Journal of Nuclear Energy Science & Power Generation Technology

Research Article

A SCITECHNOL JOURNAL

Use of Knowledge-Based Data Extraction Methods for Text Extraction

Arun Prasad B[1]*, Vikas RaoVadi[2], A Prakash[3], Pavithra M[4], A Velayudham[5] and Md Sajid Anwer[6]

Abstract

Combining the available information of academic papers accessible on the Internet is a key task in academic research. The major emphasis information collecting for citation tasks is necessary for the development of content through various secondary sources of data. In this work, we use an understanding method to acquire data, with such a focus on student journal data mining algorithms. We utilize an epistemic information retrieval scheme called INFO-MAP to mechanically obtain the data source. The findings show that together we can properly obtain author, name, publication, volumes, issue, date, and pages information from several citation formats utilizing INFO-MAP. The average total area reliability of reference retrieval for a genomics database is 97.87% for 6 citation categories.

Keywords: INFO-MAP; Information; Annotation; Reference retrieval

Introduction

Combining the available information of academic papers accessible on the website is a key effort in academic research. Attractive style information extraction for research papers is necessary for such user interfaces from diverse dispersed secondary information, where data is defined as large datasets regarding data [1]. Citation article refers to mechanization comparison retrieving as a typical task owing to the variations in area separation. For example, the author and surname parts might be separated by spaces like 0-7812-87093-8/15/1350 IEEE. To split the volume and issue components, brackets can be utilized [2]. Inside-field differences in syntax and space offer substantial separator challenges [3]. Several techniques are used to extract citation information from online text citations [4,5]. Cite seer is a system for automatically classifying academic publications in electronic format [6,7]. It recognizes multiple sorts of citations from the same text using machine learning algorithms. Authors utilize a pattern matching approach to retrieve references in digital materials.

Throughout this work, researchers give expertise and knowledge of the citation metadata retrieval model for research papers. The classification we're using, INFO-MAP, is a data removal process that finds important reference concepts in complex communication literature [8,9]. A key use of INFO-MAP is its ability to describe the complicated pattern frameworks, like hierarchy comparison, information structures, concept applying processes, screen matching, and networks try to match. INFO-MAP may be used to get data about an author, name, publication, volume, issue, date, and pages from a range of different views kinds, and styles.

Related Works

To start, artificial intelligence approaches like genting berhad increase performance by using a probabilistic approximation that is reliant on training data containing annotated bibliographies data. Citeseer's reference processing technique has an accuracy of approximately 80% for detecting titles and authors, and about 40% for detecting category pages utilize deep feature modelling for removing major fields from the IT headers with lead reliability of 93%. Peng with his fellow workers to extract different common features from the headers and citations of published studies, [10] use markov random field (CRF). The total word quality for document header extraction is 98.3%. The quotation data researchers are using is the cor dataset which contains 500 quotations in 13 areas. Get an average final examination of 86%

(HMM) vs 96% (CRF) and entire occurrence reliability of 10% (HMM) vs 78% (CRF) for article citations (CRF).

Secondly, regulatory tools including Chowdhury and Ding et al. utilize a technique that matches patterns to extract quotes from digital text. Employ three templates to extract information from the articles and succeed in delivering the knowledge collected in published references by each unit (around 90%). They have the benefit of the rule-based technology to swiftly achieve obvious facts. But they only process the connections in one style from the labeling texts, while our technique treats references in plain text in a much broader variety of environmental design kinds.

Unfortunately, the RME approach to academic works suggested that information for 907 entries be available in multiple literary formats and with a high level of efficiency, unlike approaches that make use of a limited percentage of the test data set. For the six main styluses outlined in Table 1, the average total field accuracy is 98%, again 99% for MISQ and 88% for the other 30 randomly selected styles.

Table 1: Various journal citation formatting sources.

Journal reference styles	Reference style example
Bioinformaticsstyle(BIOI)	Davenport ,T.,Delong, D, Beers ,M.(1998) successful knowledge mangement project.Sloan management review,39(2),43-57.
ACM style(ACM)	1.Devenport, T.,Delong D.and Beers,M1998. successful knowledge mangement project.Sloan management review,39(2).43-57

*Corresponding author: Arun Prasad, Institute of Law, Nirma University, Assistant Professor (Economics), Ahmedabad-382481, India, E-mail: arunprasad16@gmail.com

Received date: August 31, 2021 **Accepted date:** September 15, 2021 **Published date:** September 22, 2021

IEEE style(IEEE) APA style (APA) JCB style (JCB)	[1]T.Devenport ,D.Delong , M.Beers “successful knowledge mangement project.”Sloan management review,vol.39.no.2,pp.43-57.1998. Davenport,T., Delong,D.,&Beers,M.(1998).successful knowledge management projects.Sloan management review,39(2),43-57. Davenport,T., Delong,D.,&Beers,M.1998. successful knowledge management projects.Sloan management review,39(2),43-57.
MISQ style(MISQ)	Davenport,T., Delong,D.,&Beers,M”successful knowledge management projects,”Sloan management review(39:2) 1998,pp 43-57

Proposed Methodology

The system structure of our knowledge-based RME for academic publishing comprises four phases: (1) Knowledge discovery in INFO-MAP (2) Reference metadata extraction (3) Experience and understanding reference data extraction (4) Experience and understanding references output resistance-online service the current system for experience and understanding for scholarly articles is shown in Figure 1.

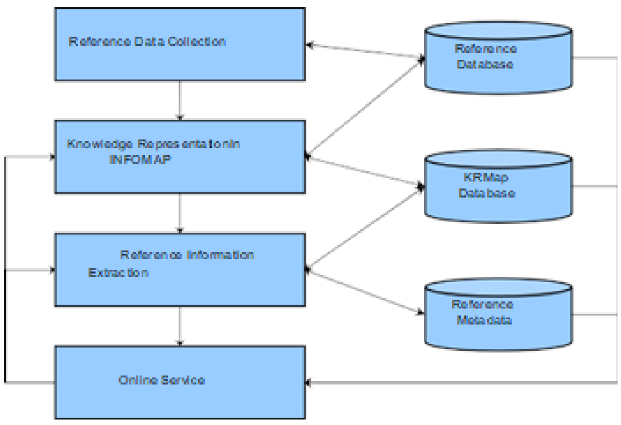

Reference Data Collection
Reference Database
Knowledge Representationin INFOMAP
KRMap Database
Reference Information Extraction
Reference Metadata
Online Service

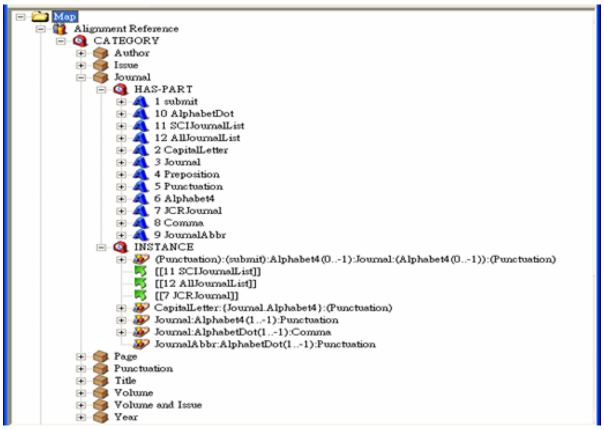
Figure 1: The experience and understanding of RME are important levels.

Information gathering for reference

During data collection, we use articles goblin to acquire publication data from the ISI-indexed data cited (JCR) and online sources upon on the Internet. The bulk of the reference information comes from various publishing houses. We retain the data in the relational database as the image retrieval test shows.

INFO-MAP knowledge discovery

Horizons is the data modification interface for RME inside the information retrieval step at INFO-MAP [11] Based on an ontology information retrieval methodology, INFO-MAP is an intermediation cost for collecting important reference concepts from a framework. The INFO-MAP approach organizes information on references concepts in a hierarchical structure. In INFO-MAP, Figure 2 depicts an example of processing information for data RME. We utilize INFO-MAP to describe how to retrieve information about personal information from a variety of citation styles.



```
Map
  Alignment Reference
    CATEGORY
      Author
      Issue
      Journal
        HAS-PART
          1 submit
          10 AlphabetDot
          11 SCIJournalList
          12 AllJournalList
          2 CapitalLetter
          3 Journal
          4 Preposition
          5 Punctuation
          6 Alphabet4
          7 JCRJournal
          8 Comma
          9 JournalAbbr
        INSTANCE
          (Punctuation):(submit):Alphabet4(0..-1):Journal:(Alphabet4(0..-1)):(Punctuation)
          [[11 SCIJournalList]]
          [[12 AllJournalList]]
          [[7 JCRJournal]]
          CapitalLetter:{Journal.Alphabet4}:(Punctuation)
          Journal:Alphabet4(1..-1):Punctuation
          Journal:AlphabetDot(1..-1):Comma
          JournalAbbr:AlphabetDot(1..-1):Punctuation
      Page
      Punctuation
      Title
      Volume
      Volume and Issue
      Year
```

Figure 2: In INFO-MAP an illustration of information processing using experience and understanding RME.

Retrieval of information from references

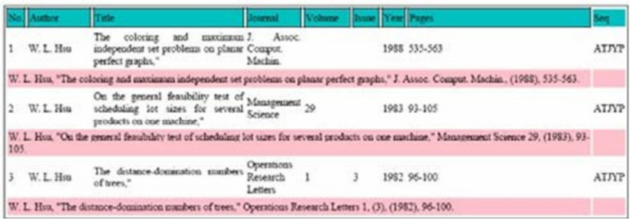
Because of so many different source formats used during scholarly journals, citations for an item can be done in a variety of methods. Table 1 displays examples of six different citation formats for the work “prolonged performance improvement system” by Davenport et al. [12-20]. An example, in Apa format, a reference reads such as:

Splitting variables having different reference kinds may be done in a variety of ways [21-25]. To distinguish the author and header parts, for example, use cycles or commas. The use of different languages, and notations, creates more separator issues inside sections [26-28].

Despite this, we can extract labeled field information from a wide range of data sources. Mostly during the original information extraction stage, we use INFO-MAP as well as the align references citations agents to obtain personal information from different kinds of citation documents.

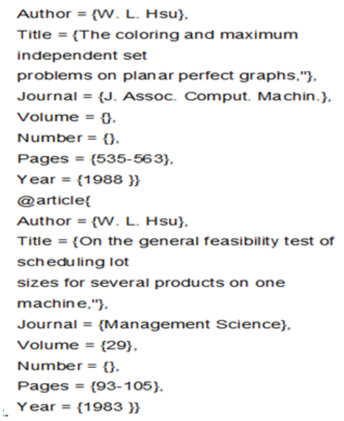
Connection to the base of knowledge web application for data retrieval

Figure 3 depicts the program’s input for extracting supporting documentation for scholarly tasks gained through experience and comprehension. In a variety of citation styles the online platform will then collect information about the article’s basic information from multiple reference forms and inclusion (Figure 4). The data RME’s web state allows is seen in Figure 5.



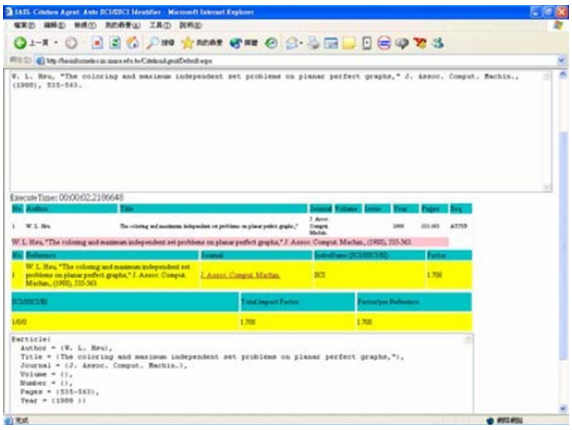
No.	Author	Title	Journal	Volume	Issue	Year	Pages	Seq
1	W. L. Hsu	The coloring and maximum independent set problems on planar perfect graphs,"	J. Assoc. Comput. Machin.			1988	535-563	ATJYP
W. L. Hsu, "The coloring and maximum independent set problems on planar perfect graphs," J. Assoc. Comput. Machin., (1988), 535-563.								
2	W. L. Hsu	On the general feasibility test of scheduling lot sizes for several products on one machine,"	Management Science	29		1983	93-105	ATJYP
W. L. Hsu, "On the general feasibility test of scheduling lot sizes for several products on one machine," Management Science 29, (1983), 93-105.								
3	W. L. Hsu	The distance-domination numbers of trees,"	Operations Research Letters	1	3	1982	96-100	ATJYP
W. L. Hsu, "The distance-domination numbers of trees," Operations Research Letters 1, (3), (1982), 96-100.								

Figure 3: RME systems output responses.



```
Author = {W. L. Hsu},
Title = {The coloring and maximum
independent set
problems on planar perfect graphs,"},
Journal = {J. Assoc. Comput. Machin.},
Volume = {},
Number = {},
Pages = {535-563},
Year = {1988 }}
@article{
Author = {W. L. Hsu},
Title = {On the general feasibility test of
scheduling lot
sizes for several products on one
machine,"},
Journal = {Management Science},
Volume = {29},
Number = {},
Pages = {93-105},
Year = {1983 }}
```

Figure 4: Bib text format systems response.



W. L. Hsu, "The coloring and maximum independent set problems on planar perfect graphs," J. Assoc. Comput. Machin., (1988), 535-563.

ExecuteTime: 00:00:02.2186648

No.	Author	Title	Journal	Volume	Issue	Year	Pages	Seq
1	W. L. Hsu	The coloring and maximum independent set problems on planar perfect graphs,"	J. Assoc. Comput. Machin.			1988	535-563	ATYP

W. L. Hsu, "The coloring and maximum independent set problems on planar perfect graphs," J. Assoc. Comput. Machin., (1988), 535-563.

No.	Reference	Journal	IndexName (SCI/SSCI/EI)	Factor
1	W. L. Hsu, "The coloring and maximum independent set problems on planar perfect graphs," J. Assoc. Comput. Machin., (1988), 535-563.	J. Assoc. Comput. Machin.	SCI	1.708

SCI/SSCI/EI	Total Impact Factor	Factor/per Reference
1/0/0	1.708	1.708

```
@article{
  Author = {W. L. Hsu},
  Title = {The coloring and maximum independent set problems on planar perfect graphs,"},
  Journal = {J. Assoc. Comput. Machin.},
  Volume = {},
  Number = {},
  Pages = {535-563},
  Year = {1988 }}
```

Figure 5: Experimental RME web services.

Results and Discussion

The following are the major findings of experience and understanding database production for academic works. EndNote has been used to extract bioinformatics written documentation from PubMed in 2014. A total of 917 pieces of bibliographic information were found in the PubMed library services on the Internet. There is a separate test dataset in each of the six reference models. Next, randomly, 600 records were picked for assessment of each of the six reference types.

Unless the field numbers in the standard test results are successfully retrieved do we considered a field to also be delivering the outcomes in this test. The below is a definition of reference extract precision:

$$Accuracy = \frac{Numberofcorrectlyextractedfields}{TotalNumberoffields}$$

The domain precision, as opposed to the phrase and occurrence precision specified is the measure of performance we describe below. Figure 6 summarizes the results of the reference retrieval experiments for the six different citation styles. Again for six styles, the overall reference retrieval reliability is 98%. With an accuracy of 99%, the MISQ type is the most accurate. The six styles' mean journaling area correctness is 99.66%, whereas the MISQ standard style's correctness is 99.97%. Based on this analysis, our method looks to be quite trustworthy.

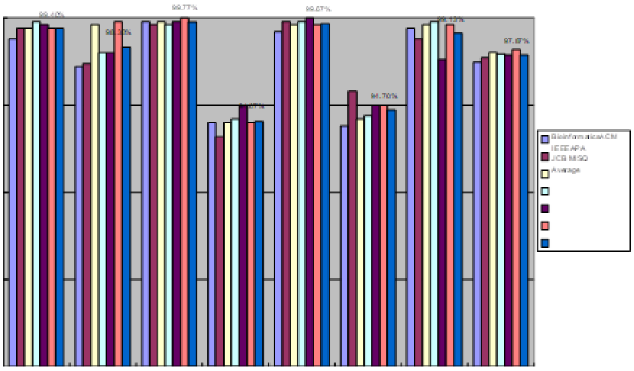
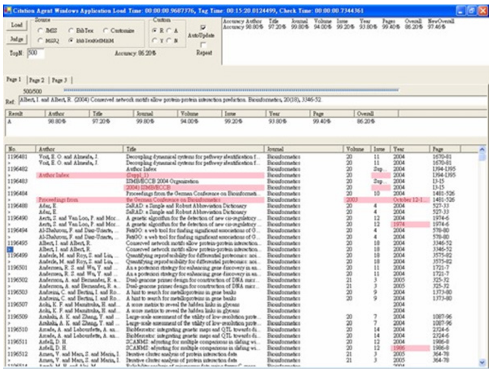


Figure 6: Experiments using reference retrieval using 6 major citation formats yielded the following findings.

Figure 7 depicts the results of the specified test database test. We can obtain article information with great accuracy using INFO-MAP from the range of reference types, according to the results.



Citation Agent Windows Application Load Time: 00:00:00.9687376, Tag Time: 00:15:20.0124499, Check Time: 00:00:00.7344361

Source: ○ IMIS ○ BibTex ○ Customize ○ MISQ ● bib TextOfMHM
Custom: ● R ○ A ○ Y ○ B
☑ AutoUpdate
□ Repeat
TopN: 500
Accuracy: 86.20%

Accuracy	Author	Title	Journal	Volume	Issue	Year	Pages	Overall	NewOverall
Accuracy	98.80%	97.20%	99.80%	94.00%	99.20%	93.80%	99.40%	86.20%	97.46%

500/500

Ref: Albert, I. and Albert, R. (2004) Conserved network motifs allow protein-protein interaction prediction. Bioinformatics, 20(18), 3346-52.

Result	Author	Title	Journal	Volume	Issue	Year	Page	Overall
A	98.80%	97.20%	99.80%	94.00%	99.20%	93.80%	99.40%	86.20%

No.	Author	Title	Journal	Volume	Issue	Year	Page
1196481	Voit, E. O. and Almeida, J.	Decoupling dynamical systems for pathway identification f...	Bioinformatics	20	11	2004	1670-81
*	Voit, E. O. and Almeida, J.	Decoupling dynamical systems for pathway identification f...	Bioinformatics	20	11	2004	1670-81
1196482		Author Index	Bioinformatics	20	Sup...	2004	1394-1395
*	Author Index	(Suppl_1)	Bioinformatics	20		2004	1394-1395
1196483		ISMB/ECCB 2004 Organization	Bioinformatics	20	Sup...	2004	13-15
*		2004) ISMB/ECCB	Bioinformatics	20		2004	13-15
1196484		Proceedings from the German Conference on Bioinformati...	Bioinformatics	20	10	2004	1481-526
*	Proceedings from	the German Conference on Bioinformatics	Bioinformatics	2003		October 12-1...	1481-526
1196488	Adar, E.	SaRAD: a Simple and Robust Abbreviation Dictionary	Bioinformatics	20	4	2004	527-33
*	Adar, E.	SaRAD: a Simple and Robust Abbreviation Dictionary	Bioinformatics	20	4	2004	527-33
1196490	Aerts, S. and Van Loo, P. and Mor...	A genetic algorithm for the detection of new cis-regulatory ...	Bioinformatics	20	12	2004	1974-6
*	Aerts, S. and Van Loo, P. and Mor...	A genetic algorithm for the detection of new cis-regulatory ...	Bioinformatics	20	12	1974	1974-6
1196494	Al-Shahrour, F. and Diaz-Uriarte, ...	FatiGO: a web tool for finding significant associations of G...	Bioinformatics	20	4	2004	578-80
*	Al-Shahrour, F. and Diaz-Uriarte, ...	FatiGO: a web tool for finding significant associations of G...	Bioinformatics	20	4	2004	578-80
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Figure 7: The test with the chosen database table yielded the following findings.

The architecture of reference models is investigated

There are semicolon and space differences amongst field separators in the creation of referencing style. Both APA and IEEE approaches use complex configurations between both the volumes and issue parts, for example. Table 2 summaries the research on region relationship patterns. There are two possible column connection types in the author column. The very first, "Writer>," stands for 55% of the designs selected, while the other, "Reader>," stands for 43%. Author- year-title-journal-volume-issue-pages-is by far the most common field relationship structure.

We performed tests on 30 randomly selected styles with no further information change and found that the total citation retrieval accuracy for those designs was about 87%. According to the data, the overall average dependability of citation retrieval for the MLA style was 88.20%. The tests' findings show that our experience and understanding method is reliable for a wide range of unidentified source styles.

For varied pendant illuminating, 13 different types of punctuation may be used in professional areas. For example, a comma might be used for the author, volumes, number, and page sections. All journal publisher standard styles might also be used as a major field divider. In JCB, on the other hand, a period was its principal field divider. We use a pattern in INFO-MAP to show the comma differences among space separators for the design citation style.

Table 2: Area relationship architecture evaluation.

Fields	Fields relation structure	Percentage%
Author	<Author><Year> <Author><Title> N/A	54.29% 42.86% 2.85%
Year	<Author><Year><Title> <Journal><Year><Volume> <Issue><Year><pages> <Author><Year><Journal> <Pages><Year> <Volume><Year><Pages> N/A	48.54% 20.00% 14.29% 5.71% 2.86% 2.86% 5.71%
Title	<Year><Title><Journal> <Author><Title><Journal> N/A	48.57% 42.86% 8.57%
Journal	<Title><Journal><Volume> <Title><Journal><Year> <Year><Title><Volume> N/A	71.43% 20.00% 5.71% 2.86%
Volume	<Journal><Volume><Pages> <Journal><Volume><Issue> <Year><Volume><Issue> <Year><Volume><Pages> <Journal><Volume><Volume> <Journal><Volume><Year> N/A	40.00% 31.43% 14.29% 5.71% 2.86% 2.86% 2.85%
Issue	<Volume><Issue><Pages> <Volume><Issue><Year> N/A	34.29% 14.29% 51.42%
Pages	<Volume><Page> <Issue><Pages> <Year><Pages> <Volume><Pages><Year> N/A	42.86% 34.29% 17.14% 2.86% 2.85%

Conclusion

The search for quotes is hard because of the number of quotations. For research papers, in this study, we suggest the informative RME technique. Analysis and results show that fundamental inputs from several reference formats may be effectively collected using INFO-MAP. For six primary quotation styles, the aggregate field accuracy in total is 97.87%. The re-training or evaluation of ontology is two significant ways in current trials to be investigated. We will use semantic and machine learning approaches to increase the effectiveness of data about reference extraction. We will also create a stronger variant that really can manage the references and incorrect data provided.

References

1. Burnett K, Ng KB, Park S0020(1999) A comparison of the two traditions of metadata development. JASIST 50:1209 -1217.
2. Bouckaert RR (2002) Low level information extraction: A bayesian network based approach workshop on text learning 10: 16-21.
3. Mishra P, Jimmy L, Ogunmola GA, Phu TV, Jayanthiladevi A et al (2020) Hydroponics cultivation using real time iot measurement system, J Phys Conf Ser 12: 012-040.
4. Rama Krishna M, Tej Kumar KR, DurgaSukumar G (2018) Antireflection nanocomposite coating on PV panel to improve power at maximum power point. Energy sources A: recovery util environ eff 40: 2407-14.
5. Sridharan K, Sivakumar P (2018) A systematic review on techniques of feature selection and classification for text mining. Int J Bus Inf Syst 28: 504-518.
6. Ding Y, Foo S (2002) Ontology research and development. Part I - a review of ontology generation, J Inf Sci 28:123-136.
7. Devaraj S, Malkapuram R, Singaravel B (2021) Performance analysis of micro textured cutting insert design parameters on machining of Al-MMC in turning process, Int J Lghtweight Mater Manu 4: 210- 7
8. Vemuri RK, Reddy PCS, Kumar BP, Ravi J, Sharma S et al. (2021) Deep learning based remote sensing technique for environmental parameter retrieval and data fusion from physical models. Arab J Geosci 14: 1-10.
9. Fleischman M, Hovy E, Echihabi A (2003) Offline strategies for online question answering: answering questions before they are asked. proceedings of ACL- 2003 conference pp: 1-7.
10. Ezhilarasi P, Kumar NS, Latchoumi TP, Balayesu N (2021) A Secure data sharing using IDSS CP-ABE in cloud storage. In Advances In Industrial Automation and Smart Manufacturing pp: 1073-1085.
11. Peng F ,McCallum A (2004) Accurate information extraction from research papers using conditional random fields, HLT- NAACL 21: 329-336.
12. Davenport T, DeLong D, Beers M (1998) Successful knowledge management projects. MIT Sloan Manag Rev 39: 43-57.
13. Yarlagaddaa J, Malkapuram R, Balamurugan K (2021) Machining studies on various ply orientations of glass fiber composite. In Advances In Industrial Automation And Smart Manufacturing pp: 753-769.
14. Garigipati RK, Malkapuram R (2020) Characterization of novel composites from polybenzoxazine and granite powder. SN applied sciences 2: 1-9.
15. Chowdhury G (1999) Template mining for information extraction from digital documents, Library Trends 48: 182-208.
16. Ding Y, Chowdhury G, Foo S (1999) Template mining for the extraction of citation from digital documents. Proceedings of the second asian digital library conference 31: 47-62.
17. Deepthi T, Balamurugan K, Uthayakumar M (2021) Simulation and experimental analysis on cast metal runs behaviour rate at different gating models. Int J Eng Syst Model Simul 12: 156-64.
18. Giles CL, Bollacker KD, Lawrence S (1998) CiteSeer: An automatic citation indexing system digital libraries. The third ACM conference on digital libraries pp: 89-98.
19. Goodrum A, McCain K, Lawrence S, Giles C (2001) Scholarly publishing in

the Internet age: A citation analysis of computer science literature Inf process 37: 661-675.

20. Han H, Giles CL, Manavoglu E, Zha H, Zhang Z et al. (2003) Automatic document metadata extraction using support vector machines JCDL 03: proceedings of the 3rd ACM/IEEE-CS joint conference on digital libraries 12: 37-48.
21. Jacso P (2004) The future of citation indexing: An interview with eugene garfield 28: 38-40.
22. Lawrence S, Giles CL, Bollacker K (1999) Digital libraries and autonomous citation indexing computer 32: 67-71.
23. Maedche A, Staab S (2001) Ontology learning from text natural language processing and information systems pp: 364-364.
24. Yarlagaddaa. J, Malkapuram R (2020) Influence of carbon nanotubes/ graphene nanoparticles on the mechanical and morphological properties of glass woven fabric epoxy composites. INCAS Bull pp: 12: 209-18.
25. Seymore K, McCallum A, Rosenfeld R (1999) Learning hidden markov model structure for information extraction AAAI-99 workshop on machine learning for information extraction pp: 37-42.
26. Takasu A (2003) Bibliographic attribute extraction from erroneous references based on a statistical model JCDL 03: Proceedings of the 3rd ACM/IEEE-CS joint conference on digital libraries pp: 49-60.
27. Wu SH, Day MY, Hsu WL (2001) FAQ-centered organizational memory proceeding of the knowledge management and organizational memory workshop on the seventeenth, IJCAI 1: 112-120.
28. Wu SH, Sai TH, Hsu WL (2003) Domain event extraction and representation with domain ontology proceedings of the IJCAI-03 workshop on information integration on the web acapulco Mexico pp: 33- 38.

Author Affiliations

Top

[1]Institute of Law, Nirma University, Economics, Ahmedabad, India

[2]Bosco Technical Training Society, Don Bosco Technical School, New Delhi, India

[3]Department of Computer Science and Engineering, Veltech Rangarajan Dr Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, India

[4]Department of Computer Science and Engineering, Jansons Institute of Technology, Karumathampatti, Coimbatore, India

[5]Department of Computer Science and Engineering, Jansons Institute of Technology, Karumathampatti, Coimbatore, India

[6]Department of Computer Science, RNT University, Bhopal, Madhya Pradesh, India