Fachbereich 4: Web and Data Science
Informationstechnik für Führungssysteme (ITF)

Sub-topic modeling and ranking analysis in document retrieval systems

Master Thesis Proposal

submitted by

Gadiyaram, Sri Sai Praveen

Web and Data Science
gsspraveen@uni-koblenz.de
219203192

January 17, 2023

Supervisor: Prof. Dr. Jan Jürjens
Institute for Software Technology

2nd Supervisor: MSc. Katharina Großer
Institute for Software Technology

Contents

1 Introduction and Motivation

Retrieving highly relevant documents in the top results for a given user query is one of the challenging tasks in Information Retrieval (IR). This challenge is amplified when the user has a specific intention, and the search query lacks the context of their intention. For example, the user query *"Robotics"* can retrieve documents related to many domains such as manufacturing, agriculture, military, etc,. A simple keyword search can overwhelm the user with many false positives when the user wants to explore the innovation documents only related to a specific domain, such as *"Military"*. To fulfill the user intent and missing context in the user query, a novel document modeling approach is proposed to extract highly coherent query-specific contexts (sub-topics) from the top retrieved documents, which helps the user immensely narrow down the search space. Furthermore, the proposed approach will be evaluated using precision and survey analysis.

1.1 Fraunhofer FKIE

Fraunhofer FKIE (Fraunhofer-Institut für Kommunikation, Informationsverarbeitung und Ergonomie) is a leading research institute for providing innovative solutions in information and communications technology, and their main focus is on developing effective and efficient human-machine systems[1]. The users at FKIE are especially interested in reading news articles related to innovation and breakthroughs in *Technology and Military*. The below image, Figure 1, shows an example of areas of interest to the FKIE users, and this list is not bounded and can include more domains.

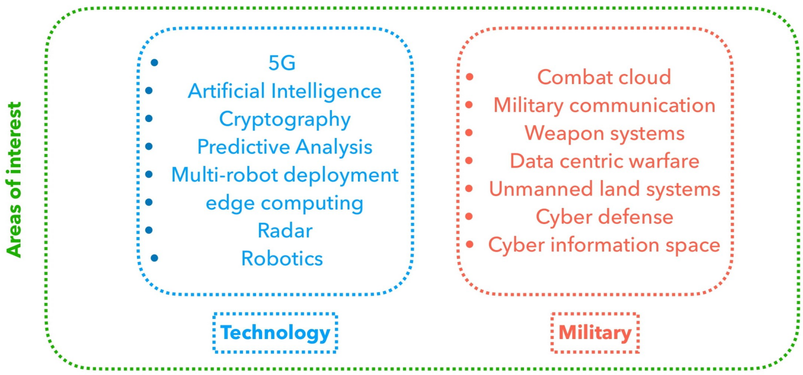

Figure 1: Areas of interest for the users at FKIE

1.2 Background

A document retrieval system was developed to support users at FKIE in retrieving news articles related to technology and military topics. The retrieval setup contains three primary components: *Web scraper, Document filter*, and *Retriever*, as shown in Figure 2. The first component, the *Web scraper*, downloads news articles (HTML pages) from a list of URLs and cleans the raw HTML data from advertisements and noise. Each cleaned news article is considered as a single entity, namely a *Document*. The majority of downloaded documents are a mixture

[1] `https://www.fkie.fraunhofer.de/en/about-fkie.html`

of topics such as military, technology, artificial intelligence, etc., and also contain a small number of typical news topics, namely politics, sports, advertisements, etc. The downloaded news articles are in *German* and *English*.

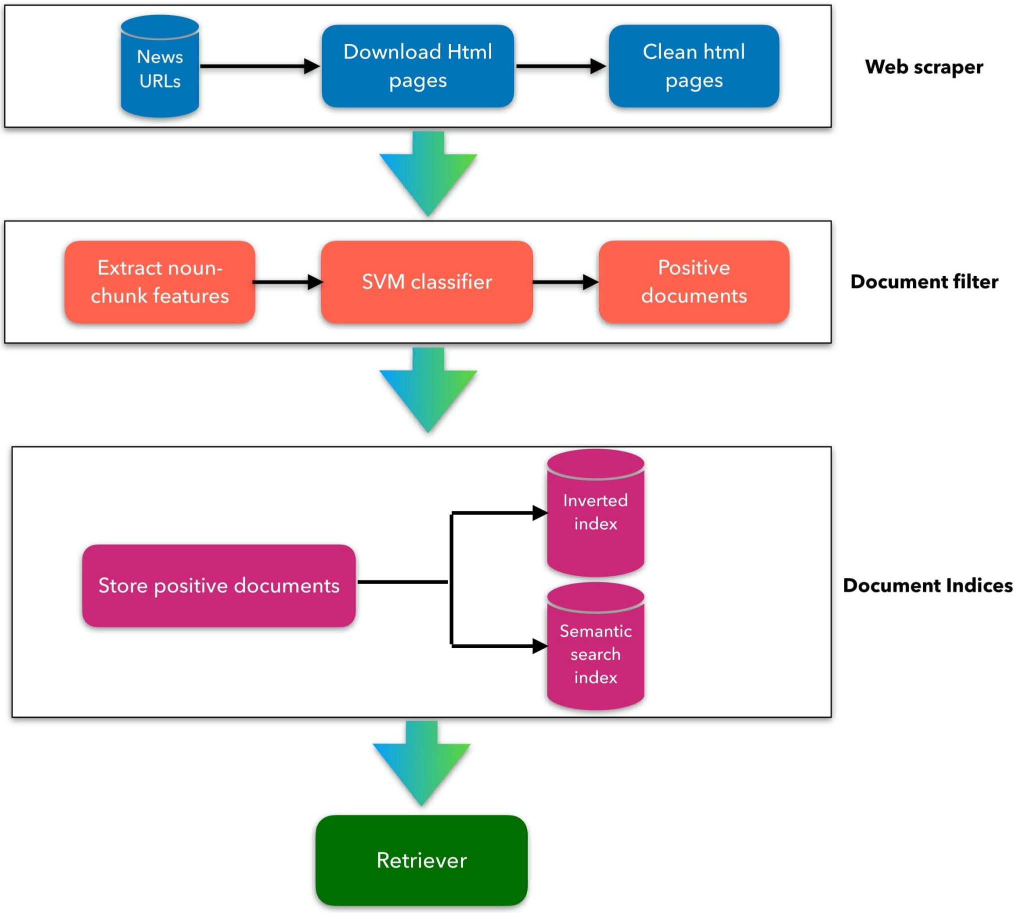

Figure 2: Document retrieval system designed at Fraunhofer FKIE

The second component, the *Document filter*, is a Support Vector Machine (SVM) classifier that filters most of the irrelevant documents related to specific news topics. The documents are classified into two classes namely, *Positive* and *Negative*. *Positive* documents are documents related to technology and military, and *Negative* documents consist of everything else. Mean noun chunk vectors from pre-trained multi-lingual Universal Sentence Encoder (USE) [21] model from Tensorflow Hub[2]. In order to facilitate positive documents to the FKIE users, a *Retriever* component is designed to retrieve documents for a given user query using lexical and semantic matching techniques. Therefore, the positive documents from the *Document filter* stage are stored in two different document indices namely *Inverted index* and Semantic search index.

Document indexing or compression is a technique to store documents in an optimized way on the disk for efficient retrieval. The *Inverted index* is a data structure that contains every word in the corpus and the respective list of documents where the word occurs [24]. The *Semantic search index* stores the distributed vectors of the documents on the disk and uses them later for retrieval. Finally, the component *Retriever* uses both the indices and retrieves documents according to the user request through a web UI, as shown in Figure 3.

[2]`https://tfhub.dev/google/universal-sentence-encoder-multilingual-large/3`

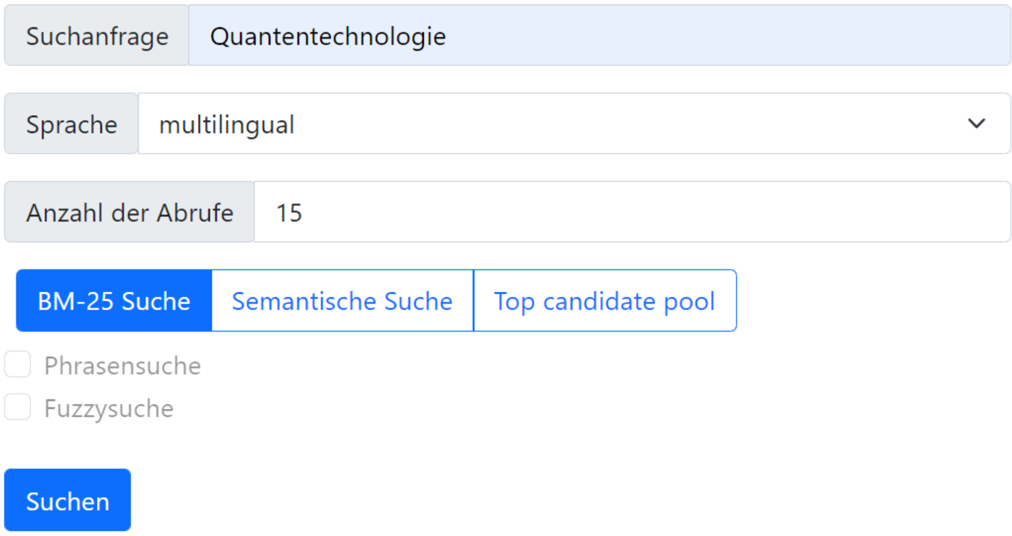

Figure 3: User interface to retrieve documents for FKIE users

1.3 Challenges

All possible user queries can be categorized into two major query types, namely phrase (three words or less) and sentence queries. Semantic matching of query and documents is better suited when the user query is a long sentence query due to the context embedded in the search query. For example, the user query *"What are the technological advancements in Robotics related to Unmanned Weapon Systems?"* provide high-quality results in the top results, as the information request is detailed in the query. Consequently, the search query *"Robotics"* results are mapped to multiple domains and lead to many false positives (according to the user's intention). In the case of keyword queries, it was observed that semantic and lexical matching are prone to high false positives and have no unique advantage. A manual observation of retrieved results is carried out with a set of sample queries to evaluate the retrieval algorithms, and the results are shared in Table 1 on page 6.

Now consider the case for the users at FKIE, the user provides only one or two phrase queries, and his or her intention is particular to specific topics such as *Technology and military*. The main challenge here is integrating the known user intention with the current document-matching algorithms with out labeled data. One further challenge is that a wide variety of sources can also result in high noise or false positives, and the user is less likely to find the relevant documents in the top results. Unlike tweets or requirements, news articles are long documents with an average token length of around 650 and consist of keywords from multiple domains or fields. An example of a news article (this is only a part of the original article) is shown in Figure 4

Table 1: Retrieval algorithms comparison on different query types

S No.	Query type	Better retrieval algorithm (comparatively)	Queries used
1	Simple single word	BM-25	Radar, Waffen
2	Phrase query	BM-25	Big Data, Unbemannten Systemen
3	Abbreviations	BM-25	FPGA, LTE, AI
4	Prefix/Suffix queries	BM-25	Datenfusion, Sensordatenfusion, Wetterdatenfusion
5	No meaning	BM-25	Person or object names*
6	Multi-lingual queries	Semantic search	Artificial Intelligence vs Künstliche Intelligenz
7	German composite words	Semantic search	Quantentech- nologie
8	Spelling mistakes	Semantic search	Kryptografy, Rbot
9	Polysemy	Semantic search	Combat Cloud, Cloud computing
10	Sentence/long phrase queries	Semantic search	Schwachstell- enanalyse eigene Waffen-Systeme

Titel: US Army Project May Improve Military Communications by Boosting 5G Technology
Veröffentlicht am: 2019-11-24 20:00:32

RESEARCH TRIANGLE PARK, N.C. (Nov. 21, 2019) — An Army-funded project may boost 5G and mm-Wave technologies, improving military communications and sensing equipment. Carbonics, Inc., partnered with the University of Southern California to develop a carbon nanotube technology that, for the first time, achieved speeds exceeding 100GHz in radio frequency applications. The milestone eclipses the performance — and efficiency — of traditional Radio Frequency Complementary Metal-Oxide Semiconductor, known as RF-CMOS technology, that is ubiquitous in modern consumer electronics, including cell phones. “This milestone shows that carbon nanotubes, long thought to be a promising communications chip technology, can deliver,” said Dr. Joe Qiu, program manager, solid state and electromagnetics at the Army Research Office. “The next step is scaling this technology, proving that it can work in high-volume manufacturing. Ultimately, this technology could help the Army meet its needs in communications, radar, electronic warfare and other sensing applications.” The research was published in the journal Nature Electronics . The work, funded

Figure 4: A sample news article from the document database [1]

Information related to Innovation and technological breakthroughs is hard to find in the news articles.However, the probability is not zero, as positive news articles are gathered during data collection for classification. Nevertheless, their low distribution makes it challenging to create a dataset sufficient for supervised approaches. After considering the challenges with positive documents for the user intention, a supervised solution is hard to achieve, in order to match the performance of a full sentence query. Real-time user feedback and continuous reinforcing algorithms can fulfill the lack of labeled datasets, but they need feedback from diverse users regularly. Otherwise, the search results can be highly inclined to a particular user and lead to biased results.

A template-based search query is an option to improve the context of a search query. For example, we have a pre-defined template such as *Innovations in XXX related to the Military*. When the user provides a query: *Robotics*, we replace the *XXX* with the user query, and this results in the final query *Innovations in Robotics related to Military*. This approach restricts the user to having only a few sets of templates and is also inefficient when a new template needs to be added, or an existing template needs to be updated.

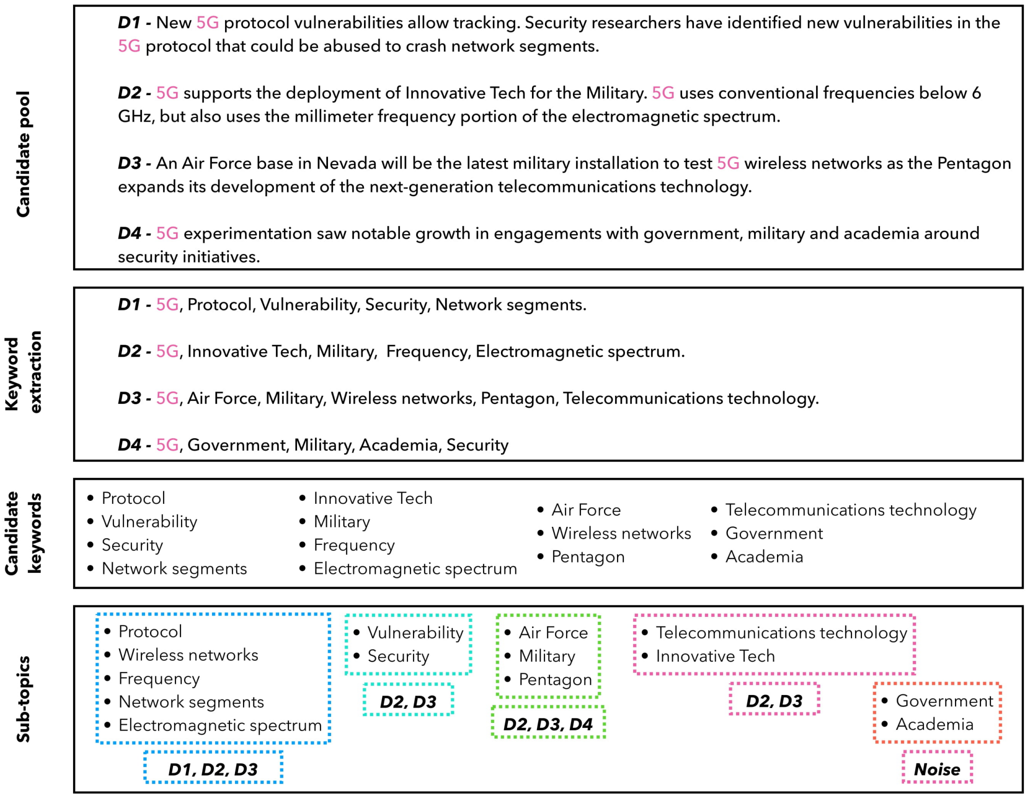

Figure 5: Sub-topic extraction for the query: 5G

After considering various approaches to fulfill the missing context, extracting the contexts from top results to the user query in an unsupervised way is more efficient and explainable. This would not only help the user to have deep insights into the results pool but also reduce the efforts to reach the highly relevant documents. A sample sub-topic extraction pipeline output is shown in Figure 5. These contexts are described as *sub-topics*. The proposed approach is aimed at handling the challenges mentioned above, and only news articles from diverse sources are considered in this experiment, and the results can be easily transferred to other data sources in the future.

2 Research questions

An unsupervised soft clustering approach is proposed to model documents (from multiple languages) as a mixture of sub-topics, which are extracted using the deep inherent information from keywords. Below are the research questions that address the problems mentioned above through a new document modeling approach.

RQ1: ***How effective is the sub-topic modeling approach in creating distinctive clusters from the news articles?***

This research question aims to test the effectiveness of the above-proposed approach. An intrinsic and extrinsic clustering evaluation techniques, and a survey is chosen to evaluate the clustering output.

RQ2: ***What is the effect of sub-topic ranking in finding the positive documents from the candidate pool?***

When a user chooses a particular sub-topic cluster, then only specific retrieval results related to the query and the sub-topic are shown. On the other hand, this can restrict the user from viewing the actual retrieved results for the given query. This research question addresses the impact of the sub-topic clustering output to find the positive documents against the baseline approach and is evaluated through an exploratory precision analysis.

3 Proposed methodology

One way to extract different contexts from the candidate pool is to perform any clustering algorithm on the retrieved documents. This results in very generic clusters closely related to a given query and not very useful to the user. To generate diverse and distinctive clusters, we need to use the latent information at the word or phrase level rather than at the document level [5]. As the documents contain multiple occurrences of the query and are also highly similar in semantic space, we need to reduce the impact of the given user query to generate a clear distinction between the documents. Figure 6 illustrates the proposed approach on an abstract level.

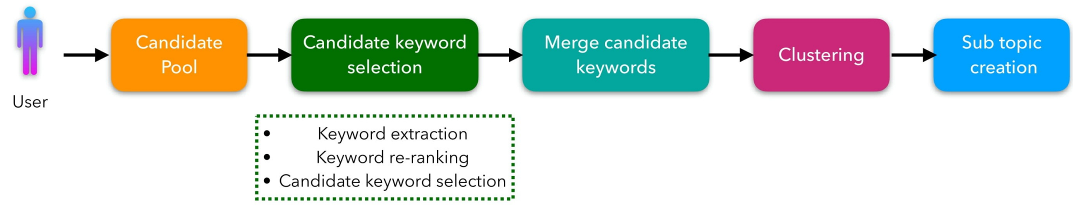

Figure 6: Proposed approach on an abstract level

The proposed approach, shown in Figure 6, does not assume fixed templates or specific user intentions. Major components in the pipeline are: *Candidate keyword selection, Merge candidate keywords, Clustering, and Sub-topic creation*. This pipeline's first step is retrieving a candidate or retrieval pool for the given query. Subsequently, to extract keywords with high diversity and low noise (stopwords), a Candidate selection module is proposed. This component consists of three significant steps namely Keyword extraction, Keyword re-ranking, and Candidate keyword selection.

Keyword extraction is extracting the most meaningful noun phrases in a text document. In the second stage, *Keyword re-ranking*, cosine similarity is calculated from the phrase embeddings between the keywords and the query. Using these similarity scores, keywords are then re-ordered in descending order. After this stage, the certain keyword, that are not similar to the query have a low cosine similarity score are precisely extracted, as they have a high potential for creating distinctive clusters or sub-topics. Specific keywords are selected and used for clustering using a threshold cut-off and this process is referred to as *Candidate keyword selection* and the resulting phrases after this stage are called Candidate keywords, shown in Figure 7.

The second component in the pipeline, *Merge candidate keywords*, merges candidate keywords from each document in the candidate pool and duplicates are removed. These keywords are then clustered semantically and modeled with the documents again. This process is also referred as Document to sub-topic modeling and is designed independent to the query and to handle multiple languages.

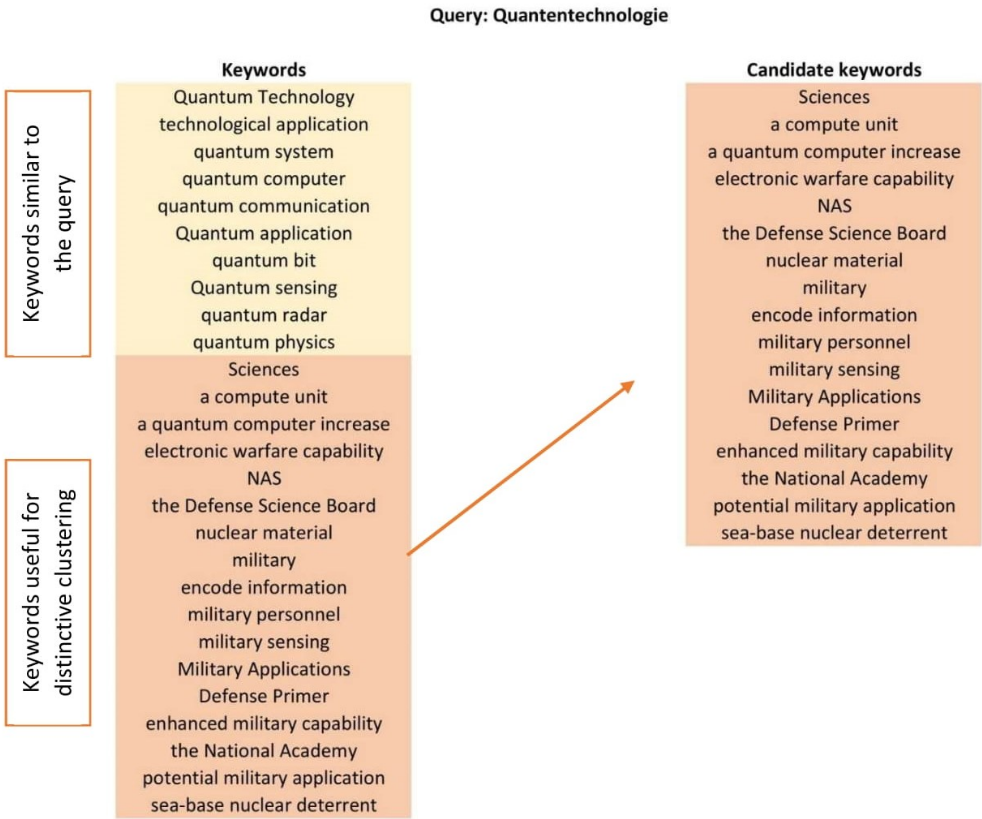

Figure 7: Candidate keyword selection step from a single document

After clustering, sub-topics are extracted using a centroid approach. A mean phrase vector (centroid vector) is calculated from all the keywords inside a cluster and the closest keyword vector to the centroid vector is considered a cluster label. This process is named *Cluster labeling* and the cluster labels are considered sub-topics. After clustering, the individual clusters are considered as sub-topics. Sub-topics and documents inside a sub-topic can be further ranked before showing to the user. The pipeline ends with this last component, *Sub-topic creation*.

4 Evaluation

4.1 Testset

For two main reasons, a dataset specific to this research problem is hard to find in the current IR data repositories. Firstly, the search query needs to be a phrase rather than a sentence. Furthermore, the documents need to be labeled with a specific intention rather than just coherence with the query. The interest at Fraunhofer FKIE is to retrieve the documents related to "Innovation and Technology", and a new testset is collected for this purpose. Below are a few specific areas of interest in news articles that describe the user intention: *Innovation, Technology breakthroughs, Future products, Applied research, New procurement strategies, Artificial Intelligence*. These topics are also described as positive document characteristics because a document is considered positive when it is strongly related to any one of the above-mentioned characteristics.

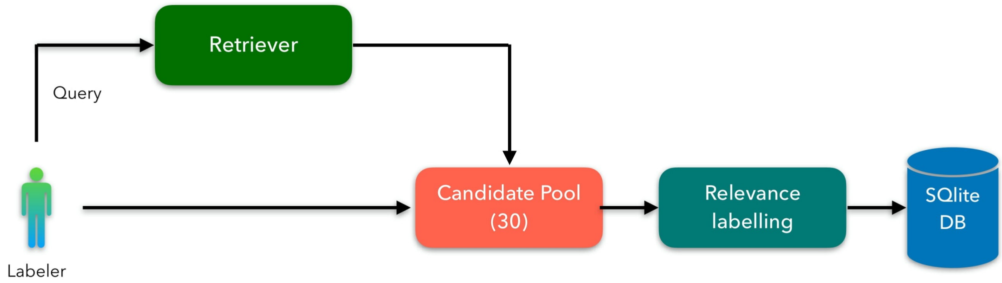

Figure 8: Testset collection strategy

The strategy for the testset collection is to consider documents from lexical and semantic matching. The results from both algorithms help find the diverse contexts related to the user query. Therefore, a candidate pool with a maximum length of 30 documents from both lexical and semantic matching is considered. Fifteen documents from the above search systems are combined, and duplicate documents are eliminated. The image Figure 8 shows the methodology followed for the testset collection. *Relevance labeling*, assigns an appropriate label to the retrieval results inside the candidate pool. Every labeler has to assign a label not only coherent to the query but also considering the FKIE user's intention, i.e., coherence with positive document characteristics mentioned above. Once the labeler assigns a particular label to a document, labeled information is stored in an SQLite DB.

Table 2: Relevance label definitions

Label id	Label name	Label definition
1	Perfect	A document that strongly matches one of the positive document characteristics.
2	Partially relevant	A document that contains keywords and seems to be relevant, but still lacks innovation or novelty.
3	Irrelevant	A document containing the given user keyword still lacks innovation and coherent discussion about the query.
4	Wrong	These are false documents and have nothing to do with the user query.

4.2 Clustering evaluation

To evaluate the quality of clustering, the *Silhouette index* is considered for the intrinsic evaluation, and a custom target function F is designed. Target function F tests the quality of clusters against the relevance labels from the dataset, as shown in Table 3 on page 11. The objective is to test whether the relevant documents are clustered into a similar cluster and the same case with irrelevant documents. Without any relation between clustering and relevance labeling, it can not be assumed that the positive and negative documents are clustered automatically because they cover a wide range of keywords in different domains. Therefore, it is more meaningful to evaluate the clustering for negative documents, i.e., Irrelevant and Wrong labeled documents. A target function is designed to address the number of negative documents isolated through sub-topic modeling.

Table 3: Label distributions in the testset

Label id	Label name	Document count
1	Perfect	78
2	Partially relevant	147
3	Irrelevant	306
4	Wrong	98

The sub-topic modeling pipeline's output is distinctive clusters with a unique context independent of relevance to user intention. However, the clusters can be divided into relevant and irrelevant clusters according to the relevance labels in the dataset. Let us consider that N_1, N_2, N_3, N_4 represent functions to get the number of documents in a single cluster with label ids $1, 2, 3, 4$ respectively, as shown in Table 2 on page 10, and C represents the cluster set.

$$C = \{c_1, c_2, c_2, \dots\}$$

Relevant clusters C_r are clusters, that contain at least one label id 1 or a majority of label id 2. This can be determined using the below expression.

$$C_r = \{c_i \in C | (N_1(c_i) > 0) \vee (2 * N_2(c_i) >= (N_3(c_i) + N_4(c_i)))$$

With this expression, relevant clusters are differentiated from others and the focus is only on labels 1 and 2. The clusters that do not satisfy the above condition are logically considered irrelevant clusters.

$$C_i = \{c_j \in C \setminus C_r\}$$

The target function assesses the clustering with a ratio of documents in irrelevant documents to the documents in the candidate pool CP_q to a given user query q. Given N queries, the target function maps the score using the below equation.

$$F = \sum_{i=1}^{N} (|C_i| / |CP_i|) * 100$$

4.3 Precision evaluation

The parameters of the sub-topic modeling pipeline are tuned using both the target functions *Silhouette index* and F. Assuming that the cluster labels are not very helpful to the user, the next evaluation technique shows that the clustering output does not deteriorate the performance of the retrieval results.

The output of clustering is hard to examine with the baseline IR systems because the order of documents is missing and the actual performance metrics related to false positives are not addressed. For this purpose, we are extending the sub-topic creation with sub-topic ranking and document ranking. These two rankings help the existing pipeline to create a sequential order of documents and facilitate the evaluation of precision against the baselines. Therefore, this experiment proposes six different retrieval systems and evaluates the ranked results.

Table 4: Proposed IR systems for evaluation

S No.	Flat clusters	Sub-topic ranking	Document ranking
1	IR0	NA	Uniform distribution
2	IR1	NA	Query similarity
3	IR2	Query similarity	Query similarity
4	IR3	Template similarity	Template similarity
5	IR4	Document cardinality	Query similarity
6	IR5	Random combinations	Query similarity

The first system, *IR0*, is an arbitrary system where the positive documents are distributed uniformly on the ranking order. *IR1* system is simple query re-ranked results based on cosine similarity between the query and documents. The systems *IR2, IR3, IR4* are results of sub-topic pipeline clustering, where the clusters are first ranked, and later the documents are re-ranked with certain criteria. These three systems simulate the user reading the results linearly or in a sequence. In *IR2*, the sub-topic clusters are ranked by the cosine similarity between the query and centroid vector of the cluster and similarly for document ranking.

The system *IR3* uses a template similarity criteria, where the similarity is calculated between a template and centroid vector rather than the query. For example, the template string can be "Innovation and Technology". In the same way, *IR4* clusters are ranked using the number of documents in the cluster. The last system, *IR5*, is an unreal system just like the *IR0*, but multiple combinations of random ranking of clusters are considered to simulate the random selection of a sub-topic by the user and reading the documents in different sub-topics.

In [16], a new evaluation measure for IR systems named expectation score is introduced. The Expectation score (E) is similar to Precision (P) but does not consider false positives. E_k represents the number of positive documents at the index k, whereas P_k represents the ratio of positive documents at the index k to k. Furthermore, Mean Average Precision (MAP) [7] is used to evaluate the ranking performance. MAP is calculated through the Average Precision (AP) metric, which is an average of precision scores only at the positive document indices. Let us consider that G is a set of all positive document indices with size g.

$$AP = (\sum_{i=1}^{G} P_i)/g$$

$$MAP = (\sum_{i=1}^{N} AP_i)/N$$

4.4 Survey evaluation

The survey proposed in this experiment is to evaluate the clustering output and also test the potential of new search query results. A new search query strategy still needs to be designed, and the number of survey inputs is still being determined, as this depends on the result of the precision evaluation.

5 Work Packages and Schedule

Pending

6 Scientific background

Many researchers have considered different techniques from Machine Learning (ML) to improve the retrieval results based on the availability of labeled data. The research can be categorized into three types: supervised and unsupervised.

6.1 Supervised approaches

Many researchers used ML algorithms with special loss functions based on relevance between the query, and documents and some of the popular pairwise ranking methods are RankBoost [8], RankNet [6], Rank-SVM [10] (using click-through data). Recent state-of-the-art supervised approaches are neural re-ranking methods and are based on complex Deep Learning (DL) architectures. Distributed word embeddings combined with the performance of non-linear neural networks have shown remarkable results in improving the performance of retrieval systems by considering semantics [17, 9, 18].

6.2 Unsupervised approaches

These approaches use no-labeled data and re-rank the retrieved results based on the user query and top retrieved documents [20, 3]. One common challenge in these approaches is the user query, which is mostly comprised of only a few keywords [3, 11]. To tackle this problem, many researchers have tested Query Expansion (QE) approaches that partially fill the missing meaning and context in the query. QE techniques include clustering search results, query filtering, word sense disambiguation, and relevance feedback, etc., [3]. Relevance Feedback is a method of retrieving search results using the original query given by the user and then using the top-k documents for query expansion [3]. Researchers have clustered search results in many different ways, such as at the document level, keyphrases, query-specific clustering, etc. [4, 12, 23, 19, 13, 14]. Typical distance-based clustering algorithms such as k-means are used in some research and also Hierarchical clustering is also tested [4, 16, 22], as it is flexible to change the threshold level for cutting the clustering dendrogram in a bottom-up approach.

A common drawback in most clustering approaches is mapping a document to a single cluster, which is not logically valid, as a document can contain keywords from different domains. The approaches based on clustering at the word level [4, 16] consider only a single language of retrieval results or corpus and hence cannot be directly implemented on a multilingual corpus. With the advantage of contextual embeddings from sentence encoders, the authors in [2] made a breakthrough in document clustering with an efficient and explainable topic-modeling approach.

6.3 Uniqueness in the proposed approach

In [15], authors have used a particular candidate selection approach to filter some phrases from the keyword extraction and a specific noun-chunks selection. This pipeline is explicitly used to extract innovation insights from research projects. As the user intention is related to *Innovation* at FKIE is proposed as a unique query-specific candidate keyword selection clustering. Moreover, the documents are semantically mapped to a specific topic, and multiple languages are modeled using a single multilingual pre-trained sentence encoder. News articles from multiple languages can be easily integrated into the document indices, and no changes are needed in the clustering pipeline.

Contextual document vectors are not well suited, as news articles are long text documents with an average text length of around 650 (text token length). Therefore, keywords are considered entities clustered into semantically similar groups rather than documents. This gives us the unique advantage of expressing each document as a combination of clusters (*sub-topics*). The sub-topics provide a detailed representation of documents and lead the user directly to the document pool according to their selection. The proposed approach can be further extended to analyze any corpus containing long text documents for a given phrase or keyword.

References

[1] U.S. Army CCDC Army Research Laboratory Public Affairs. *Army project may improve military communications by boosting 5G technology*. 2020. URL: `https://www.army.mil/article/230198/` (visited on 11/26/2019).

[2] Dimo Angelov. „Top2Vec: Distributed Representations of Topics". In: *CoRR* abs/2008.09470 (2020). arXiv: `2008.09470`. URL: `https://arxiv.org/abs/2008.09470`.

[3] Hiteshwar Kumar Azad and Akshay Deepak. „Query expansion techniques for information retrieval: a survey". In: *Information Processing & Management* 56.5 (2019), pp. 1698–1735.

[4] Andrea Bernardini, Claudio Carpineto, and Massimiliano D'Amico. „Full-subtopic retrieval with keyphrase-based search results clustering". In: *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*. Vol. 1. IEEE. 2009, pp. 206–213.

[5] David M Blei, Andrew Y Ng, and Michael I Jordan. „Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.

[6] Chris Burges et al. „Learning to rank using gradient descent". In: *Proceedings of the 22nd international conference on Machine learning*. 2005, pp. 89–96.

[7] Gordon V Cormack and Thomas R Lynam. „Statistical precision of information retrieval evaluation". In: *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. 2006, pp. 533–540.

[8] Yoav Freund et al. „An efficient boosting algorithm for combining preferences". In: *Journal of machine learning research* 4.Nov (2003), pp. 933–969.

[9] Jiafeng Guo et al. „A deep relevance matching model for ad-hoc retrieval". In: *Proceedings of the 25th ACM international on conference on information and knowledge management*. 2016, pp. 55–64.

[10] Thorsten Joachims. „Optimizing search engines using clickthrough data". In: *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2002, pp. 133–142.

[11] Ashish Kankaria. „Query Expansion techniques". In: 2015.

[12] Oren Kurland and Carmel Domshlak. „A rank-aggregation approach to searching for optimal query-specific clusters". In: *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. 2008, pp. 547–554.

[13] Xiaoyong Liu and W Bruce Croft. „Cluster-based retrieval using language models". In: *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. 2004, pp. 186–193.

[14] Xiaoyong Liu and W Bruce Croft. „Representing clusters for retrieval“. In: *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. 2006, pp. 671–672.

[15] Francesca Maridina Malloci et al. „A text mining approach to extract and rank innovation insights from research projects“. In: *International Conference on Web Information Systems Engineering*. Springer. 2020, pp. 143–154.

[16] Martin Mehlitz et al. „A new evaluation measure for information retrieval systems“. In: *2007 IEEE International Conference on Systems, Man and Cybernetics*. IEEE. 2007, pp. 1200–1204.

[17] Bhaskar Mitra, Fernando Diaz, and Nick Craswell. „Learning to match using local and distributed representations of text for web search“. In: *Proceedings of the 26th international conference on world wide web*. 2017, pp. 1291–1299.

[18] Rodrigo Nogueira and Kyunghyun Cho. „Passage Re-ranking with BERT“. In: *arXiv preprint arXiv:1901.04085* (2019).

[19] Stanislaw Osinski and Dawid Weiss. „A concept-driven algorithm for clustering search results“. In: *IEEE Intelligent Systems* 20.3 (2005), pp. 48–54.

[20] Shaurya Rohatgi, Jian Wu, and C Lee Giles. „PSU at CLEF-2020 ARQMath Track: Unsupervised Re-ranking using Pretraining.“ In: *CLEF (Working Notes)*. 2020.

[21] Yinfei Yang et al. „Multilingual universal sentence encoder for semantic retrieval“. In: *arXiv preprint arXiv:1907.04307* (2019).

[22] Meng Yuan, Justin Zobel, and Pauline Lin. „Measurement of clustering effectiveness for document collections“. In: *Information Retrieval Journal* (2022), pp. 1–30.

[23] Oren Zamir and Oren Etzioni. „Web document clustering: A feasibility demonstration“. In: *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. 1998, pp. 46–54.

[24] Nivio Ziviani et al. „Compression: A key for next-generation text retrieval systems“. In: *Computer* 33.11 (2000), pp. 37–44.