

# Domain Context-centered Retrieval for the Content Selection task in the Simplification of Scientific Literature

Notebook for the Simple Text Lab at CLEF 2023

Óscar E. Mendoza, Gabriella Pasi

*University of Milano Bicocca*

## Abstract

The DoSSIER team from the University of Milano Bicocca participated in the Simple Text lab from Clef 2023. We focus on analysing the distribution of topics (or themes) for the content selection task and its potential effects in retrieval. We introduce Domain-knowledge through an ontology with different classification algorithms for annotating documents. This allows us to study the diversity in retrieval results and show how constraining them positively impacts the information-gathering task. We constrain results to concentrate them to specific sets of themes built using pseudo-relevance feedback.

## Keywords

Retrieval, Topic analysis, Information gathering, Document classification

## 1. Introduction

The DoSSIER team from the University of Milano Bicocca participated in the Simple Text lab of Clef 2023. The lab develops around promoting scientific information access through retrieval, mining, and simplification of scientific literature [1]. We focus on analysing the distribution of topics in the collection for the content selection retrieval task.

In practical terms, the task has been defined as retrieving all passages from a large corpus of scientific abstracts and bibliographic metadata relevant to given topics [2]. Relevant abstracts should be related to specific themes of the topics. Since this task deals with scientific literature, the retrieved passages are likely to be complex and may require further simplification, covered in the other instances of the Simple Text lab.

The topics for this task are a selection of press articles from a major international newspaper for a general audience and from Tech Xplore. Queries are keyword-based, manually extracted from topics based on the fact they allow retrieving some relevant passages from the given collection that could be inserted as citations in the press article.

This task has relevant nuances that we take into account. Based on the previous motivation, queries are given under an informational intent [3], i.e., by submitting them to the search system, we want to obtain some related information which is assumed to be available in the

---


*CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece*

✉ o.espitiamentodoza@campus.unimib.it (Ó. E. Mendoza); gabriella.pasi@unimib.it (G. Pasi)

🆔 0000-0003-2725-2972 (Ó. E. Mendoza); 0000-0002-6080-8170 (G. Pasi)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

collection. Additionally, we aim to solve a task of information gathering, which involves collecting information from multiple sources [4].

The task also exhibits domain-specific properties, it gives the opportunity to explore domain-based information sources. In particular, domain-specific knowledge is often encoded in ontologies that can be used for document annotation and, in the context of search, it has the potential to constrain the information space to find relevant documents more effectively than in open-domain applications.

Because of the nature of the task, we focus on broadly analyzing document themes. We introduce contextual knowledge to the task through a large-scale ontology of research topics. This allows us to have feedback on how the information available, and presented as an answer to a query, distributes according to the themes, and then to constrain the results based on these distributions. We specifically investigate the following research questions:

RQ1 How a collection of abstracts responds to specific information needs for the content selection task?

RQ2 To what extent can the topic the analysis enhance retrieval for the content selection task?

RQ3 Can the constraining of topics influence the performance of the retrieval system?

## 2. Materials and Methods

This section describes the methods used to perform our experiments. It mainly consists of a probabilistic lexical ranking, hierarchical document classification, and pseudo-relevance feedback (PRF).

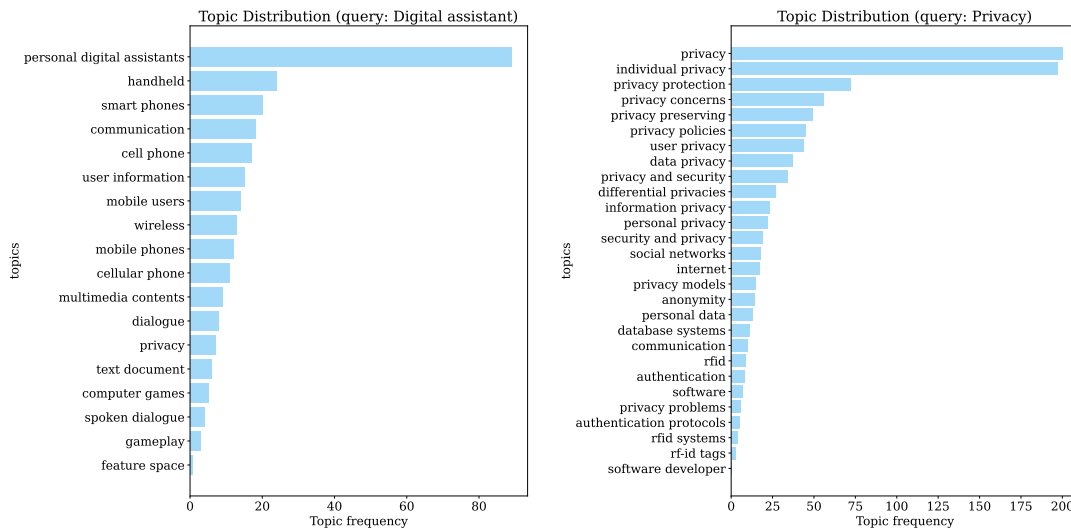
### 2.1. Ad-hoc Retrieval

In order to study how the information from the collection can meet the information needs, we first retrieved the documents following a probabilistic lexical ranking stage. We retrieve a number of documents from the collection using the given queries. Each set of retrieved documents allows us to study the distribution of topics that the collection gathers, and specifically those that are selected as a response to a given query. To offer a description of our approach, we will use two queries from the given set: “*Digital assistance*” corresponding to the topic titled “*Digital assistants like Siri and Alexa entrench gender biases says UN*”, and “*privacy*” corresponding to the topic titled “*Apple contractors ‘regularly hear confidential details’ on Siri recordings*.”

### 2.2. Hierarchical Classification

We incorporate domain knowledge through a large-scale ontology and use it for representing the retrieved documents and for studying the themes they are centered on.

On the one hand, we use the classifier proposed by [5], which uses lexical matching between terms in the candidate documents and the ontology, and trained static embeddings to infer semantically related topics from terms identified as informative, using part-of-speech tagging. Figure 1 shows the frequency of themes given a set of retrieved documents with the example queries mentioned previously. We can observe, for instance, that the query “*privacy*” matches



**Figure 1:** Frequency of themes retrieved by the query.

documents mostly about “*personal digital assistants*,” which is a potential keyword-based query for searching within the collection.

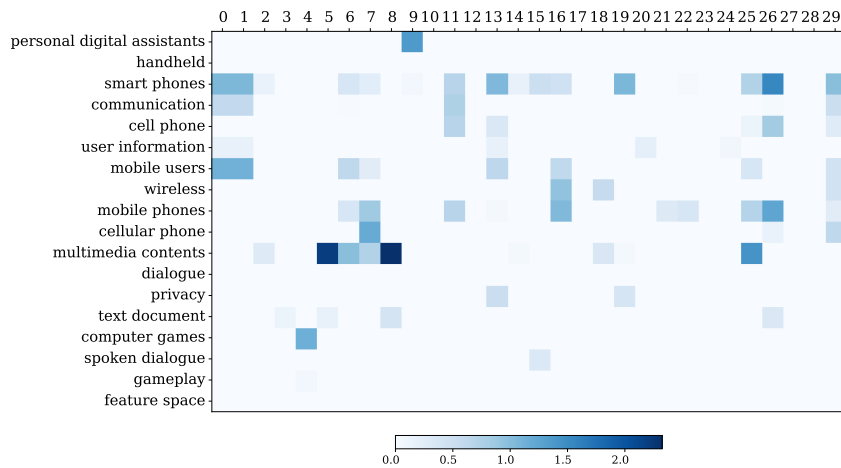
On the other hand, we train a sequence-to-sequence model [6] to generate themes from the ontology in a weakly supervised fashion and then use beam search to generate multiple themes so that we can consider the distribution themes described above. We use the score assigned by the decoding algorithm to the generated themes to display how likely a retrieved document will be assigned to a specific set of themes. In our case study, Figures 2 and 3 are heatmaps for the topic distribution of the first 30 ranked documents retrieved using the examples queries (each row then shows how likely each document is to be annotated with a theme). These two examples show the contrast of possible results the collection can offer to specific queries.

We aim to analyze the distribution of themes the collection can provide for a given query and study the effect of tailoring the search results according to specific patterns. Being able to assign granular topics to the documents, we exploit this tool in further steps.

### 2.3. Contextual PRF and Content Selection

The distribution and the probabilities of themes for the retrieved documents could benefit the content selection process. From the retrieval side, we try a simple instance for exploiting this information. Specifically, given the distribution of topics from the candidates, we first use the most frequent topics to expand the queries. Such that. e.g., “*digital assistance*” is extended by “*personal digital assistants*” and “*privacy*” is extended by “*individual privacy*.” This is an instance of PRF, which traditionally aims to leverage the most relevant documents retrieved by the initial query to improve the subsequent query. Here the difference is that we do not use the documents directly but their themes.

Furthermore, we use the distribution of topics to re-rank the documents and select specific



**Figure 2:** Theme prediction scores for the query “personal assistant.”

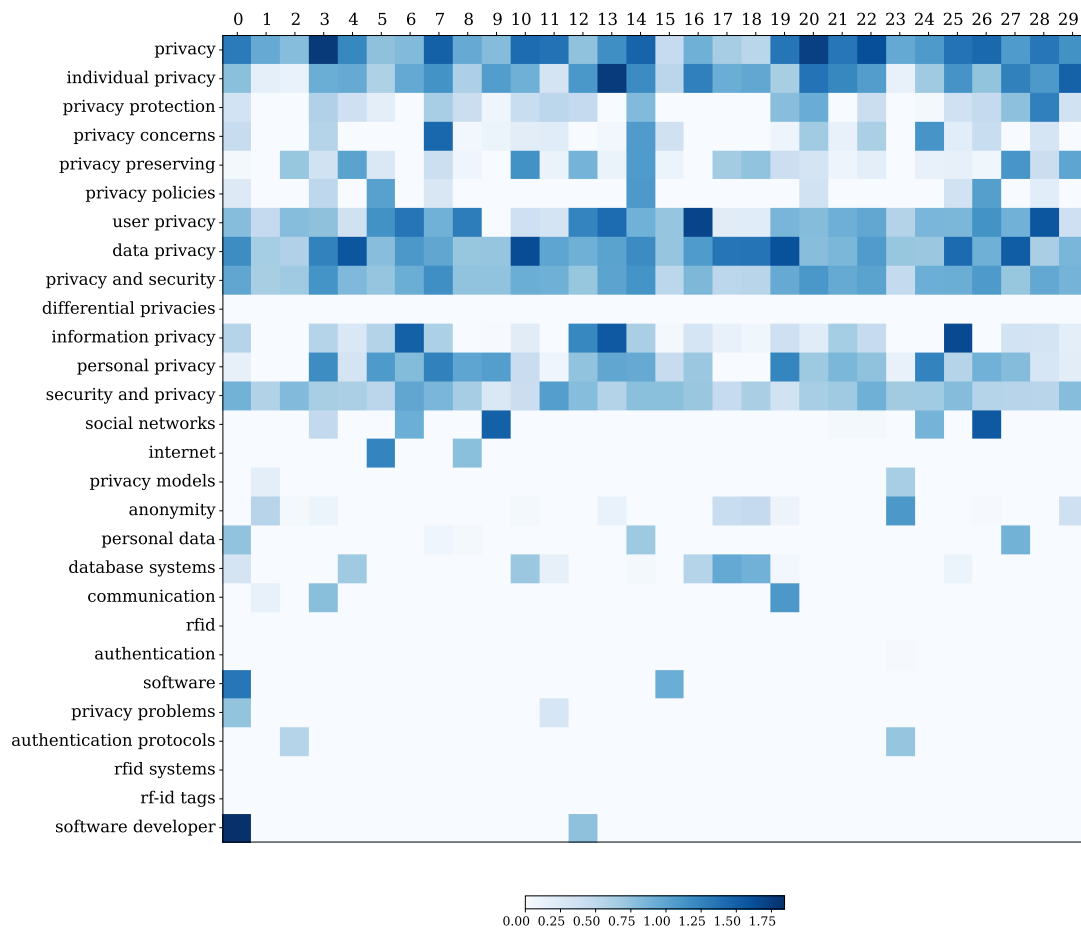
sentences from each document based on the diversity of the themes. We hypothesize that establishing relevant themes, i.e., the most frequent among the candidates retrieved by the search system, the task of content selection could benefit from trimming down to a more focused or concentrated set of documents. Thus, we penalize documents based on the diversity of themes they exhibit and re-rank them based on the same criterion.

We can see how documents are relatively diverse by looking at the different distributions, such as those described for the example queries. They are ranked around a specific keyword but still exhibit diversity that does not necessarily help to achieve the task of gathering supporting content. Considering this, we then look at themes at a different level of granularity of the retrieved documents to decide whether specific snippets are even more concentrated in the relevant themes.

Figure 4 shows examples of content selection based on the diversity of themes. For the example queries, we infer the topics highlighted are relevant from its themes distribution feedback. Ideally, the selected content should focus on those topics since dispersed content does not show a clear importance to the task.

### 3. Experiments and Results

In this section, we present the results of the experiments described previously. We are mainly interested in analyzing the distribution of themes in the collection, and in understanding retrieval results and how to make them useful for the aim of content selection for simplification rather than focusing on retrieval performance. However, state-of-the-art retrieval models still rely on ad-hoc retrieval models to limit the number of documents to be processed in later stages (pipeline-based models). Thus, it makes sense to study the appropriateness of the set of documents retrieved first to be processed and re-ranked in the later stages.



**Figure 3:** Theme prediction scores for the query “privacy.”

We use BM25 for retrieval, the CS ontology as domain knowledge, and the different strategies described in section 2.2 for classification.

We analyze retrieved sets of documents in terms of diversity of themes and entropy of their distribution. Intuitively, a higher entropy, which directly indicates more uncertain, implies a wider range of possible themes [7]. From the size of the ontology, we derived that the entropy of the distribution of themes in the collection is around 13. We limit the experiments to the top 200 retrieved documents, and the average entropy of the distribution of themes is 6.83 on these sets.

As discussed before, in general, we are interested in concentrated results, implying lower entropy values and lower diversity.

Identifying the relevant themes seems to have a positive effect. There is a clear difference between retrieving documents with the original queries (Retrieval in Table 1) and using extensions from relevant themes (PRF in Table 1). On average, documents are annotated with 8

**Topic:** Digital assistants like Siri and Alexa entrench gender biases says UN  
**Query:** Digital assistance

**Candidate:** Mobile devices are significantly changing the human-computer interaction. In particular, the ubiquitous access to remote resources is one of the most interesting characteristics achievable by using mobile devices such as Personal Digital Assistants, cellular phones and tablets. This paper presents an architecture that allows users to search and visualize complex 3D models over Personal Digital Assistants. A peer-to-peer network of brokers manages queries for searching objects among several data providers. The object selected for visualization is forwarded to a specialized graphics provider; this

Personal Digital Assistants  
 Mobile Devices  
 Smart Phones  
 Cellular Phone  
 Human Computer Interaction  
 3d Modelling  
 Visualization  
 Personal Digital Assistants  
 Cell PhonePeer-To-Peer

**Topic:** Apple contractors 'regularly hear confidential details' on Siri recordings.  
**Query:** Privacy

**Candidate:** Privacy awareness is a core determinant of the success or failure of privacy infrastructures: if systems and users are not aware of potential privacy concerns, they cannot effectively discover, use or judge the effectiveness of privacy management capabilities. Yet, privacy awareness is only implicitly described or implemented during the privacy engineering of software systems. In this paper, the author advocates a systematic approach to considering privacy awareness. He characterizes privacy awareness and illustrate its benefits to preserving privacy in a smart mobile environment. The author proposes privacy awareness requirements to anchor the consideration of privacy awareness needs of software systems...

Privacy  
 Individual Privacy  
 Privacy Concerns  
 Privacy Management  
 Privacy And Security  
 Privacy  
 Individual Privacy  
 Mobile Environments  
 Software  
 Engineering  
 Software Systems

■ Selected content

**Figure 4:** Example of content selection at document level using theme diversity.

**Table 1**

Theme concentration analysis.

Experiment	Diversity		Entropy	
	@20	@50	@20	@50
Retrieval	0.4740	0.3222	5.0224	5.5152
PRF	0.4213	0.2812	4.7444	5.1556
Re-ranking	<b>0.4145</b>	<b>0.2791</b>	<b>4.6496</b>	<b>5.0942</b>

themes, so it is difficult to measure the effect of constraining the search based on diversity and we can perceive only slight change, but positive, when trying to constrain results by penalizing diversity on documents (re-ranking in Table 1).

In terms of retrieval performance, Analyzing topics also exhibits a positive impact. Table 2 show the retrieval evaluation in terms of multiple metrics, including additional metrics regarding the content selection in terms of the length of the text and the readability (FKGL).

**Table 2**

Retrieval evaluation.

	MRR	P@10	NDCG	Bpref	MAP	Length	FKGL
Retrieval	0.4536	0.1912	0.2192	0.1384	0.0515		
Re-ranking	0.5201	0.2853	0.2980	0.1898	0.1141	1024.48	14.77
Content-selection	0.5202	0.2853	0.2972	0.1873	0.1111	238.63	15.11

## 4. Discussion

**Table 3**

Relevant Themes for a sample of queries.

Query	Relevant themes
Digital assistant	personal digital assistants, handheld
Biases	correlation analysis, sensors
self driving	vehicles, autonomous driving
humanoid robots	robots, humanoid robot
online safety for children	internet, education
cookies	privacy, web content
light positioning	positioning system, indoor positioning
intelligent parking	vehicles, sensors
emotional robot	robots, emotional expressions
empathy	emotional expressions, affective state
text classification	text classification, classification models
character relationship	character recognition, computer games
gene editing	http, database systems
conspiracy theories	signature scheme, facebook
healthcare	communication, information systems

- RQ1 The collection gathers a wide variety of topics. It is clear that for specific topics, retrieval can result in a very focused set of documents, which is the case exemplified in Figure 3. Whereas for other topics, the distribution does not exhibit a specific theme concentration (see Figure 2). We show a sample queries matching the themes found through the PRF. A qualitative evaluation shows that, in general, the feedback from the collection is closely related to the queries with some exceptions, such as “gene editing” and “healthcare” (see Table 3), definitely out of the domain of the task.
- RQ2 Even though the results have the potential to be tailored to different information needs, documents gather a wide variety of topics individually, so a more granular evaluation should be performed to decide how to incorporate the information from the distribution of topics into the retrieval system, supporting the idea of sentence selection.
- RQ3 As we hypothesize, constraining topics can contribute to the performance of the retrieval system. Table 2 shows results on retrieval performance, and we can see how re-ranking based on the variability of themes improves the overall performance. To gain a deeper understanding and draw more concrete conclusions, it would be beneficial to conduct additional evaluations and analyses, considering various factors such as content selection, which has no clear effects on our current results, as well as alternative retrieval and re-ranking approaches, which we will consider in our future work.

## Acknowledgements

This work was supported by the EU Horizon 2020 ITN/ETN on Domain Specific Systems for Information Extraction and Retrieval – DoSSIER (H2020-EU.1.3.1., ID: 860721).

## References

- [1] L. Ermakova, E. SanJuan, J. Kamps, S. Huet, I. Ovchinnikova, D. Nurbakova, S. Araújo, R. Hannachi, E. Mathurin, P. Bellot, Overview of the clef 2022 simpletext lab: Automatic simplification of scientific texts, in: *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5–8, 2022, Proceedings*, Springer, 2022, pp. 470–494.
- [2] E. SanJuan, S. Huet, J. Kamps, L. Ermakova, Overview of the clef 2022 simpletext task 1: Passage selection for a simplified summary (2022).
- [3] Z. Dou, J. Guo, *Query Intent Understanding*, Springer International Publishing, Cham, 2020, pp. 69–101.
- [4] M. Kellar, C. Watters, M. Shepherd, A field study characterizing web-based information-seeking tasks, *Journal of the American Society for information science and technology* 58 (2007) 999–1018.
- [5] A. Salatino, F. Osborne, E. Motta, Cso classifier 3.0: a scalable unsupervised method for classifying documents in terms of research topics, *International Journal on Digital Libraries* (2022) 1–20.
- [6] I. Sutskever, O. Vinyals, Q. V. Le, Sequence to sequence learning with neural networks, *Advances in neural information processing systems* 27 (2014).
- [7] L. M. Aiello, G. Petkos, C. Martin, D. Corney, S. Papadopoulos, R. Skraba, A. Göker, I. Kompatsiaris, A. Jaimes, Sensing trending topics in twitter, *IEEE Transactions on multimedia* 15 (2013) 1268–1282.