ISSN : 2350-0743

www.ijramr.com

InternationalJournalof Recent Advances in Multidisciplinary Research

Vol. 06, Issue 01, pp.4391-4396, January, 2019

RESEARCH ARTICLE

DATA MINING AND TEXT MINING WITH BIG DATA: REVIEW OF DIFFERENCES

***Filiz Ersoz**

Industrial Engineering, Karabük University, Karabük, Turkey

ARTICLEINFO

Article History:

Received 10th October, 2018
Received in revised form
18th November, 2018
Accepted 14th December, 2018
Published online 30th January, 2019

Keywords:

Big Data, Data Mining,
Text Mining, Knowledge Discovery.

ABSTRACT

In recent years, organizations have come across complex databases due to the development of technology, the growth of databases and information technologies, and the use of widespread information technologies. If the data found in line with the needs of the institutions are managed successfully and effectively, it is obvious that the institutions and organizations will offer great advantages and opportunities in economic terms. Each action performed in the digital environment leaves a data record behind it. For this purpose, big data, science, data mining and text mining concepts, methods and usage areas were mentioned, and similar and different aspects of data mining and text mining were determined. Through this study, it is expected that the separation of data mining and text mining methods will accelerate the process of strategic decisions of institutions and organizations and provide the required support.

INTRODUCTION

The data gathered in the quantitative and qualitative data collected in databases and the data warehouses and increasing the use of information technologies in parallel with the development of technology have increased. As a result of this situation, there is a need to reveal meaningful relationships, patterns and trends from large masses of data, and the importance of processing of data in making accurate and strategic decisions has increased.As a result of the exponential increase in the amount of data produced, stored and processed as the computer and internet technology develops, the big data concept and the new field of data science have begun to develop (Gursakal, 2014). As management information systems evolved from the 1960s and information systems emerged as a software category and field of application in the 1980s and 1990s, the numerical data stored in relational databases constituted a data mining discipline. In the same way, text mining in the processing of text in unstructured documents has also revealed the discipline of text mining. The increase in advanced technologies and the increase of work experience data or text as records have led not only to the large data mining studies but also to the work volume on large textual data. In today's world, data mining and text mining are now considered important disciplines for enhancing digitalization and discovering valuable and meaningful information.Whether it is numerical or textual, it is used to collect information from these masses and use automated systems for information discovery and management. If the data found in line with the needs of the institutions are managed successfully and effectively, it is obvious that the institutions and organizations will offer great advantages and opportunities in economic terms.

Corresponding author:* **Filiz Ersoz
Industrial Engineering, Karabük University, Karabük, Turkey.

IDC (International Data Corporation) According to the data of the large data and business analytics forum, the recorded data volume increased to 16 ZB (1 ZB 1.09 Trillion Gigabyte) in 2016; It is estimated that the record will be Zettabyte 163 (1024 ZB = 1 Yottabyte (YB)) (IDC, 2018). Such a large amount of information has brought about the emergence of the field of data science and the widespread use of it and necessitated the employment of Data Scientist. The areas related to "Big Data Science" (Analytics) which are already in the Business Intelligence that supports all kinds of business decision-making processes; Data Mining, Text Mining, Machine Learning and Artificial Intelligence are the concepts.State or private institutions/organizations can make strategic and correct decisions both in numerical data and in text data. In Turkey, it has further increased the importance of information technologies, large data and data mining (business analytics) in the rapid development of projects such as national data centers or city hospitals, 4.5G transition, the continued growth of cloud computing, and the formation of future data centers. Especially in recent years, in the financial sector, energy sector, health sector, telecom sector and public institutions, investment projects for transformational information technologies for different technologies have started to increase rapidly. There is a requirement for these technologies to reach their corporate goals, to increase customer satisfaction and to configure data centers to meet their business needs.Data mining is the process of discovering rules and patterns related to each other from large heaps of data. It is not just a technique but a data approach that hosts many techniques. Converts all information from data heaps into an easy and understandable structure. It is the extraction of valuable information from the data. Briefly, data mining can be defined as the way to convert data into valuable information. Data mining today; it is defined as "Business intelligence" and "Business analytics", "Knowledge mining in databases",

"Knowledge extraction", "Data and pattern analysis" and "Data archeology" (Ersoz, 2016).To be able to perform data mining; access to data and a clear definition of the subject or research, effective access methods and algorithms, and a high-based application server are required. Data mining is closely related to the concept of business intelligence. When you look at the common functions of business intelligence technologies, analytical solution, reporting, and most importantly, data mining is well known. Business intelligence, which is also defined as corporate intelligence, is the systems that enable employees and managers to devote time to more efficient jobs in today's competitive world. In our country, this concept has begun to be newly formed and its importance has been newly discovered.

In data mining, with a review aiming to achieve special results from large and meaningless data heaps, data are passed through many stages before modeling. In the first stage, the data is cleaned before modeling. Detection of outliers and end values ensures clean and high-quality data (Cleaning) so that the data can be spoken by combining the same language. Here, the selection of relevant and important variables for the research subject and the size reduction have been conducted. Before data mining, the transformation of the available data into a format suitable for reuse (Transformation) and the model suitable for research have been established. These stages prior to the establishment of a model in data mining are considered as data preparation. The data are analyzed by applying the most appropriate techniques to the problem and research. The data mining cycle is completed by drawing the information from the databases and the results of the analysis are completed with the interpretation of the decision maker. Data Mining process; the knowledge discovery in databases can be referred to as a step of the process, which is also called the Decision Support System (KDD) (Han, 2001,Ersoz, 2016). Information discovery in data mining is given below ingeneral.

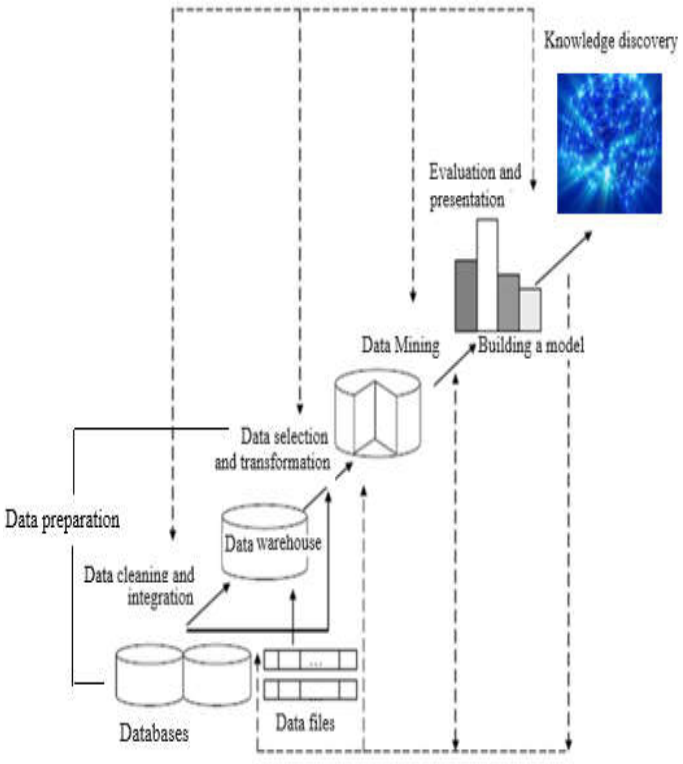

Fig. 1. Information discovery with data mining

When we look at the steps in the data mining process, it is important to understand the business problem in modeling and to understand the data. The modeling varies according to each problem or data types.

Data mining methods in general; The classifier is defined in three basic groups: Predictive, Clustering and Association rules and Sequential patterns.

Commonly known data mining stages are given below (Han, 2012, Ersoz, 2016)

- Identification of the problem or project; The purpose of research and data mining involves the determination of the planning process by evaluating the current situation.
- Understanding and preparation of data; data selection, connecting to data sources, recognizing data, understanding the quality of data. At this stage; collection of data, integration of data (Data from different databases collected in a single database), clearing data from contradictory and extreme values, transformation of data (Transformation of data into a form suitable for data mining), reduction of data (Reduction of unnecessary data likely to be introduced.
- Establishing the model
- Evaluation of the model
- The use of model results by thedecision-maker.

Data mining; it is an interdisciplinary study in which machine learning, statistics, database technology, artificial intelligence and visualization are used together. The most important of these fields is the science of statistics. For these reasons, the value of machine learning techniques has increased in parallel with the data analysis and modeling studies. In text mining, however statistics are related to data mining, machine learning, management science, artificial intelligence, computer science and other disciplines. Statistics science and machine learning can be considered as basic elements in data and text mining. Although data analysis has been carried out successfully with statistical methods in data mining, the statistical methods in text mining are insufficient in the analysis of non-structural big data.

The obtained data can also contain text, images, sound, etc. analysis of non-structural data in the form of text mining has been performed. Considering the literature on text mining, it is observed that instead of the text mining concept, text data mining (text data mining), data discovery in databases, text analysis and text data mining are used (Yıldız *et al.,* 2018).

Text mining studies is a data mining study that considers the text as a data source. Another definition aims to obtain structured data (Seker, 2015). The discovery of meaningful and important information within textual data can be defined as text mining. Text mining is used to extract facts and relationships in a structured form to support operational and strategic decision-making processes in business intelligence by considering the relevance of texts to private databases. In short, text mining is an approach to identify and extract information in unstructured text. Text mining can be defined as the information-intensive processes that provide the interaction of users with the data collected over time with the special tools it uses (Feldman, 2007). Miner *et al.* (2012) defined text mining as transforming texts into numbers technologies. Also, the importance of text mining is increasing with the increasing

use of social media tools, such as Facebook, Instagram, Twitter etc. (Oguzlar, 2011) defined text mining as the process of extracting confidential information from the data in the text and formatting the irregular data with no specific format. Hearst has described text mining as a subfield of data mining (Hearst, 1999). Yıldız (2018) showed that the use of text mining methods in the accounting field accelerated and facilitated the processes.With text mining technology, all texts can be scanned, and similar studies can be evaluated together, and comparisons can be made easily. It is integrated into a much broader analytical market, business intelligence solutions, and ready to enable semantic search. Google platform and Amazon are one of the most important examples of text mining.

Text mining can be defined as text data mining. Text mining is a process required to obtain meaningful structural information from sources that are not structured. It requires advanced linguistic and statistical methods that can analyze unstructured text formats and techniques that combine each document with actionable metadata. Once the content is released, it can be directly summarized, visualized and classified by link mapping, and making it easier to search.Text mining can be defined as transforming texts into a single form, pre-processing texts and transforming data into a structural form.Text mining usually consists of four stages (Faro *et al.,* 2011) (Zhu *et al.,* 2013):

- Information retrieval (IR), A series of text materials are collected for a specific topic and the collected texts are textual features.
- Information extraction (IE). Uncovering meaningful relationships between texts with similar features.
- Knowledge discovery (KD). The resulting meaningful relationships are the definition of particular patterns and trends.

Application Areas: The use of data mining is quite wide. Since meaningless data is rendered meaningful by processing information, it is used in many fields. Data mining; in addition to banking, biology, finance, marketing, insurance and medicine, it is also used in text and web mining. Text mining practices are often used in text analysis studies, in the classification of texts and the problems of word analysis. Web text mining, on the other hand, examines the content of the websites that constitute the basic structure of the Internet. Headings on sites, words on the pages, menus, subject structure, images and so on. The content information is examined and the relationships between the sites are determined. According to this finding, websites can be divided into classes and categories (Dolgun *et al.,* 2010).

Considering the studies conducted with data mining and text mining methods, especially in the field of text mining, many studies about the health sector can be encountered. When the literature is examined, it is observed that text mining is more widely used in banking and finance (Yıldız, 2018) and electronic commerce (Altan, 2018) sectors. Karami *et al.* (2018) and Solloum *et al.* (2017), while analyzing text mining methods by making use of the general perspective and feedback of society, worked on the impact of Altan (2018) news on the corporate image and the hospital sample, Ye *et al.* (2016), McTaggart *et al.* (2018) and Delespierre(2017) analyzed the data by using the large data defined in the system. In addition, Lammey (2015) has revealed the current status of launching the service among CrossRef member publishers by means of text mining. According to the results of all these studies, data were classified by using various text mining techniques and their relationship was determined. In addition, a new system was developed in line with the analysis results or additions were made to the system and the existing system was developed.

When we look at the data mining literature, it is observed that it is more widely used in the finance and banking sector (BHambri, 2011) (Bhasin, 2006). In addition, there are studies in the areas of health (Lopez-Pineda *et al.,* 2018) (Shameer *et al.,* 2018) (Yaday *et al.,* 2018) management and informatics to reveal customer satisfaction and profile (Minghetti, 2003). According to the credit card expenditures, the data mining application was determined by Parman. Parman(2003) conducted customer segmentation and risk valuation analyzes on the data of 1704 credit card customers belonging to a single branch of a medium-sized bank from private banks. It has been tried to determine the areas where credit card holders differ from each other according to their spending characteristics (Parman, 2003). In their study, Hsia *et al.* (2008) used a data mining technique to analyze course preferences and course completion rates at a university in Taiwan. The aim of the studies is to use the data mining technique to determine the course preferences of the students and the course preferences of the students who are continuing their education. Decision trees were used to find the students' course preferences, to determine the correlation between the link analysis course category and the participant profession, and the decision forest was used to find the participants the possibility of completing their preferred course. Chaid was used as a decision tree. As a result, high estimation success was achieved.Farquad, Ravi and Raju propose a hybrid method that uses a combination of Support Vector Machine and Naive Bayes methods to estimate loss rates of bank credit card customer. A data set consisting of 6.89% of lost customers has been used in the study. 68.52% of the correct classification rules have been obtained (Farquad *et al.,* 2009).

Comparison of Data Mining and Text Mining: The main difference in data mining and text mining is the homogeneous and heterogeneity of the structures used. It analyzes the information in a homogeneous structure format in data mining. It extracts, converts and loads data into a data warehouse. Business analysts use data mining software applications to analyze data and present data in easily understandable forms such as tables or charts. In text mining, it generates valuable information by using multilingual texts and abbreviations such as heterogeneous structure format (text documents, e-mails, social media writings, verbatim texts, etc.).

Data mining analyzes on digital data stacks. Information is easily accessed and homogeneous. Rapid solutions are obtained with the algorithms used. In text mining, the complexity of the processed data is quite high, and the solution lasts longer. Text mining needs several intermediate linguistic analysis stages before it can enrich the content. After linguistic analysis, the metadata association steps address the configuration of unstructured content and domain-specific applications.

Data mining uses only structural data as opposed to text mining, and hence meaningful relationships and trends can be found in text data using data mining techniques and

algorithms. With text mining, non-structural data can be transformed to be structured and data mining can be done.

In text mining, as in data mining, the investigator explores real usable information from data sources through investigations and descriptions of the area of interest. However, text mining does not include unstructured textual data, but there are no database records formulated in data sources. Besides, text mining has a lot in common with data mining. For example, most systems are based on preliminary processes, pattern discovery algorithms, and presentation layer elements such as visualization tools (Feldman *et al.,* 2007). The summary of data mining and text mining is given in Table 1 in below.

Table 1.Comparison of data mining and text mining

Data Mining	Text Mining
It targets the process of obtaining high-quality information from large data heaps in data mining.	In text mining, it targets the process of obtaining high-quality information from a large mass of textual data.
Data mining illustrates the meaningful relationships in data with modeling approaches.	Text processed by text mining; language, religion, race, as well as ethnic factors such as depending on the method applied according to the method and linguistic approaches may vary.
Data Mining is a necessary process to convert meaningful information into structured data.	Text mining is a necessary process to convert an unstructured text document into valuable structured information.
Data mining is related to statistics, machine learning, optimization, data warehouse, expert systems, pattern recognition, artificial intelligence and algorithm concepts in computer science.	Text mining is associated with Statistics, Machine Learning, Management Science, Artificial Intelligence, Computer Science, Natural Language Processing (NLP) and other disciplines.
In data mining, business models are generated using numerical data and using data mining models.	Text mining is the use of text methods to discover a lexical, syntactic and semantic feature in the text.
Data mining is the process of discovery of information from structural data that is homogeneous and easy to access.	Text mining is a process of text discovery from heterogeneous data and non-structural data.
Data mining enables the discovery of information from large data with algorithms for modern machine learning and data mining models (classifiers, clusters, and association rules).	In text mining, valuable information is discovered through classification, clustering, association, information extraction and summarization. It does this by using machine learning, artificial intelligence and algorithms.
It enables the determination and analysis of the appropriate method for collecting and using the data.	Multiple steps and methods may need to be applied to reach meaningful data.
Data mining works better with large amounts of data.	Text data size in text mining can vary depending on the area studied. As the area of interest grows, the results may be complicated. The best result arises from the large areas of text where the relevant fields are small.

Conclusion

With the development of technology, the importance of collecting and evaluating the data has increased and this increase has caused the accumulation of big data. For this reason, it has become increasingly important to distinguish the data stacks stored in databases and data warehouses according to the nature of the data and to establish meaningful relationships between the data. In order to increase the use of information technologies and to meet the increasing need, large data phenomena and data science have begun to develop. The data mining discipline has emerged in order to analyze large amounts of numerical data stored as data science and large data building systems develop. A text mining discipline has also emerged for the processing of texts in unstructured documents.It is a known fact that the efficient management of data and text mining will bring financial and managerial benefits to institutions and organizations. In recent years, especially in the finance sector, energy sector, health sector, energy sector, telecom sector and public institutions, investment projects for information technologies have started to increase rapidly. With these techniques, which can be used in every field, the institutions can classify their data according to the determined criteria, analyze the determined elements together, establish infrastructures that will facilitate the operation of the system, and allow the data to be clustered. In this way, it can be ensured that institutions and organizations receive their strategic decisions effectively and quickly.There are also organizations such as CrossRef which are established to provide the mentioned benefits. These organizations are aimed at gathering the work done under the discipline of data mining and text mining, to facilitate the easy access to the necessary data and to identify the relevant methods (Lammey, 2015).In this respect, it is important to understand the data mining that converts data into qualified information and to determine its usability. The use of data mining techniques depends on the availability of data, a clear definition of the subject, the definition of access methods and algorithms, and the provision of technological competences. If these conditions are met, the appropriate modeling method is chosen according to the problem or data type. These methods are generally grouped as predictive, clustering, association rules, and sequential patterns.Text mining, which is one of the sub-branches of data mining, can reach the desired self-knowledge in high-capacity texts. Most of the data mining applications are carried out by text mining since the data is mostly textual. According to Miner *et al.,* text mining is very general in terms of its applications and is very diverse in terms of its objectives. Compared to other well-established methods, text mining is thought to be a relatively new and non-standardized analytical method for discovery of information (Miner *et al.,* 2012).

The main differences between data mining and text mining are the structure of data; According to the text mining of data mining method of data complexes because of the complexity of the solution produces faster; While data mining techniques are used in data mining, it is revealed that significant relations and trends in data mining are used.Data Mining is based on Knowledge Discovery (KDD) in its databases and it is the process of finding valuable information from the information contained in the databases. Data mining is often used interchangeably with KDD.As a result, text mining, such as data mining, is now becoming increasingly widespread. Especially through the use of text mining methods that fill the missing aspect of data mining regarding the processing of linguistic expressions, it becomes easier to establish effective and fast systems and to analyze the data which is more complex than data mining.Data science includes many disciplines, many of which are data mining and text data mining. These include large data analytics, estimated modeling, data visualization, natural language processing (NLP), statistics, mathematics and artificial intelligence. Data science discovers knowledge through machine learning

through data mining and modeling techniques with text data mining.The data is a valuable product for information. Data mining and text mining combined with large data analytics helps to solve problems or make better decisions and contribute to reducing time and effort. With the increase in the use of such methods in institutions and organizations, rapid and effective improvements in many sectors, the accurate analysis of the system and meaningful data can be provided.

REFERENCES

Altan S. 2018. "Analysis of the Impact of News Stories on the Organizational Image and a Text Mining Analysis on the News Stories about Hospitals in Turkey", Journal of Communication Theory and Research, vol 46: p. 222-240.

BHambri V. 2011. "Application of Data Mining in Banking Sector", IJCST 2(2), 199-202.

Bhasin M.L. 2006. "Data Mining: A Competitive Tool in the Banking and Retail Industries", The Chartered Account, 588-594.

Delespierre T., Denormandie P., Barhen A., Josseran L. 2017. "Empirical Advances with Text Mining of Electronic Health Records", BMC Medical Informatics and Decision Making, 17 (1), 127.

Dolgun, O.M., Ozdemir, T.G. and Oguz D. 2009. Analysis of Non-Structural Data in Data Mining: Text and Web Mining. *The Journalists' Journal* 2 (2), pp.48-58.

Ersoz F. 2016. Data Mining Techniques and Applications, 72 Digital Printing, Ankara.

Faro A., Giordano D., Spampinato C. 2011. Combining Literature Text Mining with Microarray Data: Advances for System Biology Modeling. *Brief Bioinform*. 2011;13(1):61–82.

Farquad M. A. H., Ravi V., Raju S.B. 2009. Data Mining Using Rules Extracted from SVM: An Application to Churn Prediction in Bank Credit Cards, Springer-Verlag, Berlin Heidelberg, 390 – 397.

Feldman R., Sanger J. 2007. "The Text Mining Handbook – Advanced Approaches in Analyzing Unstructured Data", Cambridge University Press.

Gursakal, N. 2014. Big Data, Dora Publications, Bursa.

Hearst MA. 1999. Untangling Text Data Mining. In: Proceedings of 37th Annual Meeting of the Association for Computational Linguistics;3–10.

Han, Kamber, Han, J. ve Kamber M., 2001. "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers.

Hsia T. C., A. Shie J., Chen L. C. 2008. Course Planning of Extension Education to Meet Market Demand by Using Data Mining Techniques – An Example ofChinkuo Technology University in Taiwan, Expert Systems with Applications, 34: 596–602.

Karami A., Dahl A., Turner-McGrievry G., Kharrazi H, Shaw G. 2018. "Characterizing Diabetes, Diet, Exercise, and Obesity Comments on Twitter", *International Journal of Information Management*, Vol 38:1, p1-6.

Lammey R. 2015. "CrossRef Text and Data Mining Services", Insights, 28(2), July.

López-Pineda A., Rodríguez-Moran M. F., Álvarez-Aguilar C, Fuentes Valle S. M., Acosta-Rosales R., Bhatt A.S., Sheth S.N. and Bustamante C. D. 2018. "Data Mining of Digitized Health Records in a Resource-Constrained Setting Reveals That Timely Immunophenotyping is Associated with Improved Breast Cancer Outcomes", BMC Cancer, Vol 18:933.

Mc Taggart S., Nangle C., Caldwell J., Alvarez-Madrazo S., Colhoun H., Bennie M. 2018. "Use of Text Mining Methods to Improve Efficiency in the Calculation of Drug Exposure to Support Pharmacoepidemiology Studies", International Journal of Epidemiology 47 (2), 617-624.

Miner, G., D. Delen, A. Fast, T. Hill, J. Elder ve B. Nisbet. 2012. Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications, Elsevier, USA

Minghetti V. 2003. "Building Customer Value in the Hospitality Industry: Towards the Definition of Customer-Centric Information System" Information Technology & Tourism 6 (2), 141-152.

Oguzlar, A. 2011. Basic Text Mining, Dora Publications, Bursa.

Parman D. 2003. Data Mining Method for Customer Relationship Management in Banking Sector: Implementation in a Private Bank. PhD Thesis, Marmara University, Institute of Banking and Insurance, Istanbul.

Shameer K., Perez-Rodriguez M. M., Bachar R., Li L., Johnson A., Johnson K. W., Glucksberg B. S., Smith M.R., Redhead B., Scarpa Jhf, Kebakaran J., Kovatch P., Lim S., Goodman W., Reich D.L., Kasarskis A., Tatonetti N.P. and Dudley J.T. 2018. "Pharmacological Risk Factors Associated with Hospital Readmission Rates in a Psychiatric Cohort Identified Using Prescriptome Data Mining", BMC Medical Informatics and Decision Making 18(Suppl 3):79.

Solloum S.A., Al-Emran M., Monem A. and Shaalan K. 2017. "A Survey of Text Mining in Social Media: Facebook and Twitter Perspectives", Adv. Sci. Technol. Eng. Syst. J 2(1), 127-133.

Yadav P., Steinbach M., Kumar V. And Simon G. 2018. "Mining Electronic Health Records (EHRs): A Survey", ACM Computing Surveys, Vol. 50, No. 6, Article 85. Publication date: January.

Ye Z., Tafti A.P., He K. T., Wang K., He M. M., "Spartext: Biomedical Text Mining on Big Data Framework", Journal of Plos, 11(9), 2016.

Yıldız D., Agdeniz S. 2018. "Text Mining as an Analyzing Method in Accounting", Journal of The World of Accounting Science, 20(2); 286-315.

Zhu F., Patumcharoenpol P., Zhang C., Yang Y., Chan J., Meechai A., Vongsangnak W., Shen B. 2013. Biomedical Text Mining and Its Applications in Cancer Research. *J Biomed Inform.*,46(2):200–11.

IDC, https://www.idc.com

****