Analyzing Structures and Evolution of Digital Humanities Based on Correspondence Analysis and Co-word Analysis

Xiaoguang Wang
Digital Humanities Center for Japanese Arts and Cultures, Ritsumeikan University, Kyoto, Japan
School of Information Management, Wuhan University, Wuhan, China P.R.
E－MAIL：whu_wxg@126.com

Mitsuyuki Inaba
Digital Humanities Center for Japanese Arts and Cultures, Ritsumeikan University, Kyoto, Japan
E－MAIL：inabam@sps.ritsumei.ac.jp

Abstract:

Digital humanities is a new interdisciplinary study concerned with the intersection of computing and traditional arts and humanities disciplines. Based on two journals and four annual conference proceedings that advocate the study, we performed a correspondence analysis and a co-word analysis to understand the structures and evolution of digital humanities. The results show that there is no clear subdiscipline in digital humanities and the disciplinary representative nomenclature is changing from humanities computing to digital humanities, which implies digital humanities is still expanding its research domain.

Keywords:

digital humanities, humanities computing, co-word analysis, structure of science

1. Introduction

Over the last decades, digital humanities has been studied increasingly. As an interdisciplinary study concerned with the intersection of computing and traditional humanities disciplines, digital humanities researchers come from various fields such as history, philosophy, linguistics, literature, arts, and archaeology. Although the multi-aspects of digital humanities is obvious, the interdisciplinary structure and the development state of digital humanities has not been established yet.

Citation and keyword are two regular indicators to analyze and map the structures and evolution of science domains in scientometrics. As digital humanities is a new interdiscipline without long history and rich citation clues, keyword based analysis is more suitable and efficient than citation based analysis.

Co-word analysis and correspondence analysis are two practical keyword based analysis techniques. The former one uses patterns of co-occurrence of pairs of items in a corpus to identify the relationships between words or phrases, the extent to which these items are central to the

whole area, and the degree to which these items are internally structured (He, 1999). The latter is one of the principal statistical methods in traditional humanities, which assists in picturing the structure of categorical variables. For detecting topics, mapping the structures and tracking the evolution of digital humanities, we combined coword analysis and correspondence analysis in this paper.

This paper is organized as follows: the next section briefly describes the history of digital humanities. Afterwards, two types of research methods are given followed by multiple perspective analysis of the word data from two journals and four conference proceedings. After a description of analysis and a discussion about the combined method in this research, the result will be given out.

2. Research purposes

Although the application of computer to the humanities has last about 50 years (McCarty, 2002), it's still hard to give a clear definition to digital humanities as it remains an emergent discipline and is continually changing, developing, and redefining itself. Even the term "digital humanities" is also a new popular parlance in recent years.

On the "what is digital humanities?" McCarty said "It is methodological in nature and interdisciplinary in scope. It works at the intersection of computing with the arts and humanities, focusing both on the pragmatic issues of how computing assists scholarship and teaching in the disciplines and on the theoretical problems of shift in perspective brought about by computing. Like comparative literature it takes its subject matter from other disciplines and is guided by their concerns, but it returns to them ever more challenging questions and new ways of thinking through old problems". According to McCarty's discussion, digital humanities has possibilities to change traditional humanities scholarship not only on methodological focal points, but also on the "ways of thinking" toward their theoretical problems.

However, even the first issue of Digital Humanities Quarterly published in 2007 (Flanders, Piez, & Terras, 2007) had to differ to give a definition of digital humanities, and addressed their goal is to answer an alternative question, "How can we shape the digital humanities?" Therefore, this paper aims to get a clear view of the development state of digital humanities from bibliometric perspective and illustrate how its discipline has been shaped. It will be benefit for cognizing research directions in the future and scholarly communication in digital humanities community.

3. Method description

3.1 Correspondence analysis

Correspondence analysis is an exploratory tool commonly used to analyze and visualize simple two-way and multi-way tables containing some measures of correspondence between the rows and columns. The results provide information which is similar in nature to those produced by factor analysis techniques, and they allow one to explore the structure of categorical variables.

Correspondence analysis has several advantages: it is specifically designed to compare profiles or patterns; it is a multidimensional method that achieves appropriate data reduction, filters out noise, and objectifies correlations among variables; it is a method that provides graphic output such as maps that are easier to grasp than series of numbers (Benzecri, 1992). As a result of this feature, correspondence analysis has gained a positive reputation as a recommendable tool for the data analysts in many disciplines (Beh, 1999).

As an important analysis method, correspon-

dence analysis has been applied to explore structures of categorical variables in many research fields, such as the changes in physics (Bhattacharya, Singh, & Sudhakar, 1997), spices research in Asian countries (Senthilkumaran & Amudhavalli, 2007), and Indian research collaboration patterns (Anuradha & Shalini, 2007) .

In this paper, correspondence analysis is carried out using R language (available at http://www.r-project.org) which is a free software package for statistical computing and graphics, and is suitable to exploratory analysis of multivariate numerical and textual data.

3.2 Co-word analysis

Co-word analysis is an objective and quantitative methodology. It is based on the nature of words, which are the important carrier of scientific concepts, idea and knowledge. This enables us to follow actors objectively and detect the structures of science without reducing them to the extremes of either internalism or externalism (Callon, Law, & Rip, 1986). Co-word analysis reveals patterns and trends in a specific discipline by measuring the association strengths of terms representative of relevant publications produced in this area. The main feature of co-word analysis is that it visualizes the intellectual structure of one specific discipline into maps of the conceptual space of this field, and that a time-series of such maps produce a trace of the changes in this conceptual space.

Many researchers have used co-word analysis as an important method to explore concept networks in different fields, such as artificial intelligence (Coutial & Law, 1989), acidification research (Law & Whittaker, 1992), scientometrics (Courtial, 1994), software engineering (Coulter, Monarch, & Konda, 1998), information retrieval (Ding, Chowdhury, & Foo, 2001), and so on. Digital humanities is a developing discipline, so co-word analysis is quite suitable for the exploration of its structure and evolution.

Co-word analysis has four steps. The first step is data collection. Words are the most important elements in co-word analysis. There are two ways to extract keywords from journal articles, conference papers or technology reports. One is from titles or keyword lists. The other one is from full-text. In either way, only the words or phrases with proper frequency will be chosen as the objects of co-word analysis to denote the central topics in a specific domain. Since digital humanities is an evolving interdiscipline and growing its vocabulary, it is difficult to have specific criteria to appropriately choose keywords and eliminate noises from full-text. Therefore, this study adopts the former way to pick out representative keywords efficiently.

The second step is data standardization. There are many similar concepts which appear as different words or phrases in word collection. For standardizing those words, researchers have to consider synonyms, antonyms, ambiguity, broad term/narrow term and general terms, such as knowledge, theories, influence, projects, development, applications, production, implementation, definition and so on.

The third step is matrix construction. Once a research subject is selected, a matrix based on the word co-occurrence data is built. The higher co-occurrence frequency of the two words is, the closer relationship between them is. The matrix is then transformed into a correlation matrix by using a specific correlation coefficient.

The fourth step is data analysis and mapping. The most popular method is multidimensional scaling (MDS). MDS represents all high-dimensional points in a two- or three-dimensional space in a way that the pairwise distances between points approximate the original high-dimensional

distances as precisely as possible. The results of this kind of analysis, an ordered map of the concept space, display directly the similarity relations of different topics.

In order to detect the changing of keywords space, we apply a new research method named co-word network analysis. Actor network theory is adopted in co-word network analysis. We use the degree centrality to express the relationships owned by one word. A similar mapping method was used by Onyancha and Ocholla (2008).

4. Data

4.1 Data collection

Our research collection is consisted of two journals and four conference proceedings (see Table 1): *Literary and Linguistic Computing* from Year 2005 to 2008, *Digital Humanities Quarterly* from Year 2007 to 2008, and proceedings of *DH 2005, DH 2006, DH 2007,* and *DH 2008*. We chose 548 papers written in English from the collections.

Table 1. Papers distribution from 2005 to 2008

Year	*Number of papers*	*%*
2005	153	29.9
2006	114	20.8
2007	139	25.4
2008	142	25.9
Total	548	100

4.2 Vocabulary extracting and standardizing

Since there is no keywords list in these papers, we manually extracted keywords from the titles of these papers. 1,219 distinct keywords left after being extracted and standardized, which appeared 2,394 times in total.

Our research goal focuses on two facets: one is to detect the structure of a research field, the other is to detect the transformation of a research field. To accomplish these two aims, we picked out those keywords whose frequency is higher than 3 in respective years, and got 82 highly frequented keywords in total. These keywords' total frequency is 781 (32.6%). Then we counted the frequency of every keyword in the past four years (see Table 2).

The frequency of keywords discovered its influence on research community: the higher it is, the more influences it is on research community. From Table 1, we can find researchers' attention focus is changing every year. This means digital humanities is an unstable discipline.

4.3 Matrix Constructing

The 82 highly frequented words are distributed among 424 papers. Based on the co-occurrence relationships of the 82 words in the past four years, we constructed a unitary co-word matrix, which is a symmetrical adjacency matrix. We calculated the association values between any word pairs with Equivalence Coefficient index (E) which is defined as the following formula:

$$E_{ij} = \frac{C^2_{ij}}{C_i \times C_j}$$

Cij is the number of documents in which the keyword pair (i and j) appears. Ci and Cj are the occurrence frequencies of keyword i and keyword j in the set of articles. Eij has a value between 0 and 1. Eij measures the probability of word i appearing simultaneously in a document set indexed by word j and, inversely, the probability of word j if word i appears, given the respective collection frequencies of the two words.

The raw data matrix was recalculated (Pearson correlation coefficient) in order to find proximity on the basis of the 82-vector. In other words, the similarity between two words was calculated on

Table 2. Top 82 high-frequency keywords

Items	*05*	*06*	*07*	*08*	Items	*05*	*06*	*07*	*08*
American	1	1	3	1	History	4	1	2	2
analysis	0	1	5	5	humanities	8	3	8	13
archive	6	3	3	3	humanities computing	13	5	1	1
archway	3	0	0	0	information technology	0	0	0	3
attribution	4	2	2	2	interactive	3	1	2	1
Australian	1	3	0	0	language	2	1	7	2
author	0	1	4	1	learning	5	0	1	2
authorship	7	2	4	3	linguistic	1	1	4	0
case study	4	1	2	2	linguistics	2	2	3	2
classification	0	1	3	1	literary	3	4	3	4
collaborative	2	5	0	4	literature	0	1	3	0
collection	1	0	5	1	manuscript	1	2	4	3
comparison	3	2	1	0	markup	5	2	1	2
computational	5	2	0	0	meaning	3	3	1	0
computer	1	3	2	0	mining	1	0	2	3
computer assisted	0	1	0	3	model	4	2	2	2
corpora	0	1	3	1	modeling	4	3	2	8
corpus	5	4	5	7	music	1	2	5	1
cultural	0	0	3	1	online	6	1	5	5
database	5	3	3	2	panel	0	0	5	1
dialect	0	4	1	2	poetry	3	2	3	2
difference	1	1	3	0	reading	2	0	4	0
digital	5	8	7	10	resources	4	0	2	3
digital edition	1	4	1	2	scholarship	2	5	2	2
digital humanities	3	4	7	12	semantic	1	3	1	0
digital library	3	0	0	2	speech	1	0	1	3
document	5	2	6	2	system	2	0	1	3
DTD	3	0	0	0	teaching	6	0	1	0
early modern	2	0	0	3	TEI	5	3	4	8
editions	0	0	0	4	text	12	9	16	11
electronic	5	0	1	0	text analysis	3	1	3	1
electronic edition	4	0	1	2	text mining	0	2	4	1
encoding	4	1	4	2	textual	3	2	0	0
English	3	4	3	2	time	2	0	3	0
environment	0	0	0	3	timeline	0	1	1	3
experience	3	1	0	1	variation	2	3	1	1
France	1	3	3	0	virtual	2	1	1	3
gender	0	0	4	3	visual	3	0	0	1
generation	3	0	0	0	visualization	3	1	3	2
German	0	1	0	3	web	4	2	4	2
historical	3	3	6	2	XML	6	1	4	2

the basis of all co-occurrence frequency that these two words have with all the other items in the same matrix. So the words with high Pearson correlation coefficient are located together in the map, and those words located together in the map have high similarity in terms of co-occurrence profile within the whole matrix.

5. Result

5.1 Correspondence analysis

In order to grasp the overall topic distribution and its changing in the period (2005-2008), a correspondence analysis was applied to the raw frequency data of the 82 words (see Fig. 1).

With Table 2 and Fig.1, we can find that the hot topics in digital humanities change every year. In 2005, the hot topics include themes related to electronic, humanities computing, XML, computational, authorship, learning, and teaching. 2006 includes topics relating to dialect, semantic, collaborative and scholarship. 2007 includes topics relating to corpora, author, gender, linguistic, language, music and text mining. 2008 includes topics relating to TEI, timeline and mining. In addition, we can find some continuous topics in the past four years, which are relating to corpus, database, historical, poetry, online and Web.

5.2 Multidimensional scaling analysis

In order to get a macro view of digital humani-

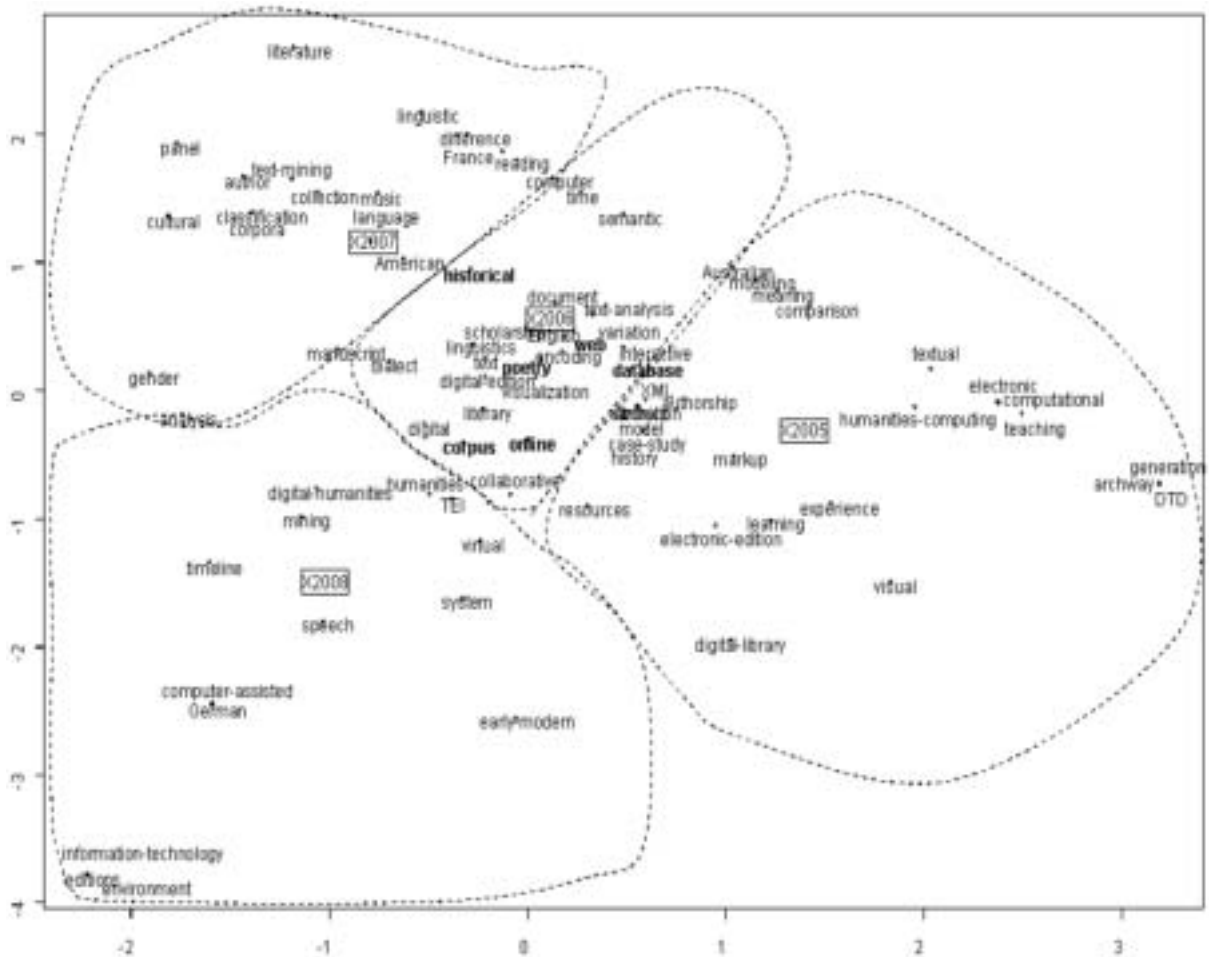

Fig. 1 Plotting map of keywords correspondence analysis

ties, we also made a multidimensional scaling analysis (see Fig. 2) based on the unitary matrix. As Fig.2 shows, the 82 words can be divided into four clusters by and large. Cluster A consists of fundamental concepts relating to information technologies and methods, such as TEI, XML, database, corpus, encoding, visualization, case study, and so on. Cluster B and cluster C represent some special application research domains, such as literature, speech, dialect, poetry, history, gender, authorship in cluster B, and music, archive, scholarship in cluster C. Cluster D consists of general words, such as digital humanities and humanities computing. Though the clusters are divided, the division is not significant and exclusive, such as text, text analysis, text mining and textual. They spread in different clusters, but their semantic relationships are close and strong.

What should be particularly noticed is that English and French have been studied more than other languages. This indicates that the current digital humanities research is unbalance in different language contexts.

5.3 Co-word network analysis

For detecting the dynamics of scientific concepts in digital humanities, we disassembled the unitary co-word matrix, constructed four co-word matrixes based on annual co-occurrence relationships, then we plotted four co-word network figures (see Fig. 3, Fig. 4, Fig. 5 and Fig. 6). The results are visualized using UCINET which is a software used for social network analysis.

In actor network theory, a node's degree centrality is defined as the number of links a node has. The higher degree centrality of the node is, the more important the node is. In the next four fig-

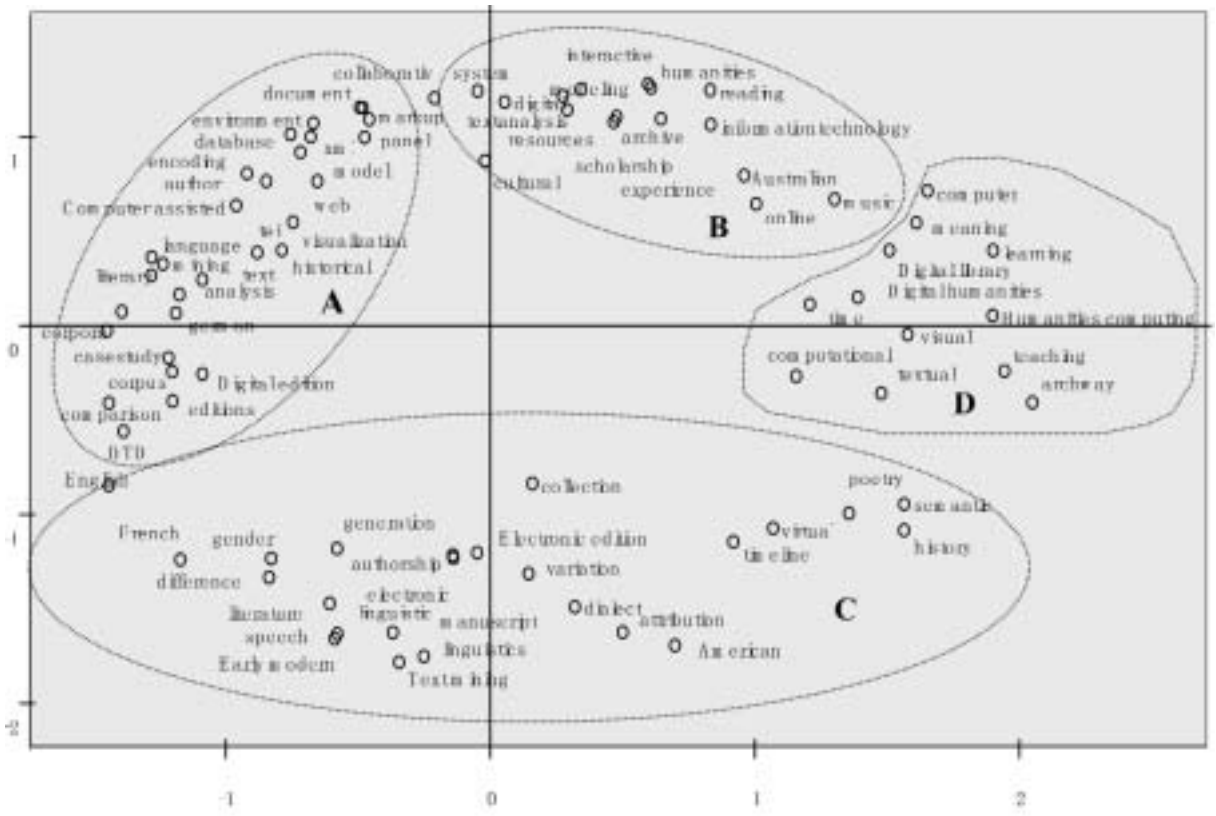

Fig. 2 Multidimensional scaling plotting map of 82 words

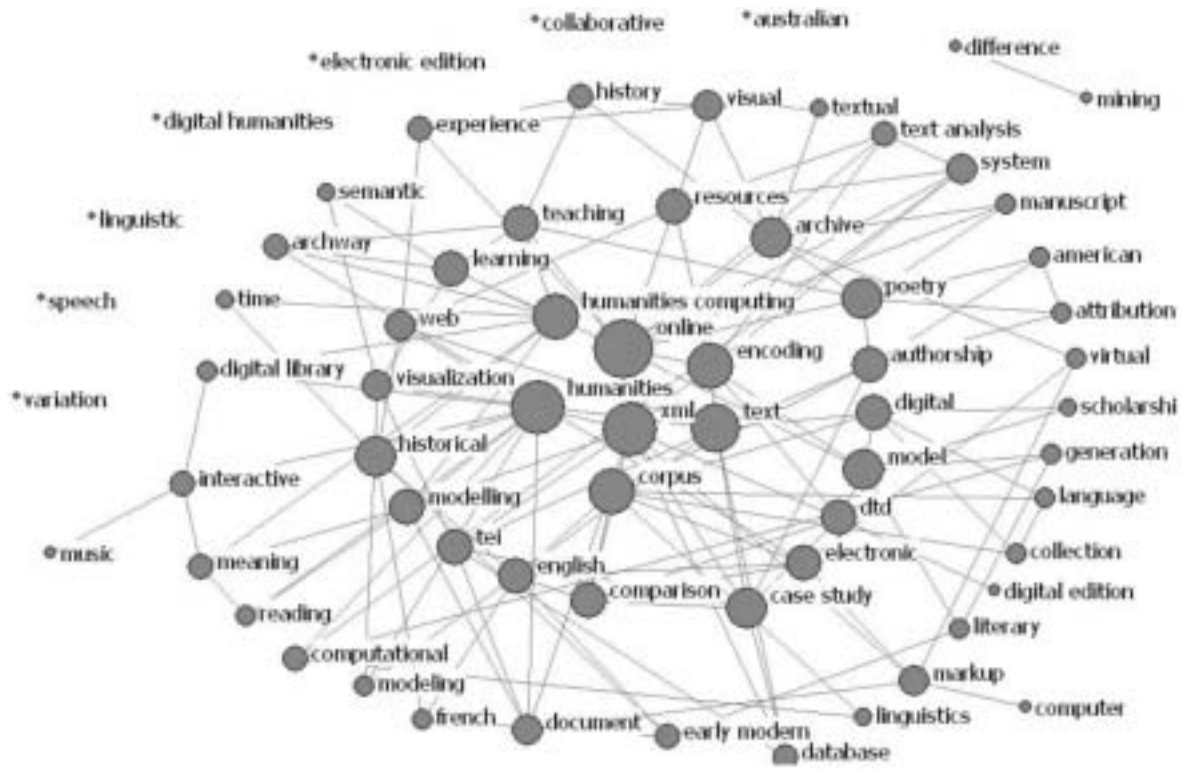

Fig. 3 The co-occurrence network of high frequency words in 2005. Node size represents degree centrality.

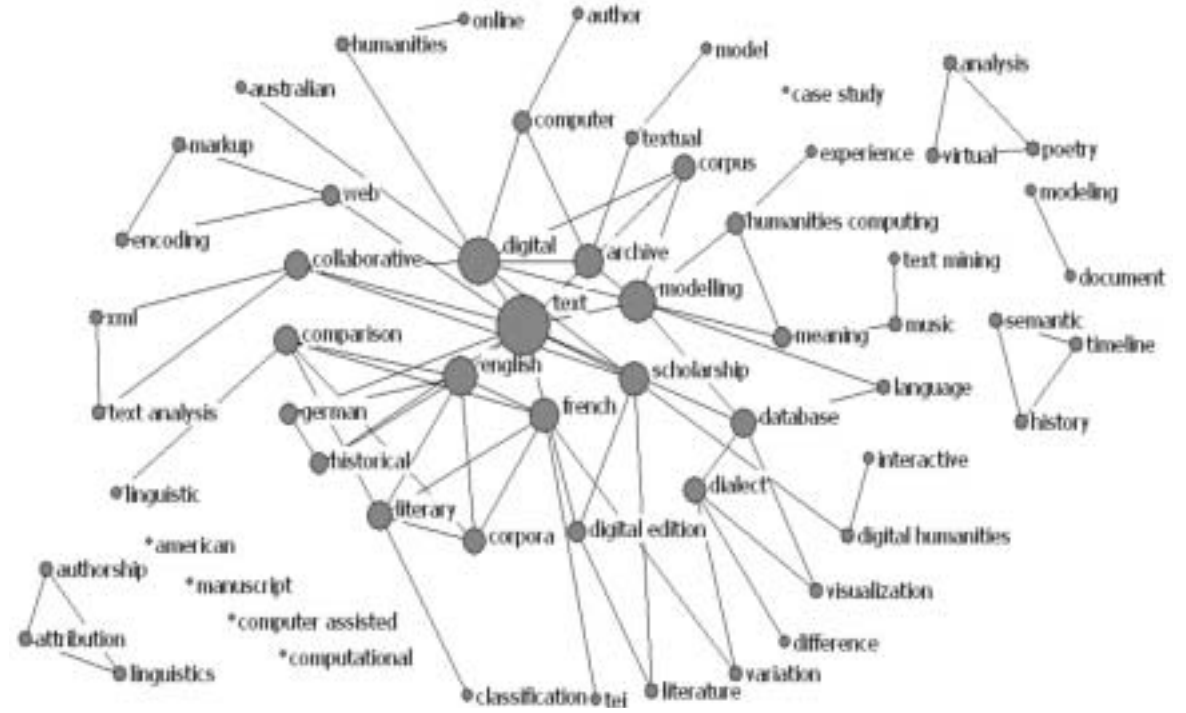

Fig. 4 The co-occurrence network of high frequency words in 2006. Node size represents degree centrality.

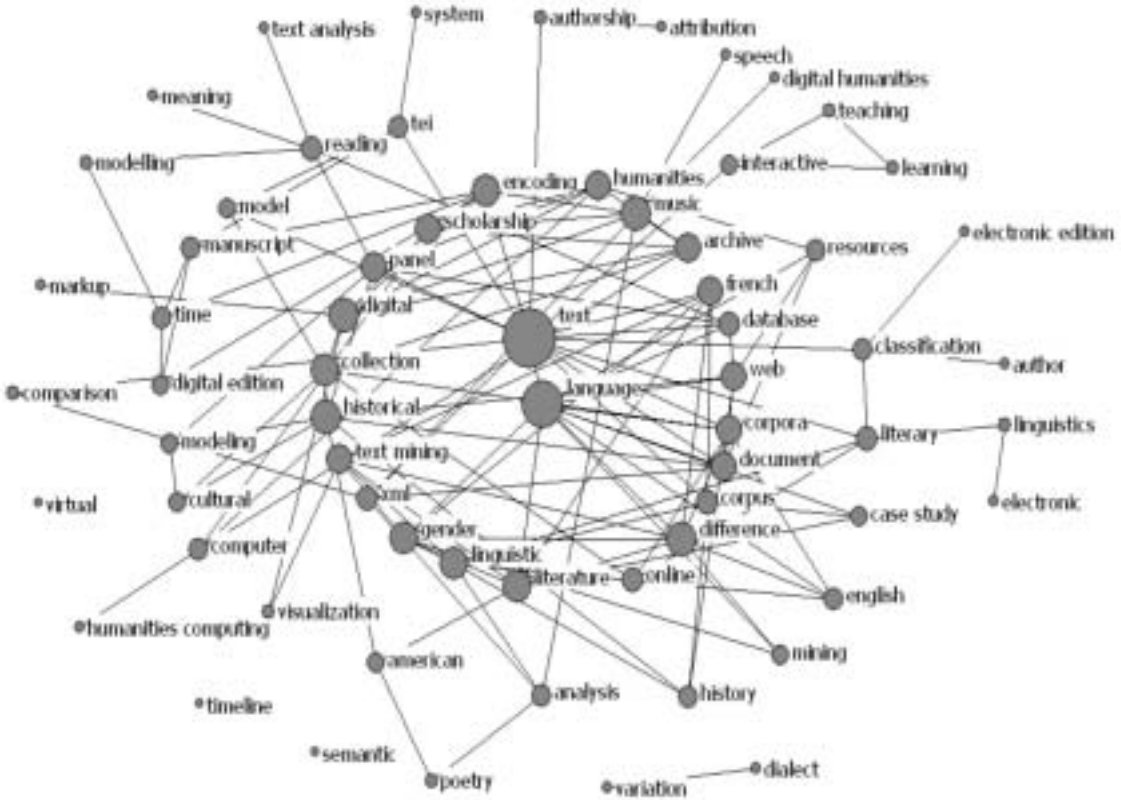

Fig. 5 The co-occurrence network of high frequency words in 2007. Node size represents degree centrality.

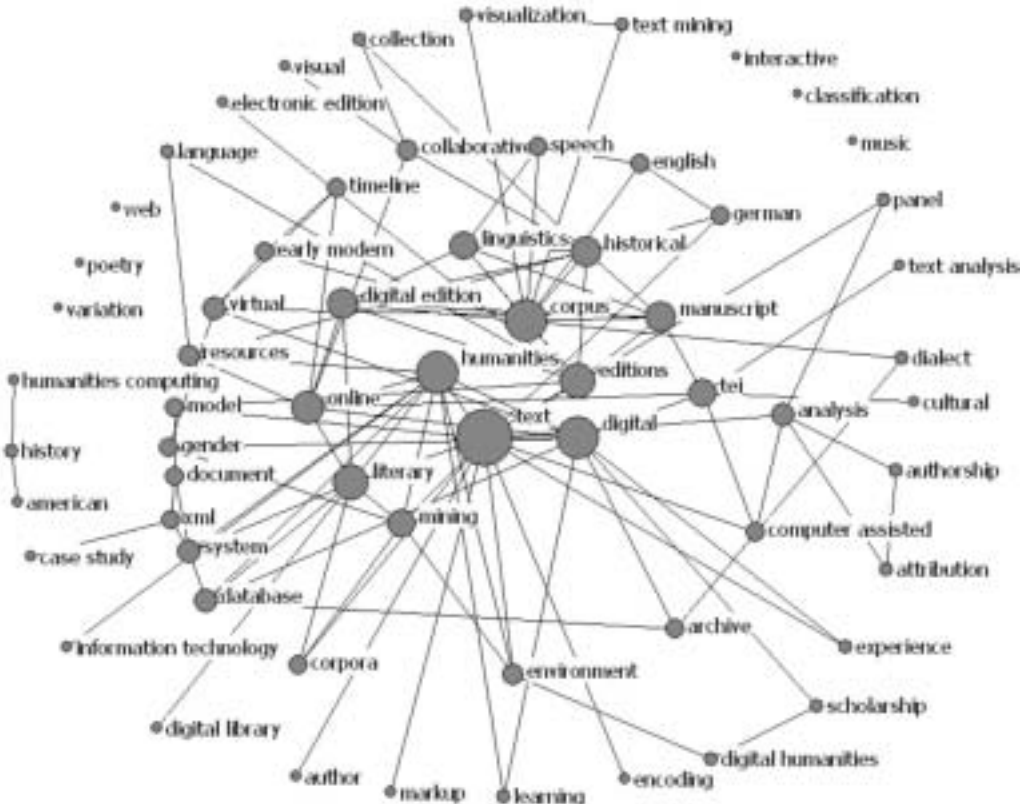

Fig. 6 The co-occurrence network of high frequency words in 2008. Node size represents degree centrality.

ures, we laid nodes (keywords) according to their degree centralities. The nodes with high degree centralities are laid in the central part, while the lower ones are laid in the peripheral part.

In co-word network analysis, frequency and degree centrality are two different indexes which represent different meanings. We laid keywords according to their degree centralities in the four co-word networks. Two interesting phenomena in these networks are the changes of degree centrality of "digital humanities" and "humanities computing". One is that the frequency and degree centrality of humanities computing all decreased in the past four years. The other one is that though the frequency of digital humanities increased continually, its degree centrality was low at all times. These phenomena indicated humanities computing and digital humanities all have not many co-occurrence relationships with high frequency words, which mean that digital humanities has passed its infancy, more and more scholars from computer science and library and information science have been involved in digital humanities community, and "humanities computing" became less satisfying as a disciplinary representative word, and had been gradually replaced by "digital humanities". However, digital humanities is still far from maturity, so "digital humanities" always occurs with low frequency words which represent recently emerged topics such as geographical information system, interactive games, timeline and virtual reality.

Besides the two phenomena, the boom of topics related to data mining and text mining is also obvious and remarkable. With more and more text digitalization projects are implemented, the research infrastructure becomes better than before, so the application of text mining becomes easier and broader.

6. Discussion

One of the key issues in the present study is the way we selected data source. Kostoff et al. (1997) claim that one of the many advantages of using full-text over keywords or index words in analysis is its ability to retain phrases with low frequency but high importance, which may be overlooked with the keyword approach due to their low frequency. Furthermore, Kostoff et al. (2001) discuss that if the analysis is targeted at disparate disciplines, experts who have diverse backgrounds are needed in order to conduct a fully credible analysis of phrases/keywords in full-text of papers. With this argument in mind, it is plausible to assume that keywords lists or paper titles identified by authors are reliable data source for the analysis of a newly emerged interdisciplinary research, such as digital humanity research, that is still in the process of building its vocabulary and methodology.

Besides the above, whether the phrases should be split to isolated words is another problem. For avoiding this problem and representing some new specific concepts, phrases are always preferentially picked out and retained. Although there is a little difference between word lists extracted by different experts, they don't impact significantly to the final results and conclusions.

In addition, factor analysis and cluster analysis are two popular approaches for mapping the structures and evolution in traditional co-word analysis. As digital humanities is a new interdiscipline, there is no clear clusters inside, so we make a co-word network analysis with the raw co-occurrence matrix. Co-word network analysis successfully visualizes the inter-relations of keywords, potential structures and evolution of digital humanities. The result is understandable and reliable, which demonstrates that co-word network is a viable

research approach. This maybe provides a new opportunity for mapping science domain.

7. Conclusion

Correspondence analysis and co-word analysis are two different content analysis techniques. In this paper, we analyzed the structures and evolution of digital humanities in the past four years with correspondence analysis and co-word analysis. In correspondence analysis, we focused on the time dimension. Based on the frequencies of keywords, we discovered the evolution of digital humanities over the past four years. Then, we focused on semantic dimension and studied the disciplinary structures of digital humanities with multidimensional scaling analysis. For getting more details on the changes of the whole discipline, we mapped four co-word network sociograms.

As a result, we found that although the hot topics related to corpus, database, historical, poetry, online and Web have lasted for four years, and there are still no clear subdisciplines in digital humanities. The utilization of specialty nomenclature in digital humanities community has been changed. Digital humanities replaced humanities computing, which indicated the extension and development of digital humanities research. In addition, some promising research methods and topics have been found, such as data mining, text mining, German, French, virtual system, GIS and so on. The empirical and visual results made in this paper furthered our understanding of the definition and development direction of digital humanities.

In the future, we will improve co-word analysis in terms of philosophical methodology, and develop an integrated software for its common application in digital humanities for text mining.

References

Anuradha, K. T., & Shalini, R. U. (2007), Bibliometric indicators of Indian research collaboration patterns: A correspondence analysis. *Scientometrics,* 71(2): 179-189.

Beh, E. J. (1999), Correspondence analysis of ranked data. *Communication in Statistics - Theory and Methods,* 28 : 1511-1533.

Benzecri, J. P. Correspondence Analysis Handbook. New York: Marcel Dekker, Inc, 1992.

Bhattacharya, S., Singh, S. P., & Sudhakar, P. (1997), Tracking changes in research priorities in physics: a macro level analysis. *Scientometrics,* 40 (1): 57-82.

Callon, M., Law, J., & Rip, A. (1986) How to study the force of science. In Callon, M., Law, J., and Rip, A. (eds.), *Mapping the dynamics of science and technology: Sociology of science in the real world.* London: The Macmillan Press Ltd, pp. 3-15.

Courtial, J.P., & Law, J. (1989), A co-word study of artificial intelligence. *Social Studies of Science,* 19(2): 301-311.

Coulter, N., Monarch, I. & Konda, S. (1998) Software engineering as seen through its research literature: A study in co-word analysis. *Journal of the American Society for Information Science,* 49(13):1206-1223.

Courtial, J.P. (1994) A coword analysis of scientometrics. *Scientometrics*, 31(3): 251-260.

Ding, Y., Chowdhury, G. & Foo, S. (2001) Bibliometric cartography of information retrieval research by using co-word analysis. *Information Processing & Management,* 37(6): 817-842

Flanders, J., Piez, W., & Terras, M. (2007) Welcome to Digital Humanities Quarterly. *Digital Humanities Quarterly,* 1(1), http://www.digitalhumanities.org/dhq/vol/001/1/000007.html

He, Q. (1999) Knowledge discovery through co-word

analysis. *Library Trends,* 48(1):133-159.

Law, J., & Whittaker, J. (1992) Mapping acidification research: A test of the co-word method. *Scientometrics,* 23(3):417-461.

Kostoff, R. N., Toothman, D. R., Eberhart, H. J., & Humenik, J. A. (2001) Text mining using database tomography and bibliometrics: A review. *Technological Forecasting & Social Change,* 68(2001): 223-253.

Kostoff, R. N., Eberhart, H.J., Toothman, D. R., & Pellenbarg, R. (1997). Database Tomography for technical intelligence: Comparative roadmaps of the research impact assessment literature and the Journal of the American Chemical Society. *Scientometrics*, 40(1): 103-138.

McCarty, W. (2002) Humanities computing, http://www.digitalhumanities.org/view/Essays/WillardMcCartyHumanitiesComputing

Onyancha, O. B., & Ocholla, D. (2008), Is HIV/AIDS in Africa distinct? What can we learn from an analysis of the literature? *Scientometrics.* DOI: 10.1007/s11192-009-0418-y

Senthilkumaran, P., & Amudhavalli, A. (2007), Mapping of spices research in Asian countries. *Scientometrics,* 73(2): 149-159.