Apply Text Mining to Advance Cancer Research

Si Yan
Stuart School of Business, Illinois Institute of Technology, Chicago, IL USA
Email: syan3@iit.edu

Yanliang Qi
Department of Information Systems, New Jersey Institute of Technology, Newark, NJ USA
Email: yq9@njit.edu

***Abstract*—Cancer is a malignant disease that has caused millions of human deaths. Text mining can help researchers discover hidden rules and relationships between documents so advanced cancer research. In this paper, we analyze the properties of text mining and cancer research documents. We discussed the research directions of text mining in cancer research with examples of systems and tools. In conclusion part, we talked about future way the text mining development.**

***Index Terms*—text mining, cancer risk assessment, hypothesis generation, knowledge discovery**

I. Introduction

Cancer is a malignant disease that has caused millions of human deaths. In 2012, about 14.1 million new cancers occurred globally, and caused about 8.2 million deaths [1], which is equivalent 14.6% of all human death [2]. Biomedical researchers spent a lot time and effort and time trying to find the way to cure cancer. Also, pharmaceutical and biopharmaceuticals companies invested heavily on the oncology studies. In each year, numerous research publications have been published. New research development in biomedical and biomedicine reply on discover hidden knowledge and relationships efficiently. Figure 1 shows the exploding number of articles available from Medline over the past 65 years (data retrieved from the SRS server at the European Bioinformatics Institute; www.ebi.ac.uk/) [3]

It is obviously the problem faced by the biomedical researchers is that how to effectively find out the useful and needed documents in such an information-overload environment. Traditional manual retrieval method is impractical. Furthermore, online biomedical information exists in a combination of different forms, including structured, semi-structured and unstructured forms [4]. It is impossible to keep abreast of all developments. Computational methodologies increasingly become important for research [5]. Text mining techniques, which involve the process of information retrieval, information extraction and data mining, provide a means of solving this [6].

Manuscript received May 28, 2015; revised July 19, 2015

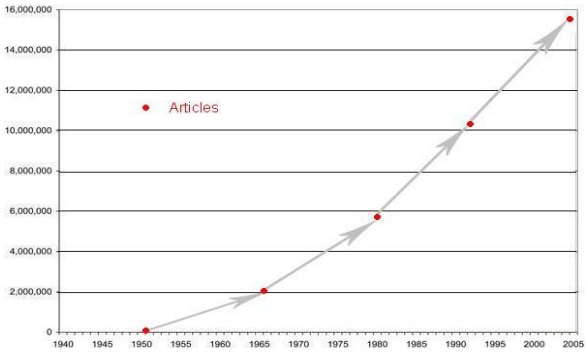

Figure 1. Medline increase.

The goal of text mining is to help researchers to identify the needed information more efficiently, uncover relationships and hidden rules from the huge amount information.

This paper is distinguished from other survey papers in the following several aspects

1) It discusses the difference between text mining, information retrieval and information extraction.
2) The property of text mining and special characters of biomedical documents is also discussed.
3) This paper also presents the research directions of biomedical text mining with examples.
4) Finally, the problems and future ways are discussed.

II. Features of Text Mining and Cancer Research Documents

Text mining is the technology which discovers patterns and trends semi-automatically from huge collections of unstructured text. It is based on several following technologies, Natural Language Processing (NLP), Information Retrieval (IR), Information Extraction (IE), and Data Mining (DM) [7].

From the search-centric point, information retrieval, information extraction and data mining play the similar role, as text mining--getting knowledge from datasets. But actually, there is a big difference between text mining and other four techniques. The main difference is that whether there is novel produced in the process [8]:

doi: 10.18178/ijpmbs.4.2.132-135

The goal of information retrieval is to search the metadata which describe documents, information in documents, documents themselves within the database, whether relational stand-alone databases or hyper-textually-networked databases such as the World Wide Web [9]. The process of IR is to find out the needed information in database which is already exists, so in this process, there is no novel produced.

From the point of Mooney and Bunescu [10], Information Extraction "distills structured data or knowledge from unstructured text by identifying references to named entities as well as stated relationships between such entities". It is one type of Information Retrieval. The specialty of IE is to extract structured information automatically. In this process, there is also no novelty produced [8].

Data mining mainly focus on the structured data while text mining mainly focuses on semi-structured and unstructured data. Data mining attempts to find the patterns and relationships from the large datasets for the analyst and decision-makers [11]. From the argument of Hearst [8], data mining is not "mining" but simply a (semi) automated discovery of patterns/trends across large database and no new facts was created in this discovery process.

Natural Language Processing is the study which focuses on automated generation and understanding of natural human languages. It is a research area that applies computer technique to understand and manipulate natural language text or speech to do useful things [12]. Text mining is also different from NLP. In the survey of biomedical text mining by Cohen and Hersh [13], the difference was fully discussed. From the viewpoint of Cohen and Hersh [13], NLP tries to understand the meaning of the text as a whole, while text mining focus on the specific problem in a specific domain and tries to solve it (possibly using some NLP techniques in the process). For example, For example, by using text mining technique, a database administrator can select the articles with specific common interest.

From the above discussion, we can know the features of text mining, data mining, information retrieval, information extraction, knowledge discovery and NLP. Their properties different on the several aspects, like data source (data base or text source), data type (structured or unstructured), question scope (broad or specific) and knowledge discovered (find existed, hidden knowledge or newly, never encountered information)

In the following table (Table I), the comparison summary was presented of text mining with information retrieval, information extraction, data mining and natural language processing.

Most cancer research documents are stored as semi-unstructured model and unstructured model, text mining can play a crucial role in helping researchers find out the hidden, existed knowledge and also discover the new relationship between those documents. Compared to normal documents, these documents have some special characters [14]:

- First, it emerged many domains, such as biology, computer science, medicine, clinical trials, artificial intelligence, applied mathematics, statistics etc.
- Due to domain specific, cancer research documents are very ambiguity and variability. Domain specific terminology is heavily used in biomedical language. Polysemic words with acronyms are widely used, which cause a lot ambiguity, like APC may stands for Argon Plasma Coagulation or Activated Protein C;

Data sparseness are also existed which means most words with low frequencies.

TABLE I COMPARISON BETWEEN TEXT MINING, INFORMATION RETRIEVAL, INFORMATION EXTRACTION AND DATA MINING

Terms	Features	Different from Text Mining
Text Mining	Discovering heretofore unknown information from a text source	N/A
Information Retrieval	Finding the required information which already in databases	Extracting new discovery and never-before encountered information
Information Extraction	Extraction of structured data/knowledge from unrestricted text sources.	Extracting new discovery and never-before encountered information
Data Mining	Mainly focus on strucutred data	Mainly forcus on unstructured data
National Language Processing	Understand the meaning of text as a whole	Concentrate on solving a specific problem in a specific domain

III. RESEARCH DIRECTIONS

There are several directions about text mining applied in cancer research. In this paper, we focus on three directions (Fig. 2):

- Hypothesis Generation, which was important in drug discovery to explore the possible solution the existed problems.
- Cancer Risk Assessment (also famous known as CRA), to evaluate the environment influence such as chemical and exposure.
- Knowledge Gathering and Discovery, to find the hidden knowledge and rules behind huge amount of information, which will accelerate the research process.

A. Hypothesis Generation

A hypothesis is a proposed explanation of phenomenon. In scientific field, it is a trial solution to a problem. In statistical testing, it is widely used in drug discovery. There are two types of hypotheses: *alternative* hypotheses and *null* hypotheses. The alternative hypothesis is the initial research hypothesis. The null is the logical opposite of the alternative hypothesis. The hypothesis test typically including four steps: generate hypothesis, set significant level, collect evidence and final make decision. We will

use a legal example to show how hypothesis works. Fig. 3 is the judicial analogy of hypothesis testing:

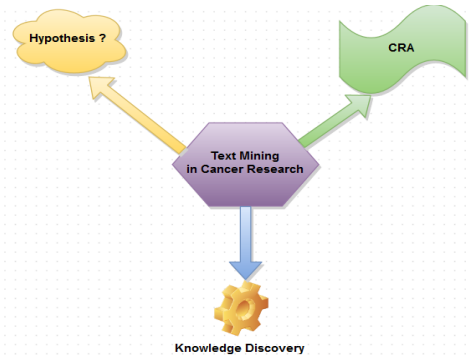

Figure 2. Research directions.

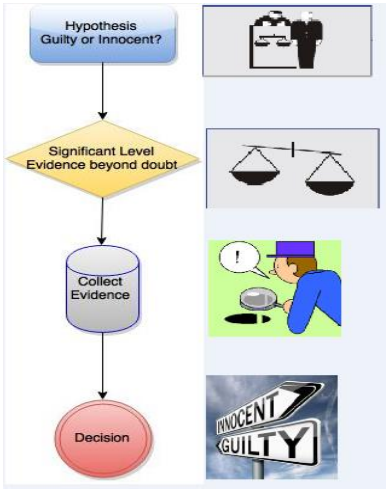

Figure 3. Analogy of hypothesis testing.

Consider this as a judicial analogy; alternative hypothesis is "the defendant is guilty", while null hypothesis is "defendant is not guilty". In clinical trial development, researchers are also selecting the significance level, typically as 5% to test the hypothesis, which means "beyond a reasonable doubt" to make decision to reject null hypothesis or fail to null hypothesis. In judicial analogy, the significant level is the amount of evidence needed to convict. In legal example, as the court of the law, the evidence must prove guilt "beyond a reasonable doubt", means the probability to conclude suspect is guilty when he or she is innocent. In clinical trial, this is used to avoid type I error, reject null hypothesis while null hypothesis is true. After make full investigation, the decision could be made based on the evidence and significant level. If the evidence is strong enough (the very low possibility of type I error), then reject the null hypothesis, means conclude the defendant is guilty. While if the evidence is not strong enough, fail to reject the null hypothesis, means can't conclude if the defendant is guilty; Also note, failing to prove guilty not means the defendant is innocent.

Previously, biomedical researchers set hypothesis based on the experience and knowledge from literatures. With the help of text mining, biomedical researchers get the summary and key points of literatures fast and accurately. So the hidden relationships and bridges between biomedical relationships and facts can be discovered. The researchers can test those hypothesis and the test results can guide the next step of research and experiments.

There are already several applications to generate hypothesis automatically. For example, Arrowsmith [15] is a basic discovery tool that identified the meaningful links between two sets of literatures from PubMed, even they share no articles or authors in common and represent disparate topics or disciplines. The software is designed based on transitive rules, which model can be simply stated as 'If A influences B, and B influences C, therefore A may influence C'. The working process is a two-node search function. First, the user needs to input two sets of articles, literatures to software, like set A and set B. Then Arrowsmith will remove common articles in both set A and set B, which will guarantee to compare only indirect articles. After that, the software identifies all words, two or three words phrase presented in the title of articles. Those words or phrases are called B-term. Finally, those outputs will be displayed based on predictive probabilities and the relevance to the user. User can click the B-term to show up related articles [16]. The working flow can be shown as Fig. 4.

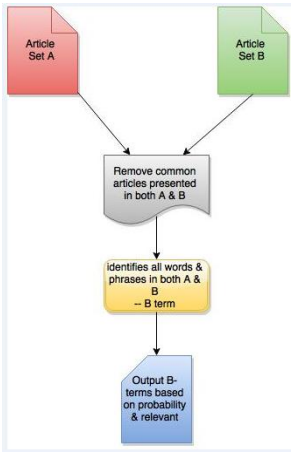

Figure 4. Arrowsmith two-node search.

You *et al.* [17] presented a text mining based system which can apply multiple classifier technique to analyze the diagnostic terms across clinical records.

B. Cancer Risk Assessment

Cancer risk assessment is the process to identify the relationship between chemical and exposure from existing published evidence [18]. In worldwide, more and more evidence shows the link between environmental chemicals and cancer and government legislations becomes tight. This trend makes CRA research increasingly important than before [19].

Lewin *et al.* [20] made an exploratory test on CRA. They used PubMed publications abstracts as sample and manually annotated according to relevance and evidence of cancer RA and selected test chemical. Their results show supervised learning technique can yield high accuracy and help finding the related tasks.

C. Knowledge Gathering and Discovery

Free text is widely used in medical and clinical records. Text mining can help researchers gather information and discover knowledge quickly.

CaFE [21] is a registry Case Finding Engine developed by University of Michigan, USA. By using CaFE, cancer patient identification can be automatically pulled from unstructured, free-text pathology report.

MedLEE (Medical Language Extraction and Encoding) [22] is a system developed by Columbia university to automated encoding of the information from text documents like discharge summary, radiology, pathology, etc. Consists of the following modules: pre-processor, parser, phrase regularization and encoding [23]. It can be used to retrieve the reportable diagnosed from medical reports.

HITEx (Health Information Text Extraction) [24] is a text mining software developed by Harvard University to extract medical information and diagnoses from discharge summary and pathology reports.

caTIES (cancer Text Information Extraction System) is a text mining tool to automatically extract free text pathology reports then stored as structured data, which will be used to facilitate advanced query and analysis [25].

IV. Conclusion

Due to its powerful information retrieval, knowledge discovery ability from unstructured text, Text mining has already played a big role for cancer research. Many researches and systems have been developed to help biomedical and biomedicine researchers. While the specific characteristics of cancer research document, ambiguous and domain specific, still made this is not an easy task. In the future, we should focus on increasing the accuracy of the system and expand the application area, make the system as domain specific, for specific type cancer. And also, we will focus on making system automated and user friendly to get more feedback from biomedical and biomedicine researchers, through this way to improve the performance of Text mining and fulfill its huge potential.

References

[1] World Health Organization, "World Cancer Report 2014," 2014.

[2] WHO, "The top 10 causes of death," May 2014 [Online]. Available: http://www.who.int/mediacentre/factsheets/fs310/en/. [Accessed 12 May 2015].

[3] D. Rebholz-Schuhmann, H. Kirsch, and F. Couto, "Facts from text — Is text mining ready to deliver?" *Plos Biology*, vol. 3, no. 2, p. 65, 2009.

[4] M. Ghanem, Y. Guo, A. Rowe, A. Chortaras, and J. Ratcliffe, "A grid infrastructure for mixed bioinformatics data and text mining," in *Proceedings of the ACS/IEEE 2005 International Conference on Computer Systems and Applications*, *IEEE Computer Society*, 2005, p. 41.

[5] G. Black and D. Stephan, "Bioinformatics: Recent trends in programs, placements and job opportunities," Report to the Alfred P. Sloan Foundation, New York, 2004.

[6] S. Ananiadou, D. B. Kell, and J. Tsujii, "Text mining and its potential applications in systems biology," *Trends in Biotechnology*, vol. 44, pp. 571-579, 2006.

[7] N. Uramoto, H. Matsuzawa, T. Nagano, A. Murakami, H. Takeuchi, and K. Takeda, "A text-mining system for knowledge discovery from biomedical documents," *IBM Systems Journal*, vol. 43, no. 3, pp. 516-533, 2004.

[8] M. Hearst, "Untangling text data mining," In *Proceedings of ACL'99: The 37th Annual Meeting of the Association for Computational Linguistics*, 1999.

[9] A. Singhal, "Modern Information Retrieval: A Brief Overview," *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, vol. 24, no. 4, pp. 35-43, 2001.

[10] R. J. Mooney and R. C. Bunescu, "Mining knowledge from text using information extraction," *SIGKDD Explorations*, vol. 7, no. 1, pp. 3-10, 2005.

[11] A. M. Cohen and W. R. Hersh, "A survey of current work in biomedical text mining," *Briefings in Bioinformatics*, vol. 6, no. 1, pp. 57-71, 2005.

[12] G. Chowdhury, "Natural language processing," *Annual Review of Information Science and Technology*, vol. 37, pp. 51-89, 2003.

[13] A. M. Cohen and W. R. Hersh, "A survey of current work in biomedical text mining," *Briefings in Bioinformatics*, vol. 6, no. 1, pp. 57-71, 2005.

[14] M. Krallinger, "Current trends in biomedical text mining," [Online]. Available: http://www.mavir.net/docs/MKrallinger.pdf. [Accessed 7 May 2015].

[15] N. R. Smalheiser, V. I. Torvik, and W. Zhou, "Arrowsmith two-node search interface: A tutorial on finding meaningful links between two disparate sets of articles in Medline," *Comput MethProgram Biomed*, vol. 94, pp. 190-197, 2009.

[16] N. R. Smalheiser, V. I. Torvik, W. Zhou, "Arrowsmith two-node search interface: A tutorial on finding meaningful links between two disparate sets of articles in Medline," *Computer Methods and Programs in Biomedicine*, vol. 94, no. 2, pp. 190-197, 2009.

[17] M. You R. W. Zhao, G. Z. Li, X. Hu, "MAPLSC: A novel multi-class classifier for medical diagnosis," *International Journal Data Min Bioinformatics*, vol. 5, pp. 383-401, 2011.

[18] U. E. P. A., "Guidelines for carcinogen risk assessment 2005," [Online]. Available: http://www.epa.gov/iris/cancer032505.pdf. [Accessed 12 May 2015].

[19] A. Korhonen, I. Silins, L. Sun, and U. Stenius, "The first step in the development of text mining technology for cancer risk assessment: Identifying and organizing scientific evidence in risk assessment literature," *BMC Bioinformatics*, vol. 10, no. 22, p. 415, 2009.

[20] I. Lewin, I. Silins, A. Korhonen, J. Hogberg, and U. Stenius, "A new challenge for text mining: Cancer risk assessment," in *ISMB BioLINK Special Interest Group on Text Data Mining*, 2008.

[21] D. C. Danauer, "Registry case finding engine (caFE): An informatics tool to identify cancer patients in electronic pathology reports," in *Frontiers in Oncology and Pathology*, Vancouver, BC, Canada, 2006.

[22] H. Xu, K. Anderson, V. R. Grann, and C. Friedman, "Facilitating cancer research using natural language processing of pathology reports," in *MEDINFO*, 2004.

[23] W. Scharbe, "Evaluation of Open Source Text Mining Tools for Cancer Surveillance," vol. 14, pp. 4, 2009. [Online]. Available: http://www.cdc.gov/cancer/npcr/pdf/aerro/text_mining_tools.pdf. [Accessed 12 May 2015].

[24] Q. Zeng, "Extracting Principal diagnosis, co-morbidity and smoking status for asthma research: Evaluation of a natural language processing system," *BMC Medical Informatics and Decision Making*, vol. 6, no. 1, p. 30, 2006.

[25] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan, "GATE: An Architecture for Development of Robust HLT Applications," [Online]. Available: https://gate.ac.uk/sale/acl02/acl-main.pdf. [Accessed 16 June 2015]

Si Yan, graduated from Stuart School of Business, Illinois Institute of Technology, majored in mathmetical finance. She has extensive research on text mining, biomeical research, biostatistics, predictive modelling and business intelligence. She also provides consulation for world class level pharmaceutical company.

Yanliang Qi, graducated from New Jersey Institute of Technology, focusing on research on text mining, bioinformatics, biostatistics, etc. He also has several publications in International conferences and journals. He also serves as editorial board member in International joural and session chair on world famous conference.