What to read next? Challenges and Preliminary Results in Selecting Representative Documents*

Tilman Beck[1], Falk Böschen[1], and Ansgar Scherp[2]

[1] Kiel University, Department of Computer Science, 24118 Kiel, Germany
{stu127568,fboe}@informatik.uni-kiel.de
[2] University of Stirling, Computing Science and Mathematics,
Stirling FK9 4LA Scotland, UK
ansgar.scherp@stir.ac.uk

Abstract. The vast amount of scientific literature poses a challenge when one is trying to understand a previously unknown topic. Selecting a representative subset of documents that covers most of the desired content can solve this challenge by presenting the user a small subset of documents. We build on existing research on representative subset extraction and apply it in an information retrieval setting. Our document selection process consists of three steps: computation of the document representations, clustering, and selection of documents. We implement and compare two different document representations, two different clustering algorithms, and three different selection methods using a coverage and a redundancy metric. We execute our 36 experiments on two datasets, with 10 sample queries each, from different domains. The results show that there is no clear favorite and that we need to ask the question whether coverage and redundancy are sufficient for evaluating representative subsets.

Keywords: Representative Document Selection · Document Clustering.

1 Introduction

As an early-stage researcher or practitioner, delving into a new, previously unknown topic usually starts with issuing broad queries to a search engine. The aim of such an introductory search is to get an overview of the different subtopics, most fundamental works in the field, and the state-of-the-art. Search engines return those documents which best match with the user query and present the results as a ranked list. Such a relevance-based ranking works well for standard information retrieval (IR) tasks, but it is not suited for finding a complete, representative, and comprehensive selection when exploring a new field. Furthermore, the large number of search results makes it very time-consuming for the user to identify the desired documents because similarly ranked documents have usually similar content. Thus, result lists are often highly redundant, especially the top

* The final authenticated version is available online at https://doi.org/10.1007/978-3-319-99133-7_19.

results (i. e., the first page) that are usually only considered by the user. It is also difficult to estimate at which position of the ranked list one has read about all aspects of a topic [3], i. e., got a representative overview, because the breakpoint differs from one topic to another.

A representative subset can address these challenges. In general, a representative subset is considered as a selection which covers most of the content, contains the least possible amount of redundant information, and is notably smaller in size than the original dataset. Zhang et al. [21] proposed such a method for text documents that uses clustering and a coverage and redundancy based selection to create a representative subset from a set of documents. We have re-implemented the work of Zhang et al. and evaluated it on search results created from queries issued to two datasets of different domains. Based on the results of preliminary experiments, we decided to investigate the following research questions:

- RQ1: What influence does the choice of a) document representation, b) clustering algorithm, and c) selection method have on the coverage and redundancy scores of the representative subset?
- RQ2: Are the evaluation metrics, coverage and redundancy, sufficient to evaluate the representativeness of a document set?

The outline of this paper is: In Section 2, we briefly discuss related work followed by an introduction of our approach in Section 3. We describe our datasets, metrics, and experiment setup in Section 4. Finally, we present our results in Section 5 and discuss them before we conclude.

2 Related Work

Zhang et al. [21] propose a selection technique which uses an unsupervised text mining approach to find a representative set of documents from a large corpus. They first cluster the documents using X-Means, an adaption of K-Means which can be used without prior specification of the cluster number k, to identify the different topics in the dataset. From each cluster, they extract the documents which maximize the content coverage and introduce as less redundant content as possible. The size of the result set is determined by the proportional sizes of the clusters. Their framework outperforms a greedy approach, which directly optimizes for coverage, and a top-n method, which selects the best n documents with regards to coverage. The authors conducted a user study with 20 participants that indicated a preference for their selection approach.

The task of generating reading lists [7] is quite similar to selecting representative documents. However, it aims more to propose subjective and expert-based reading lists rather than an objective, representative selection. Jardine and Teufel [8] proposed an adaptation to the PageRank score, called Themed-PageRank (TPR), which has an LDA-inferred topic dimension and a so-called age-tapering component to incorporate the time aspect. They compute the TPR for each document that is returned by an IR system given a specific query. They rank the documents based on TPR and return the top-20 documents as reading list. Their evaluation using expert-created reading lists showed improvements on previous state-of-the-art models based on PageRank. A recent approach by Zhang et al. [20] generates book reading lists for certain topics. They use data obtained from social media to train vector representations for each topic-book pair using content from social media for relevance, quality, timeliness, and diversity.

We base our experiments on the work by Zhang et al. [21]. However, they compared their selection to methods which optimize for coverage only, but not for redundancy, which limits the comparability. Furthermore, only X-Means was considered for clustering and no information is given about the actual number of documents retrieved from the datasets, which is an important factor when searching for a small representative document set. Thus, we extend in this paper on the work of Zhang et al. and compare different clustering algorithms, selection methods, and document representations while taking inspiration from the literature.

Please note that we decided to deliberately exclude approaches for result list re-ranking since they are not well comparable to a representative subset, due to the unknown breakpoint [3], as discussed in the introduction. Furthermore, we exclude approaches from the related area of text summarization [13] since they work on a different granularity level.

3 Document Selection

Our approach for document selection is based on Zhang et al. [21] and consists of three steps (see Fig. 1). First, we retrieve the documents that match a certain query and compute the representations of the documents in the result set. Second, we cluster the documents into topics. Third, we select documents from the clusters to form a representative subset. Thus, we extend Zhang by an initial retrieval step to address the retrieval setting. Below, we introduce for each of the three steps the different methods that we compare in our experiments.

3.1 Document Representation

We use two different document representations in our experiments, Bag-of-Words and Paragraph Vectors.

Bag-of-Words: The classical Bag-of-Words (BOW) model represents each document as a vector that contains the weighted term counts for every term of the dataset. Whissell et al. [19] have investigated the effect of different feature weighting approaches on the document clustering performance. They concluded that BM25 outperforms other feature weighting approaches and suggested to use BM25 for clustering tasks.

Paragraph Vectors: Paragraph Vectors, which were introduced by Le and Mikolov [9], are dense vector representations of text fragments of arbitrary length, i. e., paragraphs, or documents, in a significantly lower-dimensional space than the corpus' dictionary. They are generated by neural networks that are trained to predict the words surrounding a given word. Those vectors are able to

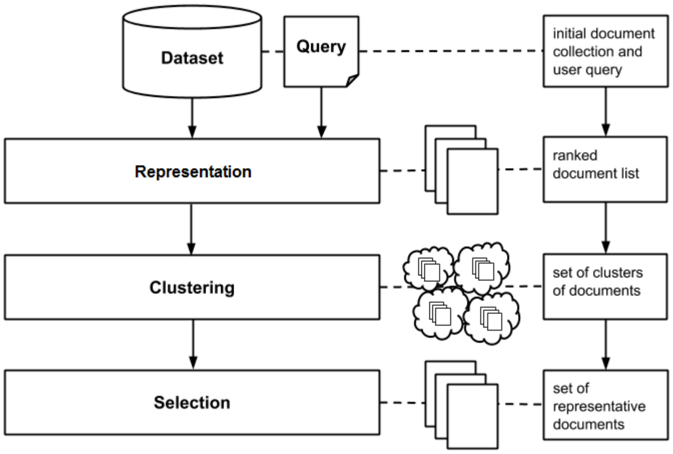
Dataset
Query
initial document collection and user query
Representation
ranked document list
Clustering
set of clusters of documents
Selection
set of representative documents

Fig. 1. Our document selection process with its representation, clustering, and selection step in a retrieval setting

carry semantic information, thus making them often superior to BOW models which ignore word ordering. This process is called doc2vec (D2V).

3.2 Clustering

We use spherical k-Means and Latent Dirichlet Allocation to cluster the documents into topics. Please note that we do not use X-Means, which was used by Zhang et al. [21], since it is too slow, as Zhang et al. stated themselves in a more recent work [22].

Spherical K-Means: K-Means [11] (KM) is a centroid-based clustering algorithm that iteratively assigns each datapoint to the nearest centroid and then recomputes the centroids from the assigned points until it converges. For a larger number of dimensions, when using the Euclidean Distance, the curse of dimensionality leads to uniform distances between data points [1], limiting the applicability. However, the distance metric can be exchanged with cosine similarity without violating the Gaussian distribution assumption underlying K-Means. This spherical K-Means [5] is more suitable for the application on textual data and thus is used in our experiments.

Latent Dirichlet Allocation: Latent Dirichlet Allocation (LDA) [4] is a probabilistic, generative model which identifies hidden topics in a corpus of documents. The basic idea of LDA is that documents are represented as a mixture of topics and each topic is identified by a distribution over words. Given a document corpus, LDA infers a model that is likely to have generated that corpus. LDA takes the term-document count matrix as input and creates a $n \times k$ document-topic matrix W with n documents and k topics. An entry $W_{i,j}$ describes the probability that topic j is contained in document i. We can use these probabilities to cluster the documents into topic clusters. To adhere to our goal of clustering the documents into topics and then selecting representatives per topic, we decided to keep it simple and cluster the documents by assigning them to the topic for which they have the highest probability.

3.3 Selection

We consider two selection strategies and a baseline. The number of selected documents results from the cluster proportions as proposed by Zhang et al. [21]. This means that we select twice as many documents from a cluster if it is twice as big as another cluster. Thus, the smallest cluster, and the number of documents selected from it, determines the size of the result set.

Selection by coverage and redundancy: Zhang et al. [21] proposed a coverage and redundancy based selection (CR). First, from each cluster, the document that is closest to the centroid is selected since it has the highest coverage inside the cluster. Subsequently, the documents with the lowest similarity to the previously selected documents are extracted to minimize redundant content inside the selected set. For all clustering algorithms, except LDA-based clustering, cosine similarity is used to compute the similarity between documents. In the case of LDA, the Jensen-Shannon divergence [10] is used because it is more suitable for probability distributions.

Selection by User Intent: The Intent Aware selection [2] (IA) of documents is based on the relevance of the documents to the query and the probability to satisfy any of the k topics. It maximizes the marginal utility, which is the sum over all topics of the product of the retrieval score and the conditional probability that the so far selected documents failed to represent that topic. The overall goal is to increase the diversity of topics among the selected documents. The original Intent Aware method does not use clustering and selects documents based on their probabilities to satisfy the query and the topics. Thus, theoretically, it can happen that certain clusters are not considered at all. We address this by taking at least one document per cluster (topic).

Random Selection: To evaluate the usefulness of the previously described selection strategies, a random selection strategy (R) is introduced as a baseline. From each cluster, based on their proportions, documents are chosen uniformly at random into the representative subset.

4 Evaluation

4.1 Datasets

We use two datasets of scientific publications for our experiments. We have sampled ten queries for each dataset from corresponding/suitable thesauri for our retrieval setting that return at least 1,000 documents (see Table 1).

ACL Anthology Network: The ACL Anthology Network [15–17] dataset is a collection of research papers of different Association for Computational Linguistics (ACL) venues. We removed all documents that did not have a full-text, leading to a dataset consisting of 22,486 English full-text papers. The query terms were selected from the ACM CCS[3] since the ACL thesaurus is still under construction. On average, the queries return 1,500 documents.

PubMed Open Access: PubMed Central[4] is a free full-text archive of biomedical and life sciences literature maintained by the United States National Institutes of Health's National Library of Medicine. From the 4.3 million publications available, about 1.5 million have an open-access license. We were able to acquire the full-text of 646,513 English documents from them. The queries for the PubMed dataset were sampled from the Medical Subject Headings[5], a hierarchically-organized medical vocabulary, each yielding on average 1,100 results.

Table 1. The queries and their corresponding number of relevant documents for both datasets. Documents are relevant if they were returned by the IR system.

ACL		PubMed	
Query terms	# rel. Docs	Query terms	# rel. Docs
cognitive science	2,066	dermatologists	1,227
supervised learning	2,035	cancellous bone	1,171
similarity measures	1,639	meniscus	1,164
bootstrapping	1,590	gastroenterologists	1,147
dynamic programming	1,497	radiation oncologists	1,084
maximum entropy modeling	1,452	endocrinologists	1,075
natural language generation	1,441	orthopedic surgeons	1,073
feature selection	1,317	surgical wound	1,048
neural networks	1,135	nephrologists	1,017
machine learning approaches	1,069	tartrate-resistant acid phosphatase	1,002

4.2 Metrics

To ensure comparability, we evaluate our approach using the metrics by Ma et al. [12] that were used by Zhang et al. [21]. Please note that $sim()$ refers to the cosine similarity, as it was chosen by Zhang et al [21].

Coverage: Coverage evaluates how much content of a dataset D is covered by a subset S:

$$\text{coverage}(S, D) = \frac{1}{|D|} \sum_{r \in D} (\max_{d \in S}(\text{sim}(d, r))) \tag{1}$$

In the case that all documents are selected, the coverage reaches its maximal value of 1. The coverage will be close to zero if the selected set of documents only resembles a minimal fraction of the complete set of documents.

[3] http://www.acm.org/about/class/class/2012

[4] https://www.ncbi.nlm.nih.gov/pmc/

[5] https://www.nlm.nih.gov/mesh/

Redundancy: The redundant information in a subset S is assessed by:

$$\text{redundancy}(S) = \sum_{d_i \in S} \left(\frac{1 - 1 \backslash \sum_{d_j \in S} sim(d_i, d_j)}{|S|} \right) \tag{2}$$

Please note that this computation also considers the size of the subset, i. e., having a subset of three duplicates and a subset of five duplicates would yield different scores. We are also aware that the metric has some short comings when used with cosine similarity. However, for the sake of comparability, we use it as it was used by Zhang et al. [21].

4.3 Experiment Setup

We indexed each dataset using the full-text of the documents and used BM25 ($k1 = 1.2$ and $b = 0.75$) as our scoring function for retrieval in Elasticsearch[6]. We retrieved the documents for each query by using exact matching and preprocessed the resulting document set using Porter stemming [14] and stop-word removal using NLTK stop-words before computing the document representation (e. g., BoW or Paragraph Vectors). To address the curse of dimensionality, all terms that appeared in more than 95% of the documents or in less than two documents were removed. We further limited our vocabulary to the remaining 50,000 most popular terms.

We calculated the document representation of the preprocessed documents using the methods described in Section 3.1. In case of the BoW model, we used BM25 with the parameters $k1 = 20$ and $b = 1$ based on a study of Whissell et al. [19]. For the paragraph vector model, we used a model that was trained on a dump of all English Wikipedia articles from December 2017 using the gensim library [18].

Under the assumption that the topical diversity is limited, we decided to cluster the documents with $k \in \{5, 10, 25, 50\}$ since the true number of clusters is unknown.

After clustering, representative documents were selected from the clusters using the proposed selection methods. For the calculation of the selected set S for the IA selection in combination with LDA, we follow the procedure described in [6]. In the case of k-Means clustering (for both BOW and D2V), we take a different approach as the necessary probability distributions are not provided by the clustering algorithm. We compute the quality value using the retrieval score, weighted by the cosine distance of a document to its corresponding cluster center. For the calculation of the conditional probability, first, the feature vector for the query is derived from the vocabulary (or pretrained model in case of paragraph vectors). Then, for each topic, the probability is the cosine distance between the topic and the query. To compute the similarity between two documents, we use the cosine similarity except for LDA-based clustering where the Jensen-Shannon divergence is used.

[6] https://www.elastic.co/

To allow for a fair comparison between the different document selection strategies, we compute the metrics using the BM25-weighted BoW model and cosine similarity (or distance, respectively) even if the clustering was using different feature vectors (e. g., LDA-based clustering).

In total, we ran 36 experiments, each using a different combination of the 2 document representations, 2 clustering algorithms, 4 different values for k, and 3 different selections methods, on 20 different document sets, which were returned by 10 queries on our two datasets. We repeated each experiment 5 times and averaged over all runs.

5 Results and Discussion

Fig. 2 and Fig. 3 show the average coverage and redundancy values for the different experiment configurations on the ACL and PubMed datasets, respectively. We analyze the results along the three components of RQ1: a) document representation, b) clustering, and c) selection. To answer RQ1 a), we look at the results for K-Means using the bag-of-words model (KM-BOW) and K-Means using document embeddings (KM-D2V). On both datasets, for k=5 and k=10, the coverage has no large difference but for the selection methods IA and R with document embeddings there is slightly less redundancy. Starting from k=25, selections based on KM-D2V have a higher coverage and a sharper increase in redundancy. The use of D2V most likely influences the representative subset selection so that for small k, slightly less redundant content is selected. For larger k, clearly, more content is covered while the increase in redundancy is neglectable.

To answer RQ1 b), we compare the results of all clustering algorithms. Except for LDA, the coverage and redundancy results for the strategies increase steadily with larger k, all achieving their maximum at k=50. For LDA, both scores are close to 1 when increasing to k=25 and above. The differences between the algorithms are more distinct at a larger k.

We compare the coverage and redundancy for the selection methods to answer RQ1 c). One can see that the CR selection generates lower redundancy scores in combination with KM-BOW clustering but the effect is diminishing with larger k. For the other clustering algorithms, one can observe that CR has equal or higher redundancy scores than the other selection methods. In terms of coverage, the choice of the selection method is less important than the clustering algorithm as the differences between CR, IA, and R are minimal.

Summarizing our results regarding RQ1, on both datasets, the best performing configurations, with respect to coverage, are those that use D2V or LDA. However, the selections based on BOW have the lowest redundancy scores.

Regarding RQ2, we make three observations with respect to coverage and redundancy. First, the scores for both metrics increase consistently for a larger number of clusters. This correlates directly with the number of documents due to the cluster proportion calculation, which defines the number of documents that are selected. In the case of more heterogeneous cluster sizes, more documents will be selected from each cluster. With larger k, it is more likely that the documents

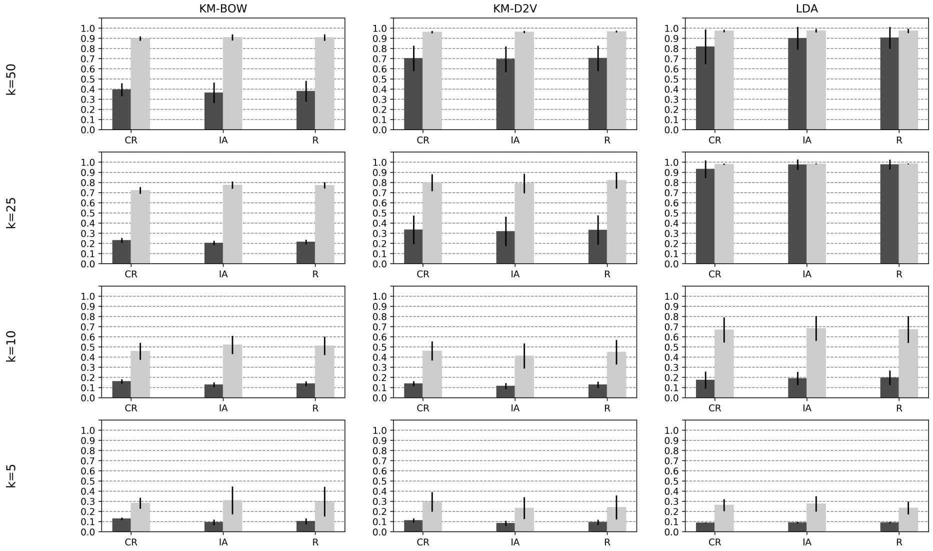

Fig. 2. Coverage (left black bars) and redundancy (right grey bars) averaged over all queries for the different document selection strategies on the ACL dataset using $k \in \{5, 10, 15, 20\}$. The standard deviation is indicated as a black line on top of each bar.

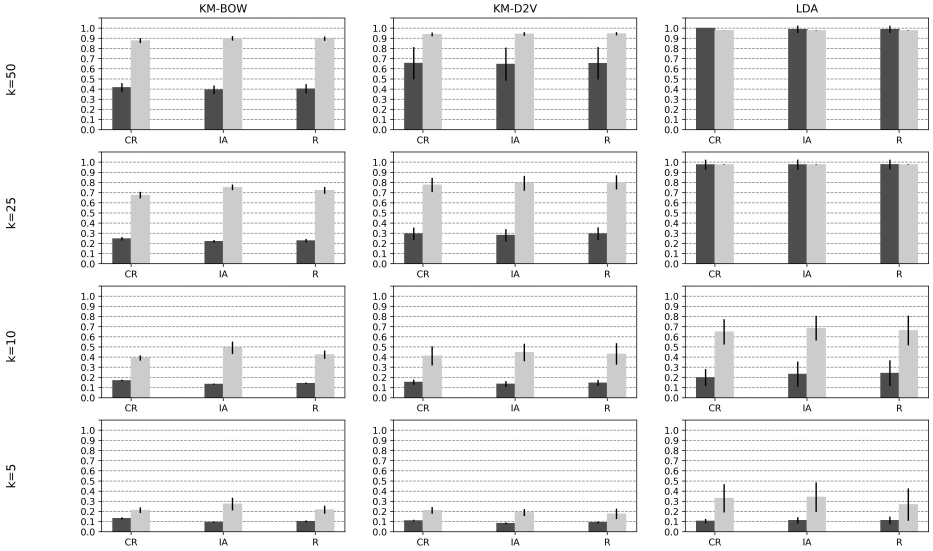

Fig. 3. Coverage (left black bars) and redundancy (right grey bars) averaged over all queries for the different document selection strategies on the PubMed dataset using $k \in \{5, 10, 15, 20\}$. The standard deviation is indicated as a black line on top of each bar.

are unevenly distributed among the clusters. One example is the LDA-based selection for $k \in \{25, 50\}$, which contains most documents and has coverage scores close to 1. Thus, coverage and redundancy are inflated by selecting most of the documents rather than being a result of a better selection.

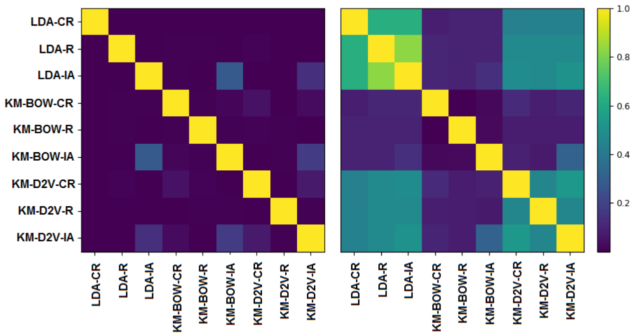

Fig. 4. Average fractions of shared documents in the representative document sets, selected by our nine different document selection strategies on the ACL dataset. On the left for k=5 and on the right for k=50.

Second, from the comparison with a random selection method, we observed an independence of the evaluation metrics from the actual choice of documents. This raises the question whether the selection methods select similar document sets. Therefore, we decided to have a closer look at the subsets. Fig. 4 shows the result of the comparison of the subsets for k=5 and k=50 at the example of ACL, i. e., the fraction of mutual documents between the selection strategies. We can see that for small k none of the selections share many similar documents, while for a larger k strategies with the same document representation and clustering algorithm (but different selection strategies) start to select similar documents. However, these document selections are more alike as more documents are selected in general. This becomes obvious, especially for D2V and LDA, since both are more susceptible to imbalanced cluster sizes. This limits the generalizability of coverage and redundancy to evaluate the representativeness of a document subset. Please note, we have omitted the analysis on the PubMed dataset since the results were similar.

Finally, in contrast to the original work of Zhang et al. [21], we observe for each strategy that the redundancy exceeds the coverage scores. We have investigated whether it results from our IR setting and hence a general higher similarity of documents. However, we achieved similar results when using an equal amount of documents randomly sampled from the full dataset. Further research needs to be conducted to explain the difference between the results on our and Zhangs' datasets.

6 Conclusion

We have proposed a document selection framework in an information retrieval context as an extension of the representative subset selection by Zhang et al. [21]. Our analysis reveals that there is no unique representative document set with regards to the evaluation metrics but instead most strategies achieve comparable results with different document subsets, even our random baseline. This raises the question whether coverage and redundancy are sufficient to evaluate the representativeness of a document set. Furthermore, we identified the size of the result set as problematic. It is often too large for a representative subset due to the selection based on the cluster proportions. Therefore, as future work, we propose to enhance the representativeness metric introduced by Ma et al. [12] with a weighting term which promotes those solutions which select fewer documents for evaluating representative subsets. Finally, we plan to further investigate the influence of different dataset characteristics and preprocessing methods on the overall document selection process.

Acknowledgment

This research was co-financed by the EU H2020 project MOVING (`http://www.moving-project.eu/`) under contract no 693092 and the EU project DigitalChampions_SH.

References

1. Aggarwal, C.C., Hinneburg, A., Keim, D.A.: On the surprising behavior of distance metrics in high dimensional spaces. In: den Bussche, J.V., Vianu, V. (eds.) Database Theory - ICDT 2001, 8th International Conference, London, UK, January 4-6, 2001, Proceedings. Lecture Notes in Computer Science, vol. 1973, pp. 420–434. Springer (2001). https://doi.org/10.1007/3-540-44503-X_27
2. Agrawal, R., Gollapudi, S., Halverson, A., Ieong, S.: Diversifying search results. In: Baeza-Yates, R.A., Boldi, P., Ribeiro-Neto, B.A., Cambazoglu, B.B. (eds.) Proceedings of the Second International Conference on Web Search and Web Data Mining, WSDM 2009, Barcelona, Spain, February 9-11, 2009. pp. 5–14. ACM (2009). https://doi.org/10.1145/1498759.1498766
3. Arampatzis, A., Kamps, J., Robertson, S.: Where to stop reading a ranked list?: threshold optimization using truncated score distributions. In: Allan, J., Aslam, J.A., Sanderson, M., Zhai, C., Zobel, J. (eds.) Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23, 2009. pp. 524–531. ACM (2009). https://doi.org/10.1145/1571941.1572031
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. In: Dietterich, T.G., Becker, S., Ghahramani, Z. (eds.) Advances in Neural Information Processing Systems 14 [Neural Information Processing Systems: Natural and Synthetic, NIPS 2001, December 3-8, 2001, Vancouver, British Columbia, Canada]. pp. 601–608. MIT Press (2001), `http://papers.nips.cc/paper/2070-latent-dirichlet-allocation`

5. Endo, Y., Miyamoto, S.: Spherical k-means++ clustering. In: Torra, V., Narukawa, Y. (eds.) Modeling Decisions for Artificial Intelligence - 12th International Conference, MDAI 2015, Skövde, Sweden, September 21-23, 2015, Proceedings. Lecture Notes in Computer Science, vol. 9321, pp. 103–114. Springer (2015). https://doi.org/10.1007/978-3-319-23240-9_9
6. He, J., Meij, E., de Rijke, M.: Result diversification based on query-specific cluster ranking. JASIST **62**(3), 550–571 (2011). https://doi.org/10.1002/asi.21468
7. Jardine, J.G.: Automatically generating reading lists. Ph.D. thesis, University of Cambridge, UK (2014), `http://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.648722`
8. Jardine, J.G., Teufel, S.: Topical pagerank: A model of scientific expertise for bibliographic search. In: Bouma, G., Parmentier, Y. (eds.) Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden. pp. 501–510. The Association for Computer Linguistics (2014), `http://aclweb.org/anthology/E/E14/E14-1053.pdf`
9. Le, Q.V., Mikolov, T.: Distributed representations of sentences and documents. In: Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014. JMLR Workshop and Conference Proceedings, vol. 32, pp. 1188–1196. JMLR.org (2014), `http://jmlr.org/proceedings/papers/v32/le14.html`
10. Lin, J.: Divergence measures based on the shannon entropy. IEEE Trans. Information Theory **37**(1), 145–151 (1991). https://doi.org/10.1109/18.61115
11. Lloyd, S.P.: Least squares quantization in PCM. IEEE Trans. Information Theory **28**(2), 129–136 (1982). https://doi.org/10.1109/TIT.1982.1056489
12. Ma, B., Wei, Q., Chen, G.: A combined measure for representative information retrieval in enterprise information systems. J. Enterprise Inf. Management **24**(4), 310–321 (2011). https://doi.org/10.1108/17410391111148567
13. Naveen, G.K.R., Nedungadi, P.: Query-based multi-document summarization by clustering of documents. In: Proceedings of the 2014 International Conference on Interdisciplinary Advances in Applied Computing. pp. 58:1–58:8. ICONIAAC '14, ACM, New York, NY, USA (2014). https://doi.org/10.1145/2660859.2660972
14. Porter, M.F.: An algorithm for suffix stripping. Program **14**(3), 130–137 (1980). https://doi.org/10.1108/eb046814
15. Radev, D.R., Joseph, M.T., Gibson, B., Muthukrishnan, P.: A Bibliometric and Network Analysis of the field of Computational Linguistics. Journal of the American Society for Information Science and Technology (2009)
16. Radev, D.R., Muthukrishnan, P., Qazvinian, V.: The ACL anthology network corpus. In: Proceedings, ACL Workshop on Natural Language Processing and Information Retrieval for Digital Libraries. Singapore (2009)
17. Radev, D.R., Muthukrishnan, P., Qazvinian, V., Abu-Jbara, A.: The acl anthology network corpus. Language Resources and Evaluation pp. 1–26 (2013). https://doi.org/10.1007/s10579-012-9211-2
18. Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. pp. 45–50. ELRA, Valletta, Malta (May 2010), `http://is.muni.cz/publication/884893/en`
19. Whissell, J.S., Clarke, C.L.A.: Improving document clustering using okapi BM25 feature weighting. Inf. Retr. **14**(5), 466–487 (2011). https://doi.org/10.1007/s10791-011-9163-y

20. Zhang, B., Yin, X., Zhou, F., Jin, J.: Building your own reading list anytime via embedding relevance, quality, timeliness and diversity. In: Kando, N., Sakai, T., Joho, H., Li, H., de Vries, A.P., White, R.W. (eds.) Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017. pp. 1109–1112. ACM (2017). https://doi.org/10.1145/3077136.3080734
21. Zhang, J., Liu, G., Ren, M.: Finding a representative subset from large-scale documents. J. Informetrics **10**(3), 762–775 (2016). https://doi.org/10.1016/j.joi.2016.05.003
22. Zhang, J., Ren, M., Xiao, X., Zhang, J.: Providing consumers with a representative subset from online reviews. Online Information Review **41**(6), 877–899 (2017). https://doi.org/10.1108/OIR-05-2016-0125