DESIDOC Journal of Library & Information Technology, Vol. 38, No. 1, January 2018, pp. 5-10, DOI : 10.14429/djlit.38.1.12307


# Rule based Text Extraction from a Bibliographic Database

Veena Makhija[*] and Swapnil Ahuja[#]

[*]*Solid State Physics Laboratory, Delhi - 110 054, India*
[#]*University of Southern California, USA*
[*]*E-mail : veena.makhija@sspl.drdo.in*

**ABSTRACT**

The emergent concept of ' Big Data' has shifted the paradigm from information retrieval to information extraction techniques. The information extraction techniques enables corpus analysis to draw useful interpretations and its possible applications. Selection of appropriate information extraction technique depends upon the type of data being dealt with and its possible applications. In an R&D environment, the published information is considered as an authenticated benchmark to study and analyse the growth pattern in that field of science, medicine, business. A rule based information extraction process, on the selected data extracted from a bibliographic database of published R&D papers is proposed in this paper. Aim of the study is to build up a database on relevant concepts, cleaning of retrieved data and automate the process of information retrieval in the local database. For this purpose, a concept based 'subject profiles' in the area of advanced semiconductors as well as the rules for text extraction from metadata retrieved from the bibliographic database was developed. This subset was used as an input to the knowledge domain to support R&D in the area of 'advanced semiconductor materials and devices' and provide information services on Intranet. Study found that concept based pattern matching on the datasets downloaded yielded better results as compared to the results by using the controlled vocabulary of the source database .

**Keywords:** Text extraction; Rule based information extraction; Knowledge domain; Semiconductors; Controlled vocabulary; Metadata extraction

## 1. INTRODUCTION

Various data models have emerged, which have changed the way, we gather, aggregate, analyse and integrate large volumes of data[1]. Data analysts visualise that a directed effort in this direction may lead to major support for scientific advances in the field of science, medicine and business. With this conceptualisation, the paradigm has slowly taken a shift from information retrieval to information extraction processes.

A data management system on 'advanced semiconductor materials and devices' with a focus on published scientific papers is presented. They tried to develop a knowledge domain on the topic using rule based text extraction system. As a source of information, they took controlled scientific corpus, the INSPEC database, a bibliographic database published by 'The Institution of Engineering and Technologies Inc,' in the area of physics, electrical engineering and electronics, computer science, etc.

The content coverage of the database support the activity of building up a knowledge base in the area of semiconductor materials and devices. The controlled indexing, uncontrolled indexing, classification code were some of the parameters which were found to be useful in defining the concepts of the knowledge domain to some extent. However, we found that in a fast developing S&T domain, where the concepts are interdisciplinary, the controlled indexing and Thesaurus of the database will not be able to support the retrieval mechanism to that depth. To use the content of the database for the effective retrieval, authors developed a semantic map of the terms required to build a knowledge domain. This semantic map or 'subject profiles' were used on the INSPEC database to extract the relevant subset for text extraction. This process made IE more effective and precise and enabled building up a knowledge domain specific for the purpose. This process also saved the effort and time of the user in identifying the pertinent information without going through the thousands of results to derive the specific information.

Received : 09 November 2017, Revised : 30 November 2017
Accepted : 30 November 2017, Online published : 01 January 2018

## 2. BACKGROUND

Information extraction is an automatic process of information retrieval (IR) by identifying patterns of information. The information might be well defined, contextually or semantically related or unstructured data from a particular domain. The goal of IE is to extract from the documents salient facts about pre-specified types of events, entities, or relationships, to build more meaningful, rich representations. IR is often used in IE for pre-filtering a very large document collection to a manageable subset, to which IE techniques could be applied. Often the output of IR is used as a structured input to IE process. To generalise, IE aims to process the textual data collections into a shape with an inbuilt IR that facilitates knowledge discovery from the collection.

Although IE systems built for different tasks may differ from each other significantly, there are certain core components shared by most IE systems. It can be categorised in two parts viz. domain-independent *vs* domain-specific components.

The domain-independent part usually consists of language-specific components that perform linguistic analysis to extract as much linguistic structure as possible. Usually, the following steps are performed[6,8]:

- Meta-data analysis: Extraction of the title, body, structure of the body (identification of paragraphs), and the date of the document.
- Tokenisation: Segmentation of the extracted text into word-like units, called tokens. This enables identification of segments of information based on occurrence of certain features such as identification of capitalised words, words written in lowercase letters, hyphenated words, punctuation signs, numbers, etc.
- Morphological analysis: Extraction of morphological information from tokens which constitute potential source of retrieved data. The morphological analysis allows tagging of data for potential use.

The domain specific information extraction is performed and supported by ontology or Thesauri like structures. In a scientific and technical domain, the development of concepts serves as an expression of growth of information in that domain.

To execute these tasks of information extraction, different algorithms are used, which are:

- Rule-based algorithms, which use patterns to extract the concepts.
- List-based algorithms, which use enumeration of words to extract concepts.
- Advanced algorithms, which use natural language processing, machine learning, statistical approaches or a combination of these to extract complex concepts.

The application of 'Rule based information extraction system' using the method of pattern identification is presented.

## 3 LITERATURE SURVEY

Information extraction based on metadata plays an important role in automating the process of information management. Various information extraction models have emerged, which follows different procedures and technologies based on natural language processing, stastastical machine learning or rule based information extraction.

It is assumed that rule-based information extraction process does not offer a platform for much R&D. It has the scope and advantage in the industry and does not offer any challenges to the academia[1]. A study was conducted by Chiticariu[1], *et al.* on the conference papers published in a period of ten years on the topics based on machine learning, rule based and a hybrid of two techniques. It was found the research papers published purely on rule based information extraction were only 3 per cent of the total output.

A number of studies have been conducted in metadata extraction using machine learning methods such as Hidden Markov Model based Viterbi Algorithm[4] on the content but in case of structured data or labelled data, rule based text extraction provides an advantage in the sense that rules can be framed and applied by using the structured or syntactic features of the content and support low level information extraction tasks such as named-entity recognition. It can also be used to do higher level tasks based on extracted concepts[10].

It is also possible to improve the accuracy of the results by implementing the validation and verification technique on the output result. As a result several automated systems have been developed. Gmbh A. described one such application called TextMarker[10]. The application applies simple rules for extracting blocks from a given semi-structured document, which can be analysed using domain specific rules.

With this background, we tried to develop a rule based text extraction system from bibliographic database INSPEC to populate the local database with a domain specific component to improve the accuracy of results.

## 4. RULE-BASED TEXT EXTRACTION SYSTEM

To develop algorithm in a rule-based system, the knowledge is represented in the form of a condition and conclusion[2]. The rule describes the action that should be taken when a condition is met and the output of the conclusion meets the specific criteria defined by the user.

The components of rule-based system are:

(i) Rule representation
(ii) Inference engine
(iii) Explanation system
(iv) Rule engineering tool
(v) User interface

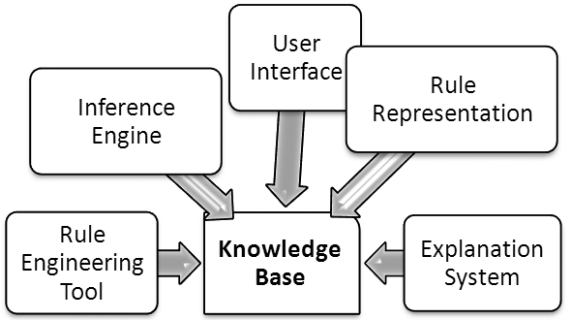


**Figure 1. Components of a rule based system.**

The Knowledge base stores all the relevant information, data, rules, relationships used in the system. A rule is a conditional statement, that links given conditions to action or outcomes. Inference engine seeks information and relationships from a knowledge base and provide results. The inference engine finds right information on the basis of facts and rules and assembles them correctly. The explanation facility allows the users to understand, how the RBS arrived at a certain result. User Interface allow the user to extract the relevant results.

A well defined rule-based system consist of complete specification of non-redundant and consistent set of rules. The verification and validation of rule-based system do not solve the problem entirely because verification is performed after the system is designed. However correction of detected errors is possible in rule-based system and then the verification cycle should be completed, which enables the management and quality enhancement of rule-based system. The result of this process is that only a small number of data is left for manual inspection and processing.

In an R&D environment, defining a knowledge base requires consideration of number of variables. One of the variable is the domain specific component, which can be defined in the form of an information metrics. The information metrics should support not only the domain knowledge but also the growth of information to satisfy the changing needs demanded by the system. It should therefore provide the user a flexibility to update it as and when required. In other words, it required a framework, which is transparent to the user and also allows a validation and verification process to judge the accuracy of the system.

Though identifying the correct information repeatedly in a changing metrics was a challenging task, we moved with the principle that 'though the R&D environment is a constantly evolving process, but the occurrence of main concept is well defined in the domain of knowledge. The occurrence and repeatability of main concept ensures that the information retrieved is relevant to the knowledge sought by the user. The occurrence of main concept is identified by a regular pattern of terms. Therefore if we defined a pattern based on the main concept and around that we tried to build up a map or subject corpus of information, it will yield relevance in the result alongwith the fact that we are not missing the nascent terminology. Further concepts were identified on the basis of exact requirement of the scientific projects, i.e. the terms defined in the Subject Profiles. This process enabled us to define the component domain knowledge in the 'Knowledge base' of the system as shown in Fig. 1.

The subject profiles of the scientists was build by identifying the main concept of the projects and identification of relevant concepts from the corpus of information retrieved from the database. For this, authors not only used the regular expression given in the indexes of INSPEC database, but also the dictionaries and ontologies on the subject. This process made the system comprehensive to define the templates for information extraction.

## 5. METHODOLOGY

This project had three levels namely, identification of corpus of information on the subject through information retrieval, linguistic analysis of record for entity identification and extraction of information on the basis of pattern matching defined in subject profiles. To build up corpus of information, we identified semi-structured data from the INSPEC database using their classification codes and index terms. Entity identification was based on the terms occurring in the indexing terminology as well as from the text of the title, abstract and corporate source.

### 5.1 Building up of Corpus of Information

We used the INSPEC database consisting of records on 'Gallium Nitride' from a period of 2005-2015. The results were transferred to a local relational database management system using MySql 5.0 as a database server.

We identified three fields from INSPEC database namely controlled terms, uncontrolled terms and classification codes. Each record contained at least one term from controlled vocabulary or classification codes, which described the main concept of the document. These index terms are assigned by the subject experts with the objective to describe the main theme of the article precisely and is controlled. The record also contains 'uncontrolled terms' an identifier field, which contains free language words or phrases assigned by human INSPEC experts. These uncontrolled index terms give exhaustive description of the content of the document taken from its title, abstract or the content of the article.

In a study conducted by Kim[7], the uncontrolled index terms reveal better Precision than the controlled index terms in their retrieval effectiveness in INSPEC database. For high precision, it is recommended to search with identifier which contains free language words and phrases assigned by the human INSPEC index experts.

We also found Controlled terms to be broad in perspective than the uncontrolled terms, For example the records of 'III V semiconductors' in 'controlled terms' contain all the narrower, and related terms in 'controlled indexing' as well as in 'uncontrolled indexing'. The uncontrolled indexing feature also gives free terms which are nascent or describe the subject content of the article as it is occurring in the content of the article. Therefore it became imperative that we select main concept from the controlled indexing in order to build up a corpus on the subject,

To identify the terms for pattern matching in the subject corpus as well as for retrieval of information, we created the 'subject profiles' on the projects as shown in Fig. 2.

'Subject profiles' incorporated all the subject terms related to our projects. A profile map was created where the main concept (A1) was linked with the specific concepts (A11). The specific concepts defined those keywords on which we want to extract the information from the source data. Where ever required, the related concepts were further explored in other

**Subject Profile**
**Project :Epitaxial Growth : Gallium Nitride**
**Activity : Epitaxial Growth – Metal Organic Chemical Vapor Deposition**
**(A1)Gallium Nitride**
**(A1)GaN**
**(A1)AlGaN**
**(A1)GaN/AlGaN**
(A11)Metal Organic Chemical Vapor Deposition
(A12)MOCVD
(A111)Epilayers
(A112)Epitaxial Growth
(A113)Substrates
(A1111)Sapphire
(A1112)Silicon
(A1113)Silicon Carbide
Characterisation
Thermal Conductivity
TEC
Thermal Expansion Coefficient
Lattice
Lattice mismatch
Dislocation density
Resistivity
Reflectivity
Nanowires

**Figure 2. Subject profile for mapping of information.**

sources of information to identify narrower terms (A111, A112, A113…..), which defines the precise requirement for the retrieval.

To create a subject profile on the project, we linked the A1 annotated term with A12 with the Boolean operator AND. If narrower terms exist to fulfil our requirement, we first connected all the narrower terms (A111, A112, A113….) with the Boolean operator 'OR', which was further connected to the resultant of (A1*(A11+A12). The resultant query was ( A1*(A11+A12)*(A111+A112+A113……) and so on all the A1 terms, describing the main concept of the project were derived from the 'Controlled Indexing'. All the A11, A12 and A111, A112, A113, etc terms were derived from the controlled indexing as well as uncontrolled indexing, title, abstracts, ontologies, dictionaries and the technical reports produced by the scientists. With these assumptions, we created the subject profiles on the projects. The records retrieved in the above process were subject to domain independent analysis of information.

### 5.2 Citation Mapping

In the downloaded text records, in order to frame the rules for information presentation, we performed the metadata analysis for data extraction[3]. This included extraction of title, author, citation on the basis of type of document, concepts and phrases, date of publication[4,5].

This phase required segmentation of information into small units called tokens on the basis of certain features such as identification of capitalised words, words written in lowercase letters, hyphenated words, punctuation signs, numbers. This phase enabled us to identify the structure of the units of information extracted, identification of paragraphs and other relevant information. Extraction of this morphological information from tokens constituted the potential source of retrieved data. The morphological analysis allowed tagging of data for potential use. Various issues were encountered, while tagging the data, which are described as follows:

*Author field :* Chinese names are usually made up of two or three parts. It is not always clear which is the family name, so the names were given in full. A definition was created in Chinese names, the punctuation mark (,) was not given

Qiuchen Lu; Sanghoon Lee

In the European country names last name separated by (,) but the first name contains more than one entry with punctuation mark like (.). Similarily in Asian names the punctuation marks contained (.) twice.

Secrest, Caleb W.; Lorenz, Robert D.

Malathi, T.; Bhuyan, M. K.

Therefore, we derived rules to consider these cases in writing the author field values.

*Type of Document :* In the database, we found following types of documents.

- Report
- Report Section
- Journal Paper
- Book
- Book Chapter
- Conference Proceedings
- Conference Proceedings in Journal
- Conference Paper
- Conference Paper in Journal
- Patent
- Standard
- Dissertation

For each type of document, we identified the mandatory fields, and other field values occurring in the database. The information extraction algorithm and the retrieval logic was based on these field values. A sample of mandatory fields for each document are tabulated in Table 1.

Table 1 provides the list of documents and the mandatory

**Table 1. List of documents with the mandatory fields**

| Type of Materials | Keywords | Title | Publication date | Subject | Author/corporate body/ publisher | Accession no. | ISBN | Docno. | Name of meeting | Journal title | url |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Books | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | | |
| Technical reports | ✓ | ✓ | ✓ | ✓ | ✓ | | | ✓ | | | |
| Annual reports | ✓ | ✓ | ✓ | ✓ | ✓ | | | ✓ | | | |
| Conferences | ✓ | ✓ | ✓ | ✓ | ✓ | | | | ✓ | | |
| S and T organisations | ✓ | ✓ | ✓ | ✓ | ✓ | | | | | | |
| Journal articles | ✓ | ✓ | ✓ | ✓ | ✓ | | | | | ✓ | |
| Presentations | ✓ | ✓ | ✓ | ✓ | ✓ | | | | | | ✓ |
| Patents | ✓ | ✓ | ✓ | ✓ | ✓ | | | ✓ | | | |
| Standards | ✓ | ✓ | ✓ | ✓ | ✓ | | | ✓ | | | |
| e-Books | ✓ | ✓ | ✓ | ✓ | ✓ | | | | | | |
| Newspaper clippings | ✓ | ✓ | ✓ | ✓ | ✓ | | | | | ✓ | |
| Thesis and dissertations | ✓ | ✓ | ✓ | ✓ | ✓ | | | | | | |
| Others | ✓ | ✓ | ✓ | ✓ | ✓ | | | | | | |

fields. However, there were many special cases encountered, which required consideration and handling of data.

As an example, for the document type : Conference proceedings. In addition to three cases identified namely Conference Proceedings, Conference proceedings in Journal, Conference Papers in Journal, we were able to identify few more cases of conference proceedings, which were identified during the iterative process[2] of mapping of records. Following cases are identified for the conference proceedings. To map these citations in the database, the rules were defined to capture the following field values.

(i) Conference Proceedings: Conference Title, Conference Location, Conference Country, Conference Date, Conference year, Pagination, (Mandatory values)

In addition to these field values, following data was also mapped for different types of publications as mentioned below:

(ii) In Journal : Journal Title, Volume, Issue no. Year of Publication, ISSN, CODEN
(iii) As paper in Journal : Journal Title, Volume, Issue no. Year of Publication , ISSN, CODEN
(iv) As book : Book Title, Editor, ISBN, Year, Pages
(v) As book chapter : Book Title, ISBN, Year
(vi) As Series : Series Title, Series Editor, Volume No., ISSN, Pagination.

In all these cases, field values for the conference details were taken as default values, which will occur in all these cases. The additional field values were added depending upon the case of conference proceedings or conference paper being published in book, Journal or Series document.

Similarly, we identified many cases in different types of documents during interactive process of validation and verification of retrieved data set. We derived rules for the above cases to write the algorithm for extraction of information from downloaded record set. We used a record set of approximately 4800 records on the topic of 'Gallium Nitride' and 'HEMT' of conference proceedings data. Validation and verification against the subset of data was performed till we were able to minimise the errors to the extent that cases could not be dealt on the basis of generalisations for defining the rules. The rules were defined on the basis of punctuation marks and field separators.

We identified several specific cases, where we find variations in the rendering of information for Publisher name, date of publication, series name, corporate source, place of publication. These variations were registered and rules were modified accordingly.

### 5.3 Variation Due to Punctuation Marks

As the entire process of rule definition was based on pattern matching, occurrence of punctuation marks and field separators, Any inconsistency in the rendering of punctuation marks or spacing in field separators in the source record affected the output. We found inconsistency in the punctuation marks defined. Which led to the error messages during the trials. Error analysis was performed and templates were created to make the rules more refined. With refinement and iterative[2] process, we were able to bring the precision to approximately 94 per cent.

### 5.4 Domain based Data Mapping

Templates were created to record the patterns defined in subject profiles. Ontologies were used to create standard entries supplemented by all the possible variations of the terms. The ontologies were to be updated by the administrator, as and when new occurrence encountered. Whenever applicable, synonyms were utilised to expand the search. It is an iterative process of template specification, evaluation and modification. Each of the extracted article was examined in the test set to derive descriptions of the various template schemes and to identify the data to be extracted from the text by the templates. The examination also resulted in the identification of additional patterns repeatable consistently and relevant to the profiles which otherwise were missed[9].

These algorithm underwent a process of iterative refinement, until their accuracy did not improve significantly between iterations and validations by the subject expert. Initially a set of 500-1000 records for each case was selected for each set of algorithms and results were verified. On the basis of results, the algorithms were modified and a new batch of articles was selected for testing purpose. Finally the algorithm accuracy was determined using precision and recall.

## 6. CONCLUSIONS

Rule-based text extraction system developed not only enabled building up an efficient knowledge domain on Semiconductor materials and Devices but also was helpful in improving the data quality of the database in terms of data cleaning during morphological analysis of data. The text mining algorithm was individually customised for each type of document in question. Some algorithms were comparatively simple and derived accurate results but some were rather complex where there were unexpected results or errors. General rules were derived for domain independent analysis of documents containg all types of documents and a number of others depending upon the specific requirements. System also allows the synchronisation and refreshment of rules for pattern matching in the knowledge domain. It allows creation of knowledge sets, elimination of duplicate information, building up of semantics using Boolean logic and translation of knowledge into rules with iterative testing.

## ACKNOWLEDGMENTS

The author would like to sincerely thank Director, SSPL for the continuous support to carry out this experimental study and giving us permission to publish these results and Dr Chandra Prakash for continuous support and encouragement.

## REFERENCES

1. Chiticariu, Laura; Yunvao, Li & Frederick, Reiss. Rule-based information extractionis dead! long live rule-based information extraction systems. *In* Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 18-21 October 2013, 827–832.
2. Bradji, Louardi & Boufaida, Mahmoud. A rule management system for knowledge based data cleaning. *Intell. Info. Manag.,* 2011, **3**, 230-239.
3. Han, Hun; Manavoglu, Eren; Zha, Hongyuan;

Tsiousiouliklis, Kostas; Giles Lee, C. & Zhang, Xiangmin. Rule-based word clustering for document metadata extraction. *In* Procedings of the 2005 ACM symposium on Applied computing*,* 2005*,* 1049-1053.

4. Cui B. Lecture Notes in Computer Science, Scientific literature metadata extraction based on HMM. Cooperative Design, visualization, and Engineering. Springer-Verlag, Berlin, 2009, **5738**, 64-68.
5. Day, Yuh-Min; Tsai R.T.; Sung, C.L.; Hsieh, C.C.; Lee, C.W.; Wu, S.H.; Wu, K.P.; Ong, C.S. & Hsu, W.L. Reference metadata extraction using a hierarchical knowledge representation framework. *Decision Support Systems,* 2007, **43**(1), 152-7.
6. Elsaeidy Asmaa & Ahmed Hamed Khalil. Survey of stages of developing the information extraction systems from the web. *Int. J. Mechanical Eng. Info. Technol.,* 2015, **3**(11), 1545-1560.
   doi: 10.18535/ijmeit/v3i11.01
7. Kim, H. Retrieval effectiveness of controlled and uncontrolled index terms in INSPEC database. *Malaysian J. Lib. Info. Sci.,* 2014, **19**(2), 103-117.
8. Piskorski, Jakub & Yangarber, Roman. Information extraction: Past, present and future. Multisource and multilingual information extraction and summarization. edited by Poibeau, T. *et al*. *Springer*, 2013, 324 p.
9. Hu, Z.Z.; Narayanaswamy, M.; Kumar, Ravi; K.E.; Shanker, Vijay K. & Wu, C. H. Literature mining and database annotation of protein phosphorylation using a rule-based system. *Bioinformatics,* 2005*,* **21**(11), 2759–275.
10. Atzmuellar, Martin, Kluegl, P. & Puppe, F. Rule based Information Extraction for Structured Data Acquisition using Textmaker.
    doi: 10.1.1.396.9396
11. Abraham, A. Rule based expert system. www.softcomputing.net/fuzzy_chapter.pdf (Accessed on 10 October 2017).

## CONTRIBUTORS

**Ms Veena Makhija** has done MSc (Physics), specialisation in 'Solid State Physics' from Delhi University, India in 1986 and Associateship in Information Science from NISCAIR, CSIR in 1989 respectively. She is presently working as Scientist 'F' and Head, Technical Information Resource Center in DRDO-Solid State Physics Laboratory, Delhi. Her current research interest include Digital reference services, knowledge management system, organisational knowledge and open access resources.

**Mr Swapnil Ahuja** received Bachelor's in Information Technology from GGSIP, Delhi, India and currently pursuing Masters in Computer Science from University of Southern California specialising in data science. He has keen interest in software development and machine learning.