Digital Library of Online PDF Sources: An ETL Approach

Gohar Zaman[1], Hairulnizam Mahdin[1], Khalid Hussain[2], Atta-ur-Rahman[3,*], Nehad Ibrahim[3] and Noor Zuraidin Mohd Safar[1]

[1]Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia
[2]Barani Institute of Sciences (Sahiwal), PMAS Arid Agriculture University, 46000, Rawalpindi, Pakistan
[3]Department of Computer Science, College of Computer Science, and Information Technology (CCSIT), Imam Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam, 31441, Saudi Arabia
*Corresponding author

Abstract: It is evident from day to day web usage experience that a huge number of PDF sources have been uploaded on daily basis. For example, there are several scientific societies that publish volumes of articles and periodicals like IEEE, ACM, Elsevier, and Springer etc. Most of these resources are unstructured or semi-structured that makes it difficult to search and retrieve information. In this paper, an effective model for digital library creation is proposed which is originally motivated by an automated ontological information extraction framework (OFIE). The framework takes a PDF published paper, extracts its structural information like title, authors, abstract, funding information, table of contents, references etc. with the help of fuzzy rule-based system (FRBS) and word sense disambiguation (WSD) approach. Consequently, this extracted information is converted to RDF triples. The proposed scheme takes this extracted information and converts into a digital library stored in MS-SQL databased by Extract, Transform and Load (ETL) process. This digital library can be an institute's library or an individual scholar's library who is interested in synthesizing his downloaded PDF files for better search and retrieve purposes. Moreover, by using the SQL queries based front-end design, the information can be searched, retrieved, and exported in the form of reports.

***Keywords**: Ontology, Digital Library, ETL, SQL, RDF, OFIE, FRBS*

1. Introduction:

World wide web is a tremendous source of a huge number of documents especially published scientific knowledge comprised of heterogeneous formats depending on several publishing houses. Although many publishing houses are switched to structured (metadata) based publications but still a significant number of publishers simply uploaded the PDF version of the published articles. Consequently, such unstructured or semi-structured documents are skipped from the search engines hence a huge source of knowledge goes undiscovered. In this regard, there are several approaches that converts these unstructured PDF sources into structured sources by extracting and appending the metadata for better indexing [1].

Information extraction is among the hottest areas of research in text mining and natural language processing (NLP). Several techniques have been investigated in the literature in this regard. It is also regarded as one of the most useful tool for information processing in case of machine readable approaches like Electronic Data Interchange (EDI) [2] in healthcare [3], sentiment analysis and text polarization [4] and semantic web [5] etc.

There are various approaches exist in the literature for automated information extraction. The well-known approaches in the literature are rule-based information extraction [5], data/text-mining and machine learning based information extraction [6], conditional random fields (CRF) based information extraction [7] and rich text properties-based information extraction [8]. Different approaches have different level of effectiveness in terms of accuracy and the type of data being extracted. Some approaches simply employed to extract some partial information like just authors' information from the PDF document automatically [9]. While other approaches aim for complete metadata extraction [5, 10-11].

In the information extraction paradigm, no approach is regarded as universal, rather approaches may be more or less effective based on the type, format and nature of document etc. It is possible that one approach which is incredibly good for one case may not be that effective or may be not effective at all in the other case. This is due to the heterogeneous nature of documents being produced online with various formats [1].

Ontology based information approaches have been gaining popularity over last decade due to the emergence of semantic web technology [5, 12]. The ontologies provide an inherent data dimensional perspective of information being extracted and later can be plugged in to the web directly or through the RDF databases [13] and relational databases like Oracle and SQL server [14]. As an institute or an individual scholar, there is a huge collection of downloaded PDF sources stored in the computer. It is tedious to search for a particular information from this collection while just relying on operating systems search options. Digital libraries have been gaining a tremendous popularity over

Manuscript received October 5, 2020
Manuscript revised October 20, 2020
https://doi.org/10.22937/IJCSNS.2020.20.10.19

the conventional libraries especially in an institute scenario. Rahman and Alhaidari (2018) proposed a model of digital library and archiving system for the institutes. The library was equipped with searching and archiving mechanisms. E-books metadata import and export engines were there however, for a physical element (book, article) or electronic element without metadata, administrator inserted the data manually. That means it was supporting the automatic information extraction from the electronic documents [15]. This is one of the basic motivations behind the proposed work to enhance a digital library by means of automatic information extraction from the semi and non-structured PDF sources. In this way, it will be two fold, firstly the structured document can immediately be transformed and import to the digital library while if the unstructured or semi-structured documents can easily be transformed into a digital library objects by means of automated information extraction (metadata). Currently, most of the digital libraries are missing this prominent feature like in [15].

The remainder of this paper is organized as follows. In section 2, system model is introduced. Performance of different codes in conjunction with different modulations is presented in section 3. The results of section 3 are used in section 4 to formulate a constrained optimization problem. In section 5 a brief introduction to Fuzzy Rule Base is given that is used to solve the optimization problem formulated in previous section. Section 6 presents the performance comparison of proposed scheme with various other famous adaptive schemes while section 7 concludes the paper.

2. System Model

Based on the above discussion on the proposed ontological framework that takes a PDF published paper, extracts its meta data and other information like title, authors, references etc. with the help of fuzzy regular expressions and word sense disambiguation and converts it into RDF triples. The proposed scheme takes this extracted information and converts into a digital library by Extract, Transform and Load (ETL) process. This digital library can be viewed as an institute library or it could be owned by an individual scholar who is interested in synthesizing his downloaded PDF files for better search, retrieve, sharing and publishing purposes. Moreover, by using the SQL queries based front-end design, the information can be searched, retrieved, shared, and exported in the form of reports. Figure 1 contains the conceptual model for the proposed digital library system. The constituent components of the proposed digital library are explained subsequently.

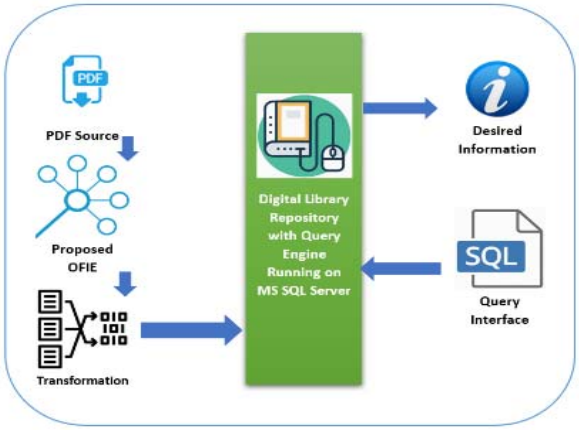

Figure 1: Digital Library Conceptual Model

3. Ontological Information Extraction Framework

The said framework is an ontological dynamic and heuristic based approach that extracts the information about logical structure and supportive materials of published research papers of diversified publication formats, styles and structures belong to various research societies like IEEE, ACM, Springer and Elsevier. The technique has an ability to learn new rules of enhanced information extraction (IE). This is the case where there is a variation among the document formats being considered for information extraction. To address this variation, the technique must be adaptive and robust. This framework is proposed in [16] by the us (same authors) and it is reported here for reference purpose.In this regard, instead of having a fixed rule-base (knowledgebase), a Fuzzy Rule Based System (FRBS) is to be investigated to learn new rules, based on the training provided by the existing rules. It is therefore, presumed that the proposed scheme will be dynamic in terms of IE from the documents with the formats different from those considered in the examples during training phase. A backend ontology, based on which the entire IE frame is defined, will work as criteria for learning new rules based on the new format presented during the testing phase. Moreover, a validator module for verification of the new formats being properly addressed will be added during training phase and removed once the system has sufficiently learnt. Further, to semantically fine tune the extraction process, semantic network-based Word Sense Disambiguation (WSD) will also be incorporated in the proposed technique to avoid possible misclassification of the extraction terms in the structure. The dataset for the undergoing research is based on the selected published articles of four scientific societies namely IEEE, Springer, ACM, and Elsevier, taken in terms of samples over last ten years. The methodology and the steps carried out in the research are shown in Figure 2.

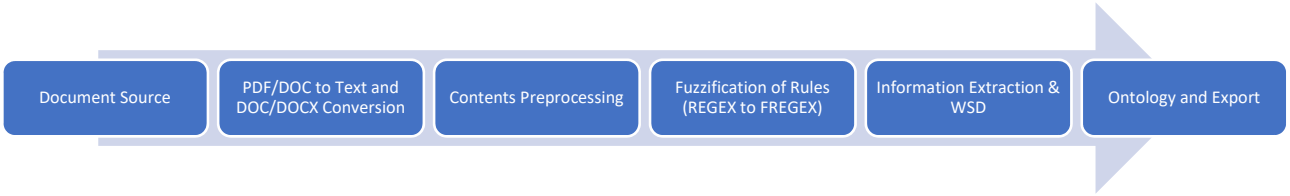

Figure 2: Methodological Process

The steps included in the proposed approach are given subsequently along with their complete detail and description. Each component is explain pertaining the undergoing research and the way it is applied in the proposed framework.

A. PDF to XML, text and doc/docx conversion

First, the given document which is supposed to be in PDF format will be converted to XML, plain text and DOC/DOCX formats. Purpose for converting the source documents (articles) into more than one formats is that each conversion has its own pros and cons and from the experiments, it is observed that some part of information may be better extracted from XML, plain text and rest from DOC and vice versa. To compliment, all the formats are used. Mainly, XML provides a better extraction support due to its tag nature, DOC/DOCX provides rich text format (RTF) features like fonts, headings and numbering etc. for a better understanding and text provide plain text in a better way like the abstract, acknowledgement part etc.

B. CONTENTS PRE-PROCESSING

It is very important phase in the information extraction. In this phase, unnecessary details omission, data cleansing and other type of pre-processing is performed. In this phase different parts of the research papers like title, authors, funding agency are identified using different rules (mainly in the form of regular expressions) and rest of the tokens are discarded. For example, in this case, we are not interested in publication year, journal ISSN and paper's main text etc. so such information can be filtered out.

C. FUZZY REGULAR EXPRESSIONS

This part is responsible for incorporating the new formatting/corpus related rules. If there is no change, then it will simply produce the output. Otherwise it will go for a rule by making it more flexible in terms of deletion, insertion, and substitution. It is mainly carried out by translating the rules which are regular expressions (REGEX) into fuzzy rules which are Fuzzy regular expressions (FREGEX) or FREJ in Java. Fuzzy regular expressions have a tolerance in terms of certain numbers of errors calculated by deletion, insertion, and substitution. The error measure is taken in terms of Levenshtein-distance of patterns with minimum number of insertions, deletions, or substitutions for pattern transformation.

D. FUZZY RULE BASED SYSTEM

As explained earlier, we have been targeting four societies mainly that contains several journals with varying formats. Moreover, in the dataset, other journals are also included for sake of testing the proposed approach. In this regard, it is very important to figure out the appropriate fuzzy regular expression (index) upon detecting a particular society (say Elsevier), the second parameter in this regard is structural index (SI) that specifies whether the given text token is title, abstract, keyword etc. The third parameter in this regard is tolerance (T), that specifies the extent to which the distance between pattern and text token can be tolerated. It is worth mentioning here, that high tolerance does not always means error is avoided. In some cases, more tolerance can result in poor detection and/or accuracy. Figure 3 shows the schematic of the FRBS.

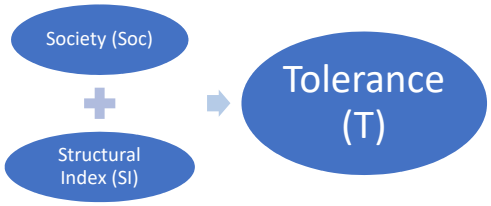

Figure 3: Schematic of Fuzzy System

To fine tune the performance by tolerating the average error, the FRBS is designed to estimate the exact tolerance being used by the fuzzy regular expression. The sample rule can be expressed as:

IF (SOC = 'index' *AND* SI = 'index') *THEN* (*AND* T = V.High)

Components of Fuzzy Rule Based System

There mainly four components of fuzzy rule-based system. Namely, fuzzifier, defuzzifier, inference engine and the rule base. In this research, we have used Triangular Fuzzifier, Center Average Defuzzifier and Mamdani Inference Engine (MIE) [17]. The rule based is created using lookup table approach [21-26]. The sample lookup table is given in Table 1.

Table 1: Lookup Table

	S. Index	0	1	2	3	4	5	6	7	8
Soc		Title	Name	Abstract	Keyword	Heading	Figure	Table	Ack	Ref
0	IEEE									
1	ACM									
2	Elsevier				Low (1)	V. Low (0)				
3	Springer									
4	Others	High (3)								

Here the input variable soc (society) is depicting the four main and one generic society ranging from [0-4]. Similarly, second input variable SI (structural index) ranging between [0-8] representing the structural index in an arbitrary paper. The output variable t (tolerance) is ranging between [0-4] against the fuzzy values (very low, low, medium, high, and very high). Figures 4 and 5 depict fuzzy input variables with their membership functions and Figure 6 shows the fuzzy output variable with its membership function mapping.

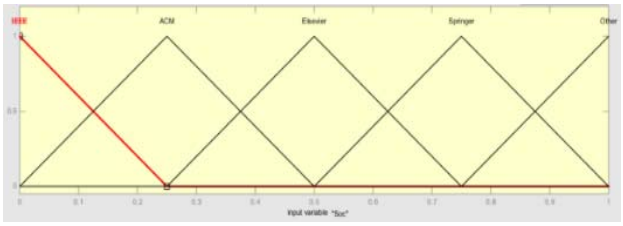

Figure 4: First Input variable Society (Soc)

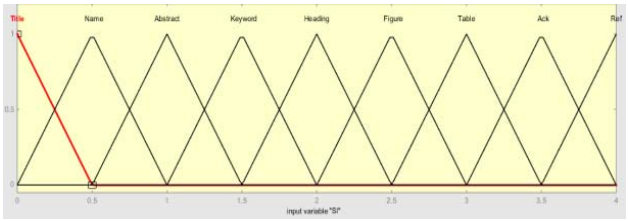

Figure 5: Second Input variable structural index (SI)

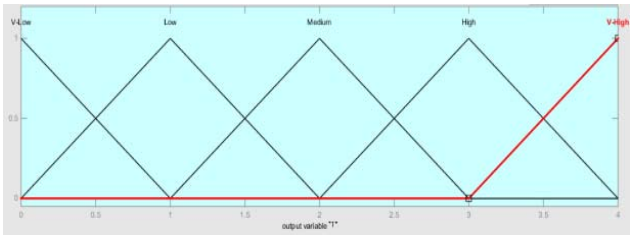

Figure 6: Output variable Tolerance (T)

E. Word Sense Disambiguation

This module is responsible for false alarms. If some word is being misinterpreted by the proposed scheme it will help suggesting the right sense. It is a value addition in the proposed technique to improve the accuracy in the information extraction process. In the implementation, it is performed by word to vector (word2vec) based similarity model [18]. The test is performed on the values for segments (author name, email etc.) obtained from XML, text, and MS word input documents, respectively. The module takes two expressions (one from word and/or text and second from XML converted document). This is because, from heuristic, it is found that some fields can be better extracted from text, doc, and XML. So, to get the right information that information is fed to WSD to obtain accurate information from any pair of strings say 'title'. Word2vec is a Neural Network based model in which the words or the concepts are represented in terms of n-dimensional vectors in a huge vector space [19]. Word2vec further finds the nearest or closest words semantically and/or syntactically. Here this approach helps us figure out the best possible outcome to be taken between two versions extracted separately that are from word/text and from XML. This is the most important phase in terms of sensing the extracted structural information in terms of removing ambiguities that are also not identified in FRBS. This part is responsible for determining the similarity between two sentences and checking the sense and context of two sentences using a natural language processing (NLP) module (library) called spaCy [20]. The term 'ambiguous' in this research refers to the meaning and sequence of sentences that is extracted from both versions i.e. text version and XML version in correct manner or not. It involves data science for checking the similarity and sensing the information. Figure 7 shows the conceptual diagram of spaCy. Similarity is determined by comparing word vectors or "word embeddings", multi-dimensional meaning representations of a word. Word vectors can be generated using an algorithm like word2vec. The words "dog", "cat" and "banana" are all common in English, so they are part of the model's vocabulary, and come with a vector. The word "afskfsd" on the other hand is a lot less common and out-of-vocabulary – so its vector representation consists of 300 dimensions of 0, which means it is practically nonexistent. If your application will benefit from a large vocabulary with more vectors, you should consider using one of the larger models or loading in a full vector package, for example, en_vectors_web_lg, which includes over 1 million unique vectors. spaCy can compare two objects and make a prediction of how similar they are. Predicting similarity is useful for building recommendation systems or flagging duplicates. For

example, you can suggest a user content that is like what they are currently looking at or label a support ticket as a duplicate if it is very similar to an already existing one.

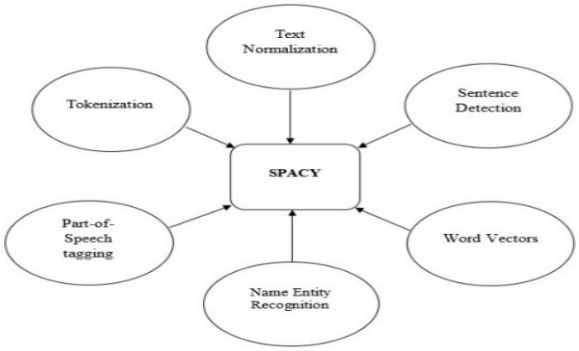

Figure 7: The conceptual diagram of spaCy library [20]

As mentioned earlier word2vec technique with the latest spacy library have been investigated in this research. Basically, we are given with the two types of data columns that are being extracted by fuzzified regular expression of two document versions that are text version and XML version. Now the aim to select the best possible extracted data value from these two versions. For this purpose, first the data is converted into individual sentences. Then the length of each sentence from both versions is checked, the adequate length will be saved. In second step, the sentences will be broken and checked word by word with the local word cloud or word net. If the words of sentence are present in word cloud then it will be saved, otherwise it will be rejected. In third and most important step the similarity based on semantically contextual information is checked. The said process for WSD is shown in Figure 8.

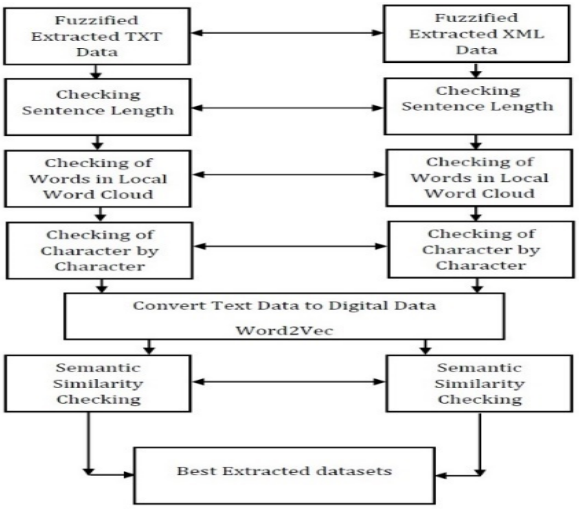

Figure 8: WSD process

For example, in the Table 2, a title extracted in txt format is “information extraction from diverse scientific sources” and the title extracted in xml format is “extraction diverse sources information from scientific”. It is apparent that there is wrong sentence sequence is in XML version and lack of sense. The word2vec converts these two sentences into vectors and then check the similarity between these two sentences and finally pick the best sentence with the right sense and right sequence and save this sentence into a resultant database column.

Table 2: WSD Example.

	Text Version	**XML Version**	**Correct Version**
Title	Information Extraction from Scientific Sources	Information from Sources Extraction Sources	Information Extraction from Scientific Sources
Section No. with Headings	2. Related Work 3. Textual Definition 5. OFIE Model	1. Introduction 2. Related Work 3. Textual Definition 4. Propose Approach 5. OFIE Model	1. Introduction 2. Related Work 3. Textual Definition 4. Propose Approach 5. OFIE Model

F. ONTOLOGY

This is the structure comprised of all the fields, subfields, and possible segments about which, data is going to be extracted. For example, authors, title, sub-title, sections and other useful and required information. The ontology is designed/engineered in Protégé as shown in Figure 9. The left side panel represent all the fields of the research paper and their logical relationships. Middle panel represent the sample extracted information against those fields.

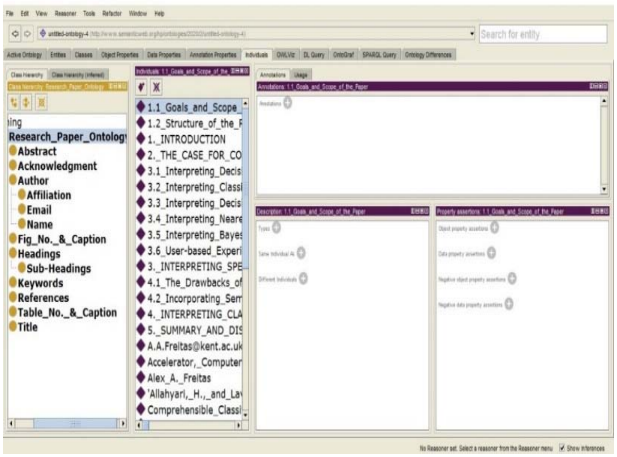

Figure 9: Ontology Implementation in Protégé

4. Proposed Digital Library

It further illustrates that the new PDF source is introduced to the system. The proposed OFIE model extracts the structural information from the source and transform it into a digital library object in the form of an excel or comma separated file. The digital library in turn, saves the information as a new record in the database running on MS SQL server. In this way we can add arbitrary number of articles. The query interface provides a graphical user interface to the user. In the interface there are several searching options like search by author name, complete title or a title word, keyword or any word being used in the abstract. As a result, all the records (saved papers) are fetched to the user. That information can be read and exported as a report because there might be several records containing the information being searched. Figure 10 shows a newly extracted record from the file that was browsed from the "Browse" button. This information can be inserted as a new record to the digital library by clicking the "Save to DB" button.

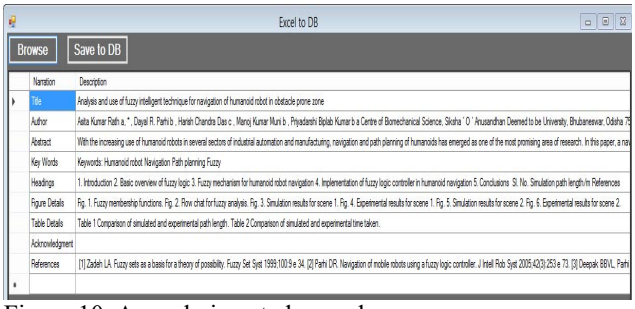

Figure 10: A newly inserted record

Figure 11 shows the search menu. That provides the means of searching in the digital library by title, author name, keyword, and any word in abstract. Consequently, all the matching records will be fetched as shown in the next figure.

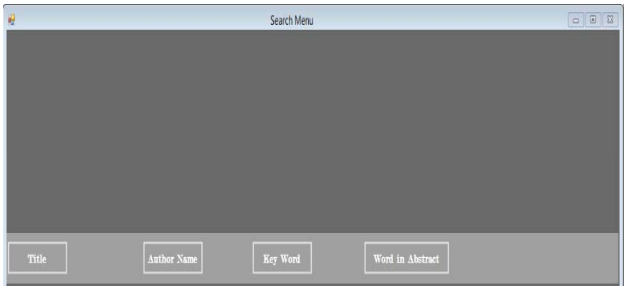

Figure 11: Search Menu

Figure 12, 13 and 14 show the sample search result after searching by title word, author name and word in the abstract. In either case user can make a choice to search as a whole word or partial word. This is helping when the user does not know the complete word or facing some spelling issues.

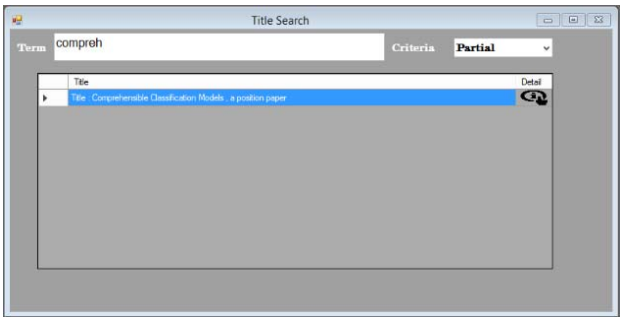

Figure 12: Search by title word

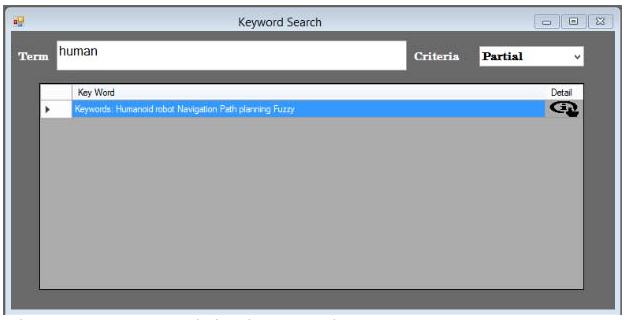

Figure 5.17: Search by keyword

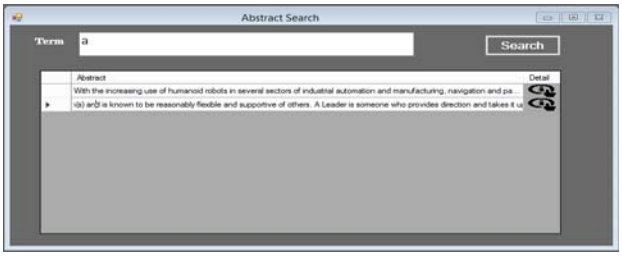

Figure 5.17: Search by a word in abstract

5. Conclusion

This research presents the creation of a digital library to archive and retrieve the information extracted from diverse scientific sources like papers published by IEEE, ACM, Springer etc. In this regard, a relational database management system (RDBMS) based digital library was built in Microsoft SQL Server. The information was extracted using OFIE framework which was available in RDF and Excel format. A routine was written to convert the Excel file to the database record. In this way, several articles' information was archived in the digital library. To search the contents of the digital library and search mechanism was develop where the user can search the designed information based on various filters like author name, title word etc. In future, this concept of digital library can be extended to semantic web and online databases can be investigated in this regard.

Acknowledgement

This research was funded by the Ministry of Education Malaysia (MOE) under the Fundamental Research Grant Scheme for Research Acculturation of Early Career Researchers (FRGS-Racer) RACER/1/2019/ICT04/UTHM/1 Vote: K154.

References

[1] G. Zaman, H. Mahdin, K. Hussain, A. Rahman, "INFORMATION EXTRACTION FROM SEMI AND UNSTRUCTURED DATA SOURCES: A SYSTEMATIC LITERATURE REVIEW", ICIC Express Letters 14(6):593-603, 2020.

[2] A. Rahman, F. Alhaidari, "An Electronic Data Interchange Framework for Education Institutes," ICIC Express Letters 13(9):831-840, 2019.

[3] A. Rahman, J. Alhiyafi, "Health Level Seven Generic Web Interface," Journal of Computational and Theoretical Nanoscience 15(4), pp. 1261-1274, 2018.

[4] S. Khan, N.M. Nawi, M. Imrona, A. Shahzad, A. Ullah, A. Rahman, "Opinion Mining Summarization and Automation Process: A Survey," International Journal on Advanced Science Engineering and Information Technology 8(5), pp. 1836-1844, 2018.

[5] R. Ahmad, M. T. Afzal, and M. A. Qadir, "Information Extraction from PDF Sources Based on Rule-Based System Using Integrated Formats," in Semantic Web Evaluation Challenge, 2016, pp. 293–308.

[6] S. A. Salloum, M. Al-Emran, A. A. Monem, and K. Shaalan, "Using text mining techniques for extracting information from research articles," in Intelligent Natural Language Processing: Trends and Applications, Springer, 2018, pp. 373–397.

[7] F. Peng and A. McCallum, "Information extraction from research papers using conditional random fields," Inf. Process. Manag., vol. 42, no. 4, pp. 963–979, 2006.

[8] M.-T. Luong, T. D. Nguyen, and M.-Y. Kan, "Logical structure recovery in scholarly articles with rich document features," in Multimedia Storage and Retrieval Innovations for Digital Library Systems, IGI Global, 2012, pp. 270–292.

[9] H. H. N. Do, M. K. Chandrasekaran, P. S. Cho, and M. Y. Kan, "Extracting and matching authors and affiliations in scholarly documents," in Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries, 2013, pp. 219–228.

[10] S. Kim, Y. Cho, and K. Ahn, "Semi-automatic metadata extraction from scientific journal article for full-text XML conversion," Proc. of the international conference on data mining (DMIN), 2014, p. 1.

[11] P. Groth, M. Lauruhn, A. Scerri, and R. Daniel, "Open Information Extraction on Scientific Text: An Evaluation," arXiv:1802.05574, pp. 3414–3423, 2018.

[12] S. T. R. Rizvi, D. Mercier, S. Agne, S. Erkel, A. Dengel, and S. Ahmed, "Ontology-based Information Extraction from Technical Documents," ICAART 2018 - 10th International Conference on Agents and Artificial Intelligence 2018.

[13] A. Rahman, F. Alhaidari, "Querying RDF Data", Journal of Theoretical and Applied Information Technology 26(22):7599-7614, 2018.

[14] M. Ahmad, M.A. Qadir, A. Rahman, T. Ali, R. Zagrouba, F. Alhaidari, F. Zahid, "Enhanced query processing over semantic cache for cloud based relational databases", Journal of Ambient Intelligence and Humanized Computing, 2020. DOI: 10.1007/s12652-020-01943-x.

[15] A. Rahman, F. Alhaidari, "The Digital Library and the Archiving System for Educational Institutes." Pakistan Journal of Library and Information Science 20(1):94-117, 2018.

[16] G. Zaman, H. Mahdin, K. Hussain, A. Rahman, J. Abawajy "An Ontological Framework for Information Extraction from Diverse Scientific Sources", Submitted to IEEE Access.

[17] A. Rahman, S. Dash, A.K. Luhach, N. Chilamkurti, S. Baek, Y. Nam, "A Neuro-Fuzzy Approach for User Behavior Classification and Prediction", Journal of Cloud Computing, 8(17), 2019.

[18] A.A. Rashid, S.A. Rahman, N.N. Yusof and A. Mohamed, "Word sense disambiguation using fuzzy semantic-based string similarity model", Malaysian Journal of Computing, 3 (2): 154–161, 2018.

[19] Mikolov, Tomas & Sutskever, Ilya & Chen, Kai & Corrado, G.s & Dean, Jeffrey. (2013). Distributed Representations of Words and Phrases and their Compositionality. Advances in Neural Information Processing Systems. 26

[20] Spacy.io, "Industrial-Strength Natural Language Processing IN PYTHON" (accessed 18 September 2020).

[21] A. Rahman, “Teacher Assessment and Profiling using Fuzzy Rule based System and Apriori Algorithm”, International Journal of Computer Applications (IJCA), Vol. 65(5), pp. 22-28, March 2013.
[22] A. Rahman, “GRBF-NN based ambient aware realtime adaptive communication in DVB-S2”, Journal of Ambient Intelligence and Humanized Computing 2020(12):1-11, 2020. DOI: 10.1007/s12652-020-02174-w
[23] A. Rahman, S. Dash, A.K. Luhanch, “Dynamic MODCOD and Power Allocation in DVB-S2: A Hybrid Intelligent Approach”, Telecommunication Systems, 2020. DOI: 10.1007/s11235-020-00700-xز
[24] A. Rahman, “Memetic Computing based Numerical Solution to Troesch Problem”, Journal of Intelligent and Fuzzy Systems, 37(1):1545-1554, 2019.
[25] A. Rahman, “Optimum Information Embedding in Digital Watermarking”, Journal of Intelligent and Fuzzy Systems, 37(1):553-564, 2019.
[26] A. Rahman, F.A. Alhaidari, D. Musleh, M. Mahmud, M.A. Khan, “Synchronization of Virtual Databases: A Case of Smartphone Contacts”, J. Comput. Theor. Nanosci., vol. 16 (3), 2019.
[27]

Authors Biography

Gohar Zaman is currently a postgraduate research student at the Faculty of Computer Science and Information Technology (FSKTM), Universiti Tun Hussein Onn Malaysia. His research interests are Information Extraction, Data Mining, Ontologies, NLP, and Automatic Text Categorization.

Hairulnizam Mahdin is an associate professor at Faculty of Computer Science and Information Technology (FSKTM), Universiti Tun Hussein Onn Malaysia. He is a member of Malaysia Board of Technologist (MBOT) and IEEE. He has been actively involved in many conferences internationally serving in various capacities including chair, general co-chair, vice-chair, best paper award chair, publication chair, session chair and program committee. He has also guest edited many special issues in journals. His current interests are in IOT and Image Processing.

Professor Dr. Khalid Hussain is working as Professor and Dean Faculty of Computing, Barani Institute of Sciences Sahiwal and in parallel he is working as Campus Director in Burewla Campus. Dr. Khalid has vast university / industry experience. During his tenure in the industry, he served in the defense related projects and in recognition of his services, he has been awarded commendation certificates by multiple government agencies. He joined academia in 2008 as full-time faculty member. In addition to his teaching role, he has been involved in numerous research projects. He helped setup a pioneer setup for information/ network security certification in Pakistan. He also introduced EC Council certification under the first academia industry partnership. He did his PhD from Malaysia, under a fully funded UTM / HEC scholarship. Up till now he published 63 papers. In which 27 are ISI Indexed Impact Factor, 13 are in HEC approved journal and 23 are in IEEE and ACM conferences. Except this he also has a book chapter and one book with the title “Information Security Handbook” is going to publish in couple of months. He has successfully completed six applied research projects in the domain of Information Security funded by NESCOM. Up till now 37 MS students completed his research thesis under his supervision. Currently he is supervising 13 MS and five PhD student in which one PhD is completing within couple of months. In reward of his contribution towards Information Security SATHA awarded him Gold Medal in 2015.

Dr.Atta-ur-Rahman is currently working at College of Computer Science and Information Technology (CCSIT), Imam Abdulrahman Bin Faisal University (IAU), Dammam, KSA, as Assistant Professor. He has completed his BS degree in Computer Science from University of The Punjab, Lahore, Pakistan; MS degree in EE from International Islamic University, Islamabad, Pakistan and PhD degree in EE from ISRA University, Islamabad Campus, Pakistan in years 2004, 2008 and 2012, respectively. He has been involved in teaching and research since 2006 and authored/co-authored more than 100 publications in conferences, books, and journals of good reputation. His research interests include Digital/Wireless Communication, DSP, Information & Coding Theory, AI, and Applied Soft computing.

Nehad M. Abdel Rahman Ibrahim is a lecturer in Computer Science and Information Technology College at Imam Abdulrahman bin Faisal University, Dammam, Saudi Arabia from 2016. He received his B.Sc. in systems and computer engineering from Al-Azhar University-Egypt in 1997. He has served as the software development manager at First Egyptian Inc. from 2008-2014. He received his M. Sc. degree in systems and computer engineering from Al-Azhar University-Egypt in 2008 (Distributed Object-Oriented Database). He received his PhD. in systems and computer engineering from Al-Azhar University-Egypt in 2020 (Text Mining of Arabic Content on The Web). He published many papers in fields such as artificial intelligence applications.

Noor Zuraidin Mohd Safar is a lecturer at Department of Information Security and Web Technology, Faculty of Computer Science, and Information Technology (FSKTM), Universiti Tun Hussein Onn Malaysia. He received BSc in Computer Science from University of Tulsa Oklahoma, MSc in Internetworking Technology from Universiti Teknikal Malaysia and Ph.D. from University of Portsmouth, United Kingdom. His research focuses on the area of machine learning, soft computing in environmental, meteorological data in tropics, computer network security and web technology.