Analytics and Data Science Institute

datascience.kennesaw.edu

Data Science Research Series

Automatic Knowledge Extraction from OCR Documents Using Hierarchical Documents Analysis

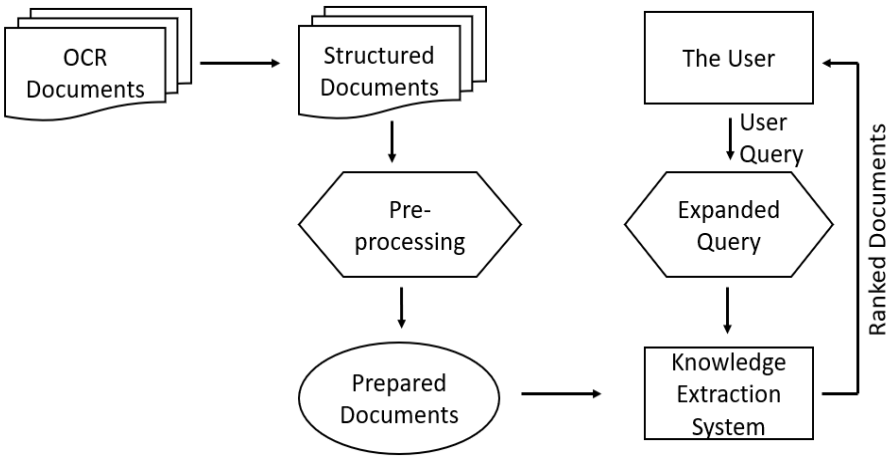
BACKGROUND

Industries can improve their business efficiency by analyzing and extracting relevant knowledge from large numbers of documents. Manual knowledge extraction from a large volume of documents is labor intensive, unscalable and challenging. Consequently, there have been several attempts to develop intelligent systems to automatically extract relevant knowledge from OCR documents. Moreover, the automatic system can improve the capability of search engines by providing application-specific domain knowledge. However, extracting the efficient information from OCR documents is challenging due to their highly unstructured format. In this paper, we propose an efficient framework for a knowledge extraction system that takes keywords-based queries and automatically extracts their most relevant knowledge from OCR documents by using text mining techniques. The framework can provide relevance ranking of knowledge to a given query. We tested the proposed framework on a corpus of documents at GE Power where documents consists of more than one-hundred pages in PDF.

APPROACH

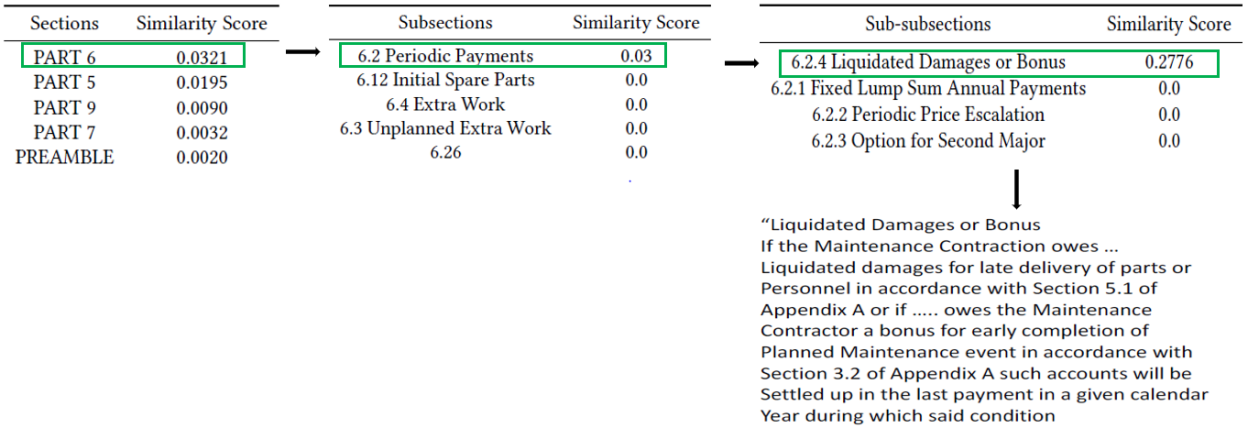
We proposed framework for knowledge extraction to extract the most relevant texts of interest from a corpus of OCR documents. The framework includes two phases. In the first phase, an unstructured OCR document is reconstructed to hierarchical structural format by analyzing the

document layout features (e.g., font size and font boldness). In the second phase, the hierarchical structured documents are then preprocessed (e.g., removing stop words and special characters, and case folding) and extended by appending tokens using N-grams. Concurrently, when a user provides a query, the query is updated by using the query expansion method. The structured data and the expanded query are then converted to a Vector Space Model (VSM) and compared for the ranking paragraph relevant to the query. The below flowchart shows the flow diagram of the framework.



RESULTS

We applied our framework to a corpus of OCR documents provided by GE Power (GEP). These documents contain multiple sections such as Appendix, Sections and Exhibits. These, in turn could be comprised of multiple layers of subsections. Our method aims at extracting the most relevant information regarding a query term that a user defines. Specifically, we demonstrate the process with the query term "Liquidated Damages" from the set of queries provided by GEP. The query is compared hierarchically with sections-subsections in the document by VSM. The most relevant section/subsection is selected based on highest similarity score. The tables below illustrate the retrieved most relevant sections/subsections and include the most relevant information within the document for the given query term "Liquidated Damages".



CONCLUSIONS

In this study, we present a knowledge extraction framework from OCR documents for a given user query with VSM. The hierarchical structure analysis of the documents provides an effective solution to fetch relevant knowledge. The extracted knowledge could be used for various applications such as automatic knowledge management and enriching the search systems. The advantage of using query expansion is to establish a correlation between query terms and document terms by analyzing provided relevant knowledge. For any new queries, expansion terms can be selected from the documents. However, this method has limitations based on rules imposed during the document reconstruction step that are dependent on the structure of the original PDF document layout features (font size and boldness) and regular expression pattern. We conducted experiments of the knowledge extraction framework with 16 queries to extract relevant knowledge from over 100 documents. The series of experiments showed performance improvement with our framework over the existing manual knowledge extraction system.

CITATION FOR FULL ARTICLE

Masum, Mohammad; Kosaraju, Sai; Bayramoglu, Tanju; Modgil, Girish; and Kang, Mingon, "Automatic Knowledge Extraction from OCR Documents Using Hierarchical Document Analysis" (2018). *Grey Literature from PhD Candidates*. 12.
https://digitalcommons.kennesaw.edu/dataphdgreylit/12