

Kennesaw State University DigitalCommons@Kennesaw State University

Grey Literature from PhD Candidates

Ph.D. in Analytics and Data Science Research
Collections

Fall 8-12-2018

Automatic Knowledge Extraction from OCR Documents Using Hierarchical Document Analysis

Mohammad Masum
mmasum@students.kennesaw.edu


Sai Kosaraju
Kennesaw State University, skosara1@students.kennesaw.edu

Tanju Bayramoglu
Georgia Electric, tanju.bayramoglu@ge.com

Girish Modgil
Georgia Electric, girish.modgil@ge.com

Mingon Kang
Kennesaw State University, mkang9@kennesaw.edu

Follow this and additional works at: <https://digitalcommons.kennesaw.edu/dataphdgreylit>

 Part of the [Computer Sciences Commons](#), [Mathematics Commons](#), and the [Statistics and Probability Commons](#)

Recommended Citation

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>

This Conference Proceeding is brought to you for free and open access by the Ph.D. in Analytics and Data Science Research Collections at DigitalCommons@Kennesaw State University. It has been accepted for inclusion in Grey Literature from PhD Candidates by an authorized administrator of DigitalCommons@Kennesaw State University. For more information, please contact digitalcommons@kennesaw.edu.

Automatic Knowledge Extraction from OCR Documents Using Hierarchical Document Analysis

Mohammad Masum
Kennesaw State University
Kennesaw, Georgia
mmasum@students.kennesaw.edu

Sai Kosaraju*
Kennesaw State University
Marietta, Georgia
skosara1@students.kennesaw.edu

Tanju Bayramoglu
GE Power
Atlanta, Georgia
tanju.bayramoglu@ge.com

Girish Modgil
GE Power
Atlanta, Georgia
girish.modgil@ge.com

Mingon Kang[†]
Kennesaw State University
Marietta, Georgia
mkang9@kennesaw.edu

ABSTRACT

Industries can improve their business efficiency by analyzing and extracting relevant knowledge from large numbers of documents. Knowledge extraction manually from large volume of documents is labor intensive, unscalable and challenging. Consequently, there have been a number of attempts to develop intelligent systems to automatically extract relevant knowledge from OCR documents. Moreover, the automatic system can improve the capability of search engine by providing application-specific domain knowledge. However, extracting the efficient information from OCR documents is challenging due to highly unstructured format [1, 11, 18, 26]. 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 corpus of documents at GE Power where document consists of more than hundred pages in PDF.

CCS CONCEPTS

• **Information systems** → **Question answering; Information extraction; Document structure;**

KEYWORDS

Ngrams, Expanded queries, Hierarchical analysis

ACM Reference Format:

Mohammad Masum, Sai Kosaraju, Tanju Bayramoglu, Girish Modgil, and Mingon Kang. 2018. Automatic Knowledge Extraction from OCR Documents

*Both Mohammad Masum and Sai Kosaraju are the first authors. They contributed equally

[†]Mingon Kang is the corresponding author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](https://permissions.acm.org).

RACS '18, October 9–12, 2018, Honolulu, HI, USA

© 2018 Association for Computing Machinery.

ACM ISBN 978-1-4503-5885-9/18/10...\$15.00

<https://doi.org/10.1145/3264746.3264793>

Using Hierarchical Document Analysis. In *International Conference on Research in Adaptive and Convergent Systems (RACS '18)*, October 9–12, 2018, Honolulu, HI, USA. ACM, New York, NY, USA, Article 4, 6 pages. <https://doi.org/10.1145/3264746.3264793>

1 INTRODUCTION

Automatic knowledge extraction from Optical Character Recognition (OCR) of documents is important for efficient and effective analysis of business-related documents [2]. With businesses adopting data digitization, documents in OCR have an advantage of archiving digital and non-editable copies of documents. The ability to analyze such documents in an automated manner improves business efficiency by enabling the information retrieval analysis and insights based on operational and domain specific key phrases, terms and items of interest. Automatic knowledge extraction can lessen human dependencies in understanding the information and add values to business by reducing costs. IBM Watson [8] is a representative example of the successful cases where an automatic knowledge extraction system is effectively utilized.

A typical approach ranks relevant texts based on a short query leading to term mismatch information retrieval. A short query may lack quantity of sufficient words to represent enough information for accurate information retrieval. Query Expansion (QE) technique have been used to address this problem. QE technique adds new tokens (words) to the existing keyword-based search terms to generate expanded queries. Local analysis [27] is one of the existing QE techniques that utilizes top ranked retrieved documents by a query. In this technique, the top ranked documents assume to be relevant to the query, and the query can be expanded based on this information. Relevance feedback [5], is the QE technique that extracts term expansion from relevant documents provided by users. However, there are limitations if users fail to provide appropriate relevant information to the given query [19].

Knowledge of interest can be extracted from text documents based on text mining techniques [9]. Bag Of Words (BOW) is a method used in text mining for computing similarity between sentences by splitting the sentence into words. However, the consideration of individual words of sentences often fails in computing similarity. For instance, the sentences, "United States Of America has Animal Kingdom" and "Animals rare in America are plentiful in United Kingdom" are not related to each other. The former describes a place called Animal Kingdom in USA, whereas the latter

is a comparative statement about animals in UK. In spite of this, the BOW approach shows a good relevance score because there are multiple words in common. N -grams [15] was proposed to tackle the problem. N -grams considers N consecutive words instead of a single word separately. For instance, 2-grams of the phrase, "United States of America", consider the following consecutive two words for BOW: "United States", "States of", and "of America". N -grams captures the context of text in a document. Lewis showed that the addition of *bi*-grams and *tri*-grams (i.e $N \leq 3$) to the BOW representation improves the performance for similarity score [15]. Grobelnik et al. also showed that the performance of relevance ranking can be improved by using N -grams technique with value of N is up to three rather than using only *uni*-grams [17]. Moreover, they reported no significant improvement in the performance of using longer sequence of words than three ($N > 3$).

The similarity of sentences are measured by Vector Space Model (VSM), where bag-of-words of sentences transferred into vector spaces by Term Frequency-Inverse Document Frequency (TF-IDF) factor of the word [25]. Term Frequency is defined as a number of times a word is shown repeatedly in a document, normalized by the total number of words in a document. Inverse Document Frequency (IDF) is value that reflect how important a word is to a document in collection of corpus.

2 RELATED WORKS

An effective knowledge extraction system can help accelerate business decisions if implemented correctly. Existing state of the art methods for knowledge extraction can be categorized into three: (1) keyword matching, (2) grammar analysis, and (3) rule based regular expression matching methods.

For keyword matching systems, knowledge can be extracted by matching a user-defined keyword to texts in documents [4, 6, 16, 28]. Texts in a document can be tokenized by a single space or a new line to match with the user-defined keywords. It considers that all words in the document are independent to each other. The performance of this approach depends on the provided keywords.

Detecting relationship between words, that have a higher probability of occurring together or have a close relationship, often plays an important role in extraction performance. Knowledge extraction via grammar analysis [7, 13, 22, 23, 29] is an approach where the relationship between the words is extracted by grammar rules. This approach is limited to finding the relation between verb-adjective or a noun-verb which follows grammar rules, but two closely related nouns or verbs cannot be related. For example, "price" and "payment" are nouns, which are closely related in a business domain, but their relationship cannot be defined by grammar analysis.

To obtain the relationship between two closely related words, rule-based regular expression matching systems have been proposed [14], where a set of rules is pre-defined by matching regular expression or searching with multiple keywords. However, this approach confines only a set of documents which follow the designed rules and involves domain experts to define rules, which can be expensive and hard to generalize the solutions.

In this paper, we propose a framework where the relationship between the words is achieved by using Query expansion (QE) technique and knowledge extraction with the user defined query

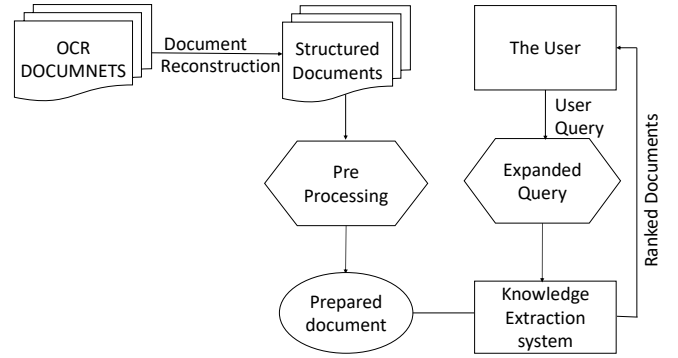


Figure 1: Process of knowledge extraction

is achieved by a vector space model and hierarchical document analysis.

3 METHODS

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 hierarchically structured data and the expanded query are then converted to a vector space model (VSM) and compared for ranking relevant paragraph to the query. Figure 1 illustrates the flow diagram of the framework

3.1 Document reconstruction

We convert OCR content into structural formatted data, by extracting document layout features and then analyzing the changes in the layout features. The document layout features include font height, font size, and boldness. Table 1 shows document text and their corresponding features of font height, boldness, font size in columns respectively. We assume block headings such as section/subsection headings follow a regular expression pattern (e.g. [a-z,A-Z]+ [0-9]+ [a-z,A-Z]+) and vary in the layout features. The section/subsection headings are then extracted by analyzing the changes of layout features and the regular expression pattern. Table 2 illustrates the headings in an OCR document with their locations in document. For example, the first row represents that the title "ARTICLE 1 DEFINITIONS" is in the seventh page and 42nd line of the document.

After extracting section/subsection headings, we can extract text present in between two headings as sections. For example, text data in between line 43 to line 284 in the document is extracted as a section with heading, "ARTICLE 1 DEFINITIONS". Algorithm 1 explains the process.

The same technique recursively can extract, subsections using the document layout feature analysis. Table 3 shows subheadings

Table 1: Feature Extraction

DATA	HEIGHT	BOLD	FONTSIZE
GECS	70	FALSE	14
GENERAL ELECTRIC INTERNATIONAL INC	14	TRUE	14
MASTER PRINTED COPY	24	TRUE	24
This copy is not be removed from its secure location...	16	TRUE	16
permission of the GECS Quality Programs Manager.	16	TRUE	16
A CONTROLLED electronic version can be viewed on the ...	13	FALSE	
Do not photocopy any pages from this	68	TRUE	22
printed copy be required an uncontrolled version may be...	13	FALSE	13
Programs Manager or printed from the electronic version...	14	FALSE	13

Table 2: Section Headings of OCR

INDEX	LOCATION	PAGE NUMBER	DATA
1	42	7	ARTICLE 1 DEFINITIONS
2	284	13	ARTICLE 2 CONTRACTOR RESPONSIBILITIES
3	490	18	ARTICLE 3 OWNER RESPONSIBILITIES
4	520	18	ARTICLE 4 TERM AND TERMINATION
5	596	20	ARTICLE 5 PRICE AND PAYMENT TERMS
6	1211	35	ARTICLE 6 DELIVERY TITLE
7	1382	39	ARTICLE 7 INSURANCE COVERAGE
8	1435	41	ARTICLE 8 WARRANTY

Algorithm 1 Section Extraction

- Step 1:** Extraction of layout features such as font-size *and* boldness
- Step 2:** Detecting the font-sizes *and* bold differentiation
- Step 3:** Filtering the text where their is a difference in font-size *or* boldness
- Step 4:** Assuming that titles have different layout features(e.g., font-sizes,boldness) and follow regular expression pattern of [a-z,A-Z]+ [0-9]+ [a-z,A-Z]+
- Step 5:** Filtering the text obtained in **step 3** with regular expression pattern
- Step 6:** Text remained after the **step 5** will be the headings of the section
- Step 7:** Text present between two concurrent headings is considering as a section

in its parent section. "6.1 Mobilization payment" in the first row is a subheading, and "2" represents the location of the subheading, "6.1 Mobilization payment". The feature analysis is repeated till all inner subsections are extracted. This process can generate hierarchically structured document from OCR documents.

The hierarchically structured data from OCR documents are stored as a dictionary format with the headings as keys and text below as values for the keys. Figure 2 illustrates the hierarchal structured headings, where "name" and "contents" represent the headings of the sections and text in the section respectively. The dictionary formatted data can take advantage of easy transition to NoSQL and TF-IDF representation for further analysis.

Table 3: Subtitles

INDEX	LOCATION	DATA
1	2	6.1 Mobilization Payment
2	4	6.2 Quarterly Payment
3	78	6.3 Parts
4	85	6.4 In addition to the price

```

"Name": "Section 8 INSURANCE REQUIREMENTS",
"contents": [ [ "INSURANCE REQUIREMENTS" ] ],
"sections": [
  [ {
    "Name": "8.1",
    "contents": [ "8.1", "Seller's Insurance", "During this term unless otherwise self insured Seller will maintain the",
    "following insurance coverage", "(a)", "During", "...." ],
    "sections": ""
  } ],
  [ {
    "Name": "8.3",
    "contents": [ "8.3", "Failure To Maintain Insurance", "Failure of any Party to maintain the insurance required under this",
    "Section 8 shall", "....." ],
    "sections": ""
  } ],
  [ {
    "Name": "8.2",
    "contents": [ "8.2", "Buyer's Insurance" ] ],
    "sections": [
      [ {
        "Name": "8.2.1 During the Term of this document Buyer shall maintain the following insurance",
        "contents": [ "8.2.1 During the Term of this document Buyer shall maintain the following insurance", "coverage",
        "(a) Workers' Compensation and any other statutory insurance required by law", "...." ],
        "sections": ""
      } ]
    ]
  } ]

```

Figure 2: Reconstructed document**3.2 Knowledge extraction**

we aims at reformulating the query to extract the most relevant knowledge from OCR documents with the help of query expansion (QE) technique. Tokens are collected by using N -grams technique (sequence of words in length N) [21] from documents and stored in bag of words (BoW) representation (Recall and subsequently

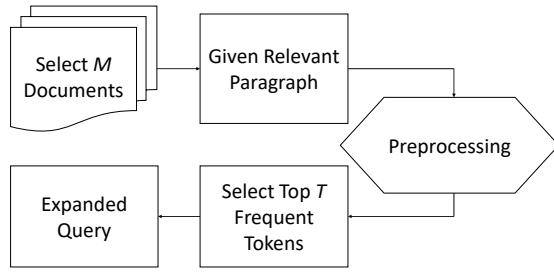


Figure 3: Process of expanded query

that $N \leq 3$). Then, TF-IDF [10, 12] is computed for the tokens. Each section/subsection of the documents is transformed into a vector space, where the entries of vector are TF-IDF values. Finally, cosine similarity [20], is used to investigate hierarchical relevance between the reformulated query (expanded query) and sections/subsections in a Vector Space Model (VSM) model.

N -grams ($N \leq 3$) is used to create new tokens from the preprocessed sections, new tokens are included with *uni*-grams in a Bag of words [24]. Initially, most relevant sections to a given query were provided from M documents manually. Preprocessing is performed on the sections. The expanded query is then generated from the BoW by selecting T most frequent of tokens where T is defined by the user. Figure 3 illustrates the flow diagram of developing expanded query.

The hierarchal structured documents after the document reconstruction process is used for further analysis. Every token in the hierarchal structured document is preprocessed and then assigned values using TF-IDF technique. In the process, each section of a document and the given query are transformed into vectors. Similarity score between the query and all sections is calculated in a pairwise manner using VSM. All of the sections are ranked by the similarity scores and the section that contains the highest similarity score is considered as the most relevant knowledge to the query. If the obtained section does not contain any subsection, the relevant section is retrieved as an output (the most relevant knowledge) for the query. If it contains subsections, the algorithm iterates over the section and determined the most relevant subsection. If the retrieved subsection contains hierarchical sub-subsection, the method repeats the process and investigates the most relevant subsubsection and so-on.

4 EXPERIMENTS AND RESULTS

In this section, we demonstrate experimental results of our proposed method.

4.1 Data

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 comprise of multiple layer of subsections. GEP provided 16 keywords (queries) for knowledge extraction. For each of the queries, they also provided five (M) relevant paragraphs (accepted answers to the

```

section described described_section January year amount shall price
january_year_thereafter upward_annual_basis upward_annual
shall_adjusted_upward adjusted_upward_annual accordance
payments shall_adjusted upward escalation basis_beginning
beginning annual adjusted thereafter fees basis
basis_beginning_january price_escalation year_thereafter
annual_basis adjusted_upward january_year payment
beginning_january annual_basis_beginning descnbedin_swtion
determined_accordance hours_adder_fees including
escalated_accordance section_shall ease_yeu
year_thereafter_greater yeu thereafter_greater_two
shall_escalated_accordance fees_termination monthly_fees ease
escalated_accordance_section payments_described_section
change_tie_conq cun index_descnbedin_swtion section_paid
formulas_described definitions_formulas_described .....
  
```

Figure 4: Expanded query

Table 4: Relevance Ranking of Sections

Sections	Similarity Score
PART 6	0.0321
PART 5	0.0195
PART 9	0.0090
PART 7	0.0032
PREAMBLE	0.0020

Table 5: Relevance Ranking of Sub Sections

Subsections	Similarity Score
6.2 Periodic Payments	0.03
6.12 Initial Spare Parts	0.0
6.4 Extra Work	0.0
6.3 Unplanned Extra Work	0.0
6.26	0.0

query) from five most relevant documents. These given answers were used to develop an expanded query.

We performed data preprocessing since stopwords contain less importance and these are common in all documents, these words are filtered out using stopwords list of python toolkit, NLTK [3]. Each of the remaining words (*uni*-gram) are considered as a token for further analysis.

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. ("Liquidated Damages" query provides information related to liabilities of industries in case of damages). First, we expanded "Liquidated Damages" query, by using query expansion technique. Figure 4 shows the expanded query of "Liquidated Damages". The expanded query is compared with sections in document by VSM. Table 4 shows the top five relevant sections in the document along with relevance scores for the given query "Liquidated damages". "PART 6" shows the highest similarity score of 0.0321, which is the most relevant section to the query.

Table 6: Relevance Ranking of Sections

Sub-subsections	Similarity Score
6.2.4 Liquidated Damages or Bonus	0.2776
6.2.1 Fixed Lump Sum Annual Payments	0.0
6.2.2 Periodic Price Escalation	0.0
6.2.3 Option for Second Major	0.0

“Liquidated Damages or Bonus
If the Maintenance Contraction owes ...
Liquidated damages for late delivery of parts or
Personnel in accordance with Section 5.1 of
Appendix A or if owes the Maintenance
Contractor a bonus for early completion of
Planned Maintenance event in accordance with
Section 3.2 of Appendix A such accounts will be
Settled up in the last payment in a given calendar
Year during which said condition

Figure 5: Most relevant paragraph to the query

The next step is to extract the most relevant portion with in Section "PART 6" to query. We compared the expanded query of "Liquidated Damages" to the subsections with in Section "PART 6". Table 5 shows the top five most relevant subsections in Section "PART 6". "6.2 Periodic Payments" is with the highest similarity score of 0.03, which is the most relevant subsection to the query. If the section "6.2 Periodic Payments" does not contain any sub-sections, "6.2 Periodic Payments" is extracted as the most relevant section for given query "Liquidated Damages". If "6.2 Periodic Payments" contains subsections with in it, we extract the most relevant section within Section "6.2 Periodic Payments". We compared the expanded query of the "Liquidated Damages" with subsections with in "6.2 Periodic Payments". Table 6 represents the most relevant sub-sections within "6.2 Periodic Payments" along with the relevance score for the query. "6.2.4 Liquidated Damages Bonus" with the relevance score of "0.2776" is the most relevant section within "6.2 Periodic payments" to the query "Liquidated Damages". Figure 5 represent the most relevant section within the document for the given query term "Liquidated Damages".

5 CONCLUSIONS

In this study, we present a knowledge extraction framework from OCR documents for 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 application such as automatic knowledge management and enrich the search systems. The advantage of using query expansion is to establish a correlations 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.

ACKNOWLEDGMENTS

We would like to thank our partners at GE Power for funding this study.

REFERENCES

- [1] H. Bast and C. Korzen. 2017. A Benchmark and Evaluation for Text Extraction from PDF. In *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. 1–10. <https://doi.org/10.1109/JCDL.2017.7991564>
- [2] Sanyam Bharara, Sai Sabitha, and Abhay Bansal. 2017. Application of learning analytics using clustering data Mining for Students' disposition analysis. *Education and Information Technologies* (2017). <https://doi.org/10.1007/s10639-017-9645-7>
- [3] Steven Bird and Edward Loper. 2004. NLTK: The Natural Language Toolkit. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*. <https://doi.org/10.3115/1118108.1118117> arXiv:cs/0205028
- [4] R. Chaniago and M. L. Khodra. 2017. Information extraction on novel text using machine learning and rule-based system. In *2017 International Conference on Innovative and Creative Information Technology (ICITech)*. 1–6. <https://doi.org/10.1109/INNOCIT.2017.8319148>
- [5] Hang Cui, Ji-Rong Wen, Jian-Yun Nie, and Wei-Ying Ma. 2002. Probabilistic query expansion using query logs. *Proceedings of the 11th international conference on World Wide Web* (2002). <https://doi.org/10.1145/511446.511489>
- [6] P. M. Darshan. 2017. Ontology based information extraction from resume. In *2017 International Conference on Trends in Electronics and Informatics (ICEI)*. 43–47. <https://doi.org/10.1109/ICOEL.2017.8300962>
- [7] T. Erekhinskaya, M. Balakrishna, M. Tatu, S. Werner, and D. Moldovan. 2016. Knowledge extraction for literature review. In *2016 IEEE/ACM Joint Conference on Digital Libraries (JCDL)*. 221–222.
- [8] J. Fan, A. Kalyanpur, D. C. Gondek, and D. A. Ferrucci. 2012. Automatic knowledge extraction from documents. *IBM Journal of Research and Development* (2012). <https://doi.org/10.1147/JRD.2012.2186519>
- [9] Alexander Gelbukh. [n. d.]. Natural Language Processing. ([n. d.]), 7695. <https://doi.org/10.1109/ICHIS.2005.79>
- [10] Vishal Gupta and Gurpreet S. Lehal. 2009. A survey of text mining techniques and applications. <https://doi.org/10.4304/jetwi.1.1.60-76>
- [11] M. Hanumanthappa and Deepa T. Nagalavi. 2015. Identification and extraction of different objects and its location from a Pdf file using efficient information retrieval tools. In *Proceedings of the IEEE International Conference on Soft-Computing and Network Security, ICSNS 2015*. <https://doi.org/10.1109/ICSNS.2015.7292375>
- [12] Siham Jabri, Azzeddine Dahbi, Taoufiq Gadi, and Abdelhak Bassir. 2018. Ranking of text documents using TF-IDF weighting and association rules mining. In *Proceedings of the 2018 International Conference on Optimization and Applications, ICOA 2018*. <https://doi.org/10.1109/ICOA.2018.8370597>
- [13] A. Kanev, S. Cunningham, and T. Valery. 2017. Application of formal grammar in text mining and construction of an ontology. In *2017 Internet Technologies and Applications (ITA)*. 53–57. <https://doi.org/10.1109/ITECHA.2017.8101910>
- [14] R. Kumaravel, S. Selvaraj, and M. C. 2018. A Multi-Domain Layered Approach in Development of Industrial Ontology to Support Domain Identification for Unstructured Text. *IEEE Transactions on Industrial Informatics* (2018), 1–1. <https://doi.org/10.1109/TII.2018.2835567>
- [15] David D. Lewis. 1992. An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval* (1992). <https://doi.org/10.1145/133160.133172>
- [16] Ahsan Mahmood, Hikmat Ullah Khan, Zahoor-Ur-Rehman, and Wahab Khan. 2018. Query based information retrieval and knowledge extraction using Hadith datasets. In *Proceedings - 2017 13th International Conference on Emerging Technologies, ICET2017*. <https://doi.org/10.1109/ICET.2017.8281714>
- [17] Andres McCallum and Kamal Nigam. 1998. A Comparison of Event Models for Naive Bayes Text Classification. *AAAI/ICML-98 Workshop on Learning for Text Categorization* (1998). <https://doi.org/10.1.1.46.1529> arXiv:0-387-31073-8
- [18] Thi Tuyet Hai Nguyen, Antoine Doucet, and Mickael Coustaty. 2018. Enhancing Table of Contents Extraction by System Aggregation. In *Proceedings*

- of the *International Conference on Document Analysis and Recognition, ICDAR*. <https://doi.org/10.1109/ICDAR.2017.48>
- [19] Gerard Salton and Chris Buckley. 1990. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science* (1990). [https://doi.org/10.1002/\(SICI\)1097-4571\(199006\)41:4<288::AID-ASI8>3.0.CO;2-H](https://doi.org/10.1002/(SICI)1097-4571(199006)41:4<288::AID-ASI8>3.0.CO;2-H) arXiv:arXiv:1011.1669v3
 - [20] M Steinbach, G Karypis, and V Kumar. 2000. A Comparison of Document Clustering Techniques. *KDD workshop on text mining* (2000). <https://doi.org/10.1109/ICCCYB.2008.4721382>
 - [21] Chade Meng Tan, Yuan Fang Wang, and Chan Do Lee. 2002. The use of bigrams to enhance text categorization. *Information Processing and Management* (2002). [https://doi.org/10.1016/S0306-4573\(01\)00045-0](https://doi.org/10.1016/S0306-4573(01)00045-0)
 - [22] R. Upadhyay and A. Fujii. 2016. Semantic knowledge extraction from research documents. In *2016 Federated Conference on Computer Science and Information Systems (FedCSIS)*. 439–445.
 - [23] D. G. Vasques, A. C. Zambon, G. B. Baioco, and P. S. Martins. 2016. An Approach to Knowledge Acquisition Based on Verbal Semantics. In *2016 49th Hawaii International Conference on System Sciences (HICSS)*. 4144–4153. <https://doi.org/10.1109/HICSS.2016.514>
 - [24] Hanna M. Wallach. 2006. Topic Modeling: Beyond Bag-of-words. In *Proceedings of the 23rd International Conference on Machine Learning (ICML '06)*. ACM, New York, NY, USA, 977–984. <https://doi.org/10.1145/1143844.1143967>
 - [25] Ho Chung Wu, Robert Wing Pong Luk, Kam Fai Wong, and Kui Lam Kwok. 2008. Interpreting TF-IDF term weights as making relevance decisions. *ACM Transactions on Information Systems* (2008). <https://doi.org/10.1145/1361684.1361686>
 - [26] Zhaohui Wu, Prasenjit Mitra, and C. Lee Giles. 2013. Table of contents recognition and extraction for heterogeneous book documents. In *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*. <https://doi.org/10.1109/ICDAR.2013.244>
 - [27] Jinxi Xu and W Bruce Croft. 1996. Query expansion using local and global document analysis. *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '96* (1996). <https://doi.org/10.1145/243199.243202>
 - [28] Wenhao Zhu, Laihu Luo, Chaoyou Ju, and Bofeng Zhang. 2012. Cross language information extraction for digitized textbooks of specific domains. In *Proceedings - 2012 IEEE 12th International Conference on Computer and Information Technology, CIT 2012*. <https://doi.org/10.1109/CIT.2012.226>
 - [29] S. T. Zuhori, M. A. Zaman, and F. Mahmud. 2017. Ontological knowledge extraction from natural language text. In *2017 20th International Conference of Computer and Information Technology (ICCTIT)*. 1–6. <https://doi.org/10.1109/ICCTECHN.2017.8281776>