universität koblenz
weiter:denken

Fraunhofer FKIE

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 24, 2023

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

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

Contents

1 Introduction

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 for retrieval is proposed to extract highly coherent query-specific contexts (sub-topics) from the top retrieved documents, which helps the user immensely to narrow down the search space. Furthermore, the proposed approach will be evaluated using precision and survey analysis.

2 Research Questions (RQ)

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 problem 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. Both intrinsic and extrinsic clustering evaluation techniques, as well as a survey, are chosen to evaluate the clustering output.

RQ2: ***Which IR system retrieves more relevant documents for a user query and a sub-topic?***

When a user chooses a particular sub-topic cluster, it is assumed that the retrieval results related to the query and the sub-topic are shown. Two different IR systems are proposed in this master thesis to retrieve documents relevant to both the given user query and the chosen sub-topic. This research question targets comparing these two retrieval systems and determining the better one. A survey will be performed, and the collected data will be analyzed to answer the *RQ2*

RQ3: ***What is the effect of sub-topic ranking in finding the positive documents from the candidate pool?***

Showing only the documents related to a specific sub-topic can restrict the user from viewing the original 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 Background and Motivation

3.1 Fraunhofer FKIE

Fraunhofer FKIE (Fraunhofer-Institut für Kommunikation, Informationsverarbeitung und Ergonomie) is a 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.

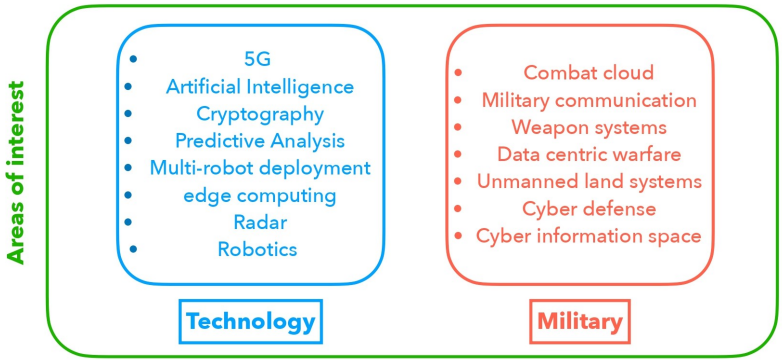


Figure 1: Areas of interest for the users at FKIE

3.2 Technical details

3.2.1 Abbreviations

1. **URL:** A URL is a short form for *Uniform Resource Locator* and is used to locate resources uniquely on the Internet [7]. Any resource on the Internet can be accessed with a unique URL. For example, the URL <https://www.linux.org/> represents a resource on the Internet.

2. **HTML:** HTML stands for *HyperText Markup Language* and is a markup language for representing documents on the World Wide Web (WWW) and links to other documents or information sources such as images, video, audio, etc [22].

3.2.2 Machine Learning (ML)

1. **Ground-truth:** Ground-truth labels are the information that is more accurate, relevant, and true than the knowledge of the system we are testing [11]. This information is critical to evaluate and compare different systems.

2. **Supervised learning:** Supervised learning algorithms are ML approaches that use labeled data [14] for training the algorithm parameters using specific criteria or a loss function. *Classification* is a supervised technique to learn patterns from the labeled data and classify the unseen data automatically into several classes.

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

3. **Unsupervised learning:** Clustering is an example of these algorithms, and similar data points are clustered into groups according to the features in the data [29]. These groups are called *Clusters*. Document clustering is a technique to group documents into topics without ground-truth information [15].

4. **Soft clustering:** In *Soft clustering*, the data points are assigned to one or more clusters by the clustering algorithm [15]. This depends mainly on the structure of the data, and especially in news articles, documents are assigned to multiple topics rather than one.

5. **Support Vector Machine (SVM):** SVM is a supervised learning algorithm used for classification and regression. Using the labeled data information, *SVM* selects a maximum margin separating hyperplane(a decision boundary) between the data points. This hyperplane is used later for classifying new data points [34].

6. **Tensorflow Hub:** Tensorflow Hub[2] is a repository of pre-trained ML models from *Tensorflow*[3]. Tensorflow is an open-source platform for ML that provides an ecosystem of tools and libraries and allows developers to build and deploy ML-powered apps and researchers to push state-of-the-art models [16].

3.2.3 Natural Language Processing (NLP)

1. **Token:** In NLP, a token is a word or basic entity in a text document, and Tokenization is the process of splitting a text document into tokens [39]. Document token length is calculated as the number of tokens present in a document.

2. **Noun chunks:** Noun chunks or phrases are the nouns and all the words that depend on these nouns [33]. Consider the sentence, *"Army project may improve military communications by boosting 5G technology"* [1]. The possible noun chunks extracted from this sentence are *"Army project"*, *"Military communications"*, *"5G technology"*.

3. **Keywords:** Keywords are the noun chunks that are highly meaningful in a text document and can best describe or summarize a document [5]. An unsupervised multi-lingual keyword extraction approach is used in the proposed approach to extract the most significant keywords from each news article.

4. **Sentence embeddings:** The distributed vector representation of a sentence or paragraph in a semantic space is generally referred to as sentence embeddings. These embeddings can also be generated with short phrases and noun chunks [12, 40].

5. **Universal sentence encoder (USE):** USE is an ML model that encodes text data such as sentences, phrases, or paragraphs into a distributed semantic vector. USE embeds text from sixteen different languages into a single semantic space [12, 40]. In this master thesis, a USE model from Tensorflow Hub[4] is used.

6. **Lexical matching:** Lexical or syntactic matching is a technique to assign a relevance score between two text data (strings) based on the terms present in the data. This matching technique is not optimal for retrieval, as it does not consider the meaning of the query [25].

7. **Semantic matching:** Semantic matching assigns a relevance score between the two text data by considering the semantic information (meaning of the terms).

[2] `https://www.tensorflow.org/hub`
[3] `https://www.tensorflow.org/`
[4] `https://tfhub.dev/google/universal-sentence-encoder-multilingual-large/3`

3.2.4 Information Retrieval (IR)

1. **Document retrieval system:** IR system specially developed to retrieve the document or text data for a given user query is generally referred to as a Document retrieval system.

2. **BM-25:** BM-25, Best Match 25, is a ranking function based on a probabilistic relevance framework that ranks documents based on the query terms occurring in each document [2]. BM-25 ranking is a lexical or syntactic matching approach and does not consider word semantics.

3. **Semantic search:** Unlike syntactic matching or calculating term frequencies, Semantic search engines try to understand the meaning of the search query and retrieve the matching documents close to the query in the semantic space [17].

3.2.5 Data storage

1. **Document index:** Document indexing or compression is a technique to store documents in an optimized way on the disk for efficient retrieval. This stored data on the disk is now referred to as *Document Index* [43].

2. **Inverted index:** The *Inverted index* is a data structure that contains every word in the corpus and the separate list of documents where the word occurs [43].

3. **Semantic search index:** The *Semantic search index* stores the distributed embedding vectors of the documents on the disk and uses them later for retrieval.

4. **SQLite DB:** SQLite is a lightweight serverless, self-contained, transactional database engine [8]. In this master thesis, labeled data are stored in *SQLite DB* using a library *sqlite3*[5].

3.2.6 Evaluation

1. **Intrinsic evaluation:** In case of no labeled data or ground-truth, the clustering output is evaluated through the methods considering only the inherent representation of clustered data [15]. These methods of evaluation are referred to as *Intrinsic evaluation*.

2. **Extrinsic evaluation:** In *Extrinsic evaluation*, the clustering output is evaluated using the external knowledge such as ground-truth or the relevance judgments [15].

3. **Silhouette index:** Irrespective of the clustering algorithm, the output is more distinctive when the distance between the data points within the cluster is minimum and the distance between the clusters is maximum. Silhouette index [38] is an intrinsic clustering evaluation measure and is calculated by using the intra-cluster and inter-cluster distances for each sample.

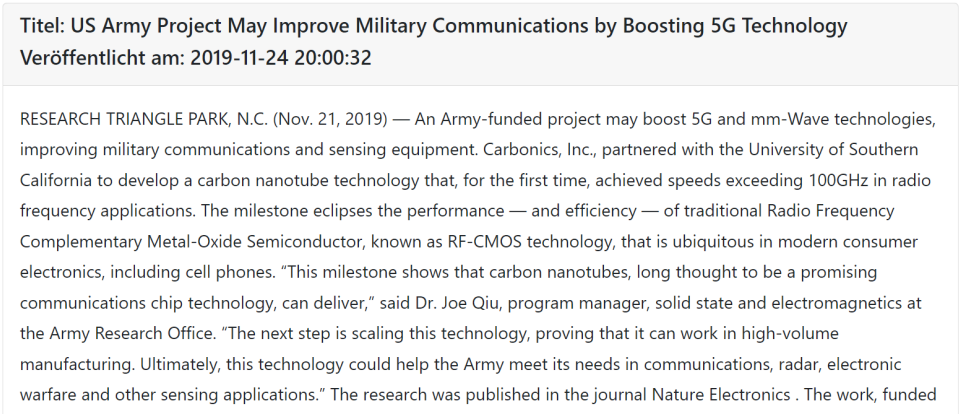
4. **Precision:** In IR system evaluation, Precision is defined as the ratio of retrieved documents that are relevant to all the retrieved documents [44]. This measure can be used to compare different IR systems and be calculated at different retrieved indices. For example, $P@5, P@10, P@15$ measures precision scores at retrieved indices $5, 10, 15$ respectively.

[5] `https://docs.python.org/3/library/sqlite3.html`

5. **Cosine similarity:** Cosine similarity is a metric to measure the degree of similarity between two vectors [26]. In the case of IR systems, the similarity is calculated between the user query and document sentence embeddings, and can be further used to rank the documents.

3.2.7 Keywords specific to this master thesis

1. **News article:** A news article is a text document published by a news website. An example of a news article (this is only a part of the original article) is shown in Figure 2.


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

2. **Web scraping:** Web scraping is a technique to automatically extraction of data from websites [23]. In the case of text data, most approaches download the structured HTML webpages and extract needed information. In this master thesis, news articles from different websites are scraped.

3. **Candidate pool:** A candidate or retrieval pool is a set of documents from lexical and semantic matching results for a given user query. These documents are very diverse, contain keywords present in the query (or semantically similar), and are further used for clustering.

4. **Sub-topic:** Sub-topics are second-level representations of a document. Generally, news articles are long text documents and can not be represented logically with a single topic or keyword. If the user query provided to the *Retriever* is considered the main topic, then the distinctive topics extracted from the candidate pool are sub-topics.

5. **Context:** In this master thesis, we define a context as a particular domain or field in which the user is interested. For example, in the user query *Cloud*, the retrieved documents are related to different domains or contexts, such as cloud computing, combat cloud, and clouds in the sky. Even though there is some syntactic and semantic matching, the user intention is still unclear from the query.

6. **Labeler:** A person who assigns an appropriate label to the data according to the labeling criteria.

7. **Query type:** Input search queries from the user can be of any form. For example ., abbreviation, single word query, etc. In this master thesis, each form of a possible user query is referred to as a query type. All possible user queries can be categorized into two major query types, namely phrase (three words or less) and sentence queries.

3.3 Background and retrieval setup

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 3. 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 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*.

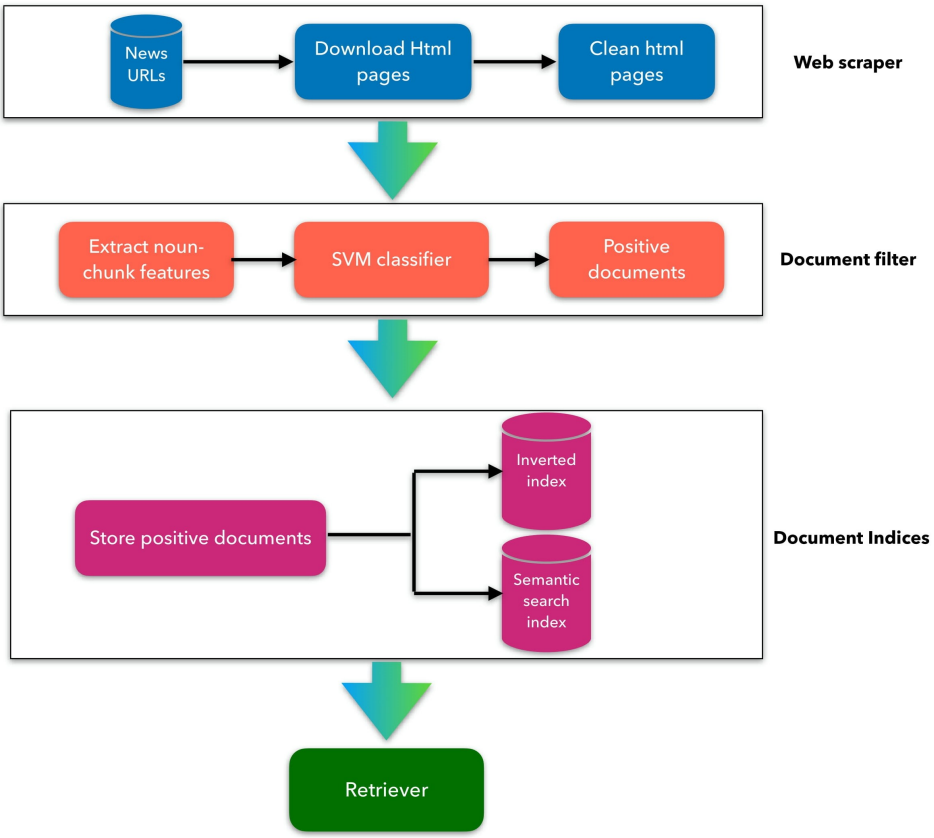


Figure 3: Document retrieval system designed at Fraunhofer FKIE

The second component, the *Document filter*, is based on 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 are related to everything else. After several tests at FKIE, it was found that features from noun-chunks in a document are performing better to differentiate *Positive* documents from the *Negative* documents. Noun chunk features based on the pre-trained multi-lingual Universal Sentence Encoder (USE) is used for the task of classification.

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*. Finally, the compo-

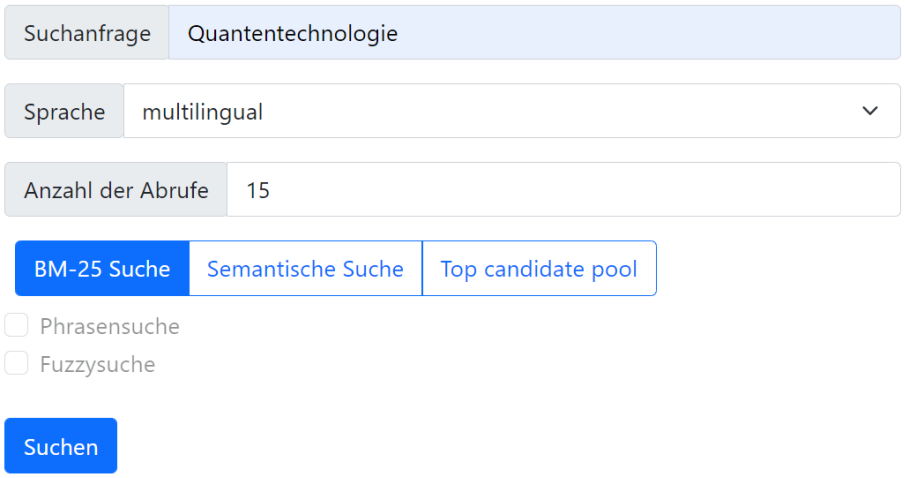


Figure 4: User interface to retrieve documents for FKIE users

nent *Retriever* uses both of the indices and retrieves documents according to the user request through a web user interface, as shown in Figure 4.

3.4 Challenges

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).

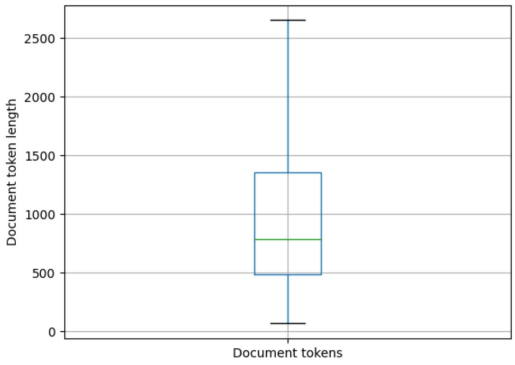
Table 1: Retrieval algorithms comparison on different query types

S No.	Query type	Better retrieval technique	Reason	Queries used
1	No meaning queries	BM-25	Lexical matching	Person or object names[6]
2	Multi-lingual queries	Semantic search	Semantic matching	Artificial Intelligence vs Künstliche Intelligenz
3	German composite words	Semantic search	Semantic matching	Quantentech-nologie
4	Spelling mistakes	Semantic search	Semantic matching	Kryptografy, Rbot
5	Polysemy	Semantic search	Semantic matching	Combat Cloud, Cloud computing
6	Sentence/long phrase queries	Semantic search	Semantic matching	Schwachstell-enanalyse eigene Waffen-Systeme

[6]User query with no innate meaning of the word namely out of vocabulary words: for example John Dowe, Wester etc.

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. In [25], the authors observed a similar challenge in their research. On the one hand, lexical matching does not consider the inherent meaning of the word causing a vocabulary mismatch problem, and semantic matching fails to retrieve the relevant documents in the top results as it matches too many keywords semantically. 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 9.

The users at FKIE provide only one or two phrase queries, and his or her intention is to explore information to specific topics such as *"Technology"* and *"Military"*. Without labeled data, learning user intention from a single word or phrase query is a huge challenge. 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 the 50% (percentile) token length of 788 and consist of keywords from multiple domains. Document token length details is shown in Figure 5.



(a)

S No	Statistical parameter	Value
1	Minimum token length	69
2	Mean token length	1280.64
3	Maximum token length	122657
4	Token length at 25%	489
5	Token length at 50%	788
6	Token length at 75%	1355

(b)

Figure 5: (a) Bloxplot showing the document token length distribution(after removing the outliers) (b) Important statistical details about the token length

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 reinforcement 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*. An option to update the template according to the user's interest from the user interface can provide tailored results without any extra training. 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.

After considering various approaches to fulfill the missing context, we believe that extract-

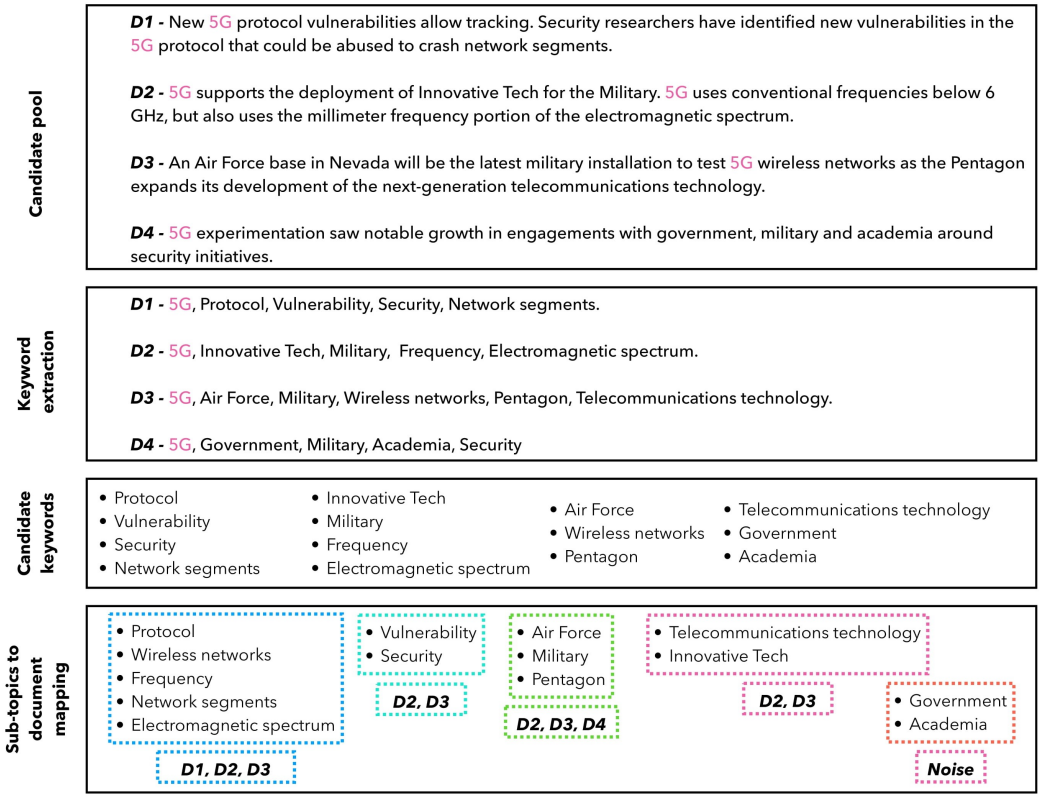


Figure 6: Expected sub-topic extraction for the query: 5G

ing the contexts from top results to the user query in an unsupervised way is more efficient, more explainable and can be better reproducible compared to supervised approaches. 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 expected sub-topic extraction pipeline output is shown in Figure 6. These contexts are described as *sub-topics*. The proposed approach in this master thesis is aimed at handling the challenges mentioned above. News articles from diverse sources are considered, and the results can be easily transferred to other data sources in the future.

4 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 does not provide any new insights 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 [9]. 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 7 illustrates the proposed approach on an abstract level.

The proposed approach, shown in Figure 7, 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

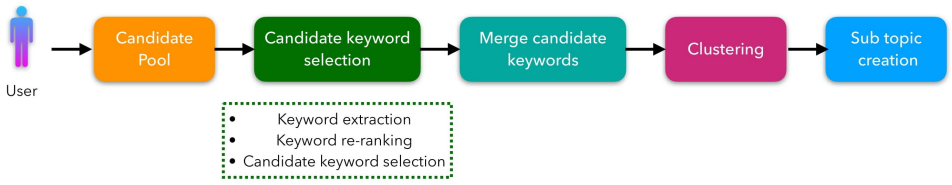


Figure 7: Proposed approach on an abstract level

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*.

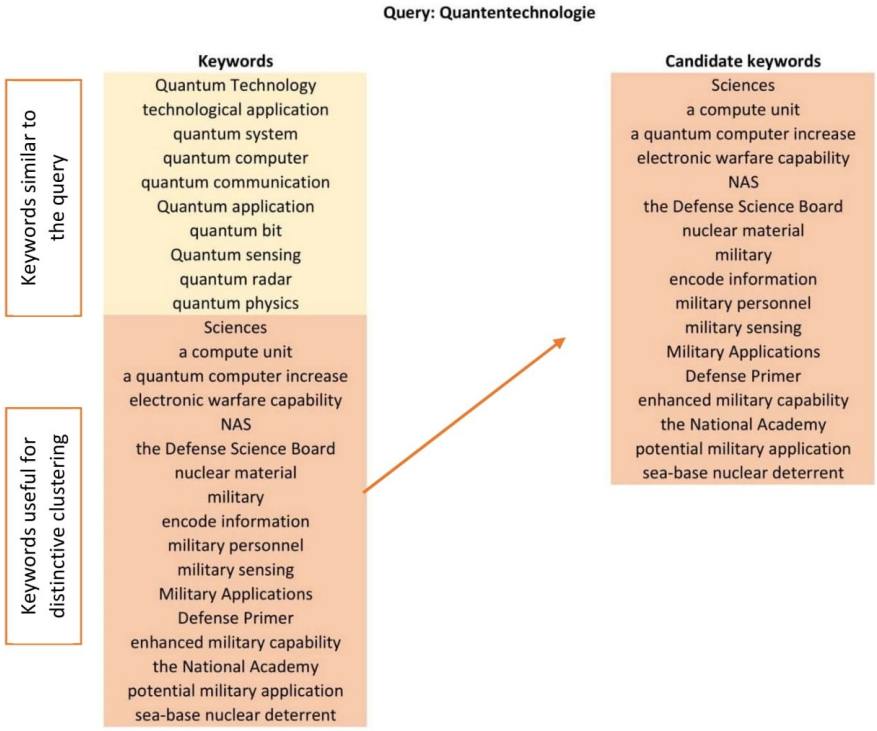


Figure 8: Candidate keyword selection step from a single document

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 cut-off threshold and this process is referred to as *Candidate keyword selection* and the resulting phrases after this stage are called Candidate keywords, shown in Figure 8.

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.

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*.

5 Evaluation

5.1 Testset

For two main reasons, a dataset specific to this research problem is hard to find in the current IR data repositories. Foremost, 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.

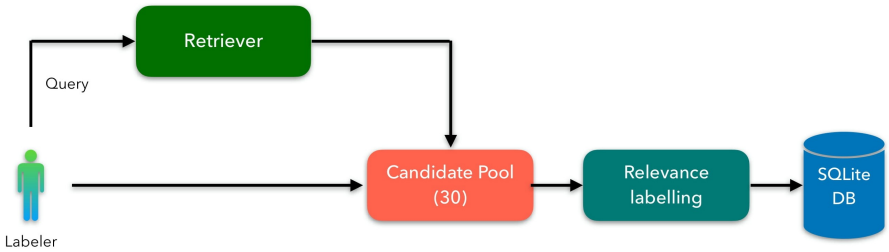


Figure 9: Testset collection strategy

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 to a merged set where duplicate documents are eliminated. The image Figure 9 shows the methodology followed for the testset collection. *Relevance labeling*, is a process to assign 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.

5.2 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. 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.

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 14, 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 document with label-id 1 or documents with 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.

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

Both the target functions *Silhouette index* and F are used to tune the parameters of the sub-topic modeling pipeline. *Silhouette index* evaluates the clustering output and the custom target function F evaluates the distribution of relevant labels. An automatic parameter tuning

can lead to very small clusters and there is a possibility of documents being marked as noise. Therefore, A manual evaluation of these two metrics will be considered to finalize the pipeline parameters.

5.3 Survey evaluation

The survey proposed evaluates the clustering output and also tests the performance of new search query results for a given query and sub-topic. Below two IR systems are proposed and further compared using the survey data.

1. **System A** is an IR system that retrieves documents using a new search query based on the original query and the chosen sub-topic.
2. **System B** is an IR system that retrieves documents from the sub-topic clusters from the sub-topic modeling output.

A new search query strategy still needs to be designed, and the number of survey inputs is still being determined.

5.4 Precision evaluation

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 evaluation approach proposes six different retrieval systems and evaluates the ranked results.

Table 3: 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 [31], 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) [13] 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 index with size g.

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

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

5.5 Evaluation summarization

The table Table 4 on page 16 shares the evaluation techniques chosen in this thesis and the respective research questions answered.

Table 4: Proposed evaluation techniques

S No.	Evaluation type	Research questions addressed
1	Clustering	**RQ1**
2	Survey	**RQ1, RQ2**
3	Precision analysis	**RQ3**

6 Work Packages with Schedule

Work packages with deadlines is shown in Figure 10

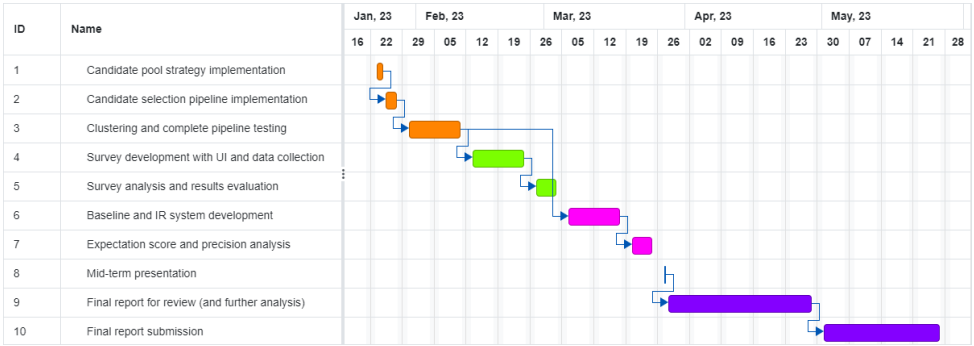


Figure 10: Work packages and schedule

7 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 two types: supervised and unsupervised.

7.1 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 [18], RankNet [10], Rank-SVM [20] (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 [32, 19, 35].

7.2 Unsupervised approaches

These approaches use no-labeled data and re-rank the retrieved results based on the user query and top retrieved documents [37, 4]. One common challenge in these approaches is the user query, which is mostly comprised of only a few keywords [4, 21]. 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., [4]. 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 [4]. Researchers have clustered search results in many different ways, such as at the document level, keyphrases, query-specific clustering, etc. [6, 24, 42, 36, 27, 28]. Typical distance-based clustering algorithms such as k-means are used in some research and also Hierarchical clustering is also tested [6, 31, 41], 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.

7.3 Uniqueness in the proposed approach

The approaches based on clustering at the word level [6, 31] consider only a single language of retrieval results or corpus and hence cannot be directly implemented on a multi-lingual corpus and does not have any special keyword selection stage. With the advantage of contextual embeddings from sentence encoders, the authors in [3] made a breakthrough in document clustering with an efficient and explainable topic-modeling approach.

In [30], 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. 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] Giambattista Amati. „BM25". en. In: *Encyclopedia of Database Systems*. Ed. by LING LIU and M. TAMER ÖZSU. Boston, MA: Springer US, 2009, pp. 257–260. ISBN: 978-0-387-39940-9. DOI: 10.1007/978-0-387-39940-9_921. URL: https://doi.org/10.1007/978-0-387-39940-9_921 (visited on 01/17/2023).

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

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

[5] Slobodan Beliga. „Keyword extraction: a review of methods and approaches". In: *University of Rijeka, Department of Informatics, Rijeka* 1.9 (2014).

[6] 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.

[7] Tim Berners-Lee, Larry Masinter, and Mark McCahill. *Uniform resource locators (URL)*. Tech. rep. 1994.

[8] ST Bhosale, Tejaswini Patil, and Pooja Patil. „Sqlite: Light database system". In: *Int. J. Comput. Sci. Mob. Comput* 44.4 (2015), pp. 882–885.

[9] 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.

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

[11] Jefferson Rosa Cardoso et al. „What is gold standard and what is ground truth?" In: *Dental press journal of orthodontics* 19 (2014), pp. 27–30.

[12] Daniel Cer et al. „Universal sentence encoder". In: *arXiv preprint arXiv:1803.11175* (2018).

[13] 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.

[14] Kushal Rashmikant Dalal. „Analysing the Role of Supervised and Unsupervised Machine Learning in IoT". In: *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*. 2020, pp. 75–79. DOI: 10.1109/ICESC48915.2020.9155761.

[15] Christopher M De Vries, Shlomo Geva, and Andrew Trotman. „Document clustering evaluation: Divergence from a random baseline". In: *arXiv preprint arXiv:1208.5654* (2012).

[16] TensorFlow Developers. *TensorFlow*. Version v2.8.2. Specific TensorFlow versions can be found in the "Versions" list on the right side of this page.
See the full list of authors on GitHub. May 2022. DOI: 10.5281/zenodo.6574269. URL: https://doi.org/10.5281/zenodo.6574269.

[17] Hai Dong, Farookh Khadeer Hussain, and Elizabeth Chang. „A survey in semantic search technologies“. In: *2008 2nd IEEE international conference on digital ecosystems and technologies*. IEEE. 2008, pp. 403–408.

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

[19] 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.

[20] 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.

[21] Ashish Kankaria. „Query Expansion techniques“. In: 2015.

[22] Eric Kasten. „HTML“. In: *Linux J.* 1995.15es (July 1995), 3–es. ISSN: 1075-3583.

[23] Moaiad Ahmad Khder. „Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application.“ In: *International Journal of Advances in Soft Computing & Its Applications* 13.3 (2021).

[24] 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.

[25] Saar Kuzi et al. „Leveraging semantic and lexical matching to improve the recall of document retrieval systems: A hybrid approach“. In: *arXiv preprint arXiv:2010.01195* (2020).

[26] Alfirna Rizqi Lahitani, Adhistya Erna Permanasari, and Noor Akhmad Setiawan. „Cosine similarity to determine similarity measure: Study case in online essay assessment“. In: *2016 4th International Conference on Cyber and IT Service Management*. IEEE. 2016, pp. 1–6.

[27] 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.

[28] 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.

[29] Batta Mahesh. „Machine learning algorithms-a review“. In: *International Journal of Science and Research (IJSR).[Internet]* 9 (2020), pp. 381–386.

[30] 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.

[31] 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.

[32] 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.

[33] nlpcloud.com. *Noun Chunks / Noun Phrases API*. 2020. URL: https://nlpcloud.com/nlp-noun-chunks-noun-phrase-extraction-api.html (visited on 12/01/2020).

[34] William S Noble. „What is a support vector machine?“ In: *Nature biotechnology* 24.12 (2006), pp. 1565–1567.

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

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

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

[38] Peter J Rousseeuw. „Silhouettes: a graphical aid to the interpretation and validation of cluster analysis“. In: *Journal of computational and applied mathematics* 20 (1987), pp. 53–65.

[39] Jonathan J Webster and Chunyu Kit. „Tokenization as the initial phase in NLP“. In: *COLING 1992 volume 4: The 14th international conference on computational linguistics*. 1992.

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

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

[42] 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.

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

[44] Keneilwe Zuva and Tranos Zuva. „Evaluation of information retrieval systems“. In: *International journal of computer science & information technology* 4.3 (2012), p. 35.