Chem Text Mining- An Outline

Shidha M.V\textsuperscript{1}, Dr.T.Mahalekshmi\textsuperscript{2}

\textsuperscript{1}Research Scholar, Bharathiar University, Asst. Professor, SNGIST, N Paravur
\textsuperscript{2}Professor & Principal, Sree Narayana Institute of Technology, Kollam

Abstract—Chem Text mining involves the application of text mining in cheminformatics that deals with the use of computer technology to process and manage chemical data. Millions of chemical syntheses like articles, patents, theses etc. are published in the public literatures and most common language used to narrate these information is English. It is extremely difficult, error-prone and time consuming to extract information from these literatures manually. This leads to the need of automated information extraction tools that employs text-mining as a key activity to process literatures in natural language with high accuracy and speed. This paper deals with the need of text mining in chemical world and gives a review of some of the existing chem text mining tools.

Keywords—Chem Text mining, Cheminformatics, OSCAR, Chemical Tagger, WENDI, Chemicalize.org.

I. INTRODUCTION

For the last few decades, there has been an explosion in the publicly available chemical information. These give significant opportunity to develop and apply computational tools to process chemical literatures to identify the behavior and structure of molecules, structure derived properties, underlying structure-activity relationships, associated biological activities etc. in a very fast and efficient manner. Chem text mining helps to mine these chemical data resources to extract chemical information results less time and avoids duplication of research efforts. But this task is not easy due to the large data volume and non linear representation of the chemical compounds. The area of information technology that deals with the manipulation of chemical information is known as cheminformatics.

The field of cheminformatics is multi-disciplinary and is an amalgam of chemistry, mathematics and computer science [1]. Cheminformatics focuses on the development of methods and tools to address problems in the management and analysis of chemical information.

Such information contain chemical structures in various formats like SMILES (Simplified Molecular-Input Line-Entry System), SDF (Structure Data Format), CML (Chemical Markup Language) and derived aspects of chemical structures in terms of its number of atoms and various descriptors. In order to mine data from these unstructured fascinated chemical literatures, many barriers needed to be overcome and these should be taken into consideration while designing text mining tools to manipulate chemical data.

The paper is arranged as follows. Section II gives the notion about the basic activities involved in the text mining process. Scope of text mining in cheminformatics is discussed in section III with the help of suitable examples. Sections IV discusses the barriers often seen in chemical literatures and which is taken into consideration while design such tools. A review of four chem text mining tools is being done in section V.

II. TEXT MINING PROCESS

Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different unstructured written resources [2]. This turns data mining to text mining as patterns are extracted from natural languages rather than databases. The details extracted from the documents, like extracting people’s name, addresses, job skills etc. cannot be taken as a text mining, rather it falls to an area called information extraction. In text mining, the goal is to discover heretofore unknown information, something that no one yet knows and so could not have yet written down [3].

A text mining process consists of several sub processes. The six commonly used sub processes are Text as input, Text preprocessing, Text transformation, Attribute selection, Pattern discovery and Evaluation [4]. These are depicted in Fig1 and that detail of each of this stage is given below.
A. Text as Input

In this phase millions of documents needed to be handled as input and may not have a clear picture of what documents will suit the application. Classification methods are needed to be applied and those methods are commonly known as Document Clustering Methods (DCM). DCM helps to automatically group the retrieved documents into a list of meaningful categories. In general, there are two common algorithms- one based on hierarchical clustering and the other based on k-means clustering. The partition style of hierarchical clustering is shown in Fig2.

This method builds the hierarchy from the individual elements by progressively merging clusters. In the case of document clustering, the process starts with the documents in separate clusters and joins the closest pair of clusters repeatedly until there exists only one cluster. This type of clustering is suitable for browsing.

K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. Fig3 describes the general procedure applied in partition clustering and the algorithm of k-means clustering. According to k-means clustering, as a first step k initial "means" are randomly generated within the data domain. After that, k clusters are created by associating every observation with the nearest mean. Next the centroid of each of the k clusters becomes the new mean [6]. These steps are repeated until convergence has been reached.

Initially, hierarchical clustering was better than K-means for building document hierarchies. But a major problem found with hierarchical clustering is the chance of occurring nearest neighbors of the documents shall belong to different classes. A hybrid method, ‘Bisecting K-means’ that combines the quality of hierarchical clustering and the efficiency of K-means clustering is popular by its better performance [7].

Text characteristics are necessary to consider in this phase. Common characters are (i) input modes of text i.e. text may be intended for human consumers or automated consumers, (ii) text dependency, so words and phrases create context for each other (iii) ambiguity may occur due to word or sentence ambiguity (iv) Noisy because of erroneous or misleading data (v) Unstructured text from chat room, normal speech etc. Document categorization is done in account of these factors.

B. Text Pre-processing

Text Pre-processing involves a sequence of processes like Text Cleanup, Tokenization, Part of Speech tagging, Word Sense Disambiguation and Semantic structure production [4].
In the Text Cleanup process, it removes advertisements and other unnecessary texts from web pages and corrects the text converted from binary formats, tables, figures and formulae. Tokenization involves the process of splitting up cleaned text into a string of valid tokens (basic atomic element represents a meaningful entity). Tokens are then associated with their corresponding parts of speech. It is done in Parts Of Speech (POS) tagging. POS tagging can be rule based or statistically based. In the prior case, tagging depends on grammatical rules and the latter relies on different word order probabilities uses a manually tagged corpus for machine learning [8].

Word Sense Disambiguation is done for determining sense of a word to identify in which sense it is used in a given sentence. For example, consider two statements - “Time flies like an arrow”; “Fruit flies like a banana” [8]. In first, ‘flies’ is being used as verb and in second, it is used as a noun. In these sentences ‘flies’ has different meanings. Such kind of word sense disambiguation is performed in this phase.

Semantic structure is produced as output for each statement. It can be either a parse tree (full parsing) or a set of syntactic constructs like Noun Phrases and Verb Groups (partial parsing). Producing a full parse tree often fails due to grammatical inaccuracy, novel words, bad tokenization, wrong sentence splits and errors in POS tagging. This makes partial parsing more popular[8].

C. Text Transformation

Text Transformation also known as attribute generation, uses two techniques- Text Representation and Feature Selection. These two techniques have many approaches. In this paper two main approaches are discussed, the details of which are given in fig4.
In Bag of words representation, each word is represented as a separate variable having numeric weight. The most popular weighting schema is normalized word frequency **tfidf**: [9]

$$\text{tfidf}(w) = \text{tf}(w) \cdot \log \left( \frac{N}{\text{df}(w)} \right)$$

**tf(w)** – term frequency (number of word occurrences in a document); The word is more important if it appears several times in a target document

**df(w)** – document frequency (number of documents containing the word); The word is more important if it appears in less documents.

**N** – Number of all documents

**‘tfidf’** is calculated for each words in the document for assessing the relative importance of the word in the document.

In Vector Space method, a classic model known as term frequency-inverse document frequency model is used to find weight vector [10] of the document d, and it can be written as

$$\mathbf{v}_d = \left[ \frac{w_1, d}{|D|}, \frac{w_2, d}{|D|}, \ldots, \frac{w_N, d}{|D|} \right]^T$$

where

$$w_{t, d} = \text{tf}_{t, d} \cdot \log \left( \frac{|D|}{|\{d' \in D | t \in d'\}|} \right)$$

and

$$\text{tf}_{t, d}$$ is term frequency of term t in document d (a local parameter)

$$\log \left( \frac{|D|}{|\{d' \in D | t \in d'\}|} \right)$$ is inverse document frequency (a global parameter). |D| is the total number of documents in the document set; |{d' \in D | t \in d'}| is the number of documents containing the term t.

Using cosine the similarity between document d_j and query q can be calculated as:

$$\text{sim}(d_j, q) = \frac{\mathbf{d}_j \cdot \mathbf{q}}{||\mathbf{d}_j|| \cdot ||\mathbf{q}||} = \frac{\sum_{i=1}^{N} w_{i, d_j} w_{i, q}}{\sqrt{\sum_{i=1}^{N} w_{i, d_j}^2} \sqrt{\sum_{i=1}^{N} w_{i, q}^2}}$$

The goal of **feature selection** is to reduce the dimensionality of the dataset by removing features which are irrelevant for the classification. Dataset restriction makes classification more efficient and improves generalization error. Stop words removal and stemming are the two sub processes held in feature selection. Stop words removal removes the most common words which are unlikely to help text mining, e.g. ‘the’, ‘a’, ‘an’, ‘you’ etc. and stemming identifies a word by its root and reduce its dimensionality, e.g. seeing, saw → see.

Attributes generated in feature selection are merely labels of the classes. The next step is to populate a database that results from the information extraction process. The output of this activity is produced as a combined effect of data mining techniques-segmentation, classification, association and clustering. Fig6 shows a database produced by the attribute generation [9].

Further reduction of dimensionality and removal of irrelevant features are done in **Attribute Selection**. Now this is the time to merge text mining process with the traditional *data mining* process. This phase is termed as Pattern Discovery. Mining techniques are applied to the structured database which is resulted from the previous stages. This is purely an application-dependent phase. Finally the output is evaluated in the *Interpretation* phase and accordingly, the process is terminated or suggested for reiteration.

### III. TEXT MINING IN CHEMINFORMATICS

Chemical information is rapidly accumulating in various types of free text documents like patents, industry reports, and scientific articles. This has motivated the development of specifically tailored text mining applications. Text mining utilities in cheminformatics help to extract chemical information from variety of documents containing chemical literatures. This is done in a very rapid and efficient manner. Some of the applications of text mining in cheminformatics are discussed in the following paragraphs.
i. Chem text mining helps to automate the process of finding chemical names based on standard chemical naming conventions from thousands of documents. In OSCAR4 (Open Source Chemistry Analysis Routines), Experimental Data Checker is used for the identification and mark up of plain text experimental data [11].

ii. Chem mining tools help to convert chemical names found in searches into live chemistry objects as shown in fig7 below.

Fig7: OSCAR3 markup displaying recognized chemical entities (CM) [11]

Fig7 shows a mouse over action on an annotated term displays the associated metadata, in this case for 2,5-dichlorobenzylamine, and displays an image representing the structure generated by the Chemistry Development Kit (CDK). The primary purpose was to identify and extract the following types of object: chemicals (CM), ontology terms (ONT; looked-up from ChEBI [12], FIX [13] and REX [14] etc.), reactions (RN; as identified by linguistic constructs, e.g. “methylated”), chemical adjectives (CJ) mainly formed from chemical nouns), enzymes (ASE) and chemical prefixes (CPR), highlighted in different colors.

iii. Chem text mining is used for creation of interactive reports to highlight chemicals found in the context of documents. Novartis’s web based cheminformatics system uses an in-house database in Java, containing all in-house and many reference structures, allows efficient data-mining, reporting and SAR (Structure Activity Relationship) analysis [15].

iv. Chem text mining tools help to manage queries on structure databases in terms of structure, substructure, structure similarity, etc.

v. Chem text mining tools can be used to combine chemical structure searches with text searches for key terms and concepts (such as disease processes or biological molecules) to perform true context-specific, chemically-intelligent searching or screening.

IV. CHALLENGES OF CHEM TEXT MINING

The text mining of chemical interests pose a number of unique challenges that are not encountered in other fields of text mining. The major challenges need to be resolved are:

- Non uniformity in universal publication standards & structure.
- Ambiguity in chemical name recognition.
- Mining of chemical patents.

Many problems in chem text mining occur due to the lack of universal publication standards and structures. These are mainly because of the terminology and indexing style used with the chemical documents. Sometimes problems may be due to formatting and copyright/licensing barriers. Chemical names may appear in different styles depends on various situations. Fig8 shows the multiple ways of representing ‘Aspirin’.

Fig8: Multiple ways of naming a simple molecule [16]

The problem said above is very clear from the figure. Another difficulty found with the documents is its different formatting styles and copyright problems. Now a day data is rich, but finding fact is difficult. Even a single web page may contain different styles of patterns and extremely tough to handle them.
An example is given in Fig 10 where the placement or misplacement of spaces between methyl, ethyl and malonate can result in four different chemical structures viz. “methyl ethyl malonate”, “methyl ethylmalonate”, “methyl ethylmalonate”, and “methylene malonate”. Misplacement of a single space results dual interpretation of the chemicals. It is very risky and make the process much tougher.

Another challenge is in the mining of chemical patents. Patents are often rich sources of information of high commercial value. News gets released earlier in patents than any other literatures. Also it contains several information that may not be found in other resources. But the chance of occurring errors is more and more in the case of chemical documents. One of the major problems faced in this case is misspelt chemical names due to the incorrect specification of chemical names that occur accidently [17]. As an example, consider the situation specified in Fig 11.

In Fig 11, A is (E and Z)-2-(5-chloro-2-phenoxy-phenyl)-3-hydroxy-acrylic acid methyl ester and B is 8-chlorodibenz[b,f]oxepin-10-carboxylic acid methyl ester. This may be a usual case, but very problematic.

The next issue encountered in mining chemical patents is the occurrence of incorrect white spaces. This leads to ambiguity in chemical names. Another problem is the occurrence of Truncated chemical names due to the splittance in chemical names as shown in Fig 12. In this figure, D represents 5-chloro-2,3-dihydro-2-methyl-1H-dibenzo[2,3;6,7]oxepino [4,5-c]pyrrol-1-one which is truncated and written in two lines.
Fig12: Truncated chemical names [17]

Sometimes labels may appear with the chemical names (Conjoined chemical names) as shown in Fig13.

Fig13: Conjoined chemical name [17]

In the above figure, label (III) is written with the chemical name. Actually it is not a part of 11-chloro-2,2-dihydro-2-methyl-1H-dibenzo[2,3:6,7]oxepino 4,5-clpyrrol-1-one [15]. Another difficulty is the identification of patent document formats and is tough to handle OCR patents (e.g. PDFs, scanned TIFFs etc.). A lot of chemical information exists as images and classical NLP tools take too much time to mine patents of hundreds of pages.

All these problems are prone to happen, need solutions through chem text mining tools.

V. A WALK THROUGH OF SOME CHEM TEXT MINING TOOLS

This section conducts a review on certain chem text mining tools giving an insight towards their function in the chemical literature world.

1. OSCAR [11]: Oscar is an Open-Source Chemistry Analysis Routines software used as a toolkit for identifying chemical entities in chemical literatures. It is developed in 2002. Initially it comes up with two software packages- the Experimental Data Checker (EDC) and OSCAR2. EDC employs regular expressions to recognize chemical terms from the plain text input. OSCAR2 is used as the library for EDC and is based on n-grams and simple grammar. The next version OSCAR3, focuses on the recognition of connection tables [18] for chemical named entities. This uses two methods – Pattern Recognizer and MEMM Recognizer. Pattern Recognizer works based on predetermined regular expressions, while MEMM Recognizer uses machine learning in the form of Maximum Entropy Markov Model (MEMM) [19]. OSCAR4 offers a domain-independent architecture upon which chemistry specific text-mining tools can be built and features a modular API that permits client programmers to easily incorporate it into external applications. Now OSCAR became common in many laboratories for the identification and extraction of chemical entities in varied chemical literatures. This software is applicable to organic, biological and atmospheric chemistry.

2. Chemical Tagger [20]: Chemical Tagger is a chemical text mining tool uses NLP (Natural Language Processing) for the automatic discovery of previously unknown information from unstructured chemical data. Chemical tagger reads chemical literature, separating and parsing these with the help of OSCAR, domain-specific regex, parts-of-speech taggers, ANTLR grammar [21]. It outputs a set of Action Phrases ( Add-phrase, Stir-phrase, Dissolve-phrase…) in XML formats. Fig14 shows the sequence of actions adopted by Chemical Tagger to accomplish its task.
This sequence consists of five major steps: text normalization, tokenization, tagging, phrase parsing and finally action phrase identification. Using Chemical Tagger, it is possible to parse chemical experimental text using rule-based techniques in conjunction with a formal grammar parser.

3. WENDI [22]: WENDI (Web Engine for Non-obvious Drug Information) is an integrative data mining tool for drug discovery. It attempts to find non-obvious relationships between a query compound and scholarly publications, biological properties, genes and diseases using multiple information sources. It is a web service that takes a query compound (usually SMILES) as input, calls multiple web services for computation and database search and returns an XML file as output. This file tells the biological properties of the compound, non-obvious relationships between the compound and assays, genes, and diseases, that spread over different types of data sources. This XML file can then be parsed by an intelligent client to extract information pertinent to compound properties.

The overall architecture uses a four layer approach that includes storage, interface, aggregation and smart interface layers as shown in Fig 15. The storage and interface layers are implemented using the Web Service Infrastructure. Each layer can be accessed directly or by higher layers. WENDI offers Web services, Database services for PubChem Compound, PubChem Bioassay, Pub3D, ChEM BL etc., and Prediction services for Tumor cell line prediction, Toxicity prediction and Gene-Disease relationship. On submission of a query, WENDI returns results within a minute.

4. Chemicalize.org [23]: Chemicalize.org is an on-line resource used for extracting and connecting chemical structures from text-entombed chemistry. This explores chemical structure connectivity between documents and databases, perform structure searches in PubChem and InChIKey searches in Google and chemicalize.org archive. Dipeptidyl Peptidase 4 (DPPIV) inhibitor is used as the common theme to perform its functionality. Fig 16 shows the activities of chemicalize.org based on DPPIV in a search-analysis-storage pattern. (Abbreviations used - Free Patents Online = FPO, Europe Pub Med Central = EPMC).
It can process IUPAC names, SMILES, InChI strings, CAS numbers and drug names from pasted text, PDFs or URLs. This tool is able to generate structures, calculate properties and launch searches and explore its utility for answering where these structures overlap with database records. It has the flexibility to extract text from any internal, external or web source. It should thus facilitate progress in medicinal chemistry, chemical biology and other bioactive chemistry domains.

VI. CONCLUSION

This paper briefs the application of text mining in the field of cheminformatics. It is designed with the intention of providing basics on chem text mining. The discussion opens the scope and working of text mining in the chemical literature world. A review of four chem text mining tools was also presented in this paper. Efforts are still continuing to discover more and more efficient mining techniques for chemical entities.

The journey of chemical reading capability started by Eugene Garfield in 1958 is in its flourished state with a wide collection of varied unimaginable, powerful, automated chemical extraction tools. And in future we can expect many more highly advanced, tremendous and hybrid tools.

REFERENCES

[1] Rajarshi Guha, Kevin Gilbert et al; Advances in Cheminformatics Methodologies and infrastructure to support the Data mining of Large Heterogeneous Chemical Databases.
[5] Clustering; Available at “cecs.wright.edu/~tkprasad/courses/cs499/Clustering.ppt” accessed on 29.06.2013.
[18] Fraction, George F, Walker, Justin C, Tauber, Stephen j; Connection tables from Wiswesser chemical structure notations - a partial algorithm.


[22] Zhu, Lajiness et al; WENDI: A tool for finding non-obvious relationships between compounds and biological properties, genes, diseases and scholarly publications.

[23] Christopher Southan, Andras Stracz; Extracting and connecting chemical structures from text sources using chemicalize.org.