International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-1, November 2019

Machine Learning and Security Applications in Digital Library

Beena A L, Humayoon Kabir S

Abstract: ***This paper presents the applications of machine learning machine learning applications in the digital library. Using machine learning it is possible to search and retrieve non-textual information. The paper also discusses the machine learning applications in security aspects. A systematic review of literature is also done and with the help of citation mapping in Web of Science citation network analysis is presented***

Keyword: machine learning, digital library, knowledge base, citation network analysis(CNA).

I. INTRODUCTION

Machine learning is the application of artificial intelligence, wherein systems automatically get data and learn from data. The machine learning algorithms may be supervised, unsupervised, semi-supervised and reinforcement machine learning algorithms. Section II of the paper gives some of the previous works in the field. Section III describes the development of a knowledge base and the applications of machine learning in the digital library. Section IV details the machine learning techniques in security applications. Section V details citation network analysis (CNA). Results and discussion are detailed in section VI.

II. RELATED WORKS

A method for the analysis and detection of the table of contents was introduced by Lin and Xiong in 2005 using Natural Language Processing (NLP). This method is most useful in the digitization of documents. Here they have used content association [1]. Intelligent analysis of unstructured data is possible through machine learning. [2] This feature is useful in information search and retrieval in the digital library. An algorithm termed Key phrase Identification Program (KIP) is developed that extracts the key phrases in the document and helps the user to build up a database of key phrases which in turn makes learning effective [3]

Revised Manuscript Received on November 30, 2019.
***** Correspondence Author
Beena A L* Department of Computer Applications, Cochin University of Science & Technology, Cochin, India. Email: beena.al@gmail.com
Dr. Humayoon Kabir S, former Head, Department of Library and Information Science, University of Kerala, Trivandrum, India. Email: humayoonkabirs@gmail.com

The internet is flooded with a lot of information and extracting the required information from the internet is a tedious task. One of the efficient mechanism for this is text summarization. It compresses a large document by summarizing and including the most relevant text. Baber and Patil in 2015 presented a fuzzy logic and semantic approach for text extraction and summarization [4].

Extraction of information from digital repositories is useful in learning. Armadillo system gives automatic domain-specific annotation on large websites by extracting information from various sources [5]. Mukherjee and others describe a technique for annotating content-rich HTML documents. Using semantic concepts and bootstrapping of annotated documents, the unlabelled concepts in all the documents are identified [6].

III. DEVELOPMENT OF KNOWLEDGE BASE

Ontology can be used for developing common vocabulary in the knowledge base [7]. For searching and browsing large images, Content Based Image Retrieval System is being used [8]. Witten (2002) described hierarchical phrase browsing, text mining and keyphrase extraction in extracting information from plain text [9]. An e-learning platform has been developed by Xu et al (2017) using machine learning and data mining techniques, [10]. This platform gives which gives a literature organization for a learner based on the knowledge map.

A. Whole Book Recognition

Image recognition is one of the techniques of machine learning. In whole book recognition, image recognition is used with automatic adaptation. The scanned image is to be OCRed for text searching. An algorithm to initialize an iconic and linguistic model for whole book recognition is proposed by Xiu and Baird [11]. In this algorithm, the disagreements between the distribution of character classes and word classes are detected. Automatic recognition of books by image recognition can be done using deep learning method and support vector algorithm [12]. Support vector machine algorithm has been applied to pattern recognition and document classification [13] [14].

B. Information Extraction

The extraction of information from raw data is one of the applications of machine learning. The extracted information will be maintained in databases. Data Fountains and iVia projects go a long way in internet resource discovery, metadata generation and data harvesting in the digital library using machine learning [15].

Support vector machine can be combined with Natural Language Processing to extract keyphrases from scientific papers [16]. Topical crawlers can be used for web mining and harvesting data. A machine learning technique to harvest information from the web and to process using lexical analysis is proposed in [17].

C. Search Pattern

Techniques of machine learning are useful for automatic document classification. A first-order logic framework for dealing with different kinds of the evolution of digital libraries is introduced by Ferilli and others [18]. Machine learning can be applied to automatically extract the title from documents which is useful in the retrieval of documents while searching [19].

D. Retrieval of Non-Textual Information

Normally search techniques can only retrieve text documents based on a particular keyword. An architecture for retrieving documents with figures using machine learning has been developed by Lu et al (2006). In this model, figures are categorized automatically and indexed, which can be used for retrieval of scientific documents [20].

E. Metadata Generation

Online citation matching can be used to generate metadata to solve the bibliographic management problem in digital libraries. Councill and others outline an algorithm for this Bayesian framework [21].

F. Document Classification

Machine learning methods can be used for classification of documents and extraction of table data. Kim and Liu (2011) categorized tables based on topic and functions and proposed an automatic classification method [22]. For archiving of documents, the separation of text and image, and extraction of characters from the image are required. A document processing system named WISDOM++ has been proposed by Esposito et al (2004) for automatic processing of documents [23]. Using Positive Unlabeled learning approach, an attempt has been made by Shirude and Kolhe (2016) for document classification [24].

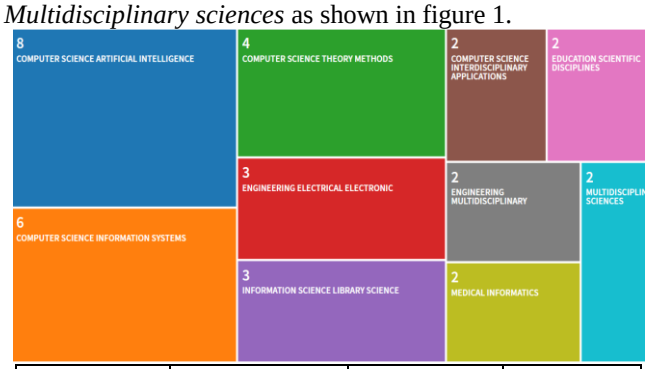
IV. SECURITY APPLICATIONS

Machine learning techniques can be used for security applications in digital libraries. The application of artificial intelligence and deep learning will improve the quality of the security system, especially in detecting malicious code [25]. Bradley (2019) proposed an artificial threat model which identifies the security vulnerabilities which cannot be resolved in the existing risk management methods. The model recognizes the malicious activity in the network and intervenes in times of vulnerability[26]. A detection framework namely, Sec-Lib has been developed by Nissim et al (2019) to detect malicious PDF documents [27]. They developed a two-layered framework which includes a deterministic layer and a machine learning based layer.

V. CITATION NETWORK ANALYSIS

Citation analysis is a method of calculating the relative impact of an author or a publication by the number of times that an author or a publication has been cited in other articles. Sources for citation analysis are Web of Science (WoS), Scopus, Google Scholar, etc.

In this paper, the researchers used WoS for citation network analysis. The citation analysis results show that there are 10 subdomains related to the study, namely, *Computer science Artificial intelligence*, *Computer science Information systems*, *Computer science theory methods*, *Engineering-electrical electronic*, *Information science Library science*, *Computer science Interdisciplinary applications*, *Engineering multidisciplinary*, *Medical Informatics*, *Education scientific disciplines* and *Multidisciplinary sciences* as shown in figure 1.



Subdomain 1	Subdomain 3	Subdomain 6	Subdomain 9
Subdomain 2	Subdomain 4	Subdomain 7	Subdomain 10
	Subdomain 5	Subdomain 8	

Fig. 1. Citation Analysis Results

The citation network diagram is drawn for group 1, i.e., Computer science Artificial intelligence (Fig. 2) and for group 5, i.e., Information science Library science, which are most relevant subdomains in the study using the tool Pajek.

A. Citation Network Analysis for the subdomain Computer Science Artificial Intelligence

Figure 2 shows that the most influential and relevant paper in the citation network for the subdomain *Computer Science Artificial Intelligence* is:

I. H. Witten, "Learning Structure from Sequences, with Applications in a Digital Library", in ALT 2002, LNAI 2533 N. C. Bianchi, Ed. Heidelberg: Springer-Verlag, 2002, pp. 42–56.

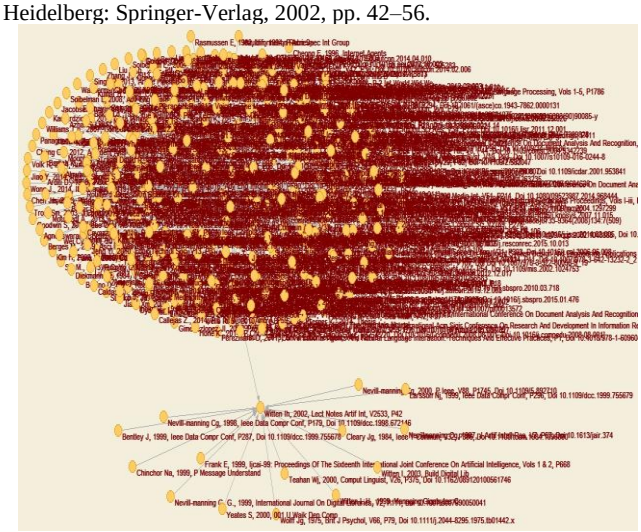


Fig. 2. Citation network diagram for the sub-domain 1

In this paper, Witten [9] gives techniques for automatically extracting information from the full text and applies in the digital library context using hierarchical phrase browsing, text mining, and key-phrase extraction.

B. Citation Network Analysis for the subdomain Information Science Library Science

Fig. 3. Citation network diagram for the subdomain 5.

Figure 3 shows that there are two most relevant and cited papers in the subdomain *Information Science Library science*:

1. Y. B. Wu, Q. Li, R. S. Bot and X. Chen, "Finding Nuggets in Documents: A Machine Learning Approach", Journal of the American Society For Information Science And Technology, vol. 57, no. 6, pp. 740–752, 2006.

2. S. Mitchell, "Machine Assistance in Collection Building: New Tools, Research, Issues, and Reflections", Information Technology And Libraries, pp. 190-216, Dec 2006.

In the first paper, Wu and others describe how keyphrases can be used for knowledge management and retrieval. In this, the researchers explain the Keyphrase Identification Program (KIP), that allow the users to build a glossary database for the area of their interest.

The second paper discusses the tools and technology for internet resource discovery and metadata generation, especially the projects Data Fountains and iVia systems.

In addition to this, the paper by Fox [2] is cited in more than one article:

R. Fox, "Digital libraries: the systems analysis perspective machine erudition", Digital Library Perspectives, vol. 32, no. 2, pp.62-67, 2016. http. s://doi.org/10.1108/DLP-02-2016-0006

C. Citation Report

Citation report (Fig. 4) shows the papers from the year 2000 to 2019 and the total number of results for the topic of the study is 26, average citations per item is 12.96 and h-index is 10 (as on May 2019).

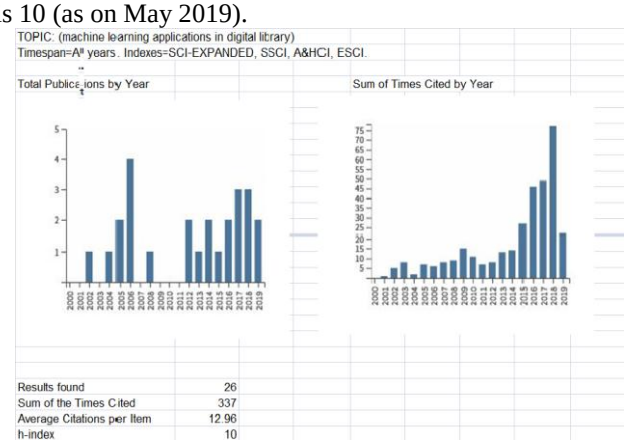


Fig. 4. Citation Report

VI. RESULTS AND DISCUSSION

The paper depicts the application of machine learning in the digital library. With the emergence of artificial intelligence, resource discovery from the web has become flexible, and it is possible to extract, classify and retrieve the information.

The citation network studies from Web of Science details the disciplines in which artificial intelligence and machine learning are applied. The citation network diagram for the most relevant disciplines in the current study are drawn, and the important publications in the domain are identified. The citation network analysis shows the relevance of machine learning applications in digital library and the citation report shows that there are more studies in this area during the past years.

REFERENCES

1. X. Lin, Y. Xiong, "Detection and analysis of table of contents based on content association", International Journal of Document Analysis, vol. 8, no. 2, pp. 132-143, 2006. DOI 10.1007/s10032-005-0149-4
2. R. Fox, "Digital libraries: the systems analysis perspective machine erudition", Digital Library Perspectives, vol. 32, no. 2, pp.62-67, 2016. http. s://doi.org/10.1108/DLP-02-2016-0006
3. Y. B. Wu, Q. Li, R. S. Bot and X. Chen, "Finding Nuggets in Documents: A Machine Learning Approach", Journal of the American Society For Information Science And Technology, vol. 57, no. 6, pp. 740–752, 2006.
4. S. A. Babar and P. D. Patil, "Improving Performance of Text Summarization", Procedia Computer Science, vol. 46 no. 2015, pp. 354 – 363, International Conference on Information and Communication Technologies(ICICT2014), doi: 10.1016/j.procs.2015.02.031.
5. F. Ciravegna, S. Chapman, A. Dingli, Y. Wilks , "Learning to Harvest Information for the Semantic Web", pp. 312-326, 2004, doi: 10.1007/978-3-540-25956-5_22.
6. S. Mukherjee, I.V. Ramakrishnan and A. Singh, "Bootstrapping Semantic Annotation for Content-Rich HTML Documents, Proceedings of the 21st International Conference on Data Engineering (ICDE 2005). 1084-4627/05.
7. Y.L. Chi, T.Y. Hsu and W.P. Yang, "Building Ontological Knowledge Bases For Sharing Knowledge In Digital Archive", Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, 18-21 August 2005, IEEE. 0-7803-9091-1/05
8. S. Wang, J. Wang, B. Wang and X. Z. Wang, "Study On The Content-Based Image Retrieval System By Unsupervised Learning", Proceedings of the Eighth International Conference on Machine Learning and Cybernetics, Baoding, 12-15 July 2009 . IEEE. 978-1-4244-3703-0/09/
9. I. H. Witten, "Learning Structure from Sequences, with Applications in a Digital Library", in ALT 2002, LNAI 2533 N. C. Bianchi, Ed. Heidelberg: Springer-Verlag, 2002, pp. 42–56.
10. X. Xu, K. Wang, Y. Deng, T. Li, "REPDL: Research-oriented e-learning platform based on digital library", 2017 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE), 12-14 December 2017, The Education University of Hong Kong, pp. 139-142. 978-1-5386-0900-2/17
11. P. Xiu and H. S, Baird, "Whole-Book Recognition", IEEE Transactions on Pattern Analysis and Machine Intelligence", vol. 34, no. 12, pp. 2467-2480, 2012.
12. B. Zhu, L. Yang, X. Wu and T. Guo, "Automatic Recognition of Books Based on Machine Learning", 3rd International Symposium on Computational and Business Intelligence, IEEE 978-1-4673-8501-5/15, 2015. DOI 10.1109/ISCBI.2015.20
13. X. Zhang, "The statistical learning theory and support vector machine," Acta Automatica Sinical, vol.26, no.1, pp.32-42, 2000.
14. F. Tang, Z. Wang and M. Chen, "The study of multi-class classification based on support vector machine algorithm," Control and Decision, vol.20, no.7, pp.746-754, 2005.

15. S. Mitchell, "Machine Assistance in Collection Building: New Tools, Research, Issues, and Reflections", Information Technology And Libraries, pp. 190-216, Dec 2006.
16. M. Krapivin, M. Marchese, A. Yadrantsau and Y. Liang, "Unsupervised Key-Phrases Extraction from Scientific Papers Using Domain and Linguistic Knowledge", *2008 Third International Conference on Digital Information Management*, London, 2008, pp. 105-112, IEEE 978-1-4244-2917-2/08,2008,doi: 10.1109/ICDIM.2008.4746749.
17. G. Pant, K. Tsioutsiouliklis, J. Johnson and C. L. Giles, "Panorama: Extending Digital Libraries with Topical Crawlers", *Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries (JCDL'04)*, ACM, 1-58113-832-6/04, 2004.
18. S. Ferilli, M. Biba, T. M. A. Basile, and F. Esposito, "Incremental Machine Learning Techniques for Document Layout Understanding", IEEE. 978-1-4244-2175-6/08, 2008.
19. Y. Hu, H. Li, Y. Cao, D. Meyerzon, Q, Zheng "Automatic Extraction of Titles from General Documents using Machine Learning", Information Processing and Management, vol.42, no. 5, pp. 1276-1293 doi: 10.1016/j.ipm.2005.12.001.
20. X. Lu, P.Mitra, J. Z. Wang, C. L. Giles, "Automatic Categorization of Figures in Scientific Documents" JCDL '06 June 11–15, 2006, Chapel Hill, North Carolina, USA, pp. 129-138.
21. I. G. Councill, H. Li, Z. Zhuang amd S. Debnath, "Learning Metadata from the Evidence in an On-Line Citation Matching Scheme", Proceedings of the ACM/IEEE Joint Conference on Digital Libraries. 2006. pp. 276 - 285, doi: 10.1145/1141753.1141817.
22. S. Kim, and Y. Liu, "Functional-based Table Category Identification in Digital Library", IEEE Computer Society, pp. 1364-1368, *2011 International Conference on Document Analysis and Recognition*. IEEE 1520-5363/11, DOI: 10.1109/ICDAR.2011.274
23. F. Esposito, D.Malerba, G.Semeraro, S.Ferilli, O. Altamura, T. M. A. Basile, M. Berardi, M. Ceci, and N. D Mauro, "Machine Learning methods for automatically processing historical documents:from paper acquisition to XML transformation, Proceedings of the First International Workshop on Document Image Analysis for Libraries DIAL'04) IEEE 0-7695-2088-X/04. IEEE Computer Society
24. S. B. Shirude and S. R. Kolhe, "Classifying library resources in Library Recommender Agent using PU learning approach," *2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)*, Ernakulam, 2016, pp. 79-83.
25. V. Mahesh, and K. A. Sumithra Devi, "Spyware Detection and Prevention using Deep Learning AI for user applications", International Journal of Recent Technology and Engineering, pp. 345-349, vol. 7, no. 5, 2019.
26. P. Bradley, Risk management standards and the active management of malicious intent in artificial superintelligence. AI & Society, Springer, DOI: https://doi.org/10.1007/s00146-019-00890-2
27. N. Nissim, A. Cohen, J. Wu, A. Lanzi, L. Rokach, Y. Elovaci, L. Giles, "Sec-Lib: Protecting Scholarly Digital Libraries From Infected Papers Using Active Machine Learning Framework," in *IEEE Access*, vol. 7, pp.110050-110073, 2019. doi: 10.1109/ACCESS.2019.2933197

AUTHORS PROFILE

Beena A L, Assistant Librarian, Cochin University of Science and Technology, Research scholar in the Department of Computer Applications. The area of interest is information security, information centres, digital resources.

Dr.Humayoon Kabir S, former Head, Department of Library and Information Science, University of Kerala, Trivandrum, Kerala. He had his Ph.D. from Bangalore University, M.Phil.(Delhi), M.Lib.Sc and M.Sc. (Aligarh),Published around 100 articles and involved in several conferences. Editor of KELPRO Bulletin. He is a research guide and life member of several professional organizations.