International Journal of Life Sciences
International Peer Reviewed Open Access Journal

Special Issue A16, March 2021
ISSN:2320-7817(p) 2320-964X(0)

Original Article Open Access

An overview of information extraction methods, techniques and tools for the contents in chemical document

Agarkar VV[1] and Ajmire PE[2]

[1]Assistant Professor, Department of Computer Science, Shri D. M. Burungale Science & Arts College, Shegaon MS, India,
[2]Professor & Head, Department of Computer Science, G. S. Science Arts & Commerce College, Khamgaon MS, India,
Email: vinodvagarkar@gmail.com, peajmire@rediffmail.com

Manuscript details:

Available online on http://www.ijlsci.in
ISSN: 2320-964X (Online)
ISSN: 2320-7817 (Print)

Cite this article as:
Agarkar VV and Ajmire PE (2021) An overview of information extraction methods, techniques and tools for the contents in chemical document, *Int. J. of. Life Sciences*, Special Issue, A16: 26-30.

Article published in Special issue of National Conference on "Recent Trends in Science and Technology-2021 (RTST-2021)" organized by Department of Environmental Science, Shri. Dnyaneshwar Maskuji Burungale Science & Arts College, Shegaon, Bhuldhana, and Department of Botany Indraraj Commerce and Science College Shillod, DIst. Aurangabad, Maharashtra, India date, February 22, 2021.

ABSTRACT

The amount of electronic documents has speedily increased and the information over the internet is increasing day by day too. The web is continuously growing because the new information is added over it a day. These massive documents contain substantial information, but it has to be retrieved and managed in a constructive and useful way. Extracting the information from these documents is useful for many applications such as text categorization, summarization, clustering, topic tracking etc. Information Extraction (IE) is the field of extracting useful information using different methods and approaches. In this paper, the concept of information extraction (IE) is discussed, as well as presents overview of techniques used for information extraction from chemical documents.

Keywords - Information Extraction, NLP, ChemEx, IR

INTRODUCTION

The World Wide Web is the largest, popular and most widely used information source. It is much popular among users because it can easily accessible and searchable. The information over the internet is increasing day by day. The web is continuously growing because the new information is added over it a day, due to the web, users not only allow to look needed information on the web, but also easily share the information and knowledge with others (Agarkar *et al.*, 2020). The information available on the web is in the form of web pages. The contents of web pages may include texts, images, audios, videos, links, lists, charts, tables etc. Analyzing such data can help to extract meaningful information from web. In fast moving academic world, new conferences, journals and other publications are rapidly comes into existence and are expanding the already vast repository of scientific knowledge (Patil and Mahajan, 2012). E-Journals are an important source for the scientific research and development. Researchers and other users are widely used to carry out day-to-day qualitative research, education and knowledge.

Scientific documents are the main source of current information for researchers. Usually, this textual data is available either in semi-structured or unstructured form. Today most of the scientific documents are available in Portable Document Format (PDF). Extracting the Knowledge from these documents is beneficial for several applications like text categorization, summarization, clustering, topic tracking etc. Information extraction (IE) is the task of automatically extracting structured information from unstructured and/or semi-structured machine-readable documents and other electronically represented sources. In past decades, IE system development has grown rapidly; gaining attention from a lot of researchers. Information extraction tools make it possible to pull information from text documents, databases, websites or multiple sources. Usually, however, IE is used in natural language processing (NLP) to extract structured from unstructured text. IE systems are developed to extract information from differing types of text such as unstructured, semi-structured and structured text. The differences between the three types of text documents are:

Unstructured data

According to Sint *et al.* (2009) unstructured data (or unstructured information) refers to information that either does not have a predefined data model or does not fit into relational tables. Unstructured information is typically text-heavy, but may contain data such as dates, numbers, and facts as well. This result in irregularities and ambiguities that make it difficult to understand using traditional computer programs as compared to data stored in fielded form in databases or annotated in documents.

Semi-structured data

The term semi-structured data is a form of structured data that does not conform to the formal structure of data models associated with relational databases or other forms of data tables, but nonetheless contains tags or other markers to separate semantic elements and enforce hierarchies of records and fields within the data (Sukanya and Biruntha, 2012).

Structured text data:

Structured data includes mainly text, these data are easily processed. These data are easily entered, stored and analyzed. Structured data are stored in the form of rows and columns which is easily managed with the language called "structured query language" (SQL). Relational model is a data model that supports structured data and manages it in the form of row and table and process the content of the table easily. XML also Support structured data. Most of the content of the web pages are in the XML forms (Praveen and Chandra, 2017).

This paper discusses some important aspects of IE concepts, together with the methods, techniques and tools which are used to extract information from chemical documents.

Information Extraction

Information Extraction (IE) is a process that analyses natural language in order to extract specific data. The process takes texts (and sometimes speech) as input and produces fixed-format, unambiguous data as output. This data may be used directly for display to users, or may be stored in a database or spreadsheet for later analysis, or may be used for indexing purposes in Information Retrieval (IR) applications such as Internet search engines like Google search engine (Cunningham, 2006). Early work in information extraction from documents is based on two major machine learning techniques. The first is Hidden Markov models (HMM) and second is Support Vector Machine (SVM).

Information extraction is an important research area, and many research efforts have been made so far. Among these research work, rule learning based method, classification -based method, and sequential labeling based method are the three state-of-the-art methods (Jie *et al.*, 2007).

1) Rule Learning based Extraction Methods

Numerous information systems have been developed based on this method, which can be grouped into three categories: dictionary-based method, rule-based method, and wrapper induction.

a) *Dictionary based method*

Traditional information extraction systems first construct a pattern (template) dictionary, and then use the dictionary to extract needed information from the new untagged text. These extraction systems are called as dictionary- based systems (also called pattern-based systems).

b) *Rule based method*

The rule based method use several general rules instead of dictionary to extract information from text. The rule based systems have been mostly used in information extraction from semi-structured web page. Two main rule learning algorithms of these systems are: bottom-up method which learns rules from special cases to general ones, and top-down method which learns rules from general cases to special ones.

c) *Wrapper induction*

Wrapper induction is another type of rule based method which is aimed at structured and semi-structured documents such as web pages. A wrapper is an extraction procedure, which consists of a set extraction rules and also program codes required to apply these rules. Wrapper induction is a technique for automatically learning the wrappers. Given a training data set, the induction algorithm learns a wrapper for extracting the target information (Jie *et al.*, 2007).

2) Classification based method

In this method, information extraction is done using supervised machine learning approach. The basic idea is to cast information extraction problem as that of classification. Support Vector Machines (SVMs) is one of the most popular methods for classification (Jie *et al.*, 2007).

3) Sequential labeling based method

Information extraction can be cast as a task of sequential labeling. In sequential labeling, a document is viewed as a sequence of tokens, and a sequence of labels are assigned to each token to indicate the property of the token. For example, consider the nature language processing task of labeling words of a sentence with their corresponding Part-Of-Speech (POS). In this task, each word is labeled with a tag indicating its appropriate POS. Hidden Markov Model, Maximum Entropy Markov Model, and Conditional Random Field are widely used sequential labeling models (Jie *et al.*, 2007).

Table 1: Comparison of some chemical related IE methods for information extraction

Approach used	Dataset tested	Extracted Information	Reference
Lexical and syntactic aspects	American Chemical Society journals	Facts about chemical reactions	Zamora and Blower (1984)
nearest neighbour KNN	300 datasets	Protein names form biological information	Mani and I Zhang (2003)
NLP	Free-text documents in a patients (EMR) Electronic Health Record	potential medical problems	Meystre and Haug (2006)
Searching with key Words and Dictionary based Systems	Scientific Literature from web	Protein	Ono *et al.*, (2001)
Rule-based IE system	Hospital records of diabetic Patients and Reports	Useful information from Polish medical texts	Mykowiecka *et al.*, (2009)
Hybrid approach that combines a Conditional Random Field (CRF) with a dictionary	any natural language texts and documents of bioinformatics	Identifying chemical names that are mentioned in natural language texts	Rocktaschel *et al.*, (2012)

Applications of IE

There are several applications of IE, such as news extraction, literature extraction, text extraction, pharmaceutical, healthcare, bioinformatics, and so forth. The development happens in Natural Language Processing and its applications increase so as to involve the extraction methods in different areas. In the areas of chemical, biomedical and other related areas, a lot of IE methods have been developed (Elsadig *et al.*, 2015). Some of well-known applications are discussed here:

IE Methods and Techniques for Chemical Documents

Recently, the method of automatically extracting knowledge and information from text data has become one of the most relevant and active fields of study. In this regards, particularly the IE techniques used to extract information from chemical and biomedical literature. The common documents that contain chemical information are Journal Literature and Conference Papers, Reports, Dissertations, Books, Research papers, patents, drug description, scientific articles, online articles, and so forth. Many IE techniques including various tools and methods have been developed for chemical document domain. Following table Table-1 shows some of the IE methods that are proposed for information extraction from chemical and related domain. This table contains four columns namely; author, approach used, dataset tested, extracted information and references.

Tools for information extraction in Chemical Documents

Following are some of the important tools that are used for information extraction from chemical documents:

1) TICA

TICA (Postma et al, 1990) is a program for the analysis of short texts, such as abstracts. It is particularly used for the extraction of factual and methodological data from abstract-like texts on analytical chemical methods. The system consists of a parser/interpreter (which performs a parallel analysis based on requests) and a frame-based reasoning system (script-applier). This program is capable of analyzing short abstract-like texts containing declarative and imperative simple sentences and complex sentences with participle clauses. Inorganic substance names are translated to their formulas and the program can handle various kinds of quantifiers.

2) Chem Data Extractor

Matthew and Jacqueline (2016) presented a complete toolkit ChemDataExtractor for the automated extraction of chemical entities and their associated properties, measurements, and relationships from scientific documents that can be used to populate structured chemical databases. This system provides an extensible, chemistry-aware natural language processing pipeline for tokenization, part-of-speech tagging, named entity recognition and phrase parsing. This toolkit uses of unsupervised word clustering based on a massive corpus of chemistry articles to improve the performance for chemical named entity recognition. Also for phrase parsing and information extraction, the multiple rule-based grammars are used in this toolkit. They also described document-level processing to resolve data interdependencies, and show that this is particularly necessary for the auto-generation of chemical databases since captions and tables commonly contain chemical identifiers and references that are defined elsewhere in the text. The performance of the toolkit to correctly extract various types of data was evaluated, a affording an F-score of 93.4%, 86.8% and 91.5% for extracting chemical identifiers, spectroscopic attributes, and chemical property attributes, respectively. All tools have been released under the MIT license and are available to download.

3) Chem Ex

Tharatipyakul *et al.* (2012) have developed ChemEx, a chemical information extraction system. ChemEx processes both text and images in publications. Text annotator is able to extract compound, organism, and assay entities from text content while structure image recognition enables translation of chemical raster images to machine readable format. A user can view annotated text along with summarized information of compounds, organism that produces those compounds, and assay tests. ChemEx facilitates and speeds up chemical data curation by extracting compounds, organisms, and assays from a large collection of publications. This software and corpus can be downloaded from web.

CONCLUSIONS

This paper discusses some of the important extraction methods, techniques and tools that are proposed for medication and chemical documents. Various researchers

extracted information by analyzing natural text, and extracted information like chemical names, facts, potential medical problems, structured medication information, protein information etc from scientific documents or reports with good accuracy. Extraction approaches are based on traditional machine learning techniques, rule based algorithm, hybrid techniques and some newly applied techniques. The automation process of IE needs to be implemented in different domains. Numerous methods and techniques are being proposed in the IE field to automate these processes.

Conflicts of interest: The authors stated that no conflicts of interest.

REFERENCES

Agarkar VV, Ajmire PE, Bodkhe PS (2020) "Web Mining: An Application of Data Mining", Aayushi International Interdisciplinary Research Journal, Special Issue No.66, pp 56-59.

Atima Tharatipyakul, Somrak Numnark, Duangdao Wichadakul, Supawadee Ingsriswang (2012) ChemEx: information extraction system for chemical data curation", From Asia Pacific Bioinformatics Network (APBioNet) Eleventh International Conference on Bioinformatics, (InCoB2012), Bangkok, Thailand.

Cunningham H (2006) Information Extraction, Automatic", Encyclopaedia of Language and Linguistics, 2nd Edition, 5:665–677.

Elsadig Muawia, Ahmed Ali & Himmat, Mubatak (2015) "Information Extraction methods and extraction techniques in the chemical document's contents: Survey", ARPN Journal of Engineering and Applied Sciences. 10. 1068-1073.

Jie Tang, Mingcai Hong, Duo Zhang, Bangyong Liang, and Juanzi Li, (2007)"Information Extraction: Methodologies and Applications", Emerging Technologies of Text Mining: Techniques and Applications. 10.4018/978-1-59904-373-9.ch001.

Mani. and I Zhang. (2003), "kNN approach to unbalanced data distributions: a case study involving information extraction," in Proceedings of Workshop on Learning from Imbalanced Datasets, 2003.

Meystre S and Haug PJ (2006), "Natural language processing to extract medical problems from electronic clinical documents: performance evaluation," Journal of biomedical informatics, vol. 39, pp. 589-599.

Mykowiecka, M. Marciniak, and A. Kupsc. (2009), "Rule-based information extraction from patients' clinical data," Journal of biomedical informatics, vol. 42, pp. 923-936.

Ono T, Hishigaki H, A. Tanigami, and T Takagi (2001) "Automated extraction of information on protein–protein interactions from the biological literature," Bioinformatics, vol. 17, pp. 155-161, 2001.

Patil SR and Mahajan SM (2012) Optimized summarization of research papers as an aid for research scholars using data mining techniques", International Conference on Radar, Communication and Computing (ICRCC), IEEE, pp 243 – 249.

Postma GJ, van der Linden B, Smits JR, Kateman G (1990) TICA: A System for the Extraction of Data from Analytical Chemical Text. Chemometrics and Intellegent Laboratory Systems 9, 65–74.

Praveen Shagufta and Chandra Umesh. (2017), "Influence of Structured, SemiStructured, Unstructured data on various data models", International Journal of Scientific & Engineering Research Volume 8, Issue 12.

Rocktaschel T, M. Weidlich, and U. Leser.(2012), "ChemSpot: a hybrid system for chemical named entity recognition," Bioinformatics, vol. 28, pp. 1633-1640, 2012.

Sint R, Shaffert S, Stroka S and Ferstl R, (2009) Combining unstructured, fully structured and semi-structured information in semantic wikis", Paper presented at the Semantic Wikis.

Sukanya M and Biruntha S (2012), "Techniques on Text Mining", IEEE International Conference on Advanced Communication Control and Computing Technologies (ICACCCT), pp: 269-271.

Swain Matthew C and Cole Jacqueline M (2016), "ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature", J. Chem. Inf. Model. 56, 1894–1904.

Zamora EM and Blower Jr PE.(1984), "Extraction of chemical reaction information from primary journal text using computational linguistics techniques. 1. Lexical and syntactic phases," Journal of chemical information and computer sciences, vol. 24, pp. 176- 181, 1984.