DATA SCIENCE JOURNAL

Zhang, H et al 2015 A Correlation Analysis Model for Multidisciplinary Data in Disaster Research. *Data Science Journal,* 14: 19, pp. 1–12, DOI: http://dx.doi.org/10.5334/dsj-2015-019

PROCEEDINGS PAPER

A Correlation Analysis Model for Multidisciplinary Data in Disaster Research

Hongyue Zhang[1], Xiuling Qing[2], Mingrui Huang[1] and Guoqing Li[1]

[1] Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, No.9 Dengzhuang South Road, Haidian District, Beijing, China,100094
ligq@radi.ac.cn

[2] National Science Library, Chinese Academy of Sciences, 33 Beisihuan Xilu, Zhongguancun, Beijing, China, 100190

Data play an important role in disaster mitigation applications, and the integrated employment of multidisciplinary data promotes the development of disaster science. Therefore it is very useful to identify the multidisciplinary data usage in the research of disaster events. In order to discover the correlation between multidisciplinary data and disaster research, three earthquake events, the Tangshan earthquake, the Wenchuan earthquake, and the Haidi earthquake were selected as typical study cases for this paper. A knowledge model for literature data mining was applied to analyze the correlation between earthquake events and multidisciplinary data types. The results indicate that high-cited papers show different data usage trends when compared with whole-set papers and also that data usage for the three earthquake events varies. According to analysis results, the factors that influence multidisciplinary data usage include the characteristics of spatial and temporal elements as well as differing interests of the data users.

Keywords: Multidisciplinary data; Literature analysis; Knowledge model; Earthquake events; Spatio-temporal elements

1 Introduction

The development of data acquisition technologies has resulted in the accumulation of massive amounts of data. How to discover knowledge from this huge amount of data is an important question. Theoretical science, experimental science, and scientific simulation are the traditional knowledge discovery methods. However with the emergence of such a large volume of data, data intensive computing is becoming the new approach for knowledge discovery (Tansley et al., 2009).

As data are the most valuable resource in scientific research and new knowledge discovery, ways to discover potential knowledge from the masses of data are attracting more and more attention (Adriaans & Zantinge, 1996). For example, in disaster mitigation research, data play a key role; the integrated use of multidisciplinary data significantly promotes the development of disaster science. According to their various research objectives, researchers select very different types of data for their work. The type of data that is suitable for researching certain disaster events has become an important question for researchers.

In order to discover the correlation between various kinds of data and specific disaster events, we propose a literature-based analysis model for domain knowledge discovery. Although several models have been previously developed and have received a good deal of attention, none of them is suitable for this study. The model used in the paper incorporates **prior knowledge** and **statistical methods**, and the overall framework consists of several steps executed in a sequence that can be used for literature data mining.

This paper is organized as follows. Background and related work are introduced in Section 2. Section 3 discusses the objectives and scientific questions of the study. The principle and research model are provided in Section 4. Section 5 gives the framework of the process model, and the showcase and conclusions are analyzed in Sections 6 and 7, respectively.

2 Background and Related Work

The research approaches and models relevant to our study include literature-based analysis, knowledge discovery, and data mining models. The content of our research is correlation analysis of data types and disaster events. The approach model employed in our study is the specific application of literature-based knowledge discovery and data mining. Many researchers have been working with similar approaches and data mining techniques to develop specific applications.

In knowledge discovery and data mining research, two major types of models are proposed in the book *Advances in Knowledge Discovery and Data Mining* (Cios & Kurgan, 2005). Fayyad and others did pioneering studies in KDDM (Fayyad et al., 1996; Fayyad et al., 1996c; Fayyad et al., 1996d; Fayyad et al., 1996e; Fayyad et al., 1996f). The human-centric model emphasizes the interactive involvement of a data analyst during the process (Brachman & Anand, 1996), and the data-centric model emphasizes the iterative and interactive nature of the data analysis tasks (Fayyad et al., 1996a). The analysis model proposed in our study is a data-centric knowledge discovery model. Data-centric models are structured as sequences of steps that focus on performing manipulation and analysis of data and information surrounding the data, such as domain knowledge and extracted results (Kurgan et al., 2006).

The models are usually defined as a fixed sequence of predefined steps (Kurgan et al., 2006). A data-centric model has been used in several domains, especially in business, banking, medical treatment (Cios & William Moore, 2002; Pratt & Yetisgen-Yildiz, 2003), biology, and so on. Over the last ten years, research efforts have been focused on proposing new models rather than improving the design of a single model or proposing a generic unifying model. Most models are not tied specifically to academic or industrial needs (Kurgan et al., 2006).

KDDM (knowledge discovery via data mining) models are employed to discover knowledge in literature, and researchers have achieved great progress in many fields, especially in biological knowledge extraction. Literature-based data mining has progressed from simple recognition of terms to extraction of interaction relationships from complex sentences (Hirschman et al., 2002). Hristovski et al. (2005) identified disease candidate genes by using a literature-based knowledge discovery approach. Meliha and Wanda (2006) incorporated knowledge-based methodologies with a statistical method to mine the biomedical literature for new, potentially causal connections between biomedical terms. Frijters et al. (2010) employed literature mining to discover hidden connections between drugs, genes, and diseases.

However, there is no such research on correlation analysis between data types and disaster exists, and most disaster researchers choose data based on their domain knowledge and experience (Yang, 2013). According to our investigation, the data used for disaster research usually depend on the research model. For example in the natural disaster evaluation model, all kinds of data, including geophysical data, socio-economic data, historical hazards data, disaster reports, etc., are needed for the evaluation. Some researchers also choose data according to disaster indicators, and in order to calculate the disaster indictors, various data need to be integrated. The most important element to researchers when choosing data is the application they are being used for, such as disaster loss prediction or flood submerged area simulation. Based on international disaster databases such as Sigma, EM-Dat, and NatCatSERVICE, Wirtz et al. (2014) concluded that a disaster event database should include: the event identification number, event categorization, geographical information, data and duration, victims, damage and affected economic sectors, losses, and scientific data as well as other information. All the important information involved is based on the foundation of a unified standard for global disaster database construction.

The question becomes how to use the knowledge discovery approach to analyze the correlation between multidisciplinary data and disaster events. Perhaps there are several methods that can be employed to answer this question, but currently relevant work in the field of disaster study is rare. Therefore, a literature-based knowledge discovery approach is proposed in our study to solve this question.

We use an interdisciplinary literature-based data mining approach for the correlation analysis of multidisciplinary data for certain disaster events. This approach is quite different from the above application areas. The research model used in our study follows the same sequence of steps as the above data-centric KDDM mode, and the specific research objectives are also integrated to build our application-oriented research model. The model used in our study follows the same sequence of steps and uses similar steps to the KDDM process model.

3 Scientific Questions

In recent years, with the development of computer science and network technology as well as the data accumulation from disasters, the sharing and application of disaster relevant data has become more and more mature, but there is still no answer to the question of which data are useful for which disaster events.

The research objects in our study include disaster events and multidisciplinary data; therefore, before implementing the study, we must investigate and understand the disaster events and the multidisciplinary data that they generate.

Because there are different types of disasters as well as countless disaster events, we need to determine the specific disaster events we wish to study. The international classification of disasters and the terminology of perils divide disasters into 6 groups: biological, geophysical, meteorological, hydrological, climatological, and extra-terrestrial. Each group covers a different type of disaster that includes many subtypes (Below, Wirtz, & GUHA-SAPIR, 2009). The unified disaster classification standard has been established by the integrated research of CRED, Munich RE, Swiss Re, Asian Disaster Reduction Center (ADRC), and the United Nations Development Program (UNDP), which is the foundation for interchange and communication among international databases.

Unfortunately, disaster data are multi-source, stored in different formats, and exist in various forms. Until now, no clear definition of multidisciplinary data has been given. The term multidisciplinary data refers to the combination of various data from the different disciplines involved in interdisciplinary research. In our research, multidisciplinary data refer to the multi-source data used in disaster research. The data used in disaster research are, for the most part, of two types - socio-economic data and natural science data. Socio-economic data relate to people, society, and economics while natural science data relate to the disaster events, such as data obtained by observation, measuring, remote sensing, etc.

Because no studies on which data are useful for which disaster events exist, there are no mature approaches to solving this scientific question. As disaster study is an interdisciplinary research, researchers choose different data to study specific disaster events, and the data usage of researchers is reflected in their literature. In this case, literature analysis can be employed as one approach to address the scientific question proposed. In order to discover the correlation between multidisciplinary data and certain disaster events, we gather prior knowledge from domain experts for our study. The aim of this study is to mine the knowledge in the literature for further analysis. The analysis results and conclusions can be used as a knowledge base for further research.

4 Research Model

So that readers will understand the model used in our study, we explain below the general scientific research process in detail.

The scientific research process consists of five phases:

First phase: Scientific question abstraction from phenomenon. Phenomenon is defined in the *New Oxford American Dictionary (*http://en.wikipedia.org/wiki/New_Oxford_American_Dictionary*)* as any observable occurrence. A disaster is a phenomenon that can cause a sequence of serious effects, such as casualties, economic loss, etc. In order to understand this phenomenon, researchers should decide what to observe, how to measure the observations, and how to define the research problem. With the solution of these questions, the characteristics of the observable phenomenon can be recorded and formed. As for specific disaster events, once researchers have understood the disaster event mechanisms and characteristics, the scientific question, which is the output of the first phase, can be proposed.

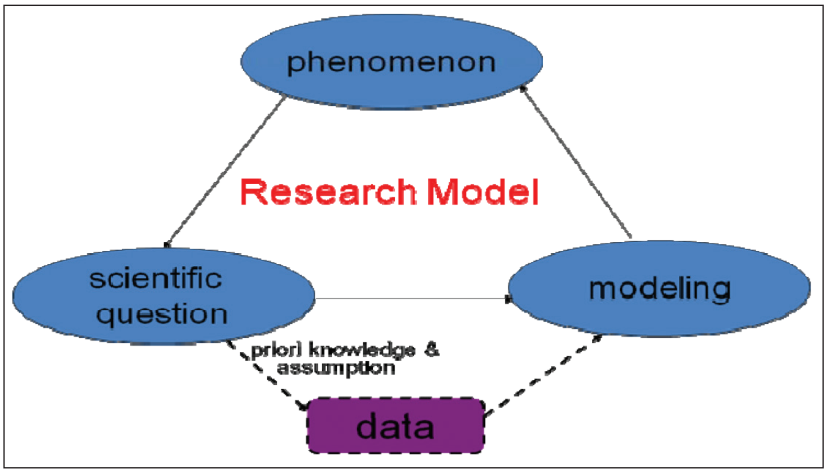

Figure 1: Scientific research process.

Second phase: Research model construction. With the formation of scientific questions, researchers must consider the solutions. In this phase, researchers decide which scientific question is the most urgent and can be modeled for a solution, what factors are involved, what factors can be observed, and what factors can be omitted. What answers do we expect? We need an hypothesis and a theory to construct the concept models. Then the logical relationships among the concepts can be extracted and a logical model can be built.

Third phase: Model simplification and rectification. Because a disaster is the complex effect of multiple factors, many factors are difficult to observe and predict. In order to make the model operable, prior knowledge and assumption are employed.

Fourth phase: Data input and solution. Observable factors can be recorded with data forms. The model's input is the observable factors while the output is the data information of the factors that we are interested in.

Fifth phase: Statistical analysis of output data. Researchers analyze the result data and discuss the information reflected by this data. Then they propose the conclusion and solution method.

The knowledge model proposed in this paper is the specific application of the general research model. In the general model, domain knowledge is formed first, and then the research model is constructed according to this domain knowledge. The knowledge model in our research is also based on the understanding of the scientific question. In our research model, the knowledge and technical levels are integrated to solve the proposed question, and the domain knowledge is acquired by a relevant technique that is also a new technique for new knowledge discovery. Knowledge discovery is an iterative process, and domain knowledge is the basis for new knowledge mining.

Domain knowledge is very useful for the preparation and recovery process in disaster management. Thus an important and challenging research direction is how to utilize the domain knowledge to guide disaster mitigation and make decisions. In the past, there were no data acquisition approaches; the only way to gather information about natural phenomena was through the naked eye. Thus, experts provided various hypothetical theories according to their daily observations.

With the development of various data acquistion approaches and observation tools, researchers began to study the mechanisms of natural disasters and perform experiments with the accumulated data. Therefore, new knowledge and discovery continued to emerge. In order to spread the new knowledge and scientific achievements, domain experts recorded knowledge in different forms. Literature and media became the knowledge carriers, and the media began to play an important role in representing domain research directions and trends. At present, literature has become the most important way to learn of research hotspots and ideas related to specific research topics.

In recent years with the development of data acquisition technologies and the accumulation of published papers, there are more and more data sources that can be used to discover domain knowledge. This has changed the traditional knowledge discovery process that depended for the most part on the human brain and naked eye observation. Thus, knowledge discovery and data mining have become important measures with which to obtain potential knowledge. Some researchers have tried to incorporate data mining techiniques and statistical approaches to discover knowledge from massive literature sources.

The knowledge model proposed in our research is the specific application of knowledge discovery and data mining in disaster literature study. The potential knowledge we expected to obtain was the correlation

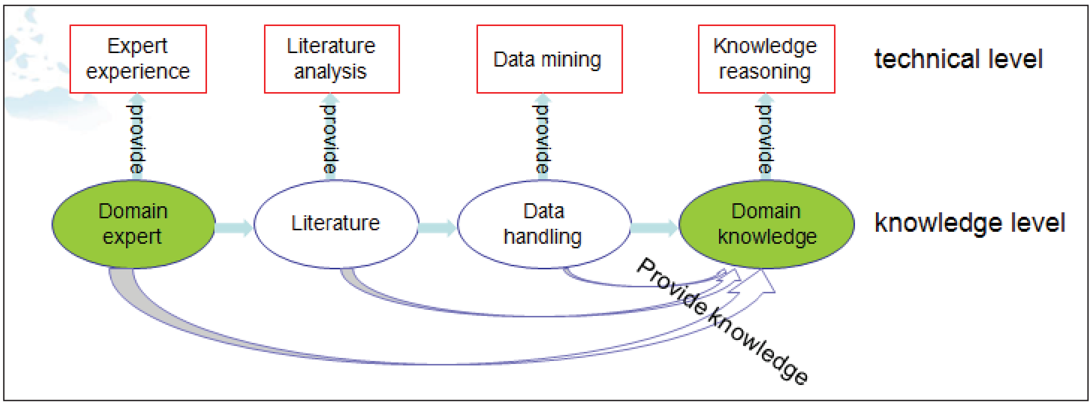

Figure 2: Two level knowledge model employed in our study.

between multidisciplinary data and certain disaster events. The model contains two levels: the lower level is knowledge about disaster taxonomy and the multidisciplinary data, and the upper level is the technical approach used to study the multidisciplinary data for a specific earthquake event.

The knowledge level is the knowledge base for implementation of the technical level. Domain experts as well as literature are the data sources used for construction of the domain knowledge. While domain experts are also the sources of the literature, data handling is the means of transferring the data source to the domain knowledge. As we can see, domain experts are the key element in the knowledge level .

The technical level provides the technical means to realize the knowledge level. Because this is a multidisciplinary research topic, there is no defined method for achieving the research aim. In order to understand how to implement the ideas obtained from the knowledge level, expert experience should first be obtained by communicating with or interviewing domain experts. Literature analysis and data mining are employed in the construction of the knowledge base. Knowledge reasoning is the most important part of the analysis and explanation. As we can see, the domain experience is the main reference for each technical process in the technical level.

The knowledge and technical levels complement each other: the knowledge level provides the knowledge base for the technical level, and the technical level implements the ideas from the knowledge level. Domain experience from domain experts is the most important source for both levels.

5 Technical Framework

The technical framework employed in this paper is presented in **Figure 3**. The simple technical framework includes three parts: input data, the knowledge model, and output of new data. We should pay attention to three questions: what data to input, what is the core of the knowledge model, and how to interpret the output data.

The general technical framework can be divided into the following technical steps, see **Table 1**.

1. The research objective is to identify the correlation between multidisciplinary data and certain disaster events as there are so many kinds of disasters and each type of disaster covers several historical disaster events. We chose earthquakes to study because they have the most destructive impact on man. The Tangshan, Wenchuan, and Haidi earthquakes are the three most famous earthquakes in recent years and have generated a great deal of relevant research. Our study compares these three disaster events.

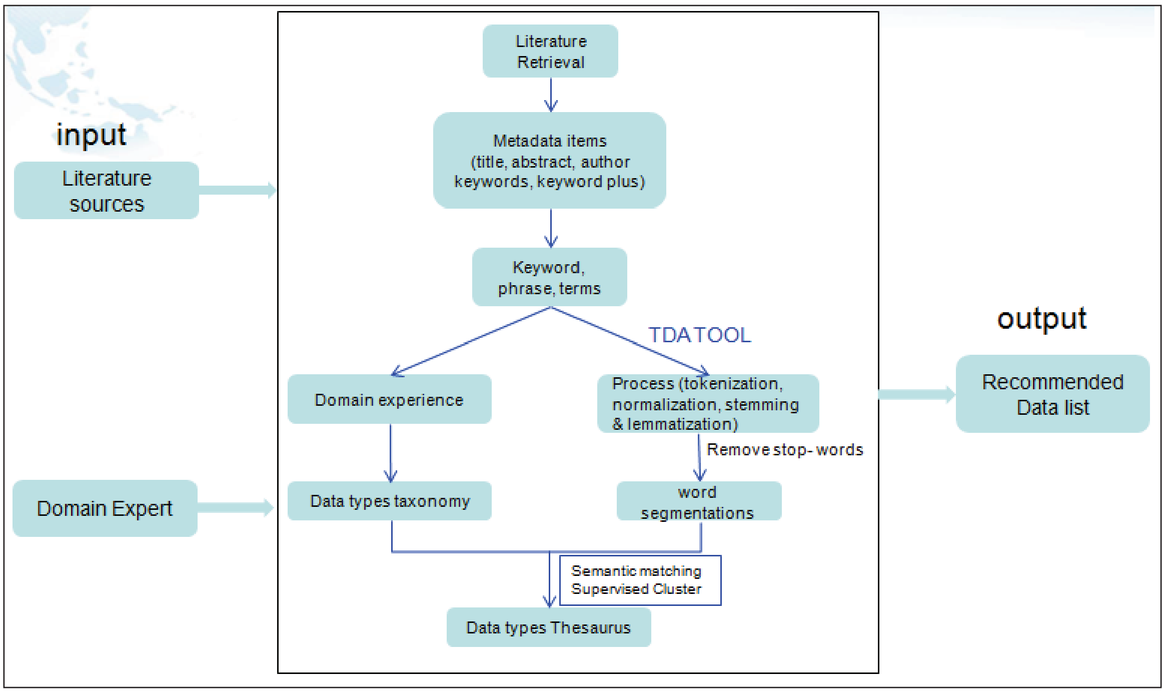

Figure 3: Technical framework of the literature-based disaster data discovery process.

Steps	Earthquake as an example
Identify scientific question	Multidisciplinary data for certain disaster events
Specify the certain disaster events	Take three earthquake events as examples.
Obtain the literature data sources	Determine the SCI articles as data sources and search keywords. Separate the obtained literature into two parts: high-cited set and whole set.
Segment words of articles	Thomson Data Analyzer is employed to segment the words and phrases.
Generalize thesaurus lists	Incorporate expert knowledge to define data type thesaurus from the segmentation words and phrases.
Cluster words and classification	Thesaurus list is classified into several groups by semantic mapping and domain knowledge.
Calculate frequency value	Thesaurus occurrence frequency is counted by statistical approach.
Analyze the statistic results	Multidisciplinary data list for earthquake events is sorted in sequence according to the frequency table.
Compare two views of disaster events	The differences of multidisciplinary data on three earthquake events are analyzed in global and local views, possible reasons are presented.
Compare two views of multidisciplinary data	The differences of multidisciplinary data on certain earthquake events are analyzed in global and local views, possible reasons are also presented.
Conclude	Conclude the analysis result and give some suggestions.

Table 1: Technical steps in the research model.

2. According to the selected disaster events, the literature retrival formulas are defined. To make the retrieved articles necessary and sufficient, synonyms should be taken into consideration, and the required items should cover certain disaster events.
3. The literature processing tool "Thomson data analyzer (TDA)" is employed to segment the metadata items: title, abstract, and author keywords. The words' segmentation and phrases should be further processed, including stop words and non-relevant word removal, word simplification, etc.
4. Thesaurus classification. A semantic similarity calculation approach is used to cluster the synonyms; then domain knowledge is employed to identify the words that have been misclassified. A data type glossary is obtained by the integrated approach of semantic mapping and domain knowledge interpretation.
5. The simple knowledge discovery model is run: the literature obtained from the first step is the input source; the data type glossary obtained in the fourth step is the knowledge base; the TDA tool is used as the frequency counter; and the word occurrence frequency of each data type is acquired. A domain expert should participate in the model execution process.
6. Statistical result analysis. The occurrence frequency of each data type phrase reflects the data usage for a specific disaster event. According to the occurrence frequency, a multidisciplinary data list for each earthquake event can be obtained. Based on the analysis results, we should think about the following questions: What information can we deduced from the results and what has affected the analysis results.

6 Earthquake Showcase

6.1 Data Sources

In our research, the SCI articles in the ISI Web of Science-Science Citation Index Expanded database are used as data sources. Three earthquake events, the Tangshan, the Wenchuan, and the Haidi earthquakes, are used as case studies. Research papers on earthquakes in the database are obtained by a literature retrieval approach. Taking the Haidi earthquake as an example, the retrieval formula is "Theme= (haiti or hayti) AND Theme=(earthquake or seism* or earthshock or earthdin or temblor)". The time span selected is 2003–2013. It is worth noting that the number of retrieved papers may vary because of time or library elements.

In order to reflect the differences between high-cited papers and whole articles, the papers are separated into two sets, WS (whole set) and HCS (high-cited set). WS refers to all the research papers we retrieved, and HCS refers to the 100 high-cited papers we selected as a comparison data set.

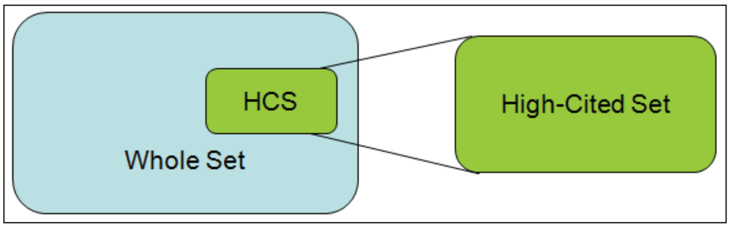

Figure 4: Relationship between high-cited set and whole set.

Because the citation index reflects the impact factor of the research papers, the relevant user community of research papers includes the authors who write the papers, the editors who review the papers, and the readers who read the papers. The published papers reflect the authors' research interests and the editors' recommendations; the high-cited papers are a group of the published papers that are most popular with readers. In this sense, in our study, the WS represents the authors' and editors' views, and the HCS represents the readers' views.

After literature retrieval, the data sources we obtained are as follows:

Tangshan earthquake: WS of 226 articles and no HCS control group.
Wenchuan earthquake: WS of 296 articles and HCS of 100 articles.
Haidi earthquake: WS of 1124 articles and HCS of 100 articles.

6.2 Knowledge Abstraction

The data classification standard is obtained from the domain knowledge of disaster experts through questionnaires and seminar discussions. According to the domain knowledge we gathered from these experts, multidisciplinary data for earthquake events include the following types.

- Geological data
- Geophysical data
- Earth observational data
- Ground observational data
- Clinical data
- Space physical data
- Basic geographical data
- Statistical data

The literature sources are processed by the TDA tool and the word list is obtained. The domain knowledge is employed to extract concepts that belong to multidisciplinary data. In this step, semantic cluster and interpretation are integrated to classify the thesaurus into specified data types.

The process of obtaining a thesaurus of disaster data types is dynamic. As new knowledge is added, the classification of data types and thesaurus may be changed. Classification results are listed in **Table 2**.

Data Types	Thesaurus
Geological data	geology, geotechnical, geological investigations, earthquake engineering, topography, topographic, geological, lithology, stratigraphic mapping, fault, thrust belt, slip zones…
Geophysical data	geophysical, geoelectric, geodynamic, isostasy, kinematic, gravitational, magnetosphere, magnetotellurics, oscillations, wave, electromagnetic, geochemical, ground-motion…
Ground observational data	atmosphere, in situ, field investigation, geodetic, geomorphologic, geodesy…

Table 2: Part of the data type thesaurus.

The occurrence frequency of each data type group is counted using a statistical approach. In this step, a knowledge model is executed: literature sources as input, the multidisciplinary data thesaurus as knowledge base, and the occurrence frequency of each data type in the thesaurus as output. Based on the results obtained, an analysis of the different aspects can be displayed.

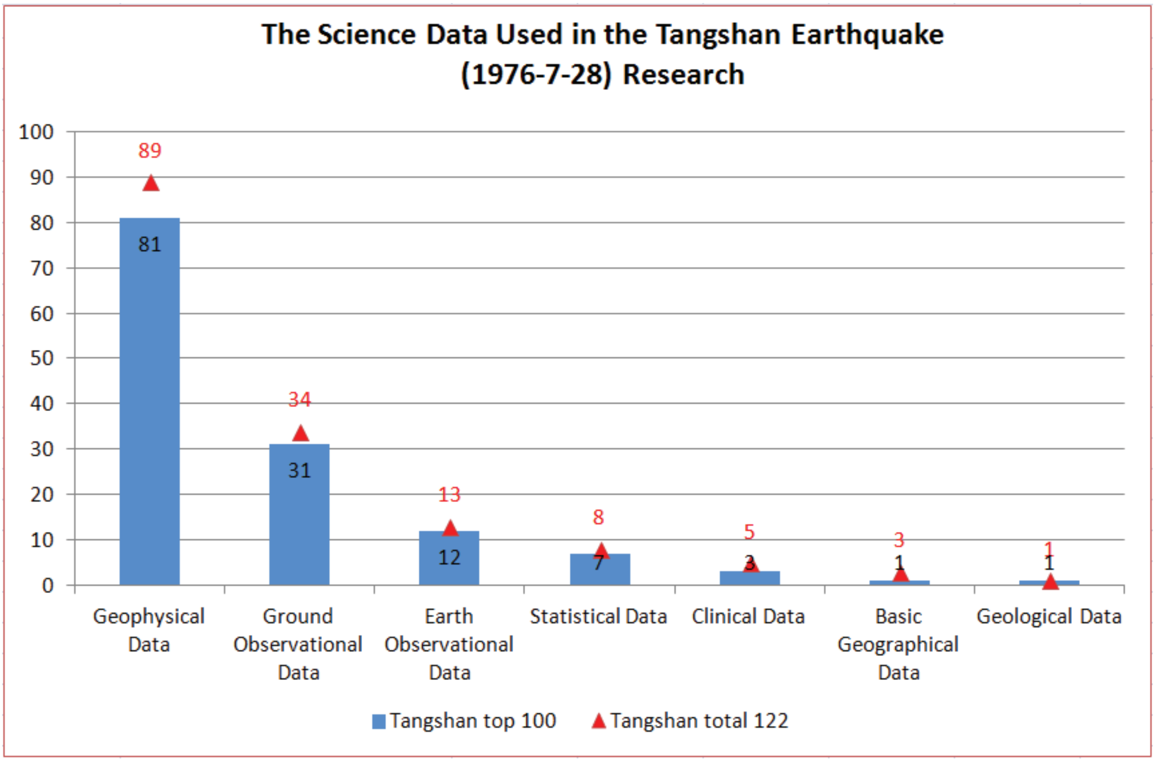

Figure 5: Frequency statistics of multidisciplinary data used in Tangshan earthquake research (WS view and HCS view).

Taking the Tangshan Earthquake as an example, the occurrence frequency of each data type is as follows:

- Geophysical data 81%
- Ground observational data 31%
- Earth observational data 12%
- Statistical data 7%
- Clinical data 3%
- Basic geographical data 1%
- Geological data 1%

6.3 Results Analysis

In the above graph, the numbers in red and the numbers in black represent the number of papers in the WS and HCS, respectively.

The Tangshan earthquake occurred on 28 July 1976. Limited to observation technology, the information about the Tangshan earthquake depends greatly on the geophysical and ground observational data reflected in the research papers. The graph above shows the similar data usage trends for Tangshan earthquake research in WS papers and HCS papers.

The Wenchuan earthquake occurred on 12 May 2008. As shown in **Figure 6**, geological and geophysical data are found widely in the WS research papers, reflecting researchers' emphasis on these two kinds of data that contain elements of topography, tectonic, lithosphere, geomorphologic, electromagnetic, etc.

In our opinion, the whole-set articles approximately reflect the data preference of authors and editors. However, the high-cited articles show the data preference of the readers. Therefore, the difference between HCS and WS papers clearly shows the distinct data preference of authors and readers. The large gap of data occurrence frequency reflected in WS and HCS shows that readers pay more attention to ground observational data, earth observational data, and statistical data while authors and editors like to use geophysical data and geological data. This gap reminds us that data users need to use multidisciplinary data in their research.

Based on a discussion seminar about the usage gap of socio-economic data in disaster events, domain experts generally thought that it was of great importance to enhance the use of socio-economical data in disaster studies. This opinion also can be confirmed by the graph in **Figure 7**. Clinical and statistics data show a large gap among different user communities. Statistics data have not been widely and scientifically used. Thus, they will be the most important potential data types for future earthquake study.

The Haidi earthquake occurred on 12 January 2010. Because observation technology had been in development for several years by that time, all kinds of data are available. As we can see in **Figure 7**, clinical data ranks first out of all data types. According to the disaster assistance situation, a U.S. hospital ship obtained much on-site clinical data during the Haidi earthquake rescue phase; thus the data usage reflected in the literature is consistent with the actual situation.

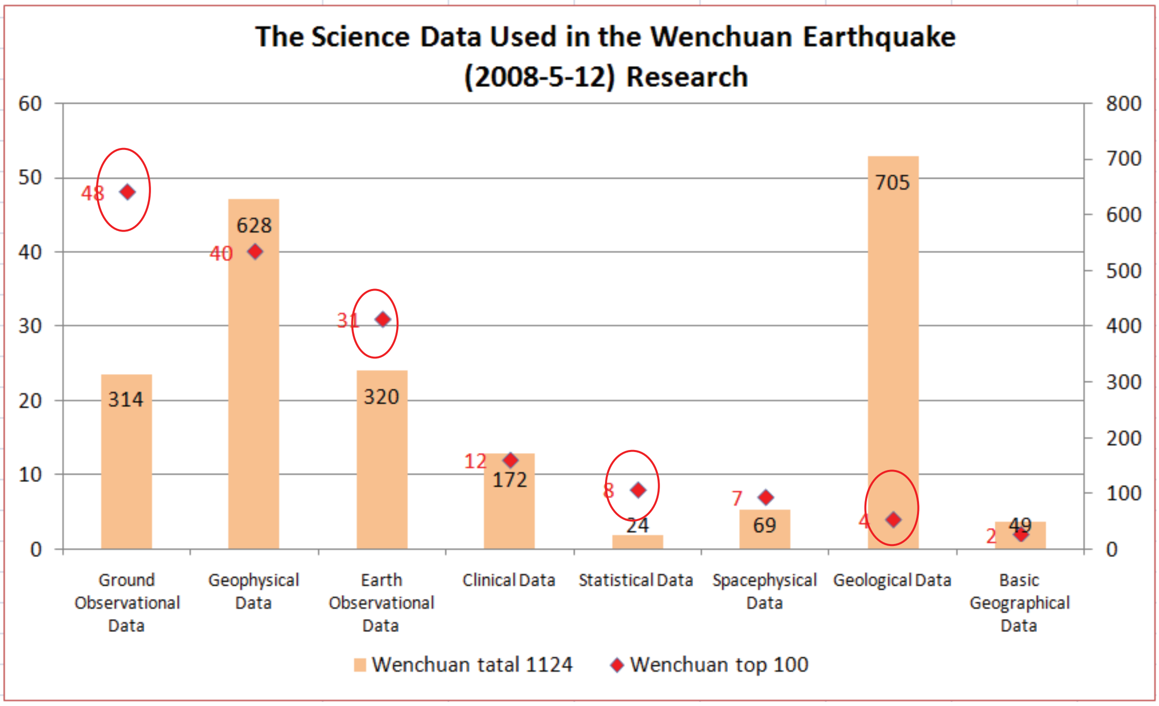

Figure 6: Frequency statistics of multidisciplinary data used in Wenchuan earthquake research (WS view and HCS view).

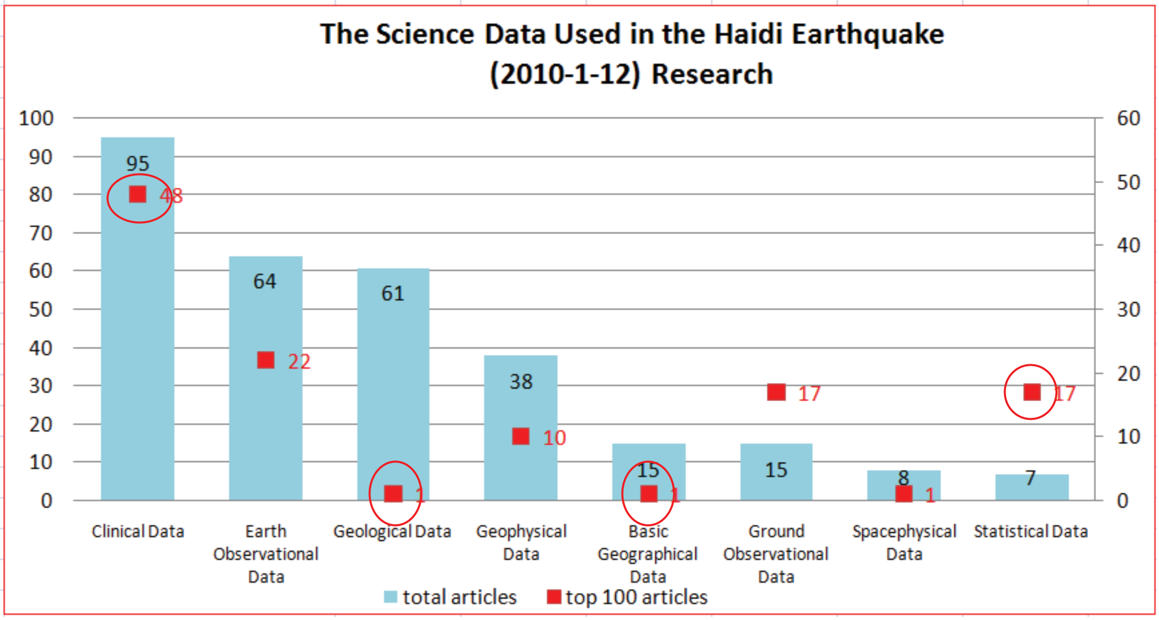

Figure 7: Frequency statistics of multidisciplinary data used in Haidi earthquake research (WS view and HCS view).

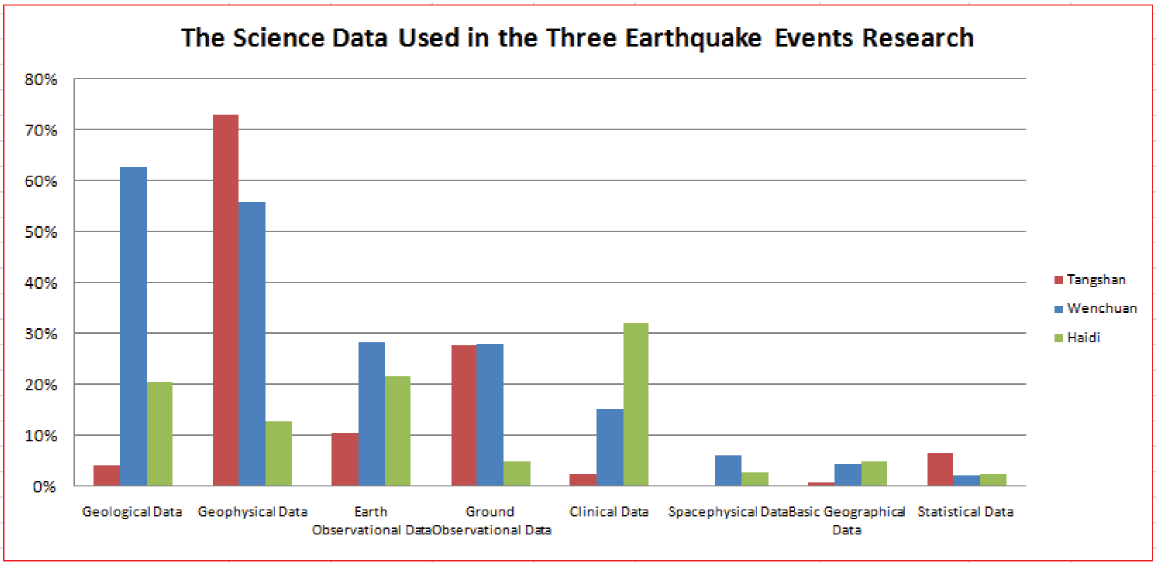

Figure 8: Comparison of multidisciplinary data used in the three earthquake events (WS view).

However, there are still usage gaps reflected in the graph. The large data usage gaps between the WS and HCS papers shown in the graph clearly demonstrate that readers need to pay more attention to clinical data, geological data, basic geographical data, and space physical data when studying the Haidi earthquake event because these data types haven't been paid enough attention to by readers. Also the usage gap of statistical data reminds authors and editors to enhance their statistical data employment in Haidi earthquake study.

In order to make the comparison clearer, **Figure 8** demonstrates the proportional differences in data usage in the literature about the three earthquake events.

With the development of technology and science, the data usage of different earthquake events has changed greatly. The comparison of data usage in the three earthquake events reflects that the data usage trend in earthquake research has changed from geophysical data and observational data to clinical data. More and more researchers have begun to pay more attention to socio-economic data. What has caused these data usage differences?

6.4 Trend analysis of data usage

From the above analysis of three earthquake events, we can see that geological data played the key role in the early years when multidisciplinary data were relatively scarce. With the development of observation technology, earth observation data and ground observational data became more and more significant in research. However, socio-economic data are not fully utilized due to the open access and open policy gaps in scientific data and the lack of website online services. The role of clinical data in disaster research is obvious because of a disaster's devastating social influence. However, the acquisition capacity of clinical data needs to be developed. According to the above analysis, the elements that affect multidisciplinary data for certain disaster events include the following four aspects:

1. Spatio-temporal factors

As we know, observation technology and capabilities have been developing over time. Because the three specific earthquake events happened in different times and locations, there is a gap in the data acquisition capabilities, which are limited by the spatio-temporal factors that are the objective reasons for data selection.

2. Disaster event characteristics

The development of a disaster event needs three elements: hazards, environment, and exposure. In our understanding, disaster type is the most important factor determining which kind of data is useful for research. Also the data usage of specific disaster events varies. In our research, all three earthquake events belong to the geological disaster type. However, their specific factors vary because of environment differences and crustal movements. Thus the characteristics of the disaster event itself are also factors affecting data usage.

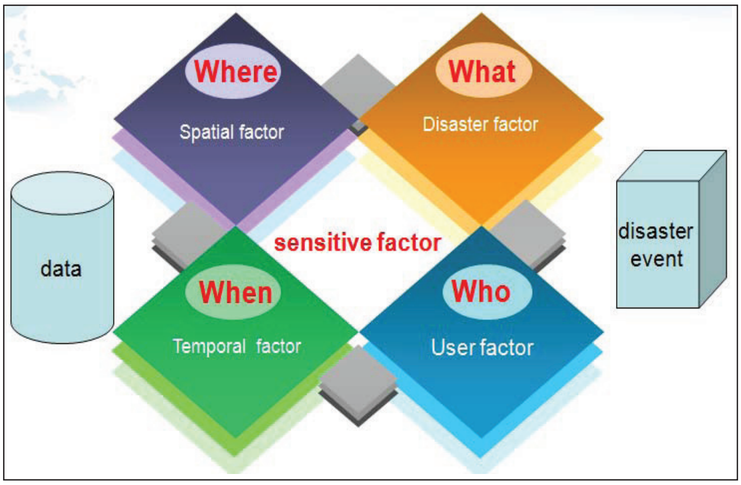

Figure 9: Elements affecting multidisciplinary data for certain disaster events.

3. User community

According to the domain experts' discussion, we have interpreted the data usage in WS and HCS as reflecting authors' opinions and readers' opinions. Thus we can obtain potential knowledge from the results given above. The gap of data usage between WS and HCS is clearly shown in the above graphs. The data concerns of the different user communities vary greatly due to disciplinary restrictions and other factors, including data acquisition capability, research objectives, etc. Therefore, the user community is the subjective element affecting data usage.

7 Conclusion and Future Work

The literature-based knowledge mining model employed in our study is a specific application of data mining in a disaster data study; the thesaurus is the core knowledge base in the study, and domain experience and a statistical approach are incorporated to extract a data type thesaurus. This model can be expanded to solve other scientific questions in disaster research, such as correlation analysis, new research directions discovery, and so on.

The experimental verification of the model justifies the reasonableness of the analysis results, and the earthquake study case shows that data selection depends heavily on four sensitive factors: spatial and temporal factors, disaster event characteristics, and user community objectives. However, the core knowledge base in our study derives from domain experts, who have the usual human frailties. Once the knowledge base changes, the output results also change. In our future research, the model needs to be modified to reduce expert participation. Also experiments on other disaster types also need to be performed.

8 Acknowledgements

This study is one research topic of "Study on Data Management Model for Historical Disaster Loss Data: Flood as Showcase" sponsored by IRDR-China. Part of this work is financially supported by the CODATA LODGD (linked open data for global disaster risk research) Task group. It is also partially supported by the 863 program "Key technology cooperation research on the application of integrated earth observation system" under grant 2012DFG21710.

9 References

Adriaans, P. & Zantinge, D. (1996) *Data Mining*. Reading, MA: Addison-Wesley.

Below, R., Wirtz, A., & Guha-Sapir, D. (2009) Disaster Category Classification and Peril Terminology for Operational Purposes (Working paper). *Centre for Research on the Epidemiology of Disasters (CRED) and Munich Reinsurance Company (Munich RE).*

Brachman, R. & Anand, T. (1996) The process of knowledge discovery in databases: a human-centered approach. In Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (Eds.) *Advances in Knowledge Discovery and Data Mining*. Menlo Park, CA: AAAI Press, pp 37–58.

Cios, K. & Kurgan, L. (2005) Trends in data mining and knowledge discovery. In Pal, N. & Jain, L. (Eds.) *Advanced Techniques in Knowledge Discovery and Data Mining,* Springer, pp 1–26.

Cios, K. J. & William Moore, G. (2002) Uniqueness of medical data mining. *Artificial Intelligence in Medicine 26*(1), pp 1–24.

Fayyad, U., Haussler, D., & Stolorz, P. (1996f) KDD for science data analysis, issues and examples. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining,* Portland, pp 50–56.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (Eds.) (1996c) The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM 39*(11), pp 27–34.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996d) Knowledge discovery and data mining: towards a unifying framework. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining,* Portland, OR, pp 82–88.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996e) From data mining to knowledge discovery in databases. *AI Magazine 17*(11), pp 37–54.

Fayyad, U., Piatesky-Shapiro, G., Smyth, P., & Uthurusamy, R. (Eds.) (1996) *Advances in Knowledge Discovery and Data Mining,* AAI Press.

Frijters, R., van Vugt, M., Smeets, R., van Schaik, R., de Vlieg, J., et al. (2010) Literature Mining for the Discovery of Hidden Connections between Drugs, Genes and Diseases. *PLoS Comput Biol 6*(9).

Hirschman, L., Park, J. C., Tsujii, J., Wong, L., & Wu, C. H. (2002) Accomplishments and challenges in literature data mining for biology. *Bioinformatics 18*(12), pp 1553–1561.

Hristovski, D., Peterlin, B., Mitchell, J. A., & Humphrey, S. M. (2005) Using literature-based discovery to identify disease candidate genes. *International Journal of Medical Informatics 74*(2), pp 289–298.

Kurgan, L. A., & Musilek, P. (2006) A survey of Knowledge Discovery and Data Mining process models. *The Knowledge Engineering Review 21*(01), pp 1–24.

Meliha, Y. & Wanda, P. (2006) Using Statistical and Knowledge-Based Approaches for Literature-Based Discovery. *Journal of Biomedical Informatics 39*(6), pp 600–611.

Pratt, W., & Yetisgen-Yildiz, M. (2003) LitLinker: capturing connections across the biomedical literature. In *Proceedings of the 2nd International Conference on Knowledge Capture*, ACM, pp 105–112.

Tansley, S. & Tolle, K. M. (Eds.) (2009) *The fourth paradigm: data-intensive scientific discovery.*

Wirtz, A., Kron, W., Löw, P., & Steuer, M. (2014) The need for data: natural disasters and the challenges of database management. *Natural Hazards 70*(1), pp 135–157.

Yang, Y. (2013) *Research on Emergency Intelligent Information Retrieval System Based on Domain Knowledge Model*, Beijing University of Posts and Telecommunications: Beijing.

How to cite this article: Zhang, H, Qing, X, Huang, M and Li, G 2015 A Correlation Analysis Model for Multidisciplinary Data in Disaster Research. *Data Science Journal,* 14: 19, pp. 1–12, DOI: http://dx.doi.org/10.5334/dsj-2015-019

Published: 22 May 2015

]u[*Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS