 10 Nr. 848, pp. 167–185, VDI-Verlag Düsseldorf, 2016.

[11] M. Kubek, T. Böhme, and H. Unger, *Empiric Experiments with Text Representing Centroids*, In Lecture Notes on Information Theory, Vol. 5, No. 1, pp. 23–28, 2017.

[12] M. M. Kubek and H. Unger, *Towards a Librarian of the Web*, In Proceedings of the 2nd International Conference on Communication and Information Processing (ICCIP 2016), pp. 70–78, ACM, New York, NY, USA, 2016.

[13] M. M. Kubek and H. Unger, *A Concept Supporting Resilient, Fault-tolerant and Decentralised Search*, In Autonomous Systems 2017, Fortschritt-Berichte VDI, Reihe 10 Nr. 857, pp. 20–31, VDI-Verlag Düsseldorf, 2017.

[14] M. M. Kubek and H. Unger, *Datasets and Analysis Results*, http://www.docanalyser.de/search-corpora.zip, 2017.

[15] L. R. Dice, *Measures of the Amount of Ecologic Association Between Species*, In Ecology, Vol. 26, No. 3, pp. 297–302, 1945.

[16] Neo4j, Inc., *Website of the Neo4j Graph Platform*, https://neo4j.com, 2017.

[17] C. Biemann, S. Bordag and U. Quasthoff, *Automatic Acquisition of Paradigmatic Relations using Iterated Co-occurrences*, In Proceedings of LREC2004, pp. 967–970, Lisboa, Portugal, 2004.

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

[19] B. H. Bloom, *Space/Time Trade-offs in Hash Coding with Allowable Errors*, In Commun. ACM, Vol. 13, No. 7, pp. 422–426, ACM, New York, NY, USA, 1970.

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.

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.