Review of Research on Biological literature text mining

Kong MeiJing1, Zhang Le2,3, Wang Jun4
1 College of Computer and Information Science, Southwest University, Chongqing 400715, China
2 Department of mathematical sciences, Michigan Technological University, Houghton, MI 49931, U.S.A.
3 College of Computer and Information Science, Southwest University, Chongqing, 400715, China.
1 a369@swu.edu.cn, 2 zhanglcq@swu.edu.cn, 4 kingjun@swu.edu.cn

Abstract - Currently, the field of biomedical research is booming, a lot of biomedical knowledge in unstructured form in all forms of text file, and now it is the exponential trend increase, how to solve the contradictions between massive growth of information and knowledge of text slowly, in a credible way to find useful patterns in the text is a challenge. In recent years, biomedical text mining technology which is one branch of an efficient automatic access to new exploration-related knowledge, has great progress. This review describes the major methods and results of biomedical text mining research, namely information retrieval and literature search tool and the main aspects of biological text mining biomedical named entity recognition, text categorization, abbreviations and synonyms of the word recognition, relationship extraction, forming hypotheses and integrated framework for the above work, and finally the recent developments in the field is summarized and discussed.

Keywords: Bioinformatics; text mining; computer applications.

1 Introduction

Text Knowledge Discovery (Knowledge Discovery in Texts, KDT) is focused on the discovery and excavation from the text inductive knowledge, such as useful models, trends, rules of knowledge and other computer procedure. KDT, namely text mining technology (Text Mining), is the products of the combining theory with technology which include artificial intelligence, machine learning, natural language processing, and data mining and related automatic text processing, such as information extraction, information retrieval, text classification[1].

With the development of computer network technology and the emergence of network version of the journal, medical bibliographic databases and other full-text database, biomedical text mining become one of the more active areas, and even some people in this area named KDiBL (Knowledge Discovery in Biomedical Literature). Using text mining techniques to deal with the massive scientific literature and biomedical text annotation data stored in databases, which found that research and innovation knowledge (such as genes, proteins, diseases, drugs and their relationship) is a hot research field of the current artificial intelligence and data mining. Text mining is a method for automatic computer processing of natural language text, it does not completely solve the pursuit and realization of natural language processing, and machine learning methods positioned to achieve limited objectives apply knowledge extraction and excavation.

The purpose of biomedical text mining is to help researchers in the massive literature more efficiently identify the information they need, find hidden relationships, and apply mathematical algorithms, statistical methods and data processing methods literature analysis and processing, so the information overload pressure from investigators passed on the computer.

Applications of the biomedical field text mining technology can improve biomedical information construction and management efficiency. Constructing biomedical databases, is the driving force to promote the advancement of biomedical text mining technology[2]. Through information extraction technology, you can build special proteins role relational database which is based on the purpose of disease diagnosis and drug design. Such as breast cancer. Alzheimer’s disease-related protein effect relationship database. Proteins database described by the network, will greatly benefit disease diagnosis, drug design, progress related to the promotion of biomedical research[3].

The greater sense of biomedical text mining, is that biomedical text mining techniques can help biomedical researchers perform faster and more effective research in several areas, such as information retrieval technology can help users quickly and efficiently find useful information in a massive collection of document information; information extraction technology can extract specific factual information from the biomedical literature, that the entire biological knowledge to build the network, the relationship between organisms forecasts, development of new drugs will have great significance; text classification techniques can be screened coarse-grained to narrow the search scope for further information processing in preparation; assume that technology can be tapped to generate experimental hypotheses and experiments suggested in the literature in order to get biologists to verify new scientific discoveries. Therefore, this research has attracted wide attention from computational linguistics, bioinformatics, machine learning and other aspects of the researchers, the paper describes the biological literature focuses on information retrieval tools and biomedical named entity recognition, abbreviations and synonyms recognition, biomedical entity relation extraction, was found to form hypotheses and integrated framework for the relationship between the above work, the other associated technical evaluation and so on.
2 Biological Literature Information

Retrieval

As one of the main achievements demonstrate the biological literature and academic exchanges, the number of large, fast growth rate of far more than the other disciplines. NCBI (National Center for Biotechnology Information), 1988 as a U.S. National Information Resource Center of Molecular Biology established, it’s main task is to establish a public database of computational biology research, and develop some software tools for analyzing genome, spread biological information. It offers 28 free database, PubMed is one of these databases which is retrieval system using a uniform interface (Entrez). As important source of literature in the field of life sciences PubMed, has been favored by the majority of medical researchers, with using the Boolean model and vector space model and a dictionary to automatic query expansion to make information more collective. Most users through Entrez database access and query the PubMed database (Entrez / PubMed), but there are some drawbacks in the search interface and functionality of the system [4]. Such as, It’s difficult for new users to grasp limited search and complex retrieval methods of Mesh and it’s quite boring for user to read these lengthy list of search results because of lacking of analytical tools to navigate the search results. To improve efficiency, many dedicated search engine came into being. So, we will briefly present some common biological literature search engines to retrieve text and non-text material.

2.1 Handling Text Material

With the rapid growth of biomedical literature, information retrieval researchers need to get the aid they need literature, following a relatively common example of HubMed and GoPubMed.

HubMed (http://www.hubmed.org/) is an optional search interface combined with external network services and provides a function to improve the literature search, browsing and retrieval efficient[5]. Users can create and visualize clusters of related articles, reference data in a variety of formats, to get updated daily publications and browse to the full-text links. HubMed is the EUtils of the NCBI Web Services, interpreting the PubMed data in another way. HubMed and PubMed use the same syntax to search and get the same result, but the expression is different between HubMed and PubMed. The result of HubMed can be sorted by date information or relevance. GoPubMed (https://www.gopubmed.org/) is a system combined with semantic network of biomedical information retrieval, to classify PubMed search results, which can help users quickly locate the most relevant literature, and base search result on visual statistical analysis from multiple perspectives[6]. GoPubMed works by using the extraction ontology terms (Go terminology, MeSH term and UniProt protein), and the search results are divided into four categories: What (ontology terms), Who (Author), Where (city and periodicals), When (publication date), which users can quickly browse required documents by classified Navigation. In addition, GoPubMed can analysis literature information.

2.2 Handling non-text material

The traditional retrieval method is based on paper level which meet the needs of most of the researchers [7]. However, it had to spend a lot of time and effort to collect relevant literature and determine the correlation, and these items embedded in the article in-depth content, such as, embedded charts, tables, diagrams, maps, photographs, etc., it is hard to be retrieved. In fact, these non-text not only to assist researchers on article content flexible and in-depth analysis, but also to promote knowledge discovery, so it has important value. In addition, charts and data are often more convincing, giving the impression that clear at a glance.

For the image retrieval, first cellular localization image search (SLIF) system[8, 9], SLIF extract and analyze specific types of images, the use of geometric moments, word processing and morphological image processing to extract all graphic images in the full text of journal articles BM and to identify these images depict fluorescence microscopy[10], and then determine the subcellular capture Location digital features (calculating SLIF6 features and converting the output of a single fraction). Although the images provide important evidence, without reference to the relevant text often cannot understand. Yu examined the three associated text: the captions of figure, the main abstract sentences and the associated sentences appearing in the full text[11]. Conclude that the summary sentences can be used to summarize the image content and other relevant text description usually does not include the indications described in experimental procedures and conclusions of the experiment.

BioText Search Engine provides free of charge biologists search engine which is based on web Service[12]. The current system indexes all Open Access articles available at PubMed Central. In order to improve the efficiency of the figures information retrieval, BioText Search Engine for articles extracted from local storage, use Lucene open source search engine indexing tool kit[1]. Currently BioText Search Engine is mainly based on keyword search, which is including keyword search the full text and abstracts, captions of figures and tables illustrate retrieval. Searcher through "view all figures and tables from this article" link associated with the article realize all charts or tables, but also through "view full article" link experiment with the full text of the link graph. BioText Search Engine’s development objective is to provide users with a more comprehensive search portal, such as author, journal name, and support like genes, proteins, organisms, species and other local features retrieval. Another search engine FigSearch can be extracted by retrieving summary chart containing the gene names [13]. Compared with the BioText Search Engine, it’s more targeted, but shows the limitations of its retrieval.
While researchers have been working on how to improve the retrieval efficiency, but there are still many urgent problems for the non-text information retrieval. First, the accuracy rate is not high, and the current precision of the system cannot meet user’s expectations. We can find three factors affecting its precision by analyzingAbove all, it’s difficult to accurately extract the target. Then, whether the target object accurately indexing, indexing words that accurately describe whether the target object topic. Finally, there is no a standard to describe metadata. Second, paying for resources lead to restrictions on access to get full literature as result of copyright issues or other issues. Third, the quality of graphics or table is not high, for example, data are incomplete or unclear, leading to its value in use not high.

Thus, we can predict the trends that how to improve the efficiency of non-text information retrieval. Following is about this.

First, the type of retrieve objects will continue to be expanded, currently the main targets for article retrieval of charts, tables and photos. In addition, the academic literature also contains a variety of formulas, flow charts, scientific data and other resources, which play an important role in research work.

Second, Content-based retrieval methods will be more and more attention. Retrieve content based on depth refers to the semantic content according to the target object, and analyze contextual understanding, in order to achieve a deeper level of technical retrieval methods, such as color, texture and shape of the images, some distribution of tones, shoot theme, scenes in the photo. This retrieval method which directly extracts content from the target object can perform similarity retrieval and meet the user’s needs to retrieve multi-level, interactive and intuitive query. Third, meta-data [14] will continue to be improved due to target objects belonging to a new resource type, and the type, format and dimensions are quite different. It’s impossible to accomplish description of all the items through the establishment of a metadata Description Standard. So, it’s difficult and complicated to build object metadata. Database developers and experts in various disciplines required more effort to construct the metadata standards, which will also become a major research focus in the future.

Fourth, to develop complex integrated retrieval will be a new direction. The current retrieval focus on the integration of text information. However, integration of non-text information retrieval, text information retrieval and cross with non-text information has not yet been solved. Meanwhile, it’s a challenge that researches encapsulate different sources with the same subject text information, scientific data, pictures, charts, maps, photographs and other information into a composite object.

Overall, the challenges facing the biological field of information retrieval is how to combine biological background knowledge, correct understanding user queries and various biological entity names appearing in the biological literature. This will be a great present or future research focus.

### 3 Text Mining

Because of the multitude of biological literature, so a lot of life science issues, and bioinformatics problems, need to use text mining methods, for example, we can find the relationship between genes and diseases, protein interactions, and so on.

The main task of text mining are term recognition, information extraction, discover relationships, and applied to the field of bioinformatics. Currently, some active area of research are named entity recognition, text categorization, relationship extraction, synonyms, abbreviations extraction, hypothesis formation, text classification and more integrated framework.

#### 3.1 Terminology Recognition

##### 3.1.1. Named entity recognition (NER)

Biological field named entity recognition is to find the name of the biological entity from the biological literature, which is marked as correct category (such as genes, proteins, disease, drugs, etc.). It is an essential part of biology literature retrieval and information extraction.

One of the basic tasks of biomedical text mining named entity recognition (Biomedical Named Entity Recognition, Biomedical NER), its purpose is to identify the name of the specified type from the collection biomedical text, such as protein, gene, ribonucleic acid, DNA and so on. This is a critical step for relationship extraction and other potential information.

However, the new named entities continue to emerge and the names of entities which has not been recognized are increasing. Currently there is not a complete biomedical named entity dictionary to contain various types and following features also resulted in the identification difficulties.

First, the length of the entity name is very long, the phrase may consist of multiple-word. For example, “normal thymic epithelial cells”.

Second, more than a common noun to show an entity. Many biomedical named entity with “and” or “or” connected parallel structure, they share a common noun, so it is difficult to correctly identify named entities, for example, “91 and 84 kDa proteins”.

Third, multiple forms of expression of the same name of the entity, specifically, a lot of biomedical named entity has written a number of different forms, for example TNFA, TNF an alpha, infalpha and TNFalPha all represent the same gene.

Fourth, different names share the same form of expression. Specifically, the same word or phrase that can represent different types of biomedical named entities, depending on the context can be inferred, for example, IL22 both expressed protein name, and said DNA name.

Fifth, the nesting phenomena of biomedical named entity still exist, e.g.<PROTEIN> <DNA> kappa 3 </DNA> binding factor </PROTEIN>, where “kappa 3” is the gene name, and "kappa 3 binding factor" is the
name of the protein, and therefore it’s necessary to solve the problem of overlapping candidate named entity.

Sixth, a large number of biological named entities using abbreviations.

Seventh, it is difficult for all the names to be enumerated and be included in the dictionary.

On the whole, these features brought great difficulties to determine the boundaries of the entity’s name, therefore, biomedical named entity recognition will be a challenging study. Currently, they are about three recognition method of biomedical named entity. Following three categories are dictionary-based approach, heuristic rule-based approach and based on machine learning methods.

First, dictionary-based approach, the natural language text are compared with dictionary entry which include a large number of pre-existing biomedical named entity name, based on the results match. The algorithm requires a larger refinement over the name of the library as well as matching strategy. Implementation of this method is more direct and easy to understand, but because of the large number of names in the biological field, frequently updated and expression are inconsistent, so have to maintain a biological dictionary library is very difficult. Second, Rule-based methods, is to identity the entity and other text which is divided into different categories by defining rules. In fact, the field of bioinformatics is complex and many varied forms of entity names, and rules-based method requires a lot of knowledge structure, but it is impossible to cover all the rules, resulting in lower recognition rate. In addition, the new entity emerging name in the dictionary cannot be updated in order to contain all the new entity name, resulting in a lack of portability. Third, Machine learning-based approach, is to transform the named entity recognition problem into classification problem. By using the classification tools, the training text for machine learning, and thus to distinguish various types of entities named in the final test text identifies named entities. This method is to extract the need to identify the name of the entity model, as the standard for classification method in a tagging corpus of training. The method reflects the greater advantage with flexibility, adaptability to a specific environment, as well as the ability to deal with small sample. But it needs a large number of corpus, for manual annotation to serve as a training set, usually due to uneven data lead to "over-learning."

3.1.2 Synonyms and acronyms

Biomedical literature growth while biomedical terminology is also growing and abbreviations occupied high proportion, for example, IFN, TPA and so on. There is no discipline for the formation of a lot of abbreviations, and acronyms also has a high degree of ambiguity. In general, the extended form of the abbreviations has more than evidence to determine its category. Anyway, to identify acronyms must largely dependent on context, and cannot rely on existing biological dictionary. Biomedical entities have multiple names and abbreviations, if there is an automated method of collecting synonyms and abbreviations to help researchers conduct literature research, it’s will be very useful. In addition, if all synonyms and abbreviations entities are mapped to a term which represent the same concept, the other text mining tasks can be completed more efficiently.

Many researchers use an online database to generate a list of synonyms gene names. Yu and Agichtein combines AbGene gene named entity recognition system, using statistical methods, SVM classifier and automatic mode extraction[15], as well as hand-generated rules algorithm to extract synonymous in the full text of journal articles. According to statistics, their system’s recall rate is about 80% and accuracy rate is about 9% then the total value of the F-measure is about 30%. Thus, the problem which automatically identify abbreviation and its full name in a single article has been basically solved, and the identification systems have achieved a higher accuracy and recall rate. Future research will combine identify abbreviations with other text mining tasks, and apply them to real biomedical text mining system.

As we all know, synonym extraction accuracy of gene and protein names is also generally lower, so there are more challenging. Although the list of synonyms automatically updated, to improve the performance of document retrieval and text mining system is valuable, but the accuracy of automatic extraction system is too low. However, the current work in progress, the official are standardizing gene names and symbols of the protein, so in the future the low accuracy problem caused by informal name may be reduced.

3.2 Information Extraction

Information extraction of the biological field is to extract the relationship which may be protein-protein interactions, or gene regulation relationship, according to user-defined interactions from biological literatures. This has important significance for creating a whole network of biological knowledge to predict the biological relationships, and development of new drugs.

There are two extraction methods which are co-occurrence and natural language processing methods (Natural Language Processing, NLP) [16]. Co-occurrence method is based on a class of entities from the literature, which assumes that the closer the biological literature entities, the higher their relevance. For example, if two entity objects appear in a sentence and have a high frequency of co-occurrence of two entities by text mining dealing with a large number of abstracts, it imply the existence of two entities associated with a high probability target. Such method is relatively easy to implement, but the assumption is not reliable. The other is natural language processing method, is based on large-scale corpus analysis and complex language model to extract relationship by grammatical structure and semantic relations. Although it’s more complex and difficult to achieve, but the association is more accurate and reliable.

On the basis of the relationship between two methods, some researchers have tried to extract the biological protein and build pathway from the literature and the network
(Network). And we all know that extracting genes, proteins or gene ontology (GO) has a direct significance. Chiang and Yu's Meke system using gene ontology (GO) as a function of the name of the dictionary coding, the genes and gene products with the name of the dictionary LocusLink combined sentence arrangement while taking advantage of the system to determine patterns associated with gene function, then the model combined with the native Bayes classifier to extract sentences containing information about the function of the gene product[17]. 2004 BioCreative Task two also extract relevant gene ontology (GO) encoding gene from the free text.

Although, in recent years, the biological literature information extraction system from a simple rule-based extraction for advanced development of the use of computer language parser, however, there are still many problems to be improved, the current text mining systems cannot meet the actual needs of researchers and performance evaluation criteria system is not yet mature, so biologists and computer scientists need to cooperate further strengthened.

3.3 Hypothesis Generation

Relationship extraction is mainly extracted relationships between entities that can clearly be found in the text, while the hypothesis tries to reveal the more explicit relationship that does not exist in the text but inferred by other. The purpose is to reveal previously undiscovered deserves further study the relationship. Text mining is defined as "by automatically extracting text from different information resources, the discovery of new information previously unknown." With extracts from the published literature to determine the information compared to pay more attention to the unknown text mining access to information. For such an experimental biological sciences, the biological text mining aims to dig out from the literature experimental hypotheses and experiments suggested that in order to get biologists to verify new scientific discoveries.

In fact, all the research hypotheses are formed idea of using "non-relevant literature complementary structure" of the 1980s, Swanson proposed is simply ABC model: A affected B, and B on C, then A may influence C in the 1980s and early 1990s Swanson ABC model gives a lot of use and tap new hypothetical example, such as: fish oil treatment of Raynaud's disease [18]; magnesium deficiency and headaches relationship[19].

Hypothesis mining system is not yet a standard tool for biologists, but someday there is a need to continue research and improve the ability of the system handle a large number of different types of data. At present these data must be manually researched by scientists and need better ways to evaluate these system, which can record the improvements of the system and to make a clear choice.

3.4 Text Categorization

Reading massive biomedical literature bring great difficulties for the researchers in recent years, but the widespread use of text mining and natural language processing techniques has solved this problem, in which text classification is an important part. Text classification attempt to automatically test document or some parts whether the document contain the interesting features of a particular subject. Interesting information is not explicitly specified by the user, but to provide positive training set (set of documents have been found to contain features of interest) and negative training set. Text classification system should be able to automatically extract features that can distinguish between positive and negative, and which are applied to the candidate documents and then to make decisions.

Accurate text classification systems are particularly useful for database administrators, who may have to browse a lot of database to find a small amount of literature which contains some valuable information, but more biomedical information is created in form of text, the database manager need to convert the information to the encoded data and there is a strong need that text classification method is applied to biomedical text.

Yeh organized the text mining contest that is the part of KDD international competition in 2002[20]. And the task is to evaluate papers FlyBase data set, and according to the Drosophila gene products to determine whether it should manage that paper. The performance of the best entries are created by manually generating a set of rules which are based on tagging, semantic dictionary and semantic restrictions and the F is 78% by detecting the raining document. Another effective method is based on the text whether there are the gene products to classify biomedical papers, and after extracting feature, to use the Naive Bayesian classifiers, have a good performance.

The study that applying text classification to the actual work of biomedical managers and indexing has only just begun. one of the tasks of 2006 TREC Genomics Track is text classification problems[21]. This task is trying to imitate manual annotation, in order to find the document experimental evidence of genetic information contained in the mouse genome informatics (MGI) system browsing, and finally completed the text of the Standard Generalized Markup Language collection (SGML) format. To evaluate the performance of the task and to meet the needs of managers and other researchers for the future. Text classification technology can improve the efficiency of biological databases were screened, so improving the text classification in biomedical research must continue.

3.5 Integration Framework

To address a wide variety of user needs, many research groups are developing integrated text mining framework. The MedScan systems of Novichkova and others developed, is a synthesis of the dictionary and syntactic, semantic template as a set of entity relationship extraction of universal biomedical text mining system[22]; Glenisson developed TXTGate, perform the text profiles and clustering of genes by using multiple online databases
[23]; Becker created the PubMatrix tool which combining multiple queries PubMed list the two-dimensional gene lists contained the names and functions of terms [24]; BioRAT system of Conney is another template-based system that combine template design tools with web spider that is used for locating and retrieving the full text of journal articles combination [25]; Textpresso system is an Ontology-Based Information Retrieval and Extraction System for Biological Literature [26].

However, there are also some problems which all of these systems are still in the research and development phase and the evaluation tends to simple. In addition, the system has not conducted a thorough evaluation of the user, and the ability to meet the needs of the biomedical research community remains to be seen, this is just to satisfy a stage biomedical researchers needs.

3.6 Evaluation Of Biological Text Mining Technology

In the field of bioinformatics, The technical evaluation of the text mining systems which were mentioned in the above has two indicators which is precision and recall rate [27]. Precision is identified by the creature in the named entity named entity of the ratio , namely accuracy rate (P) = the number of correctly predicted named entity / entities to predict the total number of names; Recall rate refers to the experiment identified by all the correct ratio of biological named entity of the experimental data, that the recall rate (R) = total number of correctly predicted named entity named entity number / the total number of named entities of a text, we can see precision and recall rate which reflects two different aspects of recognition quality. So, we can use harmonic measure as a balance recall and precision rates, namely, F-measure = 2PR / [P + R].

4 Summary and Discussion

In summary, we can find the biomedical text mining has tremendous potential. The experts in various fields especially medical researchers are also gradually realizing that the job prospects. So, the next major challenge in text mining work is to develop efficient text mining tools, making it essential for biomedical researchers so that they can continue more productive under the pressure of the rapid growth of information. Research must focus on helping biomedical researchers to solve practical problems, and less on independent systems to meet user demand output. Of course, it needs to build a bridge between biologists and computer experts in the field, because the field of researchers led by a computer background, but only biologists to propose effective evaluation methods and mining tasks. Thus, biologists and computer linguists should make further cooperation in the field of text mining show the diversity and innovation, so as to understand the actual needs, access to domain knowledge to develop an effective text mining tools.

For the future, although the biological literature mining related technologies carried out some research, but still need to continue to look at areas in-depth study, for example, efforts to improve the identification and biological information extraction performance of biological entities, in addition, in terms of literature and words must be easier to obtain in the full text of journal articles, which did not mention in the Summary and Mesh terms. Current text mining research has shifted from the title and abstract, but still get the full text of copyrighted restrictions. Therefore, the research community must cooperate with publishers to obtain a wide variety of content for text mining. Next, in addressing the specific text mining tasks, more research is needed to determine what features work and what types of features are useful. Feature space for text mining is a huge array of feature types, feature types, including (but not limited to) the words, concepts, keywords, format, author, references and links. Popular long-term bags (bag-of-words) method can be applied to a variety of texts from different sources, but ignores the location of the document and paragraph information. Finally, obviously, the main theme of the future development of the coordination and cooperation between disciplines must work together text mining researchers, publishers and biomedical researchers can produce by providing a consistent, measurable, verifiable results system to meet user needs researchers must take the lead in coordinating efforts to achieve biomedical text mining full scientific potential. Obviously, the main theme of the future development is the coordination and cooperation between disciplines must work together text mining researchers, publishers and biomedical researchers can produce a consistent, measurable, verifiable results system to meet user needs. Researchers must take the lead in coordinating efforts to achieve biomedical text mining full scientific potential.

5 Acknowledgements

I thank Mr. Zhang because of his constructed advice. And I am grateful to the authors who is quoted be on this paper.

References


