PROCEEDINGS OF THE of the 31st European Symposium on Computer Aided Process Engineering
(ESCAPE31), June 6-9, 2021, İstanbul, Turkey
M. Türkay, R. Gani (Editors)

Towards Automated Information Retrieval of Process Data and Knowledge from Academic Databases

Fabian Lechtenberg[a], Javier Farreres[b], Ana Somoza-Tornos[a,c], Adrián Pacheco-López[a], Antonio Espuña[a] and Moisès Graells[a,*]

[a]*Chemical Engineering Department, Universitat Politècnica de Catalunya, EEBE, C/ Eduard Maristany 16, Barcelona 08019, Spain*
[b]*Computer Science Department, Universitat Politècnica de Catalunya, EEBE, C/ Eduard Maristany 16, Barcelona 08019, Spain*
[c]*Renewable and Sustainable Energy Institute, University of Colorado Boulder, Boulder, Colorado 80303, United States*
**moises.graells@upc.edu*

Abstract

Process modeling requires both data (chemical reaction yields, kinetic constants, cost estimates, environmental indicators, etc.) and knowledge (operation models and formulations, alternative processes and technologies, etc.). Searching in databases and published research may provide such information, but there is a lack of systematic methods and tools guiding this procedure. The present work describes and assesses an information retrieval methodology that is part of a proposed retrieval and extraction cycle addressing this problem. Two query construction methods for sampling academic databases are proposed, assessed and compared. Departing from a seed corpus of a limited number of papers, Scopus® is used as an academic database to retrieve literature containing information associated with pyrolysis processes of waste plastic. It is found that, with minimal human intervention, the methodology is able to return a ranked list of candidate documents that have a considerable (linguistic) relevance.

Keywords: Information Retrieval, Big Data, Text Mining, Academic Databases, Waste-to-Resource

1. Introduction

Due to new communication technologies the world becomes increasingly interconnected. This enables researchers from all around the globe to publish their work and access publications from their peers. In recent years this has led to a yearly growth of about 9% in academic publications (Landhuis, 2016). The abundance of available information raises the need for strategies to handle this Big Data in order to pinpoint the truly relevant information. In this work we present a methodology to semi-automatically screen academic databases to obtain a set of promising documents related to a given research question.

In Process Systems Engineering (PSE), the modeling of technical systems is a fundamental task. Model parameters and knowledge can be obtained from publications in academic databases but with the ever increasing volume of data this becomes an increasingly difficult task. Text mining and Natural Language Processing are tools that emerged to handle

and make use of Big Data. In the field of Systematic Reviews they are being applied to find and classify relevant contributions within a (sub-) field of a discipline in order to include them in a review paper (Usai et al., 2018). A related example where this concept is applied in a natural science field is shown in the work by (Kottmann et al., 2010). They present an IRE system that retrieves, classifies and extracts papers related to metagenomics marine science from the PubMed database.

This work is motivated by the waste-to-resource route assessment work presented by (Pacheco-López et al., 2020). The authors compare the potential of transformation routes of plastic waste to valuable raw materials. By applying an IRE process it is expected to populate a process database for comparison and go beyond what a manual retrieval procedure could achieve, both in terms of retrieved volume and identification of non-intuitive relations. In order to facilitate the search process in a systematic way we propose an Information Retrieval and Extraction (IRE) cycle that departs from an initial set of relevant documents. The main novelty of this contribution is the proposal and validation of a retrieval method to screen access-limited academic databases with the goal of retrieving specific parameters (in this case the ones that characterize chemical transformation processes as part of waste-to-resource routes). Here, we demonstrate our progress in the first step of the cycle, the information retrieval.

2. Problem Statement

To address this objective the problem can be stated as follows:

Given (1) a research question, (2) a set of relevant documents (seed corpus) and (3) an academic database, find an extended corpus (new relevant documents).

The extended corpus must be similarly relevant to the research question and a design objective is to postpone and so to reduce human interaction as much as possible.

3. Methodology

The proposed IRE cycle is presented in Figure 1. First, an academic database is queried using a sampling method that is conditioned to the initial set of documents (seed corpus). Second, the retrieved documents are analysed in order of decreasing estimated relevance to find the text passages that contain the desired information. So far, all parts of the information retrieval part are implemented in Python 3.8. The database used in this study is Scopus® and is accessed via the Elsevier API. The following subsections describe the hypotheses and individual steps within the information retrieval part of the cycle.

3.1. Hypotheses

1. Linguistically similar texts contain semantically similar information.
2. A seed corpus composed of documents fitting a domain represents that domain.
3. Using relevant keywords extracted from that corpus for constructing queries to structured sets of documents, organized as databases, allows for finding documents having semantically similar information.

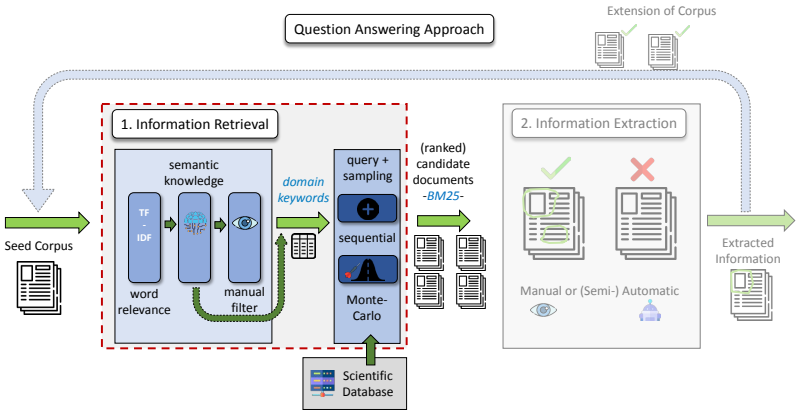

Figure 1. Proposed information retrieval and extraction cycle.

3.2. Extraction of Domain Keywords from Seed Corpus

First, a set of relevant domain keywords must be extracted from the seed corpus to characterize the domain. The Term Frequency-Inverse Document Frequency (TF-IDF) metric (Qaiser and Ali, 2018) determines the relative importance of a word in a set of documents, considering the words with high number of appearances in few documents as relevant while frequent words in all documents and non-frequent words as non-relevant.

3.3. Query Construction Methods for Sampling Academic Databases

In this work we assess and compare two methods:

- Sequential Sampling (SEQ): Add keywords to the query in decreasing relevance order until the number of search results is below a user-defined threshold.
- Monte-Carlo Sampling (MC): Build multiple queries by randomly adding keywords according to their relevance until the number of results is below a user defined threshold. An inherent ranking metric is the appearance frequency (the number of queries a given document appears in, divided by the total amount of queries).

3.4. Ranking of Retrieved Documents

The BM25 ranking function (Robertson and Zaragoza, 2009) is a popular approach in information retrieval to determine the relevance of a document to a query compared to the other documents in the retrieved corpus. It yields a metric that can be considered as relative (linguistic) relevance.

3.5. Validation and Limitations

In order to validate the retrieval method, the normalized BM25 relevance of the retrieved full text documents are compared with each other and with the seed documents. The full text information is chosen because (1) a test on using the abstracts of the documents to determine the relevance of the full documents showed an unreasonable classification and (2) the desired knowledge is contained within the full text.

4. Case Study

The methodology is applied to a set of eight seed documents that contain quantitative information about processes for the pyrolysis of plastic waste. The information extracted from these documents was used by (Somoza-Tornos et al., 2021) to assess a process screening framework for the synthesis of process networks from a circular economy perspective. The retrieved documents departing from this seed corpus are expected to help extending the study by identifying additional candidate process conditions and pathways.

4.1. Targeted Information

One example of textual information is taken from a seed document (Onwudili et al., 2009) and the keywords are highlighted by general concepts that are searched:

"The compositions of the *process* (pyrolysis) products of pure low-density *waste product* (polyethylene), *waste product* (LDPE) and *waste product* (polystyrene) and their mixtures have been investigated over a *condition* (temperature) range from 300 to 500 *condition* (°C)."

5. Results and Discussion

Table 1 shows the top 16 keywords extracted from the seed abstracts. The sequential method was applied with a threshold of 1,000 documents leading to 376 abstracts and full text documents that could be retrieved with the available licensing options. The Monte-Carlo method was performed with 1,000 iterations using 10, 20 and 30 keywords to identify three lists of candidate documents. From the top 2,500 documents of each list a total of 553, 568 and 1,110 abstracts and corresponding full documents could be retrieved respectively. Additionally, from the 30 keywords list (63,186 entries) a randomly selected subset of 2,500 papers was chosen from which a total of 972 could be retrieved and is treated as reference set. The BM25 ranking function was applied to the complete set of retrieved full text documents, as these contain the relevant information. From Figure 2 it can be seen that the documents from the 10 and 20 keywords list have a very similar distribution and a generally higher relevance compared to the 30 keywords list and the sequential method. This trend holds true when using different amounts of keywords for relevance determination. The documents from the randomly selected subset have a considerably lower relevance proving that the methods perform as expected. The number of keywords when sampling the database is a key parameter and the results show that using more "relevant" keywords does not lead to improved retrieval results. Moreover, it is evident that the Monte-Carlo method retrieved generally more relevant documents than the sequential method.

Table 1. Top 16 keywords and TF-IDF values extracted from the seed corpus abstracts.

Keyword	TF-IDF	Keyword	TF-IDF	Keyword	TF-IDF	Keyword	TF-IDF
pyrolysis	1.23	process	0.85	yield	0.73	feedstock	0.60
waste	1.12	gas	0.85	catalyst	0.68	wt	0.60
product	1.01	temperature	0.76	bed	0.64	polyethylene	0.58
oil	0.86	plastic	0.76	increase	0.64	fluidise	0.55

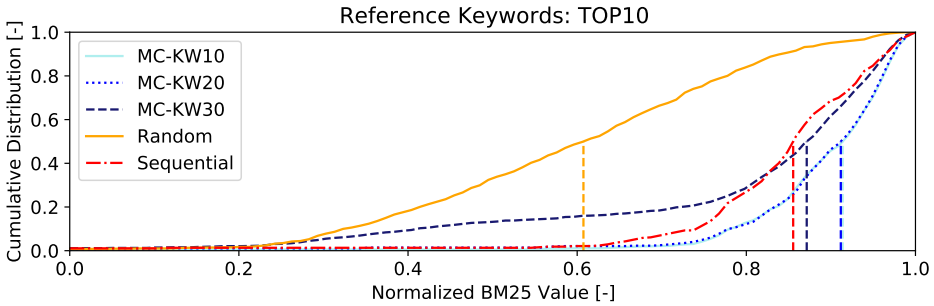

Figure 2. Cumulative distribution of normalized BM25 relevance values.

Figure 3 illustrates that the inherent ranking metric of the Monte-Carlo method correlates to some extent with the document relevance. The point clouds represent the relevance and document frequency of all retrieved documents within the 30 keywords list. It can be identified that, when only choosing documents with a normalized frequency higher than 0.1, a significant amount of irrelevant papers can be discarded (here: 80.3% of set size). Moreover, a large number of newly identified documents has an even higher relevance (22.9%) than the average relevance of the seed corpus, which is a promising starting point for the information extraction step. The sequential method, on the other hand, does not directly have an inherent ranking. Figure 3 (right) shows a derived metric that is the number of keywords contained in the downloaded papers. It appears that there is no clear correlation between the number of keywords included and the BM25 relevance.

The general impression is that the documents are indeed fitting the domain and potentially contain the desired information. A selected truly relevant document appearing in all three MC lists within the upper ranks is titled "Hydrocarbons obtained by waste plastic pyrolysis: Comparative analysis of decomposition described by different kinetic models" (Miskolczi and Nagy, 2012). A relevant passage reads as follows:

"... the *parameter* (yields) of volatile products were 15.0% [410 *condition* (°C)], 24.1% [430 *condition* (°C)] and 55.3% [450 *condition* (°C)] in case of W-2 marked sample *process* (pyrolysis), while ..."

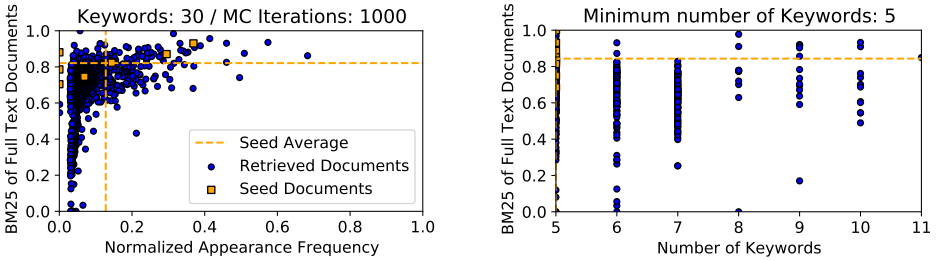

Figure 3. BM25 values plotted against model inherent ranking for Monte-Carlo (left) and sequential (right) method. Shown is the data obtained from the 30 keywords list.

6. Conclusions

Lists of (linguistically) relevant documents were identified with minimal human intervention. These lists have the characteristic of being populated by highly relevant documents in the upper ranks which allows to limit the selection of documents to download for the subsequent information extraction step. The candidate documents proved to have similar or even higher relevance to the domain than the documents in the seed corpus. A first qualitative assessment of the titles and abstracts indicates that these documents are truly relevant to the posed question. Our investigations showed that the proposed information retrieval methodology performs appropriately using the selected database and seed corpus taken from a chemical engineering field. This implies the potential of establishing a systematic machine-assisted search procedure for model parameters and knowledge, effectively reducing the workload of engineers in the PSE community and going beyond what a completely manual procedure could achieve.

As of now, the methodology assesses document relevance by means of the BM25 metric. This metric allows for a pre-selection of documents but the next necessary step in the development of the whole information retrieval and extraction cycle is to systematically classify the true relevance of the documents by a machine-assisted information extraction methodology. Moreover, the methodology has been tested using only one database. Further work is in progress to extend the search and improve its efficiency (speed and accuracy).

Acknowledgements
Financial support received from the Spanish Competitiveness, Industry and Economy Ministry and the European Regional Development Fund, both funding the research Projects AIMS (DPI2017-87435-R) is fully acknowledged. Adrián Pacheco-López thankfully acknowledges financial support received from the Spanish Ministry of Science, Innovation and Universities (grant ref. PRE2018-087135).

References

R. Kottmann, M. Radom, P. Formanowicz, F. Glöckner, A. Rybarczyk, M. Szachniuk, J. Błażewicz, 2010. Cerberus: A New Information Retrieval Tool for Marine Metagenomics. Foundations of Computing and Decision Sciences 35, 107–126.

E. Landhuis, 2016. Scientific literature: Information overload. Nature 535, 457–458.

N. Miskolczi, R. Nagy, 2012. Hydrocarbons obtained by waste plastic pyrolysis: Comparative analysis of decomposition described by different kinetic models. Fuel Processing Technology 104, 96–104.

J.A. Onwudili, N. Insura, P.T. Williams, 2009. Composition of products from the pyrolysis of polyethylene and polystyrene in a closed batch reactor: Effects of temperature and residence time. Journal of Analytical and Applied Pyrolysis 86, 293–303.

A. Pacheco-López, A. Somoza-Tornos, E. Muñoz, E. Capón-García, M. Graells, A. Espuña, 2020. Synthesis and Assessment of Waste-to-resource Routes for Circular Economy, in: 30 European Symposium on Computer Aided Process Engineering. Elsevier. volume 48, pp. 1933–1938.

S. Qaiser, R. Ali, 2018. Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. International Journal of Computer Applications 181.

S. Robertson, H. Zaragoza, 2009. The Probabilistic Relevance Framework: BM25 and Beyond. Foundations and Trends® in Information Retrieval 3, 333–389.

A. Somoza-Tornos, C. Pozo, M. Graells, A. Espuña, L. Puigjaner, 2021. Process screening framework for the synthesis of process networks from a circular economy perspective. Resources, Conservation and Recycling 164, 105147.

A. Usai, M. Pironti, M. Mital, C. Aouina Mejri, 2018. Knowledge discovery out of text data: a systematic review via text mining. Journal of Knowledge Management 22, 1471–1488.