Expert Systems With Applications 199 (2022) 116967

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Information retrieval from scientific abstract and citation databases: A query-by-documents approach based on Monte-Carlo sampling

Fabian Lechtenberg [a], Javier Farreres [b], Aldwin-Lois Galvan-Cara [a], Ana Somoza-Tornos [a,c], Antonio Espuña [a], Moisès Graells [a,*]

[a] Department of Chemical Engineering, Universitat Politècnica de Catalunya, Campus Diagonal-Besòs, Eduard Maristany, 16, 08019 Barcelona, Spain
[b] Computer Science Department, Universitat Politècnica de Catalunya, Campus Diagonal-Besòs, Eduard Maristany, 16, 08019 Barcelona, Spain
[c] Renewable and Sustainable Energy Institute (RASEI), University of Colorado Boulder, Boulder, Colorado 80303, United States

ARTICLE INFO

Keywords:
Systematic literature review
Decision-making support
Recommender system
Monte-Carlo sampling
Knowledge management

ABSTRACT

The rapidly increasing amount of information and entries in abstract and citation databases steadily complicates the information retrieval task. In this study, a novel query-by-document approach using Monte-Carlo sampling of relevant keywords is presented. From a set of input documents (seed) keywords are extracted using TF-IDF and subsequently sampled to repeatedly construct queries to the database. The occurrence of returned documents is counted and serves as a proxy relevance metric. Two case studies based on the Scopus® database are used to demonstrate the method and its key advantages. No expert knowledge and human intervention is needed to construct the final search strings which reduces the human bias. The methods practicality is supported by the high re-retrieval of seed documents of 7/8 and 26/31 in high ranks in the two presented case studies.

1. Introduction

Advances in communication technologies enable researchers from every part of the world to share information with peers. As a result, a 9% growth of yearly published articles in academic journals has been recorded (Landhuis, 2016). On one hand, this ever-increasing volume of accessible information and knowledge can be reused for solving problems and supporting decision-making. On the other hand, the higher volume of information also implies an increased effort to find and utilize it. Hence, well performing information retrieval (IR) systems are key to aid the query formulation and facilitate the search of relevant information within the big data.

Scientific abstract and citation databases, such as Scopus® or Web of Science, are large indexes of abstracts and metadata that can be sampled by user defined query strings. However, researchers report difficulties in finding appropriate combinations of keywords to construct a corpus that properly responds to their research question (Mergel et al., 2015).

Query-by-document (Yang et al., 2009) is an information retrieval approach that relies on example documents that satisfy the user's information need. While a human may have difficulties to extract and connect the most important keywords to find and retrieve further similar documents related to the topic or question, information systems can detect the most relevant keywords and connect them to adequate queries.

This work presents a novel Query-by-Document (QbD) method that can be applied to access-restricted scientific abstract and citation databases. The proposed procedure makes use of a feature vector representation of seed documents via a bag of words approach (TF-IDF). Based on this weighted feature vector, a Monte-Carlo sampling strategy is applied to repeatedly construct query strings from the previously identified keywords and automatically execute the query using the Application Programming Interface (API) of the database. This new methodology not only avoids the need of an expert decision when constructing query strings but it also avoids the possible bias that the expert could introduce. Moreover, and to the best of authors' knowledge for the first time, a query-by-document method is directly applied to an access-restricted scientific abstract and citation database.

2. Related work

Query Expansion

Query Expansion (QE) is the task of reformulating user queries, that are often too simplistic or unspecific, by adding additional meaningful

* Corresponding author.
E-mail addresses: fabian.lechtenberg@upc.edu (F. Lechtenberg), javier.farreres@upc.edu (J. Farreres), aldwin.lois.galvan@estudiantat.upc.edu (A.-L. Galvan-Cara), ana.somoza.tornos@upc.edu (A. Somoza-Tornos), antonio.espuna@upc.edu (A. Espuña), moises.graells@upc.edu (M. Graells).

https://doi.org/10.1016/j.eswa.2022.116967
Received 10 May 2021; Received in revised form 13 January 2022; Accepted 20 March 2022
Available online 29 March 2022

terms with similar significance. The target of QE and QbD is similar, that is, retrieving information that responds to a user's need. However, in QbD the user's initial query is replaced by a set of documents. Once the initial string has been extracted from the documents, QE methods can be incorporated into QbD.

For a comprehensive overview of the state-of-the-art the reader is referred to the review paper by Azad and Deepak (2019). Their review summarizes a general working methodology of QE: (1) data preprocessing & term extraction, (2) term weights & ranking, (3) term selection and (4) query reformulation. For each of these steps various methods have been proposed and evaluated. The studied works are discriminated by (1) application, (2) data source, and (3) core approaches. In the case of our proposed methodology, TF-IDF is used for term weights & ranking while Monte-Carlo sampling is used for term selection and query reformulation.

Yusuf et al. (2021) review more recent contributions focusing on query expansion in text retrieval of search engines. They conclude that semantic-ontology and pseudo-relevant feedback methods are the most studied and promising QE approaches. One recent contribution that shares an idea with the presented work is the one by Han et al. (2021). They propose a method based on Pseudo relevance feedback via text classification. The approach builds on well-known elements from the literature (BM25, LR, SVM, ensemble avg/RRF, RM3) and combines them in simple ways, arguing that, in QE, simplicity can be a virtue.

Query by Example

Query by Example targets the retrieval of elements that are similar to an example element. In order to achieve this, the main characteristics of the example element must be extracted and processed in a way that other elements of the same kind can be queried for, and ranked according to some criterion. This concept has been used in many different applications: Query by Voice (Lee et al., 2015), Query by Music (Foote, 1997), Query by Image and Videos (Araujo & Girod, 2018) and, most recently, Query by Webpage (Geng et al., 2022). A well known commercial example is the "search by image" function offered by Google that enables users to upload images and find similar images from the web.

Query by Document

Query by Document can be considered a variant of query by example. It was first introduced by Yang et al. (2009). Their methodology uses a "part-of-speech tagger" to extract some candidate phrases from the seed documents that should act as query strings. It has been demonstrated on the BlogScope search engine to retrieve similar documents to a set of 34 article from the New York Times. Weng et al. (2011) presented an approach that exploits Latent Semantic Indexing (LSI) as a strategy to project documents into a lower dimensional vector space. The focus of this work lies on efficient indexing for subsequent retrieval enhancement. The authors comment that LSI can be substituted by other dimensionality reduction techniques. Williams et al. (2014) present SimSeerX, a platform for query-by-document task that performs on the CiteSeerX database. The methodology also relies on dimensionality reduction of the seed documents and the documents in the database. Using various ranking functions the system returns ranked list of candidate documents that respond to the query documents. Chen et al. (2018) presented a strategy based on continuous active learning, a concept that is frequently implemented in other citation screening and content recommender systems (Howard et al., 2016; Wallace et al., 2010). Yang et al. (2018) use the "More Like This" function from Elasticsearch, a distributed search engine built on Lucene, in order to convert a query document into up to 25 relevant terms. Using these keywords, a disjunctive search is performed on the RCV1-v2 text categorization test collection.

Most recently, Le et al. (2021) presented a QbD method on top of a search interface. Their principled technique formulates the query selection task as an optimization problem (Docs2Query) that minimizes the position (maximizes the rank) of relevant documents. In the Docs2Query-Self problem, these relevant documents are the example documents used as a seed. Their approach makes use of statistics from the sampled corpus in order to solve the problem using their proposed "Best Position Algorithm". They find that their method outperforms state-of-the-art QbD methods on two test corpora (TREC-8 Voorhees & Harman, 1999 and TREC-9 Robertson & Hull, 2000).

These methods are either not directly applicable to abstract and citation databases (e.g. because of missing corpora statistics), must be partially adapted to comply with API requirements of the databases, or require a continuous learning and classification approach. Recently, Marcos-Pablos and García-Peñalvo (2018) published a method that solves a very similar problem statement and application as the one in the present study. A more detailed comparison of their methodology and the one presented in this work is given in subsequent sections.

3. Materials and methods

3.1. Monte-Carlo sampling

The Monte-Carlo (MC) method is a statistical approach based on repeated random sampling that is used to approximate solutions to complex or expensive to evaluate mathematical problems. It was first formulated by Metropolis and Ulam (1949) and has been applied in several research fields such as bio-/chemical and environmental systems engineering (Sin & Espuña, 2020) and statistical physics (Landau & Binder, 2014).

It has also found application in the field of information retrieval. Burgin (1999) demonstrated its use in the evaluation of information retrieval system performance (recall, precision, F-value). Through repeated random sampling of corpora of know size and number of relevant documents the statistical significance of a retrieval result can be determined, and the probability of an observation stemming from a random process can be estimated. More recently, Schnabel et al. (2016) also used Monte-Carlo based estimators to determine the performance of ranking functions in information retrieval. Their work deals with corpora of known size but unknown number of relevant documents, so expert judgement to classify the relevance of the retrieved documents is required. Their presented approach allows to choose appropriate query-results pairs in an unbiased manner for manual relevance judgement. It was shown that through this selection the number of required relevance judgements could be halved compared to other heuristic methods. Alexandrov et al. (2003) showed that Monte-Carlo algorithms can be useful in the efficient calculation of eigenvalues of sparse matrices, such as the term-by-document matrices that often appear in information retrieval tasks. A dimensionality reduction of the matrix can be achieved, which can significantly speed up the ranking function calculations.

In this study, Monte-Carlo sampling is used to formalize the implicit knowledge captured in a seed corpus, in order to support the query-construction step in information retrieval. Queries are performed on the whole scientific abstract and citation database of huge but unknown size and unknown number of relevant documents.

3.2. Citation databases

This work focuses on information retrieval from scientific abstract and citation databases. Among the largest databases are Google Scholar, Scopus®, ScienceDirect, Web of Science, PubMed, and arXiv. For the implementation and validation of the methodology we used Scopus® (Burnham, 2006) due to its large amount of entries (72.2 Mio. in 2019 according to Gusenbauer, 2019) in multi-disciplinary fields and the convenient API that allows automatic sampling of the database provided by Elsevier. It has restricted access, meaning that a subscription is necessary to use its features. A main drawback is that Scopus®, like the majority of scientific databases, does not provide free full-text information. This implies that the screening step during information retrieval can only be performed on the abstract, title and keywords information. The database used in the methodology can be exchanged but specific requirements and limitations of alternative APIs must be considered and adapted in the implementation.

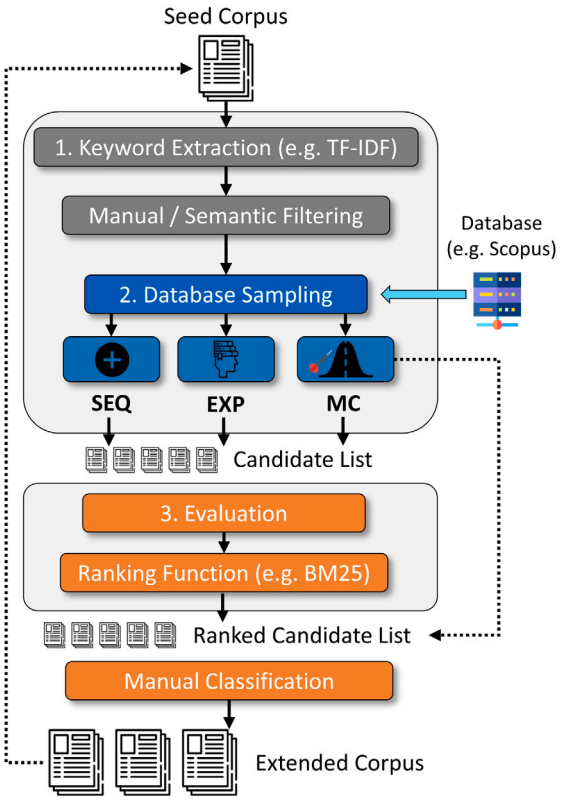


Fig. 1. Schematic representation of proposed query-by-documents approach as part of a corpus extension task. Modules for keyword extraction, database sampling, and evaluation techniques could be exchanged and applied to different databases. Sampling strategies: SEQ — Sequential, EXP — Expert, MC — Monte-Carlo.

4. Query-by-document methodology

The proposed query-by-documents approach is part of a question answering task, extending a seed corpus of already detected texts that respond to the information requirements, through inclusion of other relevant documents. Its steps are depicted in Fig. 1. The productive documents (those that provide answers to the query) may be included in the seed corpus and the cycle can be initialized again. This procedure would be repeated until no new information is found or the goal of the information search has been achieved.

4.1. Seed corpus

The methodology requires, as any query-by-documents approach, a set of seed documents. This set is used as the knowledge information repository that identifies the range of the search. Thus, it should be composed by all available documents clearly relevant to the search topic. Obviously, adding non-relevant documents will increase non-relevant results, and not incorporating documents associated to relevant research will limit the scope of the search. From that point onwards, human intervention in the retrieval process is reduced, which is important because human resources are expensive and limited. Once a Seed Corpus (SC) has been identified, the automation process will speed up the volume and the quality of information gathered, because the resulting documents after one cycle will enlarge the seed corpus, thus feeding the next iteration. Seed corpora may be obtained and provided for instance by experts in the field such as professors as starting point for a project or research line of a coworker or student. The number of documents in the corpus defines the seed corpus length (L). Another use case is the retrieval of similar documents to detect eventual plagiarism, in which L may take the value of 1. There is no generally applicable minimum number of seed documents that leads to good retrieval results.

4.2. Keyword extraction

From the seed corpus, a list of relevant domain keywords must be extracted to characterize the domain knowledge. The ranked list of keywords is obtained by computing the tfidf value of each lemmatized term in the corpus, excluding stopwords. Since the method is solely based on the occurrence of terms in a set of texts, the ranked list may contain keywords with seemingly high relevance that are not relevant in the context of investigation, or different keywords may be synonyms or related by hyponymy. Because of this, a filtering method could be applied to the list, which may be aided through the application of semantic knowledge or rules. Manual filtering of the keywords is, by no means, necessary. However, in case that the retrieval results do not match the user's expectations a (wanted) bias may be introduced by adjusting the keywords. The resulting weighted feature vector Q represents the seed corpus and the underlying domain.

4.3. Query sampling

Once the keywords are identified, the next step is to query the database. Construction of appropriate search strings is a hardship in research and investigation, and a ranked candidate list of keywords can aid in this process.

The proposed query-by-documents approach implements a Monte-Carlo (MC) sampling principle. The idea is to construct search strings by picking keywords from the ranked list of keywords with a probability distribution corresponding to their tfidf weight, applying "AND" connectors among the keywords and repeatedly query the database (the missing "OR" connector results from the addition of each new query).

The list of keywords ranked by their tfidf weight is constructed following the description and equation given in the Supplementary Material. The probability $\phi(t_i)$ of each keyword t_i being selected within the top N_{KW} keywords (where $i = 1$ has the highest weight, $i = 2$ the second highest ...) is then determined as:

$$\phi(t_i) = \frac{tfidf(t_i)}{\sum_{j=1}^{N_{KW}} tfidf(t_j)} \tag{1}$$

It should be noted that this query construction step could in principle be accompanied by the utilization of semantic knowledge (e.g. using domain ontologies) as demonstrated for instance by Amato et al. (2015). By doing so, vocabulary mismatch, also known as the vocabulary problem, can be reduced. In each performed query the occurrence of a document is registered and counted over the total amount of performed MC iterations.

This methodology comes with a few adjustable parameters.

1. The amount of MC iterations N_{MC} determines how well the relevance distribution of the keywords is captured in the sampling procedure. In the presented case studies it was found that a value between 200 and 1000 iterations is sufficient for the ranked candidate list not to change significantly anymore. See the Supplementary Material (Fig. S1) for a description of how this range was determined.
2. The upper limit for the number of documents registered in each iteration N_{it} is a parameter that controls the trade-off between exploration and exploitation of the database search space: a high value for this parameter registers many documents in each iteration, resulting in the need of more MC iterations to reach a stationary ranking. For lower values stationarity is achieved faster but relevant documents might be overlooked through the stricter cut-off. Currently, the Scopus® API imposes an upper limit of 2000 documents for this parameter.

3. The number of keywords included in the sampling procedure N_{KW} is a critical parameter with a similar trade-off characteristic as N_{it}. However, additionally to the computational trade-off, the amount of included keywords regulates how "far" from the core domain (i.e. how many "less-relevant" keywords) the sampling procedure should reach.

Tuning of these parameters, that are common to other information retrieval methods, could, in theory, be automized through a parameter sweep procedure that refines some performance metric such as seed recall or average seed position. In practice however, limits imposed on the amount of queries to the database should be taken into account and could potentially prohibit an extensive sensitivity analysis. For that reason, in this study, we limit our analysis of N_{MC} and N_{KW} to three alternative values while keeping N_{it} at the upper limit imposed by the Scopus® API.

After performing the sampling procedure, the amount of times an individual document d appeared in the query process N_d divided by the number of MC iterations N_{MC} yields the document frequency DF_d. This is an inherent relevance metric that can be directly used to rank the candidate documents and propose a reading order.

$$DF_d = \frac{N_d}{N_{MC}} \tag{2}$$

Alternatively, a naive search can be performed on the database by simply connecting the top keywords until the number of results from the database yields the amount of documents the user is willing to read. We refer to this method as the sequential sampling method (SEQ).

Finally, instead of blindly connecting the keywords an expert can use the identified terms to construct more complex strings using different combinations and connectors such as "OR" and "AND NOT". We refer to this as the expert sampling method (EXP). Compared to the MC method the user must have some sort degree of expert knowledge to apply it.

The SEQ and EXP methods do not have an inherent relevance metric and the resulting candidate documents must be ordered by other means such as the application of BM25 ranking function or naive metadata like the amount of citations.

It must be noted that in this step the database is only sampled by the information available in the abstract, title and keywords. Thus, we suggest using the abstracts of the seed corpus to obtain the set of relevant keywords based on the assumption that the language used in abstracts may be different from the full texts and, consequently, providing a fair basis for the query task.

4.4. Evaluation

Once the ranked list of references is obtained it is possible to evaluate the documents in terms of linguistic relevance. For that purpose, the freely available abstracts could be used, but we suggest to include as many full texts as possible in the evaluation step. The reasoning behind this is that abstracts only represent a very small fraction of the full-text in condensed form. Information retrieval tasks such as retrieval of parameters or experimental data will be more successful when looking into the full texts (Kottmann et al., 2010), including the Supplementary Material that often provides more quantitative information than abstracts.

This work purposely skips any discussion on publication policies and the property of the information. The general methodology developed here can be employed in public and private databases, using the total or partial information available (e.g. abstracts) according to the access rights. For more insight into the debate about Open Access (OA) the reader is referred to the review by Piwowar et al. (2018).

As previously mentioned we decided to use Scopus® for demonstration and validation purpose, which limits the sampling task to abstract, title and keyword information. Sampling and evaluating full texts instead of abstracts is debatable. The search of full texts may provide extra insight, inversely depending on the quality of the abstract, but a trade-off arises when the associated increase of computational effort is considered.

For validation purpose, after determining the ranked candidate list, the full-text information is required. It is unreasonable from and economic and computational resource point of view to download and process a huge amount of full-text information. Thus, we opt for downloading (semi-) manually a number N_{DWN} of documents. As a result, the performance of the evaluation procedure will vary depending on the institution that performs the retrieval task since the subscribed journals and databases are different for most institutions. However, future changes in publishing policies can be easily incorporated into the methodology.

Once the documents are downloaded, their relevance to the domain can be evaluated using the BM25 ranking function. The linguistic relevance is determined with respect to the weighted feature vector Q that is expected to represent the domain of interest. The user can then manually screen the resulting candidate documents in order of linguistic relevance until a threshold value $BM25_{min}$ or until he is satisfied with the retrieved information. The document frequency DF or the cosine similarity θ of the documents with the feature vector could be used as a metric for linguistic relevance alternatively (Marcos-Pablos & García-Peñalvo, 2018).

Apart from the BM25 relevance metric, the performance should be evaluated using the recall of seed documents. Le et al. (2021) verified the hypothesis that IR method performing well in re-retrieving the seed documents (Docs2Queries-Self problem) also perform well in finding similar documents (Docs2Queries-Sim problem).

4.5. Comparison with other information retrieval methods

Information retrieval procedures have been especially explored and applied in the field of systematic literature review and specialized corpora construction. Marcos-Pablos and García-Peñalvo (2018) propose an iterative methodology to construct search strings which they applied in their literature review about technological ecosystems (Marcos-Pablos & García-Peñalvo, 2019). A comparison of their approach with the one presented here is summarized in Table 1.

The objectives at the end of each iteration are different: Our approach aims at finding an extended corpus departing from a set of relevant documents (SC). The methodology by Marcos-Pablos and García-Peñalvo (2018) on the other hand, results in suggested keywords for search string construction. However, both methods follow the same main steps of keyword construction via TF-IDF, sampling and evaluation.

The main difference lies in the sampling procedure: Marcos-Pablos and García-Peñalvo (2018) procedure requires the use of expert knowledge to make the final decision on search string while the MC procedure avoids this need. Furthermore, a minor difference lies in the departing point of the methodologies that is shifted due to the different targeted endpoints (keywords vs. retrieved information/documents).

There are various other information retrieval methods that have been briefly addressed in Section 2. In this work we omit the direct comparison to these methods since their scope is not aligned with the scope of this work. On one hand, those works are applied to specialized static corpora and datasets (e.g. TREC-8 Voorhees & Harman, 1999 and TREC-9 Robertson & Hull, 2000 as used in Le et al., 2021) instead of the growing and inter-disciplinary corpora that are the academic abstract and citation databases. Our proposed methodology is designed to be applicable to corpora without the need to analyze the corpus previous to the retrieval task. Furthermore, this work goes beyond what other works are doing by evaluating the performance of the method by using the full-text information of the retrieved documents. The corpora that other works deal with are pre-classified which is not the case in the open question answering approach that is envisioned in this work and illustrated in Fig. 1.

Table 1
Comparison of information retrieval methodologies.

Marcos-Pablos and García-Peñalvo (2018)	This study
Input: Search string S; stop words vector SW; minimum cosine similarity distance θ_{min} Output: Recommended new terms T for building a new search string S1	Input: Seed Corpus SC; stop words vector SW; (optional) minimum BM25 value $BM25_{min}$ Output: Ranked list of relevant documents RL
1. Use S as input search string on academic databases and construct an abstract corpus D.	1. Project SC on vector space and compute tfidf values.
2. Project D on vector space and compute tfidf values. (corresponds to step 1 in this study)	2. Perform MC sampling method on academic databases using the filtered top keywords N_{KW}.
3. Classify documents in D as relevant (R) and non-relevant (NR) from cosine similarity.	3. Obtain a candidate list CL sorted by the document frequency DF_d.
4. Compute term weights $w_{t,D}$ in R and NR.	4. (Optional) Download the top N_{DWN} full-text documents of CL. Apply BM25 ranking function to determine linguistic relevance order.
5. Suggest new terms T based on $w_{t,D}$ sorted values.	5. Use those documents with high document frequency DF_d or higher relevance than $BM25_{min}$ for information extraction.
6. Construct a new search string S1 and repeat from step 1	6. Extend SC with newly identified truly relevant documents and repeat from step 1.

5. Case studies

The proposed methodology has been tested and illustrated on two case studies that are detailed in the following sections.

5.1. Case study I: Technological ecosystems in care and assistance

The goal of this case-study is to emulate the findings from the literature review by Marcos-Pablos and García-Peñalvo (2019) departing from a subset of the documents that have been identified as truly relevant and use them as a seed corpus in the methodology. The original systematic literature review deals with technological ecosystems in care and assistance. This topic comes with the difficulty of being based in two different fields. Therefore, the reasonable combination of suggested keywords requires a significant degree of expert knowledge. On the other hand, the proposed methodology is expected to account for, and combine, both fields implicitly in the tfidf values during the sampling procedure.

Using an initial search string on Scopus and Web of Science, Marcos-Pablos and García-Peñalvo (2019) narrowed down the candidate list of potentially relevant documents to 8394. Then, they applied a cosine similarity threshold to only consider the top 809 documents. These documents were then screened using a quality assessment checklist to further reduce the selection to 194 documents. Finally, 37 documents were included for the quantitative synthesis of the literature review. This list of relevant documents is given in Table S1 in the Supplementary Material. Note that five of these documents are not available in Scopus® and therefore cannot be retrieved with the applied methodology.

In this case-study we depart from randomly selected subsets of L documents (Table S1) taken from these 37 relevant documents and follow the steps of the proposed methodology. The chosen quality criterion for assessing the performance in this case-study is the number of relevant documents and seed documents re-retrieved by the methodology and their position in the ranked list. After selecting an appropriate seed corpus length L_{best} using 10 keywords during sampling we vary the number of included keywords N_{KW}. The tested configurations are summarized in Table 2. The number of registered documents per iteration N_{it} and the total number of MC iterations N_{MC} are both chosen to be 1000.

5.2. Case study II: Pyrolysis of plastic waste

The goal of this second case-study is to apply and compare the three presented sampling method alternatives in the domain of chemical engineering. The targeted information is the retrieval of documents containing parameters that describe pyrolysis processes of plastic waste.

Table 2
Sampling procedures tested and evaluated in case-study I.

1. Seed corpus length L ($N_{KW} = 10$)	1	8	20
Choose best performing seed corpus length L_{best}			
2. Number of keywords N_{KW} ($L = L_{best}$)	7	20	30

Other parameters: $N_{it} = N_{MC} = 1000$.

Table 3
Sampling procedures tested and evaluated in case-study II.

Method:	SEQ	EXP A (FL)	EXP B (AST)	EXP C (APL)	MC
N_{KW}	4	9	9	16	10, 15, 20, 25, 29, 30

This study is motivated by the need to populate a process ontology with information for the selection of sustainable waste-to-resource alternatives (Pacheco-López et al., 2020).

The starting point for initialization is a seed corpus consisting of eight papers. They originate from the review performed by Somoza-Tornos et al. (2021) and are given in Table S2 (Supplementary Material). After extracting the weighted feature vector for sampling, we apply the SEQ, EXP and MC sampling methods as summarized in Table 3 and compare (1) the position of the seed documents in the resulting ranked lists and (2) the linguistic relevance distributions of the identified candidate lists. As for the EXP method three of the members of the research group (FL, AST, APL) proposed search strings using the keywords from the extraction step (FL, AST) or alternative ones (APL) based on their experience in the field.

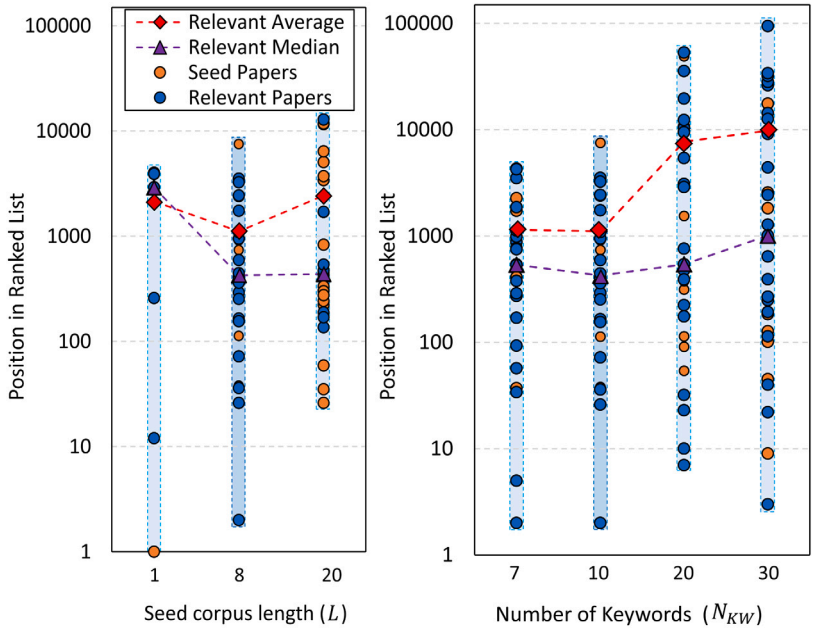
6. Results and discussion

6.1. Case study I: Technological ecosystems in care and assistance

Table S3 shows the position of all the seed papers as well as the remaining relevant papers in the ranked candidate lists. In a first step, the top ten keywords were used for MC sampling. Search was restricted to the years between 2002 and 2019 to better emulate the results of Marcos-Pablos and García-Peñalvo (2019). Fig. 2 illustrates the results.

It can be seen that using a single paper as seed corpus does not lead to satisfactory results. The seed paper itself ranks in position one with 689 appearances in 1000 iterations. Out of the remaining possible papers only 6 appear in the candidate list while only one of them ranks high (A2 in position 12). Fig. 2(a) shows the placement of the relevant papers in the candidate lists using different seed corpus length L.

Better results are obtained when using eight seed papers. In total 26 out of the 31 available relevant papers in Scopus® (83.9%) are found,



(a) Sensitivity L (b) Sensitivity N_{KW}

Fig. 2. Sensitivity analysis results for seed corpus length L and number of keywords N_{KW} in case-study I.

including seven (87.5%) of the seed papers. Moreover, these papers rank high in the list with paper A11 being the lowest one on position 7524 where the total length of the candidate list is 20,494. We consider these recall values satisfactory, having in mind that in the keyword extraction and sampling steps no expert knowledge was applied.

Increasing the number of seed papers to 20 does not improve retrieval performance. In total, 18 out of 20 seed papers appear in the candidate list (90%) and the same total number of relevant papers are found (26 out of 31 → 83.9%). Even though the seed recall value is slightly higher, the total information gain through the retrieval process is not as effective. Doubling the amount of initial information does not produce any significant effect in the number of relevant documents retrieved.

After identifying an appropriate seed corpus length (L), the influence of the number of keywords included (N_{KW}) is investigated (Fig. 2(b)). This parameter has an inherent trade-off characteristic (exploration vs. exploitation in the search space): while a low number of included keywords exploits well a limited search space, it might miss out on relevant papers associated with excluded keywords. On the other hand, a high number of included keywords will explore well the whole search space but will inevitably include more irrelevant papers. It is evident from Table S3 and Fig. 2(b) that increasing N_{KW} leads to longer candidate lists. On the other hand, no clear trend can be identified with respect to the position of the seed documents in those lists. It can be seen that using seven keywords also leads to a total recall value of 83.9% with the lowest paper being A11 in position 4299. Using 20 or 30 keywords results in higher recall values of 29 out of 31 papers (93.5%) but the seed and relevant papers are generally found in lower positions (lowest paper in rank 49,068 and 26,150 respectively).

For the final evaluation step, the seven-keyword list from the eight-paper seed corpus is chosen because of the fact that more relevant papers appear in higher positions. It is expected that this implies that the remaining highly ranked documents also have a higher linguistic relevance.

The top 5000 documents from the candidate list were checked for downloading (it includes all the found seed papers, as well as the remaining relevant documents). Using the Endnote Click (Google Chrome browser extension) a total of 2979 full-text documents could be finally downloaded, including the seed and relevant papers. These

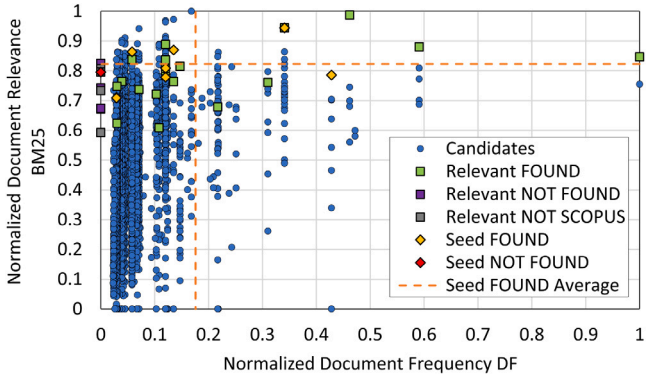


Fig. 3. Linguistic relevance of retrieved documents vs. document frequency resulting from MC sampling procedure. Shown are the position of the 37 truly relevant papers from Marcos-Pablos and García-Peñalvo (2019) and the newly identified candidates.

2979 documents form the set D and the BM25 ranking function was applied to determine their linguistic relevance.

Fig. 3 depicts the calculated normalized linguistic relevance of each document plotted against its document frequency from the MC sampling procedure. Normalization was applied by dividing the calculated BM25 values by the maximum calculated value and the document frequency was normalized by dividing the number of appearances by the number of MC iterations (1000). In this Figure, the positions of the found and not found seed and relevant documents, as well as the remaining candidate documents and the seed average values, are highlighted.

It can be seen that there is a weak correlation between BM25 and DF values. We obtain a triangular shaped cloud of points, i.e., the documents with high document frequency also have a high linguistic relevance. On the other hand, it is not ensured that documents with high linguistic relevance appear frequently in the sampling procedure. Therefore, it can be concluded that the ranking by document frequency yields good results in the higher positions but misses relevant papers in the lower ranks. When downloading the full text documents and applying the BM25 ranking function a more reliable ranking can be

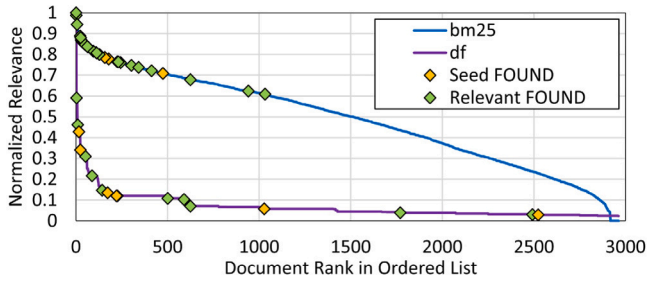


Fig. 4. Position of seed papers and relevant documents when ordering the candidate list by BM25 and DF values.

obtained. The triangular shape is a characteristic result from the MC sampling that has been found also frequently in other preliminary tests of the methodology.

Fig. 4 is an alternative representation of the data in Fig. 3. It shows the position of each document in the candidate list when ordering it by BM25 values and document frequency respectively. Apart from the fact that the relevant papers really rank comparably high in the linguistic relevance, a large number of other seemingly relevant papers are found with similar BM25 values as the seed papers. Using the lowest BM25 value of the seed papers as cut-off value ($BM25_{min} = 0.709$) the number of relevant papers to be analyzed for information extraction can be reduced by 83.5% (2979 → 475 papers). In this case, as can be seen in Fig. 4, only three relevant papers (out of the ones identified by Marcos-Pablos & García-Peñalvo, 2019) would be wrongly discarded while significantly reducing the amount of papers to be checked or read during the information extraction step.

Moreover, the BM25 metric allows sorting of the documents by relevance. The document frequency, on the other hand, shows plateaus that correspond to papers with identical amounts of appearances during the sampling procedure. This makes the determination of a cut-off value based on DF non-trivial.

The findings from the first case-study can be summarized as follows:

- The optimal seed corpus length L cannot be determined a priori. Here, eight seed papers lead to satisfactory results but for other applications this number might be too small or even less papers can be used.
- The number of keywords N_{KW} included in the sampling can be used to manage the exploration vs. exploitation trade-off.
- The proposed methodology has a strong capability of retrieving highly relevant documents and information requiring no expert knowledge in the keyword extraction, sampling and evaluation procedure.

The availability of 37 pre-classified relevant papers served as a good basis for quantifying the performance of the proposed methodology.

6.2. Case study II: Pyrolysis of plastic waste

The second case-study departs from eight seed documents (Table S2), without any available test set of other truly relevant papers. Therefore, the focus of discussion lies purely on the BM25 linguistic relevance and re-retrieval of seed documents.

Table 4 shows the keywords and associated tfidf values extracted from the seed documents. Using the top 10, 15, 20, 25, 29 and 30 chosen keywords we performed the corresponding MC sampling and observed where the seed papers rank in the candidate lists. It can be seen from Table S4 that using 20 keywords leads to the most promising results, according to amount and position of seed papers in the retrieved list. Again, there is a trade-off between exploration vs. exploitation capabilities, so the value to be chosen will depend on the importance given to these two properties.

Table 4
Case study II: Extracted keywords from the eight-paper seed corpus.

Keyword	tfidf	Keyword	tfidf	Keyword	tfidf
waste	1.12	yield	0.74	recycling	0.42
pyrolysis	1.11	plastic	0.74	gasoline	0.42
product	1.03	increase	0.66	ldpe	0.41
oil	0.90	bed[a]	0.66	char	0.40
gas	0.87	polyethylene	0.59	polymer	0.39
process	0.84	feedstock	0.55	distribution	0.36
wt	0.83	time	0.49	reactor	0.34
catalyst	0.78	residence	0.49	material	0.33
temperature	0.77	flash	0.45	recovery	0.33
monomer	0.76	hydrocarbon	0.44	work	0.32

[a]Excluded: out of scope

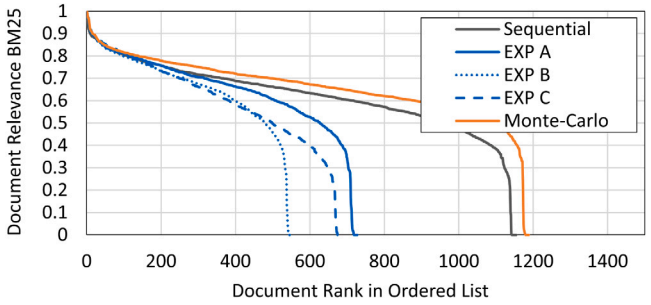


Fig. 5. Ordering candidate lists from different sampling procedures by BM25 values.

During sequential sampling (SEQ) the top four keywords are combined in a search string to yield a list of 2000 candidate documents, the upper limit for downloading references imposed by Scopus. Compared with the MC sampling, the chances are high that the results from this procedure are too general and do not correspond properly to the domain of interest.

Table 5 shows the search strings that were constructed by three member of the authors research group for the expert sampling method (EXP). While EXP A and B used only the suggested keywords from the extraction step EXP C used additional keywords that can lead to results out of the domain that is represented by the extracted keywords. On a closer inspection of the search strings, it can be seen that EXP B is a subset of EXP A, meaning that all candidate documents form search string B are included in A. Nevertheless, the documents were also retrieved and included in the evaluation step. N_{DWN} was chosen as 2000 in order to make the comparison with the Sequential procedure fair. The expert knowledge strings yield less candidates than SEQ and MC procedures.

Again, the complete set of downloaded documents (4306) was evaluated in terms of linguistic relevance using the BM25 ranking function and the 29 keywords used for MC sampling. Fig. 5 shows the relevance distribution of the three sets and the position that the seed papers rank within the MC sampling candidate list.

It is evident that in the upper ranks of each list the linguistic relevance is very similar. This is because many of those documents have been retrieved in every sampling procedure. The MC candidates are consistently more relevant than the other candidates at a given position. This means that the MC sampling procedure outperforms the other procedures in terms of retrieving a high volume of relevant documents.

Moreover, the MC sampling procedure has the highest recall value of seed documents. In fact, every seed document appears in the candidate list with B3 being the only one not included in the downloaded papers since its position (3531) is below N_{DWN}.

Based on these findings we conclude that the MC sampling procedure performs better compared to the competing procedures in all aspects of our investigation:

Table 5
Summary of sampling procedures.

Procedure	Search string	Hits	Download	Recall
SEQ	waste AND pyrolysis AND product AND oil	2015	1154	2/8
EXP A	pyrolysis AND plastic AND waste AND (gas OR oil OR product) AND (temperature OR catalyst OR yield)	1156	727	6/8
EXP B	pyrolysis AND plastic AND waste AND product AND (temperature OR catalyst OR yield)	853	548	5/8
EXP C	pyrolysis AND (plastic OR polyolefin OR polymer) AND waste AND (gas OR oil OR product OR biofuel OR chemical OR ethylene OR methane OR benzene) AND (recycling OR upcycling OR treatment)	1127	681	4/8
MC	Combinations of 29 KWs	116,435	1196	8/8[a]

[a]7/8 within top 2000.

- High amount of more relevant papers in the retrieved documents
- High recall value of seed documents

Finally, the method proved to offer a variety of benefits in terms of applicability and flexibility that can be summarized as follows:

- In principle, no need for expert knowledge
- Flexible in terms of exploration and exploitation
- Reasonable pre-download ordering based on abstracts through DF

7. Conclusions

Literature search is a specific and essential task in scientific research. The access to digital databases has boosted the search capabilities, but the scientific community worldwide still requires a lot of time and expert dedication to retrieve relevant information. This work presents a novel methodology that improves the information retrieval task from scientific abstract and citation databases via a query-by-documents approach.

The main contribution of this work consists of the inclusion of a Monte-Carlo sampling procedure during the query string construction step which leads to two desirable outcomes: (1) human expert intervention (an expensive and scarce resource) is decreased and (2) potential human bias is avoided. The proposed method has been developed, implemented and tested on the Scopus® database using two case studies.

The two case studies demonstrated the methodology's applicability to various fields of research. Remarkably, one of the studies itself is based in two distinct fields (technological ecosystems and healthcare). The retrieval results are satisfactory, i.e. high recall value of truly relevant papers declared by the reference work, considering that the authors are no experts in these fields and only a small amount of initial information (seed corpus) has been taken from the reference (Marcos-Pablos and García-Peñalvo (2019)). These results imply that corpora for multidisciplinary collaboration can be easily identified by our approach. A case-study on information retrieval of waste plastic pyrolysis processes suggests that the proposed methodology performs better in terms of number and linguistic relevance (BM25) of retrieved documents than a naive sequential sampling method as well as the query string construction by three experts.

In general, the methodology is expected to accelerate the information retrieval process through reducing the need of screening less relevant papers. Through the automatization of abstract screening using various combinations of keywords the search can go beyond what manual search could achieve, thus, finding relevant papers that could have been overlooked otherwise. Systematic literature reviews will benefit most from the methodology but really any research that starts with a literature review will find it useful.

Technical limitations like the speed of sampling and request limits using the available APIs should be addressed to further improve the performance. Furthermore, active learning strategies (Chen et al., 2018) could be integrated in the methodology to adapt the candidate ranking based on expert feedback during the manual classification step.

CRediT authorship contribution statement

Fabian Lechtenberg: Conceptualization, Methodology, Software, Data curation, Writing – original draft, Visualization. **Javier Farreres:** Conceptualization, Methodology, Software, Writing – review & editing. **Aldwin-Lois Galvan-Cara:** Methodology, Software, Data curation, Writing – original draft. **Ana Somoza-Tornos:** Conceptualization, Writing – review & editing. **Antonio Espuña:** Conceptualization, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Moisès Graells:** Validation, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Financial support received from the Spanish "Ministerio de Ciencia e Innovación" and the European Regional Development Fund, both funding the research Projects AIMS (DPI2017-87435-R) and CEPI (PID2020-116051RB-I00) is fully acknowledged. Fabian Lechtenberg gratefully acknowledges the Universitat Politècnica de Catalunya for the financial support of his predoctoral grant FPU-UPC, with the collaboration of Banco de Santander. The authors would like to thank Adrián Pacheco-López (APL) for contributing one of the expert query strings.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.eswa.2022.116967.

References

Alexandrov, V. N., Dimov, I. T., Karaivanova, A., & Tan, C. J. K. (2003). Parallel Monte Carlo algorithms for information retrieval. *Mathematics and Computers in Simulation*, *62*, 289–295. http://dx.doi.org/10.1016/S0378-4754(02)00252-5.

Amato, F., De Santo, A., Gargiulo, F., Moscato, V., Persia, F., Picariello, A., & Sperli, G. (2015). A novel approach to query expansion based on semantic similarity measures. In *DATA 2015 - 4th international conference on data management technologies and applications, proceedings* (pp. 344–353). http://dx.doi.org/10.5220/0005579703440353.

Araujo, A., & Girod, B. (2018). Large-scale video retrieval using image queries. *IEEE Transactions on Circuits and Systems for Video Technology*, *28*, 1406–1420. http://dx.doi.org/10.1109/TCSVT.2017.2667710.

Azad, H. K., & Deepak, A. (2019). Query expansion techniques for information retrieval: A survey. *Information Processing & Management*, *56*, 1698–1735. http://dx.doi.org/10.1016/j.ipm.2019.05.009.

Burgin, R. (1999). The Monte Carlo method and the evaluation of retrieval system performance. *Journal of the American Society for Information Science*, *50*, 181–191. http://dx.doi.org/10.1002/(SICI)1097-4571(1999)50:2<181::AID-ASI8>3.0.CO;2-9.

Burnham, J. F. (2006). Scopus database: A review. *Biomedical Digital Libraries*, *3*, http://dx.doi.org/10.1186/1742-5581-3-1.

Chen, S., Nourashrafeddin, S., Moh'D, A., & Milios, E. (2018). Active high-recall information retrieval from domain-specific text corpora based on query documents. In *Proceedings of the ACM symposium on document engineering 2018* (pp. 1–10). http://dx.doi.org/10.1145/3209280.3209532.

Foote, J. T. (1997). Content-based retrieval of music and audio. In *Multimedia storage and archiving systems II* (pp. 138–147). http://dx.doi.org/10.1117/12.290336.

Geng, Q., Chuai, Z., & Jin, J. (2022). Webpage retrieval based on query by example for think tank construction. *Information Processing & Management*, *59*, Article 102767. http://dx.doi.org/10.1016/j.ipm.2021.102767.

Gusenbauer, M. (2019). Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases. *Scientometrics*, *118*, 177–214. http://dx.doi.org/10.1007/s11192-018-2958-5.

Han, X., Liu, Y., & Lin, J. (2021). The simplest thing that can possibly work: (Pseudo-)relevance feedback via text classification. In *Proceedings of the 2021 ACM SIGIR international conference on theory of information retrieval* (pp. 123–129). http://dx.doi.org/10.1145/3471158.3472261.

Howard, B. E., Phillips, J., Miller, K., Tandon, A., Mav, D., Shah, M. R., Holmgren, S., Pelch, K. E., Walker, V., Rooney, A. A., Macleod, M., Shah, R. R., & Thayer, K. (2016). SWIFT-Review: A text-mining workbench for systematic review. *Systematic Reviews*, *5*, 1–16. http://dx.doi.org/10.1186/s13643-016-0263-z.

Kottmann, R., Radom, M., Formanowicz, P., Glöckner, F., Rybarczyk, A., Szachniuk, M., & Błażewicz, J. (2010). Cerberus: A new information retrieval tool for marine metagenomics. *Foundations of Computing and Decision Sciences*, *35*, 107–126.

Landau, D. P., & Binder, K. (2014). *A guide to Monte Carlo simulations in statistical physics* (4th ed.). Cambridge University Press, http://dx.doi.org/10.1017/cbo9781139696463.

Landhuis, E. (2016). Scientific literature: Information overload. *Nature*, *535*, 457–458. http://dx.doi.org/10.1038/nj7612-457a.

Le, N. X. T., Shahbazi, M., Almaslukh, A., & Hristidis, V. (2021). Query by documents on top of a search interface. *Information Systems*, *101*, Article 101793. http://dx.doi.org/10.1016/j.is.2021.101793.

Lee, L. S., Glass, J., Lee, H. Y., & Chan, C. A. (2015). Spoken content retrieval - beyond cascading speech recognition with text retrieval. *IEEE Transactions on Audio, Speech and Language Processing*, *23*, 1389–1420. http://dx.doi.org/10.1109/TASLP.2015.2438543.

Marcos-Pablos, S., & García-Peñalvo, F. J. (2018). Information retrieval methodology for aiding scientific database search. *Soft Computing*, *24*, 5551–5560. http://dx.doi.org/10.1007/s00500-018-3568-0.

Marcos-Pablos, S., & García-Peñalvo, F. J. (2019). Technological ecosystems in care and assistance: A systematic literature review. *Sensors*, *19*, 708. http://dx.doi.org/10.3390/s19030708.

Mergel, G. D., Silveira, M. S., & Da Silva, T. S. (2015). A method to support search string building in systematic literature reviews through visual text mining. In *Proceedings of the ACM symposium on applied computing* (pp. 1594–1601). http://dx.doi.org/10.1145/2695664.2695902.

Metropolis, N., & Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association*, *44*, 335–341. http://dx.doi.org/10.1080/01621459.1949.10483310.

Pacheco-López, A., Somoza-Tornos, A., Muñoz, E., Capón-García, E., Graells, M., & Espuña, A. (2020). Synthesis and assessment of waste-to-resource routes for circular economy. In *30 European symposium on computer aided process engineering* (pp. 1933–1938). http://dx.doi.org/10.1016/B978-0-12-823377-1.50323-2.

Piwowar, H., Priem, J., Larivière, V., Alperin, J. P., Matthias, L., Norlander, B., Farley, A., West, J., & Haustein, S. (2018). The state of OA: A large-scale analysis of the prevalence and impact of open access articles. *PeerJ*, *2018*, 1–23. http://dx.doi.org/10.7717/peerj.4375.

Robertson, S. E., & Hull, D. A. (2000). The TREC-9 filtering track final report. In *Proceedings of the ninth text retrieval conference*.

Schnabel, T., Frazier, P. I., Swaminathan, A., & Joachims, T. (2016). Unbiased comparative evaluation of ranking functions. In *ICTIR 2016 - proceedings of the 2016 ACM international conference on the theory of information retrieval* (pp. 109–118). http://dx.doi.org/10.1145/2970398.2970410.

Sin, G., & Espuña, A. (2020). Editorial: Applications of Monte Carlo method in chemical, biochemical and environmental engineering. *Frontiers in Energy Research*, *8*, 1–2. http://dx.doi.org/10.3389/fenrg.2020.00068.

Somoza-Tornos, A., Pozo, C., Graells, M., Espuña, A., & Puigjaner, L. (2021). Process screening framework for the synthesis of process networks from a circular economy perspective. *Resources, Conservation and Recycling*, *164*, Article 105147. http://dx.doi.org/10.1016/j.resconrec.2020.105147.

Voorhees, E. M., & Harman, D. K. (1999). Overview of the eighth text retrieval conference (TREC-8). In *Proceedings of the eighth text retrieval conference*.

Wallace, B. C., Small, K., Brodley, C. E., & Trikalinos, T. A. (2010). Active learning for biomedical citation screening categories and subject descriptors. In *Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 173–181). http://dx.doi.org/10.1145/1835804.1835829.

Weng, L., Li, Z., Cai, R., Zhang, Y., Zhou, Y., Yang, L. T., & Zhang, L. (2011). Query by document via a decomposition-based two-level retrieval approach. In *SIGIR'11 - proceedings of the 34th international ACM SIGIR conference on research and development in information retrieval* (pp. 505–514). http://dx.doi.org/10.1145/2009916.2009985.

Williams, K., Wu, J., & Giles, C. L. (2014). SimSeerX: A similar document search engine. In *DocEng 2014 - Proceedings of the 2014 ACM symposium on document engineering* (pp. 143–146). http://dx.doi.org/10.1145/2644866.2644895.

Yang, Y., Bansal, N., Dakka, W., Ipeirotis, P., Koudas, N., & Papadias, D. (2009). Query by document. In *Proceedings of the 2nd ACM international conference on web search and data mining* (pp. 34–43). http://dx.doi.org/10.1145/1498759.1498806.

Yang, E., Lewis, D. D., Frieder, O., Grossman, D., & Yurchak, R. (2018). Retrieval and richness when querying by document. In *CEUR workshop proceedings* (pp. 68–75).

Yusuf, N., Yunus, M. A. M., Wahid, N., Mustapha, A., & Salleh, M. N. M. (2021). A survey of query expansion methods to improve relevant search engine results. *International Journal on Advanced Science, Engineering and Information Technology*, *11*, 1352–1359. http://dx.doi.org/10.18517/ijaseit.11.4.8868.