Original Article

Healthc Inform Res. 2010 March;16(1):52-59.
doi: 10.4258/hir.2010.16.1.52
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

Knowledge Structure of Korean Medical Informatics: A Social Network Analysis of Articles in Journal and Proceedings

Senator Jeong, PhD, Soo Kyoung Lee, RN, Hong-Gee Kim, PhD
Biomedical Knowledge Engineering Laboratory, Seoul National University, Seoul, Korea

Objectives: This study aimed at exploring the knowledge structure of Korean medical informatics. **Methods:** We utilized the keywords, as the main variables, of the research papers that were presented in the journal and symposia of the Korean Society of Medical Informatics, and we used, as cases, the English titles and abstracts of the papers (n = 915) published from 1995 through 2008. N-grams (bigram to 5-gram) were extracted from the corpora using the BiKE Text Analyzer, and their co-occurrence networks were generated via a cosine correlation coefficient, and then the networks were analyzed and visualized using Pajek. **Results:** With the hub and authority measures, the most important research topics in Korean medical informatics were identified. Newly emerging topics by three-year period units were observed as research trends. **Conclusions:** This study provides a systematic overview on the knowledge structure of Korean medical informatics.

Keywords: Medical Informatics, Knowledge Structure, Social Network Analysis, Co-word Analysis

I. Introduction

Characterizing a domain of study presupposes the understanding of its knowledge structure. Since its inception in 1987, there have been few efforts to understand Korean medical informatics in quantitative way. The purpose of this paper was to quantitatively analyze the knowledge structure of Korean medical informatics.

There are a number of qualitative studies on biomedical informatics [1-3]. These research efforts heavily relied on experts' knowledge, experience, and intuition. Thus they may lack objective and quantitative understanding of the structure of the field. Mainstream approaches exploring the knowledge structure of a domain are quantitative studies such as co-citation analysis (CCA) and co-word analysis (CWA) which gauge topical closeness, similarity in the level of authors, papers, journals, or disciplines.

The most extensively used method, CWA, can reveal the overall picture and broad landscape and boundaries of a given field. CWA uses patterns of co-occurrence of pairs of words or phrases in a corpus of texts to identify the relation-

Received for review: November 27, 2009
Accepted for publication: March 16, 2010

Corresponding Author
Hong-Gee Kim, PhD
Biomedical Knowledge Engineering Laboratory, Seoul National University, 28-22 Yeongeon-dong, Jongno-gu, Seoul 110-749, Korea. Tel: +82-2-740-8796, Fax: +82-2-743-8706, E-mail: hgkim@snu.ac.kr

ships between ideas within the subject areas presented in texts [4]. Since their introduction [5], CWA techniques have been used to explore knowledge structure [6-8], analyze research trends [4,7,9-11], and generate hypothesis and discover knowledge [12-16]. In this study, we utilize CWA for exploring the knowledge structure of Korean medical informatics.

The following questions guided this research: 1) What are the important topics of Korean medical informatics? 2) What are the newly emerging research topics?

This paper begins with data and analysis methods to find the knowledge structure of Korean medical informatics. Then research findings such as important research topics and their contexts, and newly emerging topics are covered. Finally, we discuss the major findings and conclude the paper.

II. Methods

In this study, we adopted a well established co-word analysis protocol. It involves the following steps: 1) select the text corpus for the study; 2) extract and normalize the terms and get term weights; 3) get a term co-occurrence frequency matrix for the corpus; 4) get term-term relatedness; 5) analyze the term-term relatedness matrix, and visualize it. For this study BiKE Text Analyzer (BTA), a Java application was used.

1. Data Collection and Treatment

The time window for the target data was set as the 14 years from 1995 to 2008. We collected 1,075 papers' titles and abstracts published in the journal and symposia of Korean Society of Medical Informatics. Abstract-free papers were excluded from the corpus to have 915 for analysis. For consistency, Korean titles and abstracts of 295 papers was translated into English through Google Translator Toolkit. Then, the English terms were corrected, which have different meanings from original ("화상진료시스템" ==> "Burn care system" --> "telemedicine system"), were Romanized as pronounced ("기록지"==> "girokji" --> "record"), and have not been marked as medical terminology ("검사" ==> "inspection" --> "lab test").

In a co-word analysis, the critical step is to create a list of terms that constitute the variables for analysis. Our variables were created from a combination of the sources including: 1) symposium topic lists; 2) author keywords and biomedical informatics keywords from Thomson Web of Science; and 3) MeSH descriptors. Research topics were collected from call for paper topics and session titles in all symposia. The collected terms were appended to the Vocabulary Manager of BTA. The Vocabulary Manager automatically erases duplicate terms and manages n-grams (up to 5-grams), and it allows users to load new vocabularies and input new terms (Figure 1).

2. Term Extraction and Normalization

Topics are difficult to represent with single words because they often have more than one meaning. In most cases, topics are appropriately described in multi-word phrases, which, especially in research domains, are much more interpretable [17]. In this study, we view a topic as a multi-word phrase rather than a single word. Before we proceeded to extract

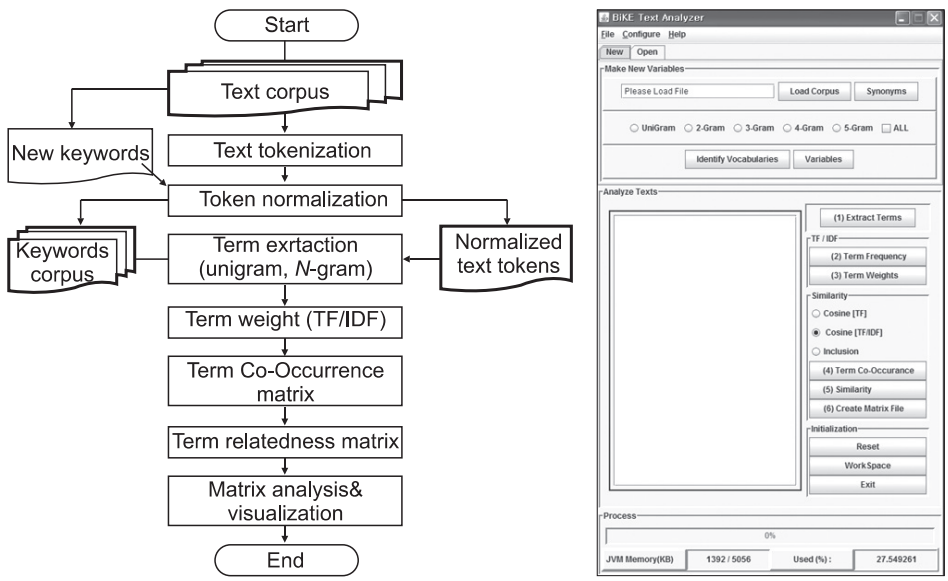


Figure 1. Analysis workflow and BiKE Text Analyzer.

phrases and obtain precise variables, tokenized words' plural forms were singularized and their synonyms controlled. We adopted a less strict normalization strategy for words: tokenized words' plural forms were singularized (eg, records to record) and abbreviations were controlled with synonym lists (eg, HER = electronic health record). From paper corpora, we extracted n-gram terms (from 2-grams to 5-grams) as variables using BTA (total number of n-gram terms = 2,954). The most frequently occurring term was "information system" (term frequency = 533). After excluding the terms that occurred less than 5 times and inappropriate to be a variable (eg, "two type"), the term variables for analysis became 748.

3. Term Weight

Since not all terms may have the same importance in a document, the weight of each term was calculated by multiplying term frequencies (TF) by the inverse document frequency (IDF) for that term.

$$W_{i,j} = TF \times IDF = \frac{f_{i,j}}{\sum_k n_{k,j}} \times log \frac{N}{n_i}$$

where $f_{i,j}$ is the number of times the term i appears in the document j, $\sum_k n_{k,j}$ is the total number of terms in the document d, N is the total number of documents, and n_i is the total number of documents containing the term i.

4. Term Co-occurrence and Closeness Matrix

The co-occurrence analysis approach quantifies term co-occurrences in documents. It assumes that the more frequently two terms appear together in the same document, the sooner they will be identified as being closely related [18]. BTA generates a term co-occurrence frequency matrix (748 × 748),

Table 1. Authority weights of top 50 important topics in Korean medical informatics

Rank	Weight	Topic	Rank	Weight	Topic
1	0.176	Information system	26	0.078	Nursing information
2	0.131	Decision support system	27	0.078	Delivery system
3	0.127	Decision support	28	0.076	Nursing process
4	0.124	Hospital information system	29	0.076	Picture archiving
5	0.119	Clinical decision support	30	0.073	Extensible markup
6	0.118	Clinical decision support system	31	0.073	Extensible markup language
7	0.118	Clinical decision	32	0.071	Nursing practice
8	0.110	Medical information	33	0.071	Nursing information system
9	0.110	Hospital information	34	0.070	Medical information system
10	0.109	Support system	35	0.070	Hospital management
11	0.108	Health care	36	0.070	Health information
12	0.105	Patient care	37	0.070	Electronic health record system
13	0.105	Medical record	38	0.069	Health information system
14	0.099	Electronic health	39	0.069	Web based
15	0.097	University hospital	40	0.069	User satisfaction
16	0.096	Electronic medical record	41	0.068	Archiving and communication system
17	0.093	Electronic health record	42	0.068	Picture archiving and communication system
18	0.091	Health record	43	0.068	System development
19	0.089	Medical record system	44	0.067	Visual basic
20	0.085	Communication system	45	0.067	Discharge summary
21	0.085	The internet	46	0.067	Care delivery
22	0.084	Electronic medical record system	47	0.066	Nursing activity
23	0.082	Management system	48	0.066	Medical service
24	0.081	Nursing care	49	0.066	Medical care
25	0.080	User interface	50	0.066	Nursing record

doi: 10.4258/hir.2010.16.1.52

and then transforms the matrix into a cosine correlation matrix, where each cell indicates the relative closeness of each term pair with a 0-1 range. The cosine measure is defined as the cosine of the angle enclosed between two term vectors x and y:

$$\text{Cosine}(x, y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{(\sum_{i=1}^{n} x_i^2) \times (\sum_{i=1}^{n} y_i^2)}}$$

5. Converting Matrix to Network, Visualization, & Analysis

The term-term closeness matrix was converted to a social network showing the binary relationships between any two terms. This network provides a useful medium for representing the topical structure of Korean medical informatics in a concise and intuitive manner. Pajek software was used for network visualization and analysis. The node size equals to the logarithm of the term frequency, and the thickness of the lines indicates the cosine value (closeness) between a pair of terms.

III. Results

1. Top Research Topics

Some terms have links with many terms; their network of co-occurrences is quite extensive and occupies a central position in a field. To identify the important research topics in Korean medical informatics, authority and hub scores were calculated for each topic. In social network analysis, if a vertex points to many good authorities, it is a good hub. And if a vertex is pointed to by many good hubs, it is a good authority [19]. The authority scores and hub scores of topics are rendered as:

$$\textit{Authority Score}(T_i) = \sum_{T_j \to T_i} \textit{Hub Score}(T_j)$$

$$\textit{Hub Score}(T_i) = \sum_{T_i \to T_j} \textit{Authority Score}(T_j)$$

The authority score of a topic i (T_i) equals the sum of the hub scores of all topics (T_j) that point to it. The hub score (T_i) of a topic i (T_i) equals to the sum of the authority scores of all topics that it points to. Authority scores mutually reinforce hub scores. As shown in Table 1, we extracted the 50 most important topics in Korean medical informatics during the past 14 years (1995-2008). Table 2 shows that top 19 topics with high authority score occupy about 3.5% of 748 topics.

As shown in Figure 2, the 50 most important topics were grouped into 12 clusters: *information system*, *decision sup-*

Table 2. Statistics of 748 research topics in Korean medical informatics

Vector values	Frequency	Freq%	CumFreq	CumFreq%
(...0.0011)	0	0.000	0	0.000
(0.0011 ...0.0449)	619	82.754	619	82.754
(0.0449 ...0.0887)	110	14.706	729	97.460
(0.0887 ...0.1326)	18	2.406	747	99.866
(0.1326 ...0.1764)	1	0.134	748	100
Total	748	100		

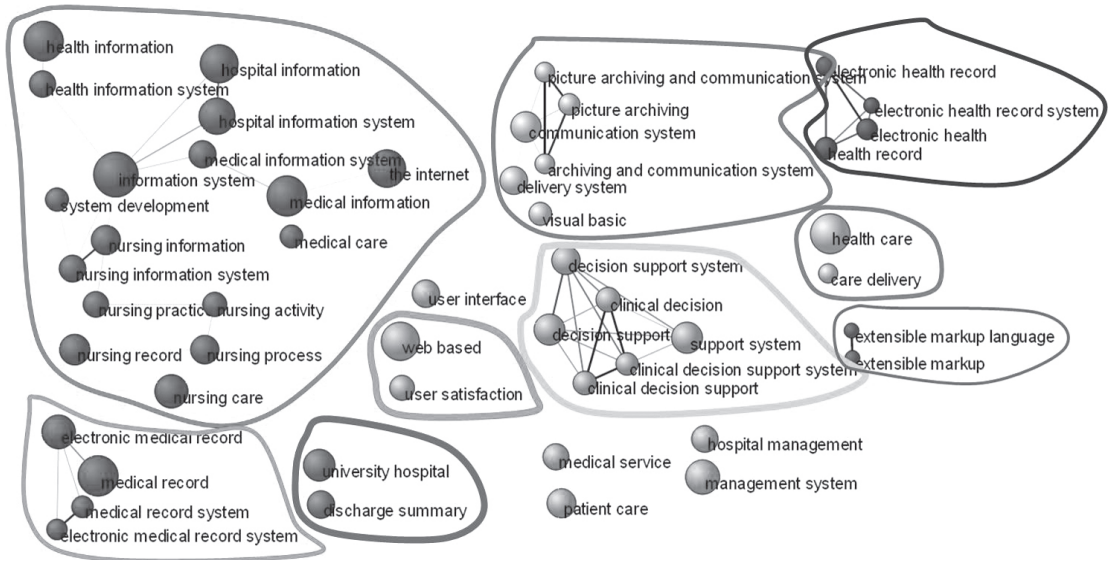


Figure 2. Top 50 important topics of the Korean medical informatics. The edge values lower than cosine 0.15 in the original network were removed and clustered with component (*tf* ≥ *5; N = 50; cosine* ≥ *0.15; k-component* ≥ *1; component = 9*). Nine components generate 12 groups. Contour lines were drawn by hand.

port system, picture archiving and communication system, electronic health record, electronic medical record, XML, university hospital, healthcare, and so on. It is comparable to the findings of a study in global scale, in which *physician order entry* and *practice guideline* are one of the major topics as shown in Figure 3. It is also interesting that the *information system* group (upper right corner of Figure 2) is closely associated with nursing science topics such as *nursing record, nursing process, nursing activity, nursing information*, and so on.

2. Research Topic Trends

To investigate newly emerging topics, for each 3-year period (2 years for 2007 and 2008), we calculated term frequencies and identified the topics which represented the lowest 10% in the low frequency group in the preceding period(s), and which also remained in the highest 10% (5% in the years 2007-2008) in the high frequency group in the following periods.

During the past 14 years (1995-1998), *information system, medical record, hospital information, management system, hospital information system, health information, web based, and information technology* have been occupying top 5% of Korean medical informatics research topics. Some of the newly emerging research topics during the years 1998-2000 are *nursing informatics* and *electronic medical record*, during the years 2001-2003 are consumer health. *Electronic health record system, information extraction,* and *ubiquitous healthcare* are newly popular topics during 2004-2006. During 2007-2008 *oriental nursing, bioinformatics, ubiquitous computing, personal health device* are some of newly emerging topics (Figures 4-6).

IV. Discussion

The social network analysis of research topics communicated through the KOSMI journal and symposia provides a systematic overview on the knowledge structure of Korean medical informatics. From our analysis it is supposed that Korean medical informatics has been paying attention to the information artifacts such as EHR (EMR), CDSS, PACS and so on, but less to methodological topics (eg, machine learning, natural language processing, support vector machine) and their applications (eg, *computerized physician order entry, clinical practice guideline*) which are some of major topics in the global scale. Since the early 2000s, bioinformatics related topics (eg, *expression data*) have been emerging in the global scale [20], whereas in Korean only since the years of 2007-2008 (eg, *bioinformatics data*). These suggest that Korean medical informatics should be equipped with more

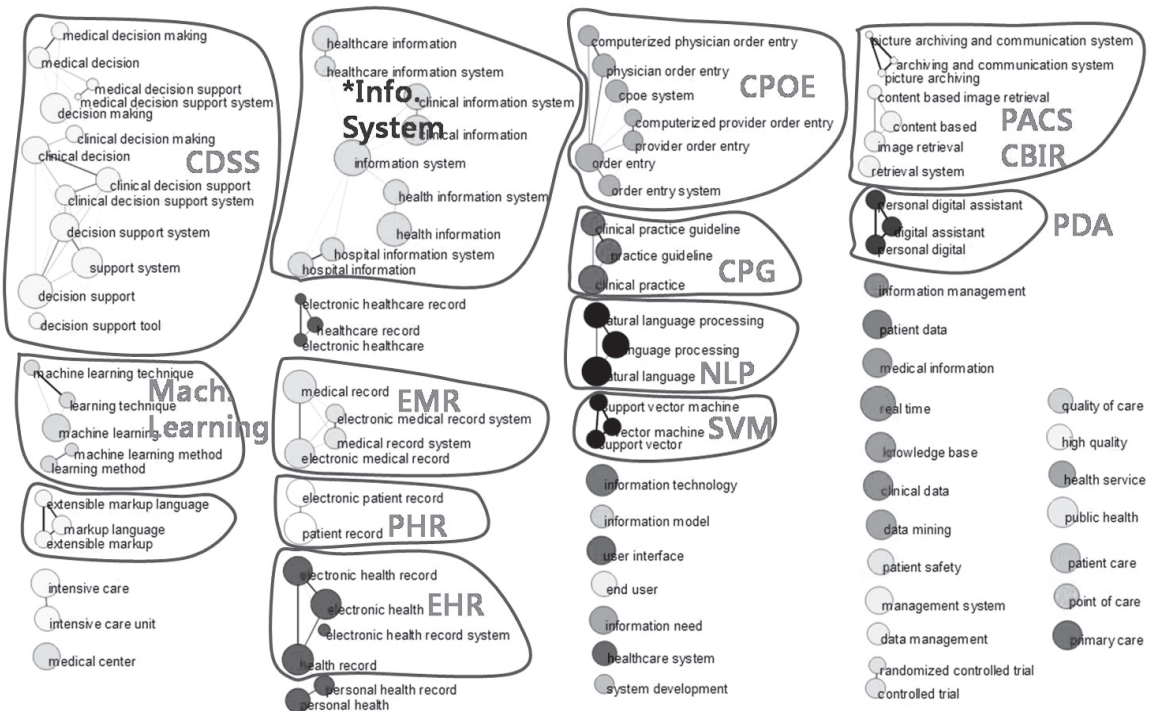
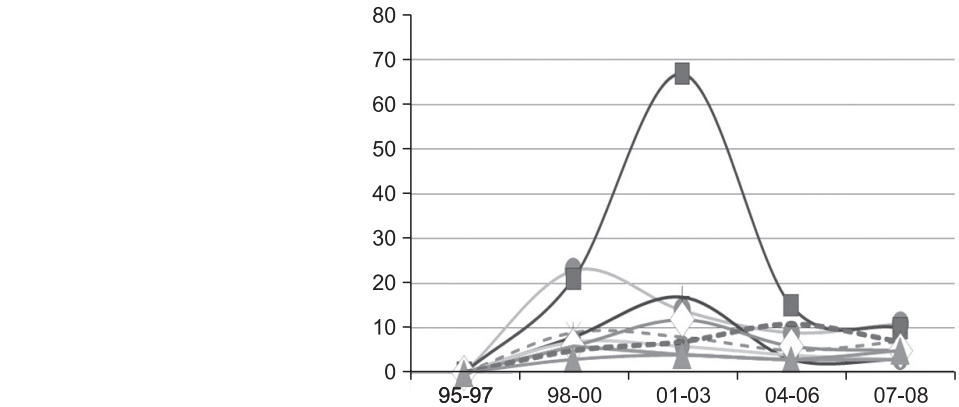


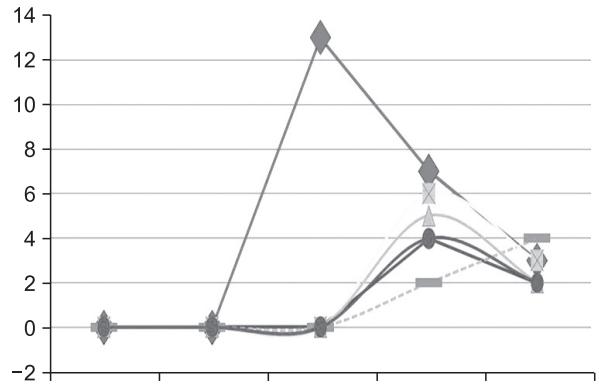
Figure 3. Top 100 important topics of medical informatics in global scale (*tf* ≥*10; N=100; cosine* ≥ *0.1;* κ*-component* ≥*1; component=40*). Adapted from [20].

 doi: 10.4258/hir.2010.16.1.52



	95-97	98-00	01-03	04-06	07-08
Nursing informatics	0	23	14	9	11
Electronic medical record	0	21	67	15	10
Logistic regression	0	9	8	5	7
Health service	0	8	17	3	3
Picture archiving	0	7	6	4	3
Oriental medicine	0	6	12	6	5
Picture archiving and communication system	0	6	4	3	3
Archiving and communication system	0	6	4	3	3
Internet health information	0	5	7	11	7
3D image	0	3	4	3	5

Figure 4. Newly emerging research topics in Korean medical informatics during the years 1998-2000.



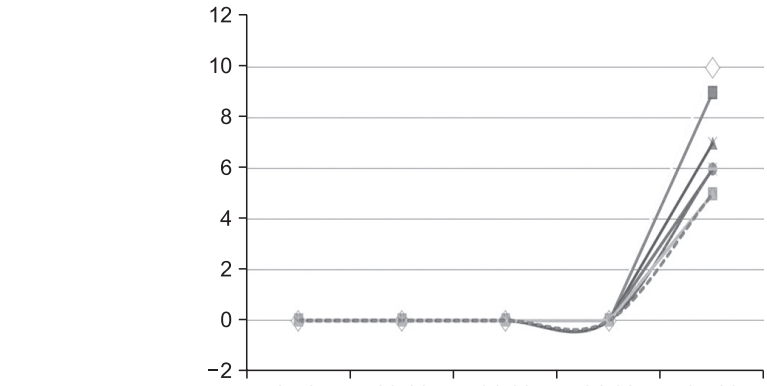
	95-97	98-00	01-03	04-06	07-08
Consumer health	0	0	13	7	3
Sysmptom management	0	0	0	4	2
Information extraction record system	0	0	0	2	4
Electronic health record system	0	0	0	6	3
Ubiquitous healthcare	0	0	0	5	2
Naive bayes	0	0	0	4	2

Figure 5. Newly emerging research topics in Korean medical informatics during the years 2001-2003 and 2004-2006.

rigorous methodologies and pay more attention to bioinformatics.

Several contributions of this study are notable. This study provides topic networks for systematic understanding of Korean medical informatics, and helps to gain a first insight into the main research interest in Korea. Our research trend analysis also helps to decide which technologies and themes should be included in medical informatics curriculum to meet ever-changing learners' needs. In addition, the methodology used for this study has implications in hypothesis generation and knowledge discovery which were demonstrated in many studies [21]. One may analyze, for example, the relationship between chief complaint and disease.

There are technical limitations to our study. Several advantage of using the N-grams as text analysis unit can also be viewed as disadvantage; N-gram may not catch important topics with single word topics (ie, *ontology*). In our study, however, even the combination of single-word and N-gram



	95-97	98-00	01-03	04-06	07-08
Oriental nursing	0	0	0	0	10
Disability and health	0	0	0	0	9
Interpretation rate	0	0	0	0	7
Risk assessment	0	0	0	0	7
Bioinformatics data	0	0	0	0	6
Bioinformatics data source	0	0	0	0	6
Medical equation	0	0	0	0	6
Surface reconstruction	0	0	0	0	6
Ubiquitous computing	0	0	0	0	6
Exercise prescription	0	0	0	0	5
Fall risk assessment	0	0	0	0	5
Health screening	0	0	0	0	5
Personal health device	0	0	0	0	5
Text learning	0	0	0	0	5

Figure 6. Newly emerging research topics in Korean medical informatics during the years 2007-2008.

topics did not result in desirable product since they hardly bear context. Sooner they loosed the fine-grained spectacles of topics, and the highly pre-coordinated meaning of the topics. We do not regard topics covered in this study as wholly definitive topics of Korean medical informatics in Korea, nor does corpus used in this study encompass all research efforts in Korea. We would simply claim that the corpus represents those of studies and should suffice for our main purpose.

Further research may include comparative analysis of Chinese, Japanese and Korean medical informatics.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgements

This paper is an extended version of our previous work [22]. This work was supported by the Korean Research Foundation Grant funded by the Korean Government (KRF-2008-562-D00035).

References

1. Hasman A, Haux R. Modeling in biomedical informatics: an exploratory analysis part 1. Methods Inf Med 2006; 45: 638-642.
2. Hasman A, Haux R. Modeling in biomedical informatics: an exploratory analysis part 2. Int J Med Inform 2007; 76: 96-102.
3. Maojo V, Kulikowski CA. Bioinformatics and medical informatics: collaboration on the road to genomic medicine? J Am Med Inform Assoc 2003; 10: 515-522.
4. He Q. Knowledge discovery through co-word analysis. Lib Trends 1999; 48: 133-159.
5. Callon M, Courtial JP, Turner WA, Bauin S. From translations to problematic networks: an introduction to co-word analysis. Soc Sci Inform 1983; 22: 191-235.
6. Morris TA. Structural relationships within medical informatics. Proc AMIA Symp 2000; 590-594.
7. Bansard JY, Rebholz-Schuhmann D, Cameron G, Clark D, van Mulligen E, Beltrame E, Barbolla E, Martin-Sanchez Fdel H, Milanesi L, Tollis I, van der Lei J, Coatrieux JL. Medical informatics and bioinformatics: a bibliometric study. IEEE Trans Inf Technol Biomed 2007; 11: 237-243.

 doi: 10.4258/hir.2010.16.1.52

8. Mane KK, Börner K. Mapping topics and topic bursts in PNAS. Proc Natl Acad Sci U S A 2004; 101 Suppl 1: 5287-5290.
9. Garfield E. Mapping the world of biomedical engineering: Alza lecture (1985). Ann Biomed Eng 1986; 14: 97-108.
10. Pickens J, MacFarlane A. Term context models for information retrieval. In: Proceedings of 15th ACM International Conference on Information and Knowledge Management; 2006 Nov 5-11; Arlington, VA. p.559-560.
11. Rebholz-Schuhman D, Cameron G, Clark D, van Mulligen E, Coatrieux JL, Del Hoyo Barbolla E, Martin-Sanchez F, Milanesi L, Porro I, Beltrame F, Tollis I, Van der Lei J. SYMBIOmatics: synergies in medical informatics and bioinformatics - exploring current scientific literature for emerging topics. BMC Bioinformatics 2007; 8(Suppl 1): S18.
12. Stegmann J, Grohmann G. Hypothesis generation guided by co-word clustering. Scientometrics 2003; 56: 111-135.
13. Swanson DR. Undiscovered public knowledge. Libr Q 1986; 56: 103-118.
14. Swanson DR. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspect Biol Med 1986; 30: 7-18.
15. Stegmann J, Grohmann G. Transitive text mining for information extraction and hypothesis generation [Internet]. 2005 [cited 2008 Jul 10]. Available from: http://arxiv.org/abs/cs/0509020.
16. Swanson DR. Migraine and magnesium: eleven neglected connections. Perspect Biol Med 1988; 31: 526-557.
17. Mann GS, Mimno D, McCallum A. Bibliometric impact measures leveraging topic analysis. In: Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries; 2006 June 11-15; Chapel Hill, NC. p65-74.
18. Noyons E. Bibliometric mapping of science in a policy context. Scientometrics 2001; 50: 83-98.
19. Kleinberg JM. Authoritative sources in a hyperlinked environment. J ACM 1999; 46: 604-632.
20. Jeong S, Kim HG. Intellectual structure of biomedical informatics reflected in scholarly events. Scientometrics. Epub 2010 Feb 11. DOI: 10.1007/s11192-010-0166-z.
21. Bekhuis T. Conceptual biology, hypothesis discovery, and text mining: Swanson's legacy. Biomed Digit Libr 2006; 3: 2.
22. Jeong S, Lee SK, Kim HG. Knowledge structure of Korean medical informatics. In: Proceedings of CJKMI Fall Conference; 2009 Oct 30-31; Daejeon, KR. p49-51.