Text Data Mining of the Electronic Medical Record of the Chronic Hepatitis Patient

Muneo Kushima, *Member, IAENG*, Kenji Araki, Muneou Suzuki, Sanae Araki and Terue Nikama

***Abstract*—This research used a text data mining technique to extract useful information from nursing records within Electronic Medical Records. Although nursing and passage records provide a complete account of a patient's information, they are not being fully utilized. Such relevant information as laboratory results and remarks made by doctors and nurses is not always considered. Knowledge concerning the condition and treatment of patients has been determined in a twofold manner: a text data mining technique identified the relations between feature vocabularies seen in past chronic hepatitis in-patient records accumulated on the University of Miyazaki Hospital's Electronic Medical Record, and extractions were made. The qualitative analysis result of in-patient nursing and passage records used a text data mining technique to achieve the initial goal: a visual record of such information. The analysis discovered vocabularies relating to proper treatment methods and concisely summarized their extracts from in-patient nursing and passage records. Important vocabularies that characterize each nursing and passage record were also revealed.**

***Index Terms*—text data mining, electronic medical record, chronic hepatitis, nursing record, passage record, KeyGraph**

I. Introduction

AN Electronic Medical Record (EMR) records information on patients by computers instead of by paper. Not only the data but also the entire management system may be called EMR. The expected effect simplifies the entire process of hospital management and improves medical care [1],[2]. Because data are managed electronically, input data can be easily managed and compared with medical records on paper [3]. Information can be easily shared electronically [4],[5]. On the other hand, falsification must be prevented and the originality of the data must be guaranteed. Data mining searches for correlations among items by analyzing a great deal of such accumulated data as sales data and telephone call histories. Text data mining resembles data mining because it extracts useful knowledge and information by analyzing the diversified viewpoints of written data [6].

Recently, interest has risen in text data mining because it uncovers useful knowledge buried in a large amount of accumulated documents [7],[8]. Research has started to apply text data mining to medicine and healing [9],[10],[11],[12],[13]. In addition, the speed of electronic medical treatment data is accelerating because of the rapid informationization of medical systems, including EMRs. Recently, research on data mining in medical treatment that aims for knowledge and pattern extraction from a huge accumulated database is increasing. However, many medical documents, including EMRs that describe the treatment information of patients, are text information. Moreover, mining such information is complicated. The data arrangement and retrieval of such text parts become difficult because they are often described in a free format; the words, phrases, and expressions are too subjective and reflect each writer [14]. Perhaps in the future, the text data mining of documents will be used for lateral retrieval, even in the medical treatment world, not only by the numerical values of the inspection data but also by computerizing documents [15],[16]. In this study, we chose in-patient nursing and passage records from among nursing and passage records preserved by the EMR system at the University of Miyazaki Hospital. Sentences were analyzed into morphemes, and the relations among feature vocabularies were analyzed using KeyGraph [17],[18],[19],[20]. Then we visualized this information.

II. Izanami(EMR)

When the medical information system was updated on May, 2006, the University of Miyazaki Hospital introduced a package version of the EMR system called Integrated Zero-Aborting NAvigation system for Medical Information (IZANAMI), which was developed in collaboration with a local IT company. The recorded main data include a patient's symptoms, laboratory results, prescribed medicines, and the tracking of the changed data. Cases that make both the images of X-rays and the appended material electronic are not infrequent either. If a network is used, EMR can be shared not only in one hospital but also among two or more hospitals. IZANAMI has a unique feature that is different from those being operated at many other university hospitals [21],[22],[23].

First, the electronic card systems used so far in university hospitals were all developed by major medical system venders, but IZANAMI was developed in collaboration with local companies. The advantages of collaboration with local companies included prompt communication and lower costs.

Second, we focused on performance, especially the speed at which the screen opens.

Third, we aimed for a useful system to improve management, reflecting a request by the University of Miyazaki Hospital after it was incorporated.

Manuscript received Dec. 8, 2011; revised Dec. 31, 2011.

M. Kushima is a doctoral candidate in the Faculty of Medicine at the University of Miyazaki Hospital. (e-mail: kushima@fc.miyazaki-u.ac.jp, (http://mit.med.miyazaki-u.ac.jp/).

K. Araki, M. Suzuki, S. Araki, and T. Nikama are with the Faculty of Medicine, University of Miyazaki Hospital.

5200, Kihara, Kiyotake-cho, Miyazaki-shi 889-1692 Japan.

Tel: +81-985-85-9057, Fax: +81-985-84-2549.

We made the medical staff concretely aware of the cost and made the management analysis system work closely with the EMR system and showed its cost when the system was ordered. In the several years since IZANAMI was introduced, there have been no big problems or confusion involved in its operation. IZANAMI has received high praise from doctors and other medical personnel and has also attracted many visitors from outside the hospital.

III. Text Datamining Application to Medicine

Text data mining is often used to analyze information hidden in the text of a document and to extract key words, phrases, and even concepts from written documents. Text data mining or data mining, which is roughly equivalent to text analytics, refers to the process of deriving high-quality information from texts. Text data mining usually structures the input text (often by parsing, adding derived linguistic features, removing others and insertion into a database), deriving patterns within the structured data, and finally evaluating and interpreting the output. Fig. 1 shows the process of text data mining. Two particular aspects should be considered when applying text data mining to a medical context. Two particular aspects should be considered when applying text data mining to a medical context. Second, final decisions can be obtained regarding courses of treatment.

One difficulty with applying text data mining to medicine is the entire process of identifying symptoms for understanding the associated risks while taking appropriate action.

IV. KeyGraph

We applied KeyGraph to the text data mining technique [25],[26],[27],[28]. We also applied it for extracting key words.

A. Example of KeyGraph Performance

Figure 2 shows an image of KeyGraph. Figure 3 shows an example when it is applied to text data.

- ・Black nodes indicate items that frequently occur in a data set.
- ・White nodes indicate the items that occur less frequently overall but frequently occur with black nodes in a data set.
- ・Double-circled nodes indicate items whose co-occurrence frequency with black nodes is especially high. Double-circled nodes are considered keywords.
- ・Links indicate that the connected item pair frequently co-occurs in a data set.
- ・Solid lines form a foundation, which dotted lines connect.

Foundations, which are circles of dotted lines, are obtained from the text data. In Fig. 3, two foundations have strong linkages with event-sets: {doctor, surgery, patient, operation}, and {cancer, medicine, injection}.

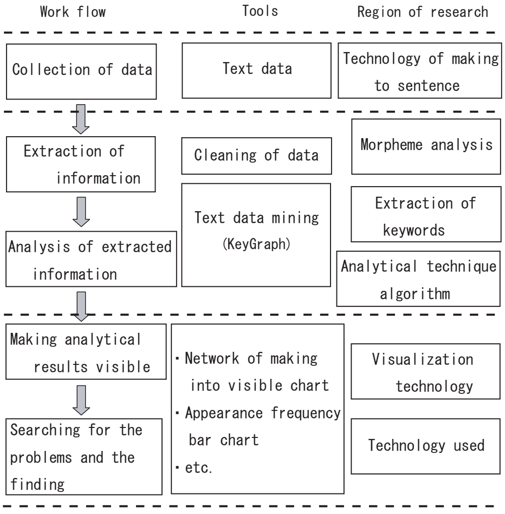
Work flow
Tools
Region of research
Collection of data
Text data
Technology of making to sentence
Extraction of information
Cleaning of data
Morpheme analysis
Text data mining (KeyGraph)
Extraction of keywords
Analysis of extracted information
Analytical technique algorithm
Making analytical results visible
・Network of making into visible chart
・Appearance frequency bar chart
・etc.
Visualization technology
Searching for the problems and the finding
Technology used

Fig.1 Process of text data mining

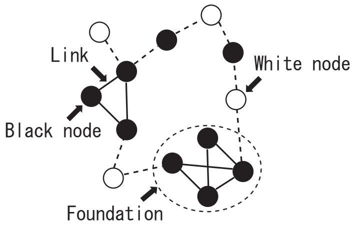

Fig.2 Image graph of KeyGraph

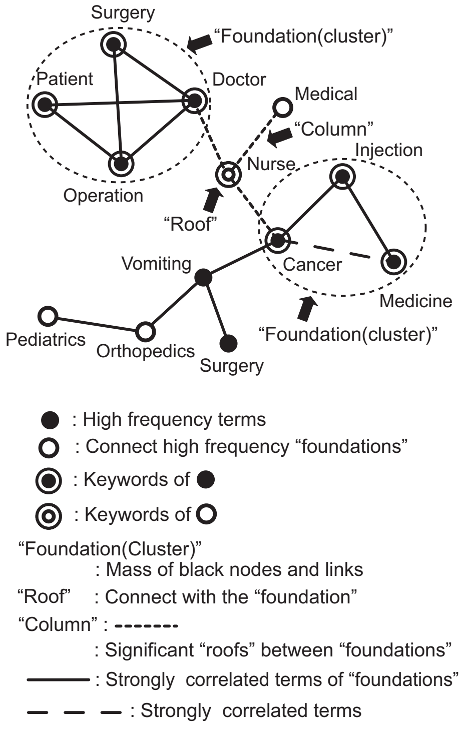
Surgery
"Foundation(cluster)"
Patient
Doctor
Medical
"Column"
Injection
Nurse
Operation
"Roof"
Vomiting
Cancer
Medicine
Pediatrics
Orthopedics
Surgery
"Foundation(cluster)"
● : High frequency terms
○ : Connect high frequency "foundations"
◉ : Keywords of ●
◎ : Keywords of ○
"Foundation(Cluster)" : Mass of black nodes and links
"Roof" : Connect with the "foundation"
"Column" : ------- : Significant "roofs" between "foundations"
——— : Strongly correlated terms of "foundations"
— — — : Strongly correlated terms

Fig.3 KeyGraph example when applied to text data

B. Outline of KeyGraph

Instead of giving a detailed explanation of KeyGraph, we briefly outline it here. KeyGraph consists of three major components derived from building construction metaphors. Each component is described as follows:

1) Foundations: sub-graphs of highly associated and

frequent terms that represent basic concepts in the data. A foundation is defined as a cluster that consists of black nodes linked by solid lines. The foundations are underlying common contexts because they are formed by a set of items that frequently co-occur in the data set.

2) Roofs: terms that are highly associated with foundations.

3) Columns: associations between foundations and roofs that are used for extracting keywords, i.e., the main concepts in the data. A column is a dotted line that connects foundations. Although the common context represented by a foundation is widely known, the context represented by a column is not. Columns are important because they connect two common contexts in items that do not frequently occur.

V. Analysis Results

In this paper, EMR data were collected the chronic hepatitis in-patient nursing and passage records from among nursing and passage records preserved by the EMR system at the University of Miyazaki Hospital, and the nursing records from August 2007 to July 2009 were used. The following analysis results are shown:
(patient 145: 1128 total document (nursing records 173 document, passing records 955 document)

1) Nursing record: Fig.4

・foundation 1,2:
Doctors observe the drug injection site at the beginning and end.

・foundation 3:
The agreement is confirmed at the time of drug administration.

The foundations are obtained from the text data with event-sets 1: { foreign, noodle, order, end, left hand, outside, administration, start }, 2: { humerus, enforcement, injection, subcutaneous, right hand , right } and 3: { abdominal, liver, hospitalization, consent, test }.

2) Passge record: Fig.5

・foundation 1:
Method and duration of drug administration

・foundation 2:
Determine whether the administration actually

・foundation 3:
A check for side effects

Nurses are being evaluated for conditions such as dizziness, the doctor can evaluate and understand the laboratory results. Doctors observe whether the drug had positive or negative effect and nurses record the administration of the drug.

The foundations are obtained from the text data with event-sets 1: { subcutaneous, week, morning, annotation, weekly }, 2:{ continuity, antigen, author, change, state } and 3:{ caution, anemia, decrease, leukocyte, individual, platelet, sphere }.

Table I shows the relationship between the top ten words and their frequency in the University of Miyazaki Hospital chronic hepatitis in-patient nursing and passage records. For example, in nursing record, injection, appeared 19 times. Subcutaneous, humerus, test, symptom, mood, please, liver, failure, schedule, outside, complaint, safe, tomorrow, hospitalization, oral, right hand, before, left hand and start also appear often.

VI. Consideration

The following is an overall evaluation of this research:

・Due to KeyGraph analysis, characteristic keywords were extracted from nursing and passage records and the formation of foundations was confirmed.

・Grasping the characteristics of foundations makes it possible to classify keywords and medical practices in each department.

・Note that many keywords are ambiguous and difficult to associate with a certain word.

・It may be possible to visualize the work practiced by nurses and doctors.

・No words or terms considered discriminative or derogatory were found.

・The information selected from the notes written by experienced staff could be made available for education.

Text data mining in general and data analysis of EMRs in particular remains a relatively unexplored field. Greater collaboration between medical and information sectors will improve the technology so that it can be applied in clinical practice.

VII. Conclusion

In the present study, the chronic hepatitis in-patient nursing and passage records were chosen from among the nursing passage records preserved by the EMR of the University of Miyazaki Hospital. Sentences were analyzed into morphemes, and the relations between feature vocabularies were analyzed by a text data mining technique to visualize this information. The result analyzed the qualitative in-patient nursing and passage records using the text data mining technique and achieved our initial goal: a visual record of this information. In addition, this result identified vocabularies relating to the proper methods of treatment, resulting in a concise summary of the vocabularies extracted from the in-patient nursing and passage records. We constructed important vocabularies characterizing each nursing and passage record. This research suggests the fruitful possibility of automatically detecting a disease and classifying it from documents used at medical treatment sites. In the future, using a text data mining approach and laterally processing medical documents will support disease classification by retrieving the examples of similar syndromes. Our approach can also be applied to the discovery of new medical knowledge for new syndrome extraction. Text data mining is expected to become a valuable technique in the analysis of medical documents in the future. We intend to accumulate clinical research data from care cards that evaluate the prognosis, the prognostic factors, the treatment results, and the safety of Medical Treatment Technologies. In the future, this information will be related to cost reduction and improvements in the efficiency and quality of clinical research.

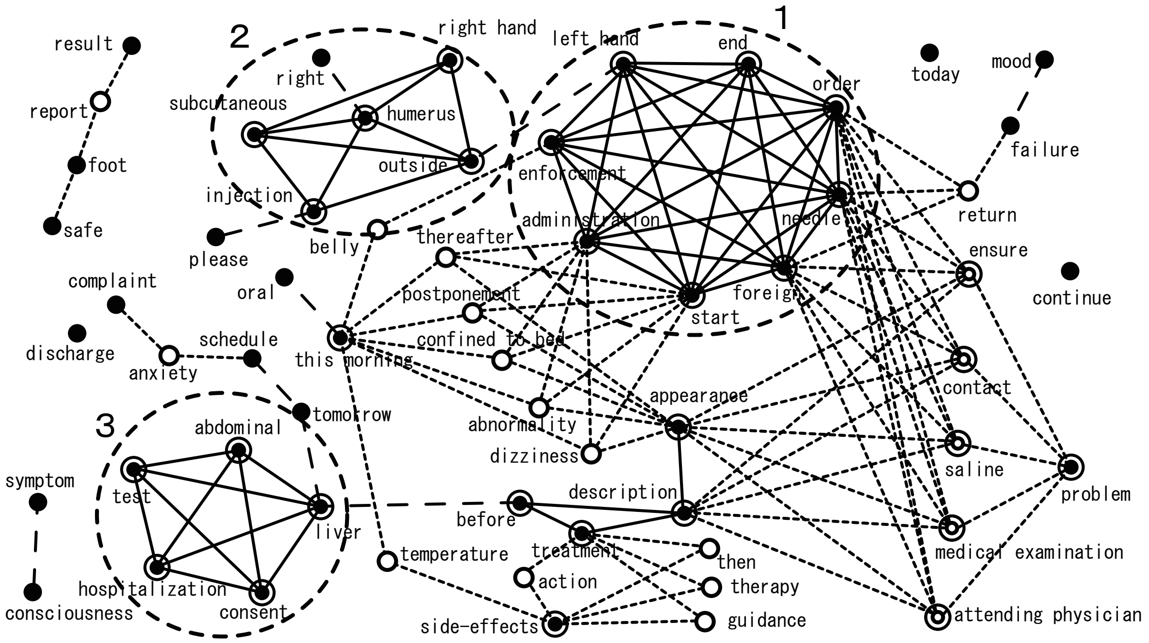

Fig.4 Nursing record

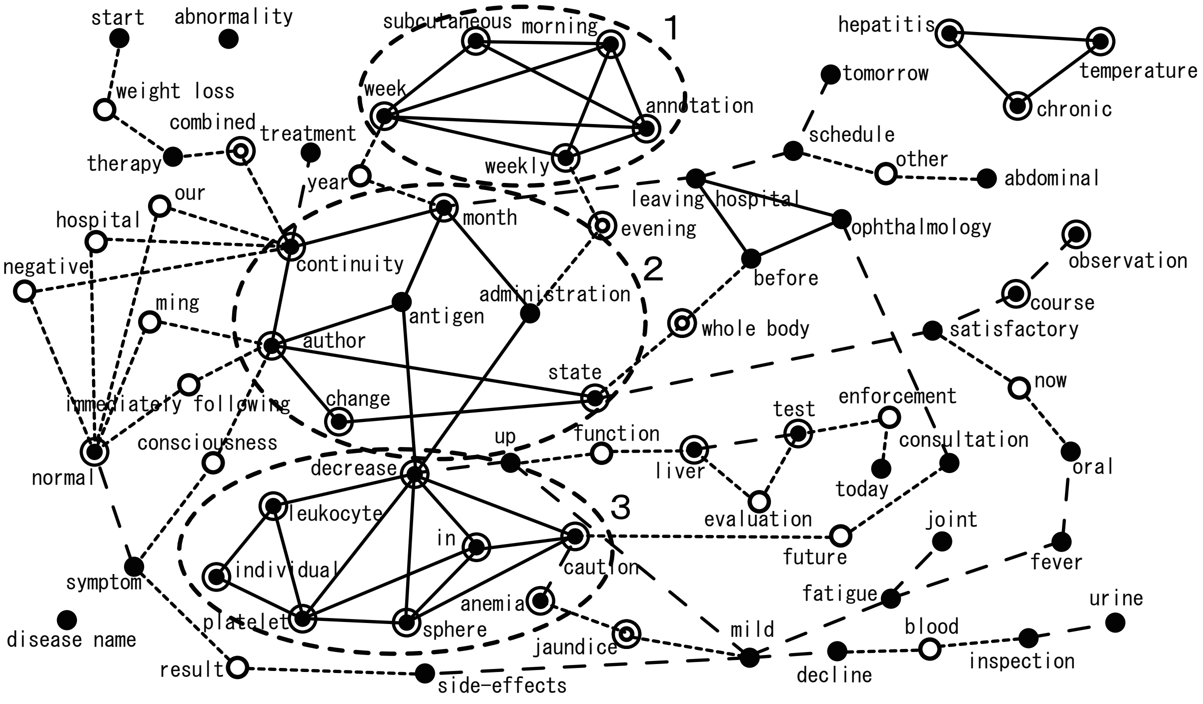

Fig.5 Passage record

Acknowledgment

The authors thank the members of the Medical Information Department at the University of Miyazaki Hospital.

References

[1] D.A. Ludwick and J. Doucette, “Adopting electronic medical records in primary care: Lessons learned from health information systems implementation experience in seven countries,” *International Journal of Medical Information*, no. 78, pp. 22-31, 2009.

[2] Y. Kinosada, “Bio Medical Informatics in the Era of Electronic Patient Record System,” *Japan Association for medical Informatics*, vol. 23, no. 5, pp. 397-405, 2003.

Table I Relationship between top 20 words and frequency

(a) nursing record

Rank	Word	Frequency	Rank	Word	Frequency
1	injection	19	11	outside	6
2	subcutaneous	12	12	complaint	6
3	humerus	11	13	safe	6
4	test	10	14	tomorrow	6
5	symptom	10	15	hospitalization	6
6	mood	9	16	oral	5
7	please	8	17	right hand	5
8	liver	8	18	before	5
9	failure	7	19	left hand	5
10	schedule	7	20	start	4

(b) passage record

Rank	Word	Frequency	Rank	Word	Frequency
1	administration	339	11	platelet	167
2	today	225	12	leaving hospital	156
3	hepatitis	190	13	chronic	154
4	start	189	14	schedule	147
5	fever	188	15	symptom	147
6	liver	182	16	abdominal	146
7	inspection	177	17	oral	136
8	decrease	173	18	ophthalmology	133
9	individual	169	19	side-effects	131
10	course	168	20	sphere	129

[3] M. Suzuki, K. Araki and H. Yoshihara, “Audit Report of the Medical Information System,” *Japan Association for medical Informatics*, vol. 22, no. 4, pp. 347-353, 2002.

[4] K. Yamamoto, S. Matsumoto, H. Matsuda, H. Tada, A. Matsuyama, K. Yanagihara, S. Teramukai and M. Fukushima, “Development and Prospects of Data Capture System for Clinical Study by the Secondary Use of Electronic Medical Records,” *Japan Association for medical Informatics*, vol. 27, no. 2, pp. 211-218, 2007.

[5] Y. Matsumura, H. Nakano, H. Kusuoka, K. Park, M. Matsuoka, H. Oshima, M. Hayakawa and H. Takeda, “Clinic Hospital Cooperation System Based on The Network Type Electronic Patient Record,” *Japan Association for medical Informatics*, vol. 22, no. 1, pp. 19-26, 2002.

[6] S. Murayama, K. Okuhara and H. Ishii, “Innovation in Manufacturing Premise by New Finding Obtained from Accident Relapse Prevention Report,” *Proceedings of The 13th Asia Pacific Management Conference*, Melbourne, Australia, vol. 13, pp. 1124-1129, 2007.

[7] Y. Takahashi, K. Miyaki, T. Shimbo and T. Nakayama, “Text-mining with Network Analysis of News About Asbestos Panic,” *Japan Association for medical Informatics*, vol. 27, no. 1, pp. 83-89, 2007.

[8] M. Usui and Y. Ohsawa, “Chance Discovery for Decision Consensus in Company - Touchable, Visible, and Sharable for a Textile Manufacture Company -,” *Japan Society for Fuzzy Theory and Intelligent Informatics*, vol. 15, no. 3, pp. 275-285, 2004.

[9] H. Abe, S. Hirano and S. Tsumoto, “Mining a Clinical Course Knowledge from Summary Text,” *The 19th Annual Conference of the Japanese Society for Artificial Intelligence*, vol. 1F4-04, pp. 1-2, 2005.

[10] Y. Kinosada, T. Umemoto, A. Inokuchi, K. Takeda and N. Inaoka, “Challenge to Quantitative Analysis for Clinical Processes by Using Mining Technology,” *Data Science Journal*, vol. 26, no. 3, pp. 191-199, 2006.

[11] H. Ono, K. Takabayashi, T. Suzuki, H. Yokoi, A. Imiya and Y. Satomura, “Classification of Discharge Summaries by Text Mining,” *Japan Association for medical Informatics*, vol. 24, no. 1, pp. 35-44, 2004.

[12] Y. Sato, H. Takeuchi K. Hoshi, N. Uramoto, T. Satoh, N. Inaoka, K. Takeda and N. Yamaguchi, “The Effectiveness of the Text Mining and Similar Document Search System for Evidence-Based Guideline Development,” *Japan Association for medical Informatics*, vol. 24, no. 2, pp. 315-322, 2004.

[13] H. Abe, S. Hirano and S. Tsumoto, “Evaluation Temporal Models Based on Association Mining From Medical Documents,” *Japan Association for Medical Informatics*, vol. 27, no. 1, pp. 33-38, 2007.

[14] K. Izumi, K. Goto and M. Matsui, “Analysis of Financial Markets' Fluctuation by Textual Information,” *The 23rd Annual Conference of the Japanese Society for Artificial Intelligence*, 2009.

[15] M. Kushima, K. Araki, M. Suzuki, S. Araki and T. Nikama, “Graphic Visualization of the Co-occurrence Analysis Network of Lung Cancer in-patient nursing record,” *The International Conference on Information Science and Applications (ICISA 2010)*, pp. 686-693, Seoul, Korea, April. 2010.

[16] M. Kushima, K. Araki, M. Suzuki, S. Araki and T. Nikama, “Analysis and Visualization of In-patients' Nursing Record Using Text Mining Technique,” *International MultiConference of Engineers and Computer Scientists 2011 (IMECS2011)*, vol.2188, pp. 436-441, Hong Kong, March, 2011.

[17] Y. Ohsawa, Benson Nels E. and M. Yachida, “KeyGraph: Automatic Indexing by Segmenting and Unifing Co-occurrence Graphs,” *The Institute of Electronics, Information and Communication Engineers*, vol. J82-D-I(2), pp. 391-400, 1998.

[18] K. Horie, Y. Ohsawa and N. Okazaki, “Products Designed on Scenario Maps using Pictorial KeyGraph,” *Proceedings of the 5th WSEAS International Confernce on Applied Computer Science*, pp. 801-808, Hangzhou, China, April, 2006.

[19] Kyung-Joong. Kim, Myung-Chul. Jung and Sung-Bae. Cho, “KeyGraph-based chance discovery for mobil contents management system," {¥it International Journal of Knowledge and Intelligent Engineering System}, Vol.11, pp. 313-320, 2007.

[20] M. Kushima, K. Araki, M. Suzuki, S. Araki, H. Tamura, K. Tanno, T. Toyama, O. Ishizuka and M. Ikeda, “Network Visualization of Electronic Medical Record (IZANAMI) by Using Mining Technology,” *Technical report of IEICE*, vol. 108, no. 388, CAS2008-97, pp. 187-192, 2009.

[21] Y. Ohsawa, Benson Nels E and M. Yachida, “KeyGraph: Automatic Indexing by Co-occurrence Graph based on Building Construction Metaphor,” *Fifth International Forum on Research and Technology Advances in Digital Libraries (ADL'98)*, 1998.

[22] M. Kushima, K. Araki, M. Suzuki, S. Araki, H. Tamura, K. Tanno, T. Toyama, O. Ishizuka and M. Ikeda, “Analysis of Nurse's Medical Record in the IZANAMI Using Text Mining Method,” *18th International Workshop on Post-Binary ULSI Systems*, Okinawa, Japan, 2009.

[23] M. Kushima, K. Araki, S. Suzuki and S. Araki, “Challenge to Qualitative Analysis for Electronic Medical Record (IZANAMI) Using KeyGraph,” *Asia Pacific Association for Medical Informatics (APAMI)*, Hiroshima, Japan, 2009.

[24] A. Abe, Y. Ohsawa, H. Kosaku, K. Aira, N. Kuwahara and K. Kogure, “Discovery of Hidden Factors for Incident by Kamishibai KeyGraph,” *The 22nd Annual Conference of the Japanese Society for Artificial Intelligence*, vol. 2B2-3, pp. 1-4, 2008.

[25] K. Tsuda and R. Thawonmas, “Visualization of Discussions in Comments of a Blog Entry Using KeyGraph and Comment Scores,” *Proc. Of 4th WSEAS International Conference on E-ACTIVITIES*, Florida, USA, vol. 5, pp. 21-26, 2005.

[26] R. Thawonmas and K. Hata, “Aggregation of Action Symbol Sub-sequences for Discovery of Online-game Player Characteristics Using KeyGraph,” *Proc. of IFIP 4th International Conference on Entertainment Computing (ICEC 2005)*, Sanda, Japan, vol. 3711, pp. 126-135, 2005.

[27] Y. Seo, Y. Iwase and Y. Takama, “KeyGraph-based BBS for Supporting Online Chance Discovery Process,” *SCIS ISIS2006*, vol. FR-I3-2, pp. 1210-1214, 2006.

[28] K. Tsuda and R. Thawonmas, “KeyGraph for Visualization of Discussions in Comments of a Blog Entry with Comment Scores,” *World Scientific and Engineering Academy and Society(WSEAS) Trans, Computers*, vol. 12, no. 4, pp. 1794-1801, 2005.

[29] G. Salton and M. J. McGill, “Introduction to Modern Information Retrieval,” *McGraw-Hill*, 1983.

[30] M. F. Porter, “An Algorithm for Suffix Stripping,” *Automated Library and Informations Systems*, vol. 14, no. 3, pp. 130-137, 1980.

[31] J. Cohen, “Language and Document, Automatic Indexing Terms for Abstracting,” *Journal of American Society for Information Science*, vol. 46, pp. 162-174, 1995.