Adaptive Search Engine for Heterogeneous Documents

Oussama Ayoub, Christophe Rodrigues and Nicolas Travers

December 1, 2020

Adaptive Search Engine for Heterogeneous Documents

Oussama Ayoub
oayoub@sevillemore.com
SMH & Léonard de Vinci Pôle
Universitaire, Research Center
Paris La Défense, France

Christophe Rodrigues
christophe.rodrigues@devinci.fr
Léonard de Vinci Pôle Universitaire,
Research Center
Paris La Défense, France

Nicolas Travers
nicolas.travers@devinci.fr
Léonard de Vinci Pôle Universitaire,
Research Center
Paris La Défense, France

ABSTRACT

Providing an efficient search engine for legal actors querying for textual documents is a challenging objective. Nowadays most engines target semantic analysis on top of text queries to enhance the relevance. But the legal context relies mainly on heterogeneous data in terms of both queries and documents length, structural complexity, and queries context. This combination makes standard solutions hardly scalable or adaptable. The proposed solution is an adaptive approach that aims to be applied to any textual database establishing a search engine. Its peculiarity is to normalize documents by producing fragments, enriching them with word embedding, here summarizing, and rebuilding documents through similarity aggregations on either enriched content, structure and context. By integrating our solution in Elasticsearch we ensure the flexibility and the fine-tuning of both words embedding and similarities.

KEYWORDS

Natural Language Processing, Information Retrieval, Search Engine

1 INTRODUCTION

The legal context relies on several types of documents which makes the comparison between them a real issue and consequently complexifies the establishment of a search engine. Moreover, several major constraints must be considered to deliver a relevant service that is fully integrated into the legal environment: professional secrecy, the user's legal context and the consideration of case law.

Providing dedicated solutions for each data type is counterproductive. Thus, the issue is to propose a unique solution which makes all data comparable regardless of its size (documents and queries), structure (contracts, profiles), richness (from simple to complex documents), or context of use (simple search to recommendations. On top of that, relevance must be tuned according to a given context, here the legal one. The main issue to tackle is to deal with really heterogeneous documents length from a single paragraph to hundreds-page documents with a complex structure and consequently its correlation with relevance and similarity measures. We will detail this problematic in the following.

Our approach combines two different domains, linked to Natural Language Processing (NLP) and a mix between Information Retrieval (RI) and Databases (DB), into an elaborate architecture that provides access to an enhanced search engine. It benefits from the NLP modeling in order to provide summaries on data according to the context (*i.e.,* local vs. global corpus content) and from the IR/DB capabilities to manipulate results to reconstitute documents and to make them comparable and scalable.

This paper aims at giving the subject's relation to the literature and an overview on the approach's architecture. Section 2 describes the main issues related to the subject and the related work on those topics. Section 3 presents the approach and the global architecture of the search engines designed to answer the preceding issues. Finally, we conclude in Section 4 with current paths of research.

2 PROBLEMATIC AND STATE-OF-THE-ART

The goal of this research topic is to create an adaptive search engine that can be applied on dedicated environments (*i.e.,* legal "clusters"), associated with external resources and knowledge. We are facing issues: (1) various documents length (not compatible with the state-of-the-art on preprocessed representations of documents), (2) queries length varies from keywords to complex documents (relevance is barely dealt with content inclusion), (3) target context-aware relevance (*i.e.*, legal data). Thus, the global approach mainly targets this goal by trying to make various documents and queries comparable. Consequently, our work meets two research domains and mix them in a relevant and efficient way: Natural Language Processing (NLP) and Information Retrieval (IR).

Language processing, based on learning models, aims to categorize legal issues. The association of different legal codes makes Topic Modeling and Abstract summary generation main objectives to address. In 2013, [3] proposed the creation of word representation in vector space known as `Word2Vec`. It changed the perspective in NLP by adding the context of words in their vectorized representation. A lot of work has been done since to create more significant representations by using attention models. Encoders, autoencoders and later transformers (combining few technologies) are many ways of creating a complex and efficient vector to represent and store the data in a small dimension. Some papers are using embeddings as representation of the data object to enable searching documents by making them comparable. Jan Rygl and al. [5] propose to transform obtained embedding with Latent Semantic Analysis into strings allowing the use of full-text search using TF-IDF similarity in ElasticSearch built with the default configuration. They refined the query result by calculating the cosine similarity between vectors on remaining documents. [4] used a neural autoencoder to create representation of documents to find similarities between different types of objects in their search engine. The combination of multiple techniques including autoencoders improves the quality of results. We differ from those approaches by introducing preprocessing steps to normalize heterogeneous documents analysis and aggregations in post-processing to rebuilt documents using similarities.

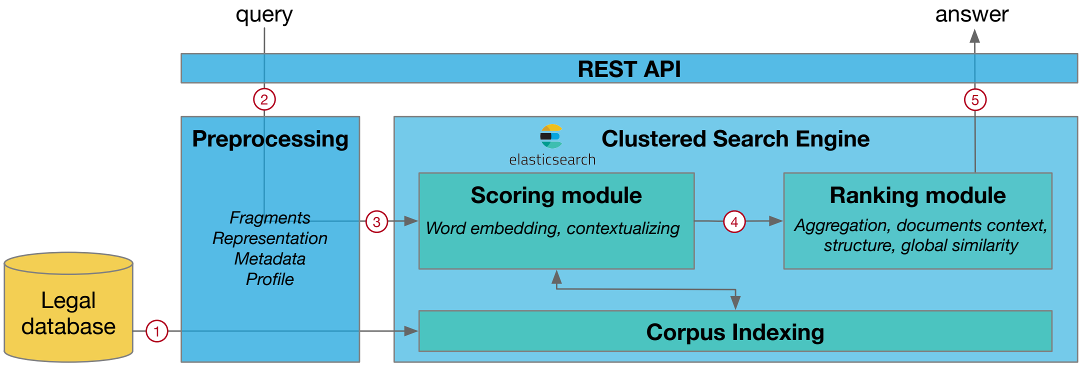

Figure 1: Architecture

Some studies combine IR with Machine Learning to facilitate the analysis of legal corpora while providing new functionalities to define a dedicated legal search engine. The issue on documents length and heterogeneity is ignored or avoided when training the main Machine Learning and Deep Learning solutions nowadays [1, 2] while documents in legal domain have a much higher length. This fact was recently raised [1, 6] where well-known machine learning methods do not perform as well as expected for long-length documents. [6] tries to solve this problem by separating the documents in chunks and merge their vectors to create the document's vector. Moreover, the conversion process is applied to queries as well as for documents without taking the length into account [4]. Those solutions inspired to build our approach by generalizing the concepts of embedding on various documents and queries adaptable on dedicated contexts.

3 APPROACH

We propose a full architecture that solves the issues mentioned. Figure 1 shows the mandatory modules for the operation of the search engine. The processing steps are detailed in the following:

`Step 1` - The contextual database is pre-processed by a refinement module in order to compute AI models and document transformations. This crucial step normalizes documents' size either by splitting in nominal sizes called "*fragments*" as well as enriching its content with "*word embedding*" according to the context of use. We intend to combine fragmentation and summarize processes to enhance the quality of relevance at various scales. Moreover, structural and context information of documents enrich the embedding, which constitutes another dimension of relevance in a hyper-connected environment where most of the documents are linked.

`Step 2` - Every query is pre-processed to be converted similarly as the refined stored documents. This request can be of various lengths, type or linked to a user's profile.

`Step 3` - Document fragments are projected in a vector space model relying on both stored data and the IA model in order to give a score for each fragment. The scoring is also based on plain text similarity to maximize the relevance of the results. By integrating the approach into `Elasticsearh`, it allows to make the computation scalable and tunable. Thus, the tuned scoring computation is distributed on each fragment by reducing the information of a document to one vector. It returns a detailed explanation of the result used to recompose documents.

`Step 4` - Relevant fragments are grouped together to give a global and contextual answer. This aggregation step is a major issue of the overall system since it allows targeting various types of use cases by mixing the combination of aspects in the vector space.

`Step 5` - The answer is a set of documents composed of linked fragments with enriched explanations of provided scores.

4 CONCLUSION

We introduced a new flexible architecture establishing a powerful search engine relying on semantic, structural and context analysis. Our work focuses on solving the heterogeneity of databases' documents, especially in terms of length, using the segmentation of the documents into fragments to enhance the relevance of results.

The design of this solution represents the initial step of our work. As future work, we will focus on confirming the key assumptions regarding the benefits of word embedding, segmentation and aggregation in search engines as well as combining multiple scoring criteria on different datasets (*e.g.,* legal, tourism, Kaggle).

ACKNOWLEDGMENT

This work has been funded by the research agreement "Smart Legal" in collaboration with *Seville More Helory.*

REFERENCES

[1] Robert Keeling, Rishi Chhatwal, Nathaniel Huber-Fliflet, Jianping Zhang, Fusheng Wei, Haozhen Zhao, Ye Shi, and Han Qin. 2019. Empirical Comparisons of CNN with Other Learning Algorithms for Text Classification in Legal Document Review. *2019 IEEE International Conference on Big Data (Big Data)* (2019).

[2] Aurelie Mascio, Zeljko Kraljevic, Daniel Bean, Richard Dobson, Robert Stewart, Rebecca Bendayan, and Angus Roberts. 2020. Comparative Analysis of Text Classification Approaches in Electronic Health Records. *CoRR* (2020).

[3] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *ICLR'13 Workshop*.

[4] Jonas Pfeiffer, Samuel Broscheit, Rainer Gemulla, and Mathias Göschl. 2018. A Neural Autoencoder Approach for Document Ranking and Query Refinement in Pharmacogenomic Information Retrieval. In *BioNLP workshop of ACL*.

[5] Jan Rygl, Jan Pomikálek, Radim Rehurek, Michal Ruzicka, Vít Novotný, and Petr Sojka. 2017. Semantic Vector Encoding and Similarity Search Using Fulltext Search Engines. *CoRR* abs/1706.00957 (2017).

[6] Lulu Wan, George Papageorgiou, Michael Seddon, and Mirko Bernardoni. 2019. Long-length Legal Document Classification. arXiv:1912.06905 [cs.CL]