CSCanada
**Cross-Cultural Communication**
Vol. 12, No. 8, 2016, pp. 31-35
**DOI:**10.3968/8722

ISSN 1712-8358[Print]
ISSN 1923-6700[Online]
www.cscanada.net
www.cscanada.org

# Analysis of Hot Points on Data Mining Research of Medical in Foreign Countries

SHI Ximin[a]; ZHAO Wenlong[a],*; CHEN Juan[a]; YANG Junxue[a]

[a]College of Medical Information, Chongqing Medical University, Chongqing, China.
*Corresponding author.

**Supported by** the National Social Science Program Planning of China (13BTQ004); Chongqing Science and Technology Commission, China (cstc2015shmszx10004).

Received 19 May 2016; accepted 14 July 2016
Published online 26 August 2016

## Abstract
To promote the current development of medical data mining research, a quantitative statistics and qualitative analysis of the papers in the field of medical data mining technologies were made with the methodology of bibliometric and knowledge mapping, which were enlisted in the database of Web of Science analyzing the general situation of the papers about data mining from several aspects: period sequences, subject funds, countries and regions, core authors and research institutions, the hotspots and research frontiers. Our analysis exposed that the research of data mining in medical showed a multi-disciplinary integration of the development trend, but high-yield leading author group has not yet formed. It is important to note that scholars should raise awareness of clinical medical data mining as well as explore new research directions for further studying.

**Key words:** Medical; Data mining; Research hot points; Knowledge mapping

Shi, X. M., Zhao, W. L., Chen, J., & Yang, J. X. (2016). Analysis of Hot Points on Data Mining Research of Medical in Foreign Countries. *Cross-Cultural Communication, 12*(8), 31-35. Available from: http//www.cscanada.net/index.php/ccc/article/view/8722
DOI: http://dx.doi.org/10.3968/8722

## INTRODUCTION
With the development of science and technology as well as the construction of hospital information, how to explore the law of medical data effectively using data analysis methods, and make full use of massive medical data resources were what researchers concerned about. A rapid and accurate diagnosis and optimal treatment plan for the medical staff was needed so as to support the scientific decision-making service (Sun, Huang, & Zhu, 2015), which play a vital role in promoting the development of medicine.

Data mining was also known as knowledge discovery, which involved in a number of areas of interdisciplinary, extracting implicit valuable information and knowledge from the vast amounts of data with algorithms and rules (Han & Kambr, 2001). The main technology of data mining include classification, prediction, clustering, outlier detection, association rules, sequence analysis, time series analysis and text mining, and also include some new techniques, such as social network analysis and emotion analysis (Yan, 2015).

## 1 DATA SOURCES AND RESEARCH METHODS

### 1.1 Data Sources
The periodical literature of the corresponding medical field data mining was selected from the Web of Science of core collection database (BIOSIS Previews, MEDLINE, SciELO Citation Index). In order to meet the need of search efficiency, theme retrieval was utilized, and retrieval expression was "data mining" AND "medical*", we defined the type of literature as a journal, and retrieval time span was from January 1998 to April 2016. Then we download all of its records and reference citation format data and stored in TXT format as a sample set. Retrieval time was May 4, 2016.

### 1.2 Research Methods
In this paper, the method of bibliometrics and CitespaceIII (Chen, 2006) knowledge mapping software were used

to analyze hotspots of data mining research in foreign medical fields, and to grasp the development trend of international medical data mining.

With the rapid growth of visualization and data mining technology, knowledge mapping as a new method raised and became an important method in scientometrics now. It was a graph that shown the development process and structure of scientific knowledge (Chen et al., 2009), which would make tacit and explicit knowledge visual and help researchers understand the development trends as well as new developments in the field of science.

## 2. RESULTS

According to retrieval conditions, 2621 papers were retrieved from the database of Web of Science. The number of foreign medical data mining research papers on the overall shown a growth trend, among of which the United States issued a maximum of 808 articles, accounting for 30.83% of all published papers; China ranked second only to the United States, accounting for 14.04%, compared with the United States, there is still a certain gap. The following section will give a detailed account of the results.

## 3. DATA ANALYSIS

### 3.1 Number of Annual Publications

From the time series point of view, the foreign medical field of data mining research appeared an overall growth trend during 1998 to 2016. The number of subjects in the field of science could reflect the development degree and the level of the subject in a certain degree, which showed that foreign data mining technology in the field of medicine is in a rapid development stage.

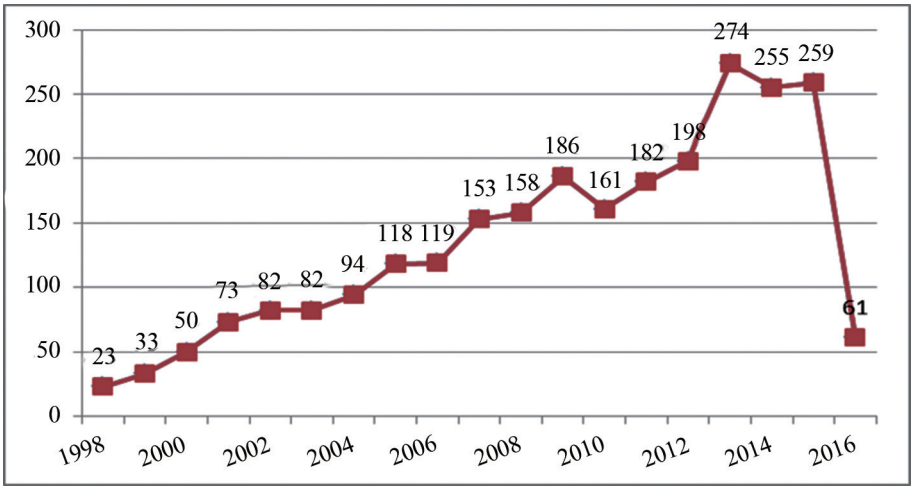


**Figure 1**
**Development of the Number of Annual Publications in Data Mining of Medical According to Web of Science (WOS) Database**
*Note.* Data searched on May 4, 2016.

### 3.2 Distribution of Authors and Institutions

Statistical analysis of high yield authors showed that TSUMOTOS ranking the first, published 60 papers. In general, there were 58 authors that published more than six articles of medical data mining research respectively, accounting for 19.19% of the total number of 503 papers. According to Preiss's law (the author published a total of 60 articles, 60 square root multiplied by 0.749, that is, 5.8, take an integer is six), it can be seen that the leading role of the medical field data mining related research high-yield author has not been formed, for not reached 50%. Table 1 listed the top ten high yield authors issued a number of statistics, a total of 170 articles published in the WOS. Statistical analysis of institutions that published in the field of data mining research shown that Univ Shimane issued a maximum amount of Stanford, followed by Univ Harvard, Univ Shimane, Med Univ, FDA US, Iowa Univ, etc. ( Figure 2).

**Table 1**
**The Top Ten Prolific Authors With the Largest Number of Papers About Medical Data-Mining Indexed in Web of Science**

| Author | Frequency | Percentage (%) |
|---|---|---|
| Tsumoto, S. | 60 | 2.289 |
| Hirano, S. | 27 | 1.030 |
| Beuscart, R. | 11 | 0.420 |
| Iwata, H. | 11 | 0.420 |
| Shah, N. H. | 11 | 0.420 |
| Zhang, J. | 11 | 0.420 |
| Liu, B. Y. | 10 | 0.382 |
| Reiner, B. I. | 10 | 0.382 |
| Zhou, X. Z. | 10 | 0.382 |
| Kusiak, A. | 9 | 0.343 |
| Sum | 170 | 6.488 |


**Figure 2**
**Knowledge Mapping of Affiliations in Web of Science**

### 3.3 Distribution of Countries /Regions

The quantity of academic papers that the top 10 countries / regions published in web of science were analysed (Table 2), which account for 78.02% of the total number of published papers. Among them, the United States issued a maximum amount, accounting for 30.83%, while China ranked second only to the United States, accounting for 9.08%. To a certain extent, it reflected that China as a developing country took a leading position in the field of medical data mining research, but compared with developed countries, such as the United States, there are still certain gap, and also showed that China has great potential for development in this field.

**Table 2**
**The Top Ten Countries/Regions With the Largest Number of Papers about Medical Data-Mining Research Indexed in Web of Science**

| Country/region | Number | Percentage (%) |
|---|---|---|
| USA | 808 | 30.828 |
| China | 368 | 14.041 |
| India | 168 | 6.410 |
| Japan | 154 | 5.876 |
| Germany | 127 | 4.845 |
| England | 114 | 4.349 |
| France | 105 | 4.006 |
| Australia | 101 | 3.853 |
| Italy | 100 | 3.815 |
| Canada | 97 | 3.701 |
| Sum | 2045 | 81.724 |

## 4. DISCUSSION

The leading edge of research often comes from the new scientific discovery or scientific development. It is the most advanced and the most development potential research topic or research field in the scientific study (Chen, 2009). CitespaceⅢ can be used in the identification of relevant research literature, and shows a new trend of scientific development. CitespaceⅢ software was choosed to study the "key words" in the field of medical data mining, and by setting the relevant attributes of CitespaceⅢ, selecting MST algorithm, setting the appropriate threshold, 172 nodes and 399 connections were presented on the CitespaceⅢ. Owing to the analysis of this paper was the research of data mining, So in the analysis we excluded the highest frequency of keyword "data mining".

### 4.1 Keywords Co-Appearance Mapping

Each of the circular nodes represents a keyword, the greater the radius of nodes, the higher frequency of keywords were. What's more, each node of the representative color keywords appear in the year, the colors of the rings tend to warm, indicating that the cited time was closer, and periphery of the purple circle nodes represents the current hot words. The size of the font and circle objectively reflects the medical field data mining research in different periods of heat (Chen, 2006). The development law of the subject field can be understood through the knowledge map of the key words. Shown as Figure 3, we got high frequency key-words such as classification, system, databases, algorithm, information, diagnosis, prediction and text mining etc..

### 4.2 Burst Terms Mapping

Burst detection can detect the research field or research topic which is obviously improved by the sudden increase in frequency or frequency of usage in a short time. CiteSpace III was used to highlight the function of the word. Through setting the relevant attributes of CiteSpace III, choosing "Keywords" as a node, selecting "citation burst", we got foreign literature high emergent word lists and keywords atlas, which help us get a clear understanding of the development trend in the field of medical data mining (Hou & Chen, 2007). As we can see from Figure 4, data mining technique, electronic health record, clinical data and data mining method were higher frequency of occurrence.

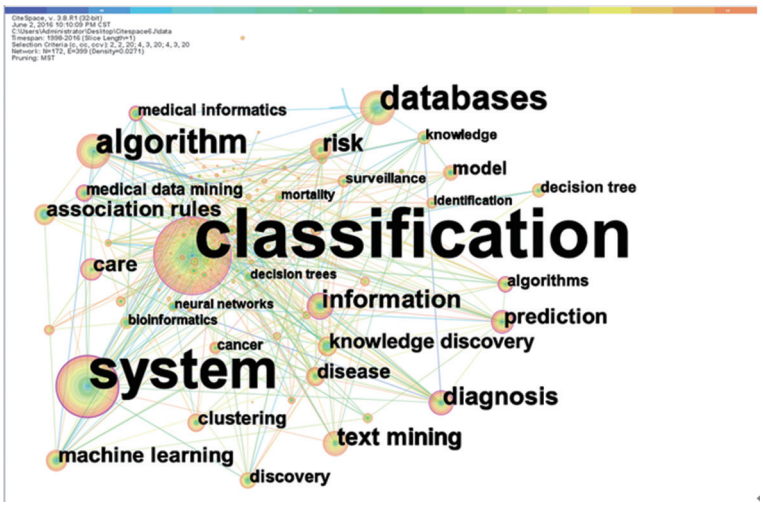


**Figure 3**
**Keywords Co-Appearance Mapping of Medical Data-Mining Research in Web of Science**

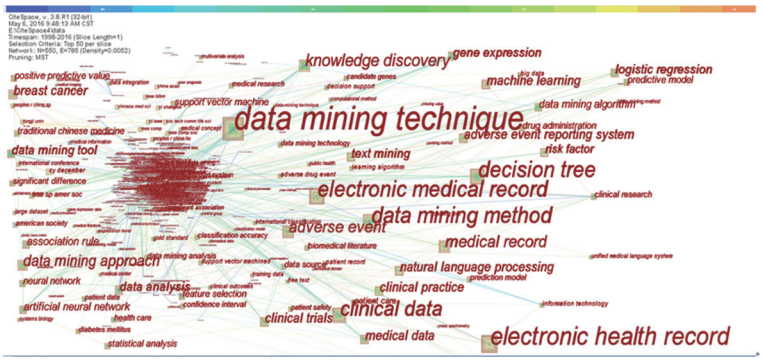


**Figure 4**
**Burst Terms Mapping of Medical Data-Mining Research in Web of Science**

**Table 3**
**Analysis of Keywords and Centrality about Medical Data-Mining Research Nearly Five Years in Web of Science Database**

| Year | Key word | Frequency | Centrality |
|---|---|---|---|
| | Risk-factors | 16 | 0 |
| 2015 | Electronic medical records | 10 | 0 |
| | Adverse drug reaction | 4 | 0.02 |
| | Immunization safety data | 4 | 0.01 |
| | Decision making | 9 | 0 |
| 2014 | Decision support systems | 8 | 0.02 |
| | Ionizing-radiation | 3 | 0.01 |
| | Outlier detection | 8 | 0 |
| 2013 | Genome-wide association | 5 | 0.01 |
| | Cardiology | 3 | 0.02 |
| | Biomedical domain | 2 | 0.01 |
| | Support vector machine | 8 | 0 |
| 2012 | Systems biology | 9 | 0 |
| | Medical-records | 12 | 0.03 |
| | Image mining | 6 | 0.01 |
| | Adverse events | 5 | 0.02 |
| | Clinical-practice | 3 | 0.01 |
| | Data integration | 8 | 0 |
| 2011 | Biomarkers | 4 | 0.01 |
| | Radiology reporting | 3 | 0 |
| | Natural language processing | 21 | 0.01 |
| 2010 | Electronic health records | 20 | 0.05 |
| | Gene | 4 | 0.02 |

### 4.3 Statistics Analysis of Key Words and Centrality

Based on the analysis of the key-words frequency and centrality in the database in recent five years, the method of data mining has diversified development trend. As seen form Table 3, decision support systems, support vector machine, data integration and natural language processing appeared gradually, which showed that the application of data mining technology in the medical field has been the deep development of the technical level.

## CONCLUSION

The amount of foreign medical data-mining research papers showed an increasing tendency, but the leading role of the research high-yield author has not been formed. Via the analysis of CiteSpaceIII, we conclude that the international researchers focused on electronic medical record of clinical data-mining research. Shown as tab3, electronic health records, risk factors, medical-records, diagnosis and natural language processing were higher frequency of occurrence. The research of data-mining in medical showed a multi-disciplinary integration of the development trend. Yamada (2014) analyzed the serum albumin level of hepatitis B virus (HBV) non related hepatocellular carcinoma in data mining research. Sudarshan (2016) studied data mining framework in recognition of the ultrasonic myocardial infarction. Teimouri (2016) used data mining tools and techniques for detecting disease medical prescription.

Different levels of hospital, clinical data acquisition of electronic medical records were not readily available, combined with the privacy of medical data mining, which increased the obstacles of medical data-mining research. Therefore, the application of data mining technology in the medical field needs to take full account of the special characteristics of medical data as well as methods of innovation.

## REFERENCES

Chen, C. (2006). Cite space Ⅱ: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology, 57*(3), 359-377.

Chen, C., Chen, Y., & Horowitz, M., et al. (2009). Towards an explanatory and computational theory of scientific discovery. *Journal of Informetrics, 3*(3), 191-209.

Chen, S. J. (2009). Survey of approaches to research front detection. *New Technology of Library and Information Service,* (9), 28-33.

Han, J. W., & Kambr, M. (2001). *Data mining: Concepts and techniques* (pp.1-18). Beijing Higher Education Press.

Hou, J. H., & Chen, Y. (2007). Research on the visualization of the evolution of strategic management science. *Studies in Science of Science, 25*(A01), 15-21.

Song, J. E. (1982). *Little science and big science* (pp.10-25). China: World Knowledge Publishing House.

Sudarshan, V. K., & Vidya, K. (2016). Data mining framework for identification of myocardial infarction stages in ultrasound: A hybrid feature extraction paradigm (PART 2). *Computers in Biology & Medicine, 71*(C), 241-251.

Sun, X. D., Huang, X. Q., & Zhu, C. L. (2015). Research on massive medical data mining analysis method based on evidence-based medicine. *Journal of Medical Informatics, 36*(3), 11-16.

Teimouri, M., Farzadfar, F., & Alamdari. M. S., et al. (2016). Detecting diseases in medical prescriptions using data mining tools and combining techniques. *Iranian Journal of Pharmaceutical Research, 15,* 113-123.

Yamada, S., Kawaguchi, A., & Kawaguchi, T., et al. (2014), Serum albumin level is a notable profiling factor for non-B, non-C hepatitis virus-related hepatocellular carcinoma: A data-mining analysis. *Hepatology Research the Official Journal of the Japan Society of Hepatology, 44*(8), 837-45.

Zhao, Y. C. (2015). *Data mining with R*. China Machine Press.