to effective use of resources as well as improve patient safety as mentioned above. We can certificate each medication and capture those data at the same time, contribute to patient safety and improve health care delivery. To be a trusted system, the systems have to use right information and consider securities of people. Trusted IT system can contribute to patient safety, effective use of blood products, reducing waste that might be essential factors for trusted health care system.

ACKNOWLEDGMENT

This work was supported by Health Labour Sciences Research Grant of Japan.

REFERENCES

[1] J.C. Chan, R.W Chu, and B.W Young, "Use of an electronic barcode system for patient identification during blood transfusion: 3-year experience in a regional hospital," Hong Kong Medical Journal, vol.10, pp166-171, 2004.

[2] A. Davies, J. Staves, J. Kay, A. Casbard, and M.F. Murphy, "End-to-end electronic control of the hospital transfusion process to increase the safety of blood transfusion: strengths and weaknesses," Transfusion, vol.46, pp352-364, 2006.

[3] M.F. Murphy, "Application of bar code technology at the bedside: the Oxford experience." Transfusion, vol.47, pp120-124, 2007.

[4] A. Ohsaka, K. Abe, T. Ohsawa, N. Miyake, S. Sugita, and I. Tojima, "A computer-assisted transfusion management system and changing transfusion practices contribute to appropriate management of blood components," Transfusion, vol.48: pp1730, 2008.

[5] R.W. Askeland, S. McGrane, J.S. Levitt, S.K. Dane, D.L. Greene, J.A. VandeBerg, K.Walker, A. Porcella,L.A. Herwaldt, and L.T. Carmen, "Kemp JD. Improving transfusion safety: implementation of a comprehensive computerized bar code–based tracking system for detecting and preventing errors," Transfusion, vol.48, pp1308, 2008.

[6] R. Davis, B. Geiger, A. Guitierrez, J. Heaser, and D. Veeramani, "Tracking blood products in blood centers using radio frequency identification: a comprehensive assessment," Vox Sanguinis vol.97, pp50-60, 2009.

[7] M. Akiyama, "Risk Management and Measuring Productivity with POAS-Point of Act System-A Medical Information System as ERP (Enterprise Resource Planning) for Hospital Management," Method Inf Med., vol.46, pp686-693, 2007.

[8] M. Akiyama and T. Kondo, "Risk Management and Measuring Productivity with POAS--point of act system," Stud Health Technol Inform. Vol.129, pp208-12, 2007.

[9] S.G. Sandler, A. Langeberq, and L. Dohnalek L, "Bar code technology improves positive patient identification and transfusion safety," Adv Transfusion Safety, vol.120, pp19–24, 2005.

[10] S. Allen, "System Targets Blood-Type Mix-Ups", February 24, 2005. Boston Globe Health/Science; available at http://www.boston.com/news/globe/health_science/articles/2005/02/24/system_targets_blood_type_mix_ups/

[11] R.W. Askeland, S. McGrane, J.S. Levitt, S.K. Dane, D.L. Greene, J.A. VandeBerg, K. Walker, A. Porcella, L.A. Herwaldt, L.T. Carmen, and J.D. Kemp, "Improving transfusion safety: implementation of a comprehensive computerized bar code–based tracking system for detecting and preventing errors," Transfusion, vol.48, pp.1308-1317, 2008.

[12] D. Watson, J. Murdock, C. Doree, M. Murphy, M. Roberts, A. Blest, and S. Brunskill, Blood transfusion administration one- or two- person checks, "which is the safest method? " Transfusion. vol.48, pp783-789, 2008.

[13] E. J. Thomas , D. M. Studdert, H.R.Burstin, E.J. Orav, T. Zeena, and E.J. Williams, "Incidence and types of adverse events and negligent care in Utah and Colorado," Med Care, vol. 38, pp261–271, 2000.

[14] R. Sharma, S. Kumar, and S.K. Agnihotri, "Sources of preventable errors related to transfusion," Transfus Med Prod, vol.81, pp37–41, 2001.

[15] R. Davis, B. Geiger, A. Guitierrez, J. Heaser, and D. Veeramani, "Tracking blood products in blood centers using radio frequency identification: a comprehensive assessment," Vox Sanguinis, vol. 97, pp50-60, 2009.

[16] S.G. Sandler, A. Langeberq, and L. Dohnalek, "Bar code technology improves positive patient identification and transfusion safety," Adv Transfusion Safety vol.120, pp19–24, 2005.

[17] M. Akiyama, A. Koshio, and N. Kaihotsu, "Analysis of data captured by barcode medication administration system using a PDA; aiming at reducing medication errors at point of care in Japanese Red Cross Kochi Hospital," Stud Health Technol Inform. Vol.160, pp774-8, 2010.

[18] C. Huckvale, J. Carl, M. Akiyama, S. Jaafar, T. Khoja, A. B. Khalid, A. Sheikh, and A. Majeed, "Information technology for patient safety," Qual Saf Health Care (BMJ) vol.19, pp i25-i33, 2010.

International Association for Management of Technology
IAMOT 2011 Proceedings

Detection of Precarious Situations in Medical Care with Mining Track record of Dosing

SATORU YAMAMOTO

Policy Alternatives Research Institutes, The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan
yamamoto@biz-model.t.u-tokyo.ac.jp

YINGZI JIN

The Institute of Engineering Innovation, The University of Tokyo, 2-11-16, Yayoi, Bunkyo-ku, Tokyo, 113-8656, Japan
Yzjin2006@gmail.com

YUTAKA MATSUO

The Institute of Engineering Innovation, The University of Tokyo, 2-11-16, Yayoi, Bunkyo-ku, Tokyo, 113-8656, Japan
matsuo@biz-model.t.u-tokyo.ac.jp

ICHIRO SAKATA

Policy Alternatives Research Institutes, The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan,
isakata@ipr-ctr.t.u-tokyo.ac.jp

MASANORI AKIYAMA

Policy Alternatives Research Institutes, The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan,
makiyama@pp.u-tokyo.ac.jp, poas@mit.edu

Abstract

We propose a new approach to detect the precarious situation in medical care analyzing tracking record. Attention is being drawn to the use of incident reports as a means of increasing patient safety. Research teams are being formed across the world by WHO and in Japan by the Health, Labour and Welfare Ministry. In this instance, what is being emphasized as a major direction for future incident analysis is the assimilation of the existing top-down type class grants and bottom-down ontological construction. In this research, targeting incident case studies collected in Japan, we evaluated the degree of similarities between incident documents obtained bottom-up and the links between existing classes granted top-down. In doing so, we made it possible to evaluate overall similarities regarding incident documents through the method of network analysis. In addition, it became clear that the use of the Cos coefficient or the Jaccard coefficient is appropriate in creating networks.

As a result of this analysis, existing classes correspond comparatively well with the characteristics of reports regarding the abstract and solution; on the other hand, regarding the background, it demonstrated that existing classes are inadequate in representing the characteristics of documents and that there is a need to improve classes. By the way, we can upgrade patient safety and quality of health care service.

Key word: healthcare service, incident report ,eHealth, WHO

1. INTRODUCTION

"In the shadow of every serious accident, there exist 29 times more minor accidents and 300 times more near misses." This principle was published in 1929 by Herbert William Heinrich, an assistant manager in the technology and research division of an American insurance company (Heinrich, 1931). This principle, which hits home the nature of the occurrence of accidents, is taken up in various fields, such as the study of failure, safety engineering, ergonomics, cognitive psychology as well as the study of reliability, and the incident analysis of minor accidents associated with this is recognized as being

important in preventing accidents. Therefore, patient safety with eHealth becomes growing field of research recently (Gaudhi and Lee, 2010, Mcloughlin et al., 2006, Huckvale et al., 2010, Kaushal and Bates, 2002).

In order to eradicate medical malpractice, medical institutions break down barriers between departments, collect and analyze incidents, and work out countermeasures. With this background, the Health, Labour and Welfare Ministry in Japan started the Project to Collect Medical Near-Miss/Adverse Event Information from 2001. This project collects, analyzes, and publishes incident reports. The guidelines for analysis are to calculate the total of each class, including related medical departments, occurrence factors, and time periods, and to root out the causes of accidents.

On the other hand, regarding patient safety, guidelines for the future deployment of incident analysis are set out in WHO's International Classification of Patient Safety (ICPS). ICPS states the necessity of first investigating the adequacy of classes of incident case studies such as those mentioned above, and second, methods of expressing incidents that adequately reflect these classes, i.e., it states the necessity of ontological construction. In this research, in line with WHO guidelines, we conducted an analysis regarding the adequacy of classes in case studies collected in the Project to Collect Medical Near-Miss/Adverse Event Information and the tendencies of description that aim at ontological construction.

In the data provided in the Project to Collect Medical Near-Miss/Adverse Event Information, the abstract, background, and solution for a single case study is described using a free composition format. In addition, in each case study the class of treatment and the class of operation are granted. There is a need to investigate whether classes granted here are in accord with the characteristics of each document item. In order to achieve the above, this research used the techniques of natural language processing and network analysis.

By using natural language processing, an understanding of the tendencies of description as well as guidelines for future ontological construction can be acquired. Moreover, by networking the reports obtained from this, discoveries of overall links that could not be found from comparing only two reports are expected.

2. MEDICAL INCIDENT REPORTS

Here, we will explain approaches and issues relating to medical incidents and characteristics of the data used in this research.

2.1 Overview of Incident Reports Sought by ICPS

ICPS's general description sets out past activities and future guidelines relating to incident reports (WHO). Until now, the main work of ICPS has been the granting and maintenance of classes to accidents by specialists. By granting this kind of top-down "agreed upon class," it becomes possible to convey a summary of incidents and accidents to even those who are not medical specialists. On the other hand, top-down type of classes created from present conditions are not detailed enough to provide satisfactory explanations of the characteristics of individual incident case studies. In addition, as class is granted in advance, opportunities to find valid unknown classes for patient safety are lost. Consequently, ICPS has stated that it will introduce ontological thought as part of future guidelines. Ontology in medicine refers to conducting from the bottom up and based on actual data the construction of methods necessary for describing individual case studies without misinterpretation as well as the discovery of classes of case studies that use these methods. ICPS indicates that the granting

of top-down type categories by specialists as well as the granting of information that uses bottom-up type of ontology are necessary.

2.2 Collection of Incident Information in Japan

With increasing social demand for the prevention of medical accidents, the Health, Labour and Welfare Ministry started the Project to Collect Medical Near-Miss/Adverse Event Information from 2001 in order to collect and analyze incident case studies and to provide information conducive to medical safety, such as measures for improvements. When the project was first started, a framework was in place in which the Pharmaceuticals and Medical Devices Agency collected incident case studies from participating medical institutions and then reported these case studies to the Health, Labour and Welfare Ministry, following which a Health, Labour and Welfare Ministry study group conducted aggregate calculations and analysis. The 1^{st}–10^{th} collection of incident case studies were conducted following this framework, and information based on these collected incident case studies was provided by the Health, Labour and Welfare Ministry. From 2004, the Japan Council for Quality Health Care took over the collection of incident case studies, collecting case studies from the 11^{th} collection. The results of aggregate calculations and analysis are published on the website of this organization.

2.3. Data Sets

From among incident data provided by the Project to Collect Medical Near-Miss/Adverse Event Information, in this research, we used data relating to medical agents from 2005 to 2010 that was published on the Internet. In order to conduct a detailed analysis, from the case studies provided, we used only 1,067 documents that included the information of the abstract, background, and solution. Each case study is in a free composition format, with the abstract, background, and solution being approximately 300 characters long, respectively. In addition, the two classes of medicine and accident are granted to each case study. With regard to the class of treatment, there are six classes of general drug, preparation of drugs, drowsy of drugs, contraindicated drug, chemo treatment, and other drug; with regard to the class of operation, there are the nine classes of name of drug, amount of drug, regimen, amount and regimen, flow rate, drug sensitivity, diapedesis, forget to dose, and object person. With regard to the class of treatment, as all the classes of operation do not exist, there are 32 cross classes that cross calculate the class of treatment and the negligent class of operation.

When describing accidents in a free composition format, the reporter makes every effort to include every single circumstance. We can say that extracting important information from these circumstances means creating a foothold for a bottom-up type of ontological construction. Results obtained from this and links with classes granted top-down is in accordance with the future guidelines for incident analysis sought by ICPS.

3. NATURAL LANGUAGE PROCESSING

In this research, we extract characteristic words using natural language processing as a first step in extracting important information that characterizes each document with the aim of ontological construction. The links between each document are determined from similarities between characteristic words obtained here. As natural language processing contains a lot of noise, there is a need to conduct preprocessing in order to obtain characteristic words that can be used in determining links. Preprocessing mainly comprises three stages of "breaking down into words in reports,"

"connecting words that have been broken down too much," and "filtering the obtained words." Details of these are set out below.

3.1 Breaking Down into Words

In the first stage of preprocessing, we conducted morphological analysis in order to break down reports into words. Morphological analysis is a method used to delimit each word in the text where words are not delimited by spaces, such as in languages like Japanese (Manning and Schutze, 2002). In this research we used MeCab, one of the most common engines for conducting morphological analysis.

3.2 Connecting Phrases

There is the possibility that words obtained using MeCab are too finely classified to conduct the analysis of links. Therefore, in this research we connected words using the two methods set out below and used them as new words.

First, we connected words using information on the parts of speech. The above-mentioned MeCab not only breaks down words but also grants major classes and minor classes relating to parts of speech. In cases where the minor class of parts of speech of certain words was a suffix and the word before it was a noun, these two words were treated as one word.

Next, we connected words based on the number of word occurrences (Matsuo and Ishizuka, 2002). Let us envisage a situation in which two words—hereafter called A and B—appeared consecutively. If we designate the number of word occurrences in instances where each word is considered separately as n (A), n (B), then the number of word occurrences in which they appear consecutively is expressed as n (A∩B). In cases where n (A∩B)/min (n (A), n (B)) exceeded the threshold value (0.8 in this research) then we treated those two words as one word.

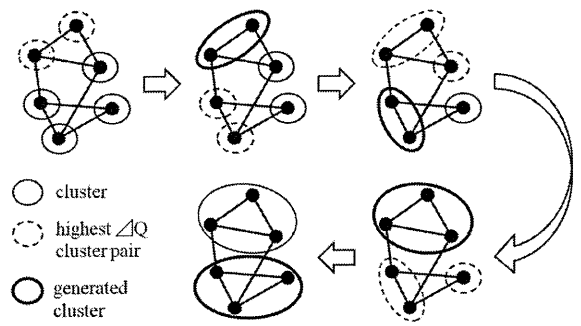

Fig1.Model of Newman clustering

3.3 Filtering Words

Words obtained through the above two processing methods still contain a lot of noise, which can be expected to exert a bad influence on the calculation of links in documents. Thus, it is necessary to select words to be used in calculating links. The following sets out details on filtering.

First, we conducted filtering using the class of parts of speech. As stated above, major and minor classes are granted to words. Nouns were the targets of research on this occasion as major classes of parts of speech. Nominalized verbs, general nouns, and proper nouns were also targeted as minor classes of parts of speech. Focusing solely on nouns is the method generally used in extracting

characteristic words. Moreover, in the case of official documents in Japanese, as many of the verbs are nominalized, a lot of information can be obtained regarding action even if using only nouns.

Next, we conducted filtering based on the frequency of occurrence. In this research, we calculated a value called tfidf from the frequency of occurrence and conducted filtering based on this values. tfidf is one of the most widely used indices in extracting characteristic words for document classes and in cases where a certain word occurs several times in a small number of documents, it is defined so as to enlarge that value (Saltin,). tfidf is calculated using the following formulas.

$$\text{tfidf} = \text{tf} \cdot \text{idf} \qquad (1)$$

$$\text{tf}_i = \frac{n_i}{\sum_k n_k} \qquad (2)$$

$$\text{idf}_i = \log \frac{|D|}{|\{d:d \ni t_i\}|} \qquad (3)$$

Here, ni is the frequency of occurrence of word i, |D| is the total number of documents, and $|\{d: d \ni t_i\}|$ is the number of documents in which word i occurs.

The tfidf of general words occurring in a large number of documents has a tendency to be of a low value, although words among even general words that have an abnormally high tf in some cases exceed the filter effect of idf and assume a high value. In this research, we set the maximum value of tf to 50 and eliminited the noise from words with an abnormally high frequency of occurrence. On the other hand, words that make a small number of appearances also have an extremely small value for idf and, as a result, the tfidf has a tendency to increase. Therefore, this time we treated all tf under 10 as 0.

4. NETWORK ANALYSIS

Network analysis is an extremely effective method of looking at the links between documents (Kajikawa et al., 2007, Uchida et al., 2009, Shibata et al., 2010, Shibata et al., 2011 as examples). By conducting network analysis, the discovery of hidden links between two nodes can be expected. In cases where links between only two documents are considered, even if there are no links, there are instances where overall links can be discovered by creating networks.

4.1 The Creation of Networks

The co-occurrence index is generally used as a method for finding links from the degree of similarities between words in documents. Here, the simplest co-occurrence index for finding links between the two documents A and B is the number of co-occurrence $|A \cap B|$ for two documents. Here, $|A \cap B|$ is the number of characteristic words that exist in A and B. If considered with only $|A \cap B|$, there are problems such as including as many characteristic words as in long texts and links with other documents being displayed as high. Consequently, a number of co-occurrence indices that improve on these points have been proposed, with representative indices including the Jaccard coefficient, the Simpson coefficient, and the Cos coefficient (Rasmussen, 1992). Each formula is shown in (4), (5), and (6) and is generally Simpson coefficient > Cos coefficient > Jaccard coefficient.

$$\text{Jaccard coeffcient: } \frac{|A \cap B|}{|A \cup B|} \qquad (4)$$

$$\text{Cos coeffcinet: } \frac{|A \cap B|}{\sqrt{|A||B|}} \qquad (5)$$

$$\text{Simpson coeffcinet: } \frac{|A \cap B|}{\min(|A|,|B|)} \qquad (6)$$

A link is established between the two documents in the event that these indices exceed the threshold value. The network created changes acutely depending on which of the above indices are selected and how the threshold value is established. As the aim of this research is to investigate the extent to which top-down type of classes are reflected in links in document content that is sought bottom-up, when forming networks, we selected the index and determined the threshold values so as to reflect most the given classes.

4.3 Network Clustering

In this research, other than given classes, we conducted clustering using the Newman method in order to carry out labeling of each document based on networks. The Newman method is a widely used method for clustering networks (Newman, 2004). It can be applied even if the number of clusters is unknown; in recent years, it is also being widely applied in large-scale network analysis, such as SNS and blog, due to it being scalable in regard to increases in node numbers. As shown in formula (7), clustering was conducted by maximizing modularity Q, an index for evaluating the modularity of defined networks.

$$Q = \sum_i (e_{ii} - a_i^2) \qquad (7)$$

Here, element e_{ij} of line e represents the fraction of the total number of edges of the number of edges that connect cluster i with cluster j, and a_i represents the sum of row i of line e. Maximizing Q corresponds with maximizing the disparity between the number of edges that exist within clusters and the number of edges that link clusters together.

5. RESULTS

5.1 Extraction of Characteristic Words

We conducted the extraction of characteristic words using the methods outlined in section 3. Table. 1 and 2 show the number of occurrences by category for the 10 words in which tfidf is at the top. We see from these tables that all characteristic words extracted here occur selectively in each class. In addition, when focusing on the combination of class and occurring word, we see that words that aptly reflect the characteristics of the class are being extracted, such as "injected part" occurring largely in the diapedesis class and "allergia" occurring largely in the drug sensitivity class. On the other hand, as the index in this research uses tfidf as an index of the degree of characteristics, general words do not appear at the top, even if the frequency of occurrence is high. For example, although "patient" had the highest number of occurrences, as this word appears in most case studies, tfidf as an index of class capability of documents takes on a low value and does not appear at the top.

Table.1characteristic word and class of operation

characteristic word	diapedesis	flow rate	object person	forget to dose	composition
insulin	0	0	4	0	0
furosemide	0	1	0	0	0
injected part	14	0	9	0	1
allergia	0	0	5	0	0
setup	0	51	0	0	0
flow rate	1	38	2	0	0
coinjection	0	3	7	0	3
anticancer	3	3	7	0	0
periphery	1	2	10	0	2
set	0	5	1	13	1
characteristic word	**regimen**	**name of drug**	**amount of drug**	**drug sensitivity**	**amount and regimen**
insulin	0	22	16	0	0
furosemide	4	13	17	0	0
injected part	0	0	1	0	0
allergia	0	1	0	19	0
setup	0	0	4	0	0
flow rate	2	0	9	0	0
coinjection	4	22	9	1	0
anticancer	0	2	5	0	7
periphery	4	4	2	3	0
set	1	9	16	0	6

Table2. Characteristic word and class of treatment

characteristic word	chemo treatment	contraindi-cated drug	dowry of drugs	preparation of drugs
insulin	0	0	3	2
furosemide	0	2	10	1
injected part	14	1	0	0
allergia	0	23	0	0
setup	14	0	0	0
flow rate	8	0	0	0
coinjection	5	7	0	0
anticancer	26	0	2	0
periphery	5	5	0	0
set	0	1	29	3

5.2 Network Analysis

In this research, we aim at uncovering links between classes granted top-down and clusters discovered bottom-up. Thus, we created a network so that documents belonging to identical classes relating to the identical treatment and operation are grouped together the most. To achieve this, we used the index of Class Closeness (CC) defined in formula (8).

$$CC = \sum_i (D_{ii} - \overline{D_i}) \qquad (8)$$

Here, D_{ij} of line D calculates the distance from all of the nodes within category i to all of the nodes within category j and takes the averages of these. Also, $\overline{D_i}$ is the average of the i row of D. The classes here refer to 32 cross classes that cross calculate the class of treatment and the class of

operation. CC taking a high value indicates that the nodes of identical classes come close to cases seen across the entire network. If the threshold value of the co-occurrence index is high, there is a tendency for CC to become high as only links with strong links remain. On the other hand, if the threshold value is made too high, a large proportion of the links are lost, the maximum number of connections (LC) of nodes within the network decreases, and the analysis of overall links becomes impossible. Thus, in this research, in order that the product of the maximum number of connections of CC and nodes (CCLC) are at the maximum, we conducted the selection of co-occurrence indices and the determination of threshold values. Table 3 shows the maximum value of CCLC for each co-occurrence index and the values of indices in these instances. By doing so, we discovered that a network that reflects given classes could be obtained by using the Cos coefficient for networking the content and the Jaccard coefficient for networking the background and solutions. Although the Cos coefficient and Jaccard coefficient demonstrate almost identical CCLC in the abstract, background, and solution, it is markedly low regarding the Simpson coefficient. This is because documents that include large numbers of characteristic words are connected with many documents and are close to documents that belong to other classes in the network.

Table 3. Class closeness and co-occurrence index

abstract

Index	Value	CC	LC	CCLC
jaccard	0.290	3.162	510	1612.4
cos	0.459	2.908	594	1727.9
simpson	1.000	0.430	750	322.5

background

Index	Value	CC	LC	CCLC
jaccard	0.334	1.441	580	835.8
cos	0.580	3.348	241	806.8
simpson	0.860	0.192	970	186.5

solution

Index	Value	CC	LC	CCLC
jaccard	0.338	1.441	550	1444.7
cos	0.544	2.955	428	1264.9
simpson	0.700	0.279	1023	285.3

Table 4.cluster and characteristic class

cluster index	characteristic class	share of the class	ratio of the class in the cluster	number of the node
1	dowry of drugs	84.80%	60.50%	147
2	preparation of drugs	58.80%	51.30%	111
3	flow rate	55.30%	70.30%	37
4	diapedesis	55%	30%	27
5	-	-	-	25

6. DISCUSSION

The characteristic words displayed in Table 1 and 2 were all selected based on actual data. Networks obtained from similarities with these characteristic words remove the effects of similarities shared in common with all documents and are formed from the overall combination of an independent degree of similarity between two documents. This kind of network is first realized by extracting characteristic words bottom-up. In cases where keywords that should be checked top-down are decided, there are instances when, after having conducted class, there is no guarantee that that keyword is not valid and a network of documents linked only by the characteristic similarities such as those described above cannot be obtained.

Figure 2 shows the abstract network from among the networks created based on the characteristic words obtained bottom-up. In addition, Table 4 shows the abstract network resulting from having conducted clustering using the Newman method. Table 4 shows the top five clusters of the number of nodes and the main classes within these. The clusters from position 1 to 4 clearly reflect the granted classes. In the cluster in position 2, the class of treatment is prioritized and collects together documents classified as drowsy of drugs and preparation of drugs; in positions 3 and 4, the class of operation is prioritized and clusters reflecting flow rate and diapedesis are created; however, in position 5, classes that clearly characterize clusters could not be seen.

Moreover, regarding the solution network, links between clusters and classes such as those described above were extracted. On the other hand, clusters that clearly reflected the given class could not be seen among the clusters of the background network. The fact that networks created through background do not reflect overall given classes is in accord with the results of analysis in which the value of CC corresponding to the co-occurrence index used in creating networks in Table 3 is around 3 in abstract and solution, while in background it is less than 1.5.

So, is the class of background not related in some way to networks created bottom-up? In order to investigate this, we conducted an investigation into how many identical classes exist within the k step from certain nodes. The results of this investigation are shown in Fig. 3. We can see that in no matter which network, the ratio of nodes with identical class decreases to the extent of the number of steps increasing. Here, attention should be drawn to the point that the proportion of identical classes existing within a single step is almost the same in each of the categories of abstract, background, and solution. This indicates that the same classification also takes place in case studies with an extremely high degree of similarity in background. On the other hand, when looking at the overall network, given classes relating to the background are scattered compared with abstract and solution. This concept is outlined in Fig. 4.

With regard to the abstract and solution, links determined from ontology reflect comparatively given categories; however, regarding the background, when looked from the perspective of ontology, there is the possibility that it is completely different from the classes of abstract and solution, even those matters in proximity to step numbers, i.e., it is difficult to predict accidents that occur from information on background. Therefore, it shows that the content of descriptions of background is inadequate from the perspective of preventing reoccurrences. On the other hand, regardless of differences between abstract, background, and solution, the reporter is likely to have taken care to use descriptions that express accurately the maximum limits and the reality of the situation. Class methods that give appropriate suggestions to on-site reporters are required. We can say that there is a strong desire for an ontological construction relating to descriptions of background.

Considering the similarities between only two documents without using network analysis is equivalent to looking at the similarities of the nodes within step one in Fig. 3. In this case, in order to

obtain almost identical values between abstract, background, and solution, differences relating to the power of expression of descriptions in the three divisions are not seen. The discovery that descriptions of background are inadequate was made possible for the first time by considering the overall similarities through the networking of case studies.

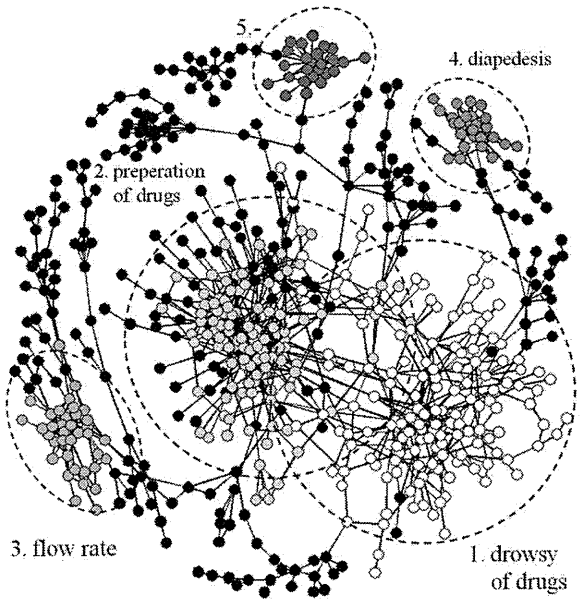

Figure 2. Network of an abstract in Near-miss/ Averse Event

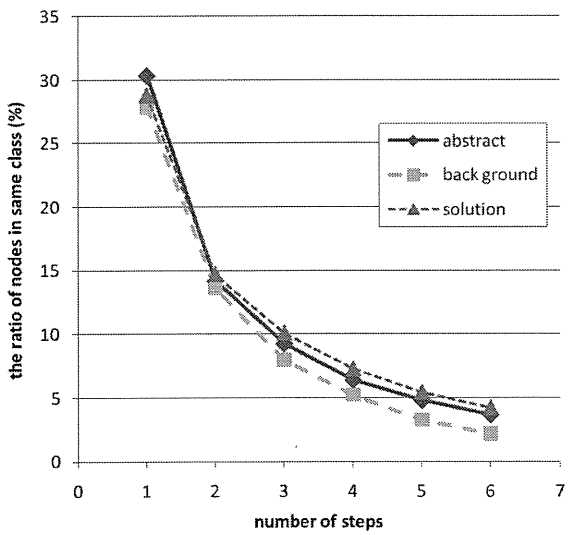

Figure 3. the ratio of nodes in same class

7. CONCLUSION

In this research, we evaluated the degree of similarities between incident documents obtained bottom-up and the links between existing classes granted top-down. In doing so, we made it possible to evaluate overall similarities regarding incident documents by using the method of network analysis. Moreover, it became clear that the use of the Cos coefficient or the Jaccard coefficient is appropriate for determining similarities in creating networks.

With regard to the background, the results of the analysis demonstrated that, compared with abstract and solution, existing classes are inadequate for representing the characteristics of documents and that there is a need to improve classes (figure 4).

By using the methods employed in this research, suggestions otherwise unobtainable through conventional methods can be made regarding the investigation of how classes should be.

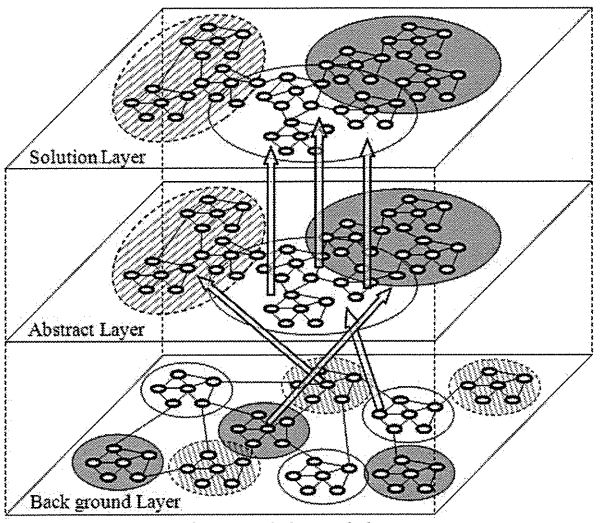

Figure 4. Structure of network in each layer

8. REFERENCES

Gandhi, T.K. and T. H. Lee (2010). Patient safety beyond the hospital. *The New England Journal of Medicine,* vol.363, pp.1000-1003.

Heinrich, H W. (1931). *Industrial accident prevention: A safety management approach.* McGraw-Hill Customer Service.

Huckvale, C., J. Car, M. Akiyama, S. Jaafar, T. Khoja, B. Farlow, S. Yaron, G. Locke and S. Whittaker (2010). Information technology for patient safety. *Quality & Safety in Health Care,* vol.19(2), pp. i25-i33.

Kajikawa, Y., J. Ohno, Y. Takeda, K. Matsushima, and H. Komiyama (2007). Creating an academic landscape of sustainability science: an analysis of the citation network. *Sustainability Science,* vol.2, pp.221–231.

Kaushal, R. and DW. (2002). Bates, Information technology and medication safety: what is the benefit ? *Quality & Safety in Health Care,* vol.11, pp.261-265.

Majeed, A., J. Car and A. Sheikh (2008). Accuracy and completeness of electronic patient records in primary care, *Fam Pract,* vol.25, pp.213-214.

Manning, C. D., and H. Schutze (2002). *Foundations of statistical natural language processing.* The MIT Press, London.

Matsuo, Y. and M. Ishizuka (2002). Keyword extraction from a document using word co-occurrence statistical information. *JSAI,* vol.17(3), pp.213-227.

Mcloughlin, V., J. Millar, S. Mattke, M. Franca, P.M. Jonsson, D. Somekh and D. Bates (2006). Selecting indicators for patient safety at the health system level in OECD countries. *International Journal for quality in Healthcare*, Sept 2006, pp.14-20.

Newman, M. E. J. (2004). Fast algorithm for detecting community structure in networks. *Physical Review* E, 69, 066133.

Rasmussen, E. (1992). *Clustering algorithms, information retrieval: data structures and algorithms.* William B. Frakes and Ricardo Baeza-Yates.

Salton, G. (1983). *Introduction to Modern Information Retrieval.* Mcgraw-Hill College.

Shibata, N., Y. Kajikawa and I. Sakata (2011). Measuring relatedness between communities in citation network. *Journal of the American Society for Information Science and Technology*, in press

Shibata, N., Y. Kajikawa and I. Sakata (2010). Extracting the commercialization gap between science and technology. *Technological Forecasting and Social Change,* vol.77(7), pp. 1147-1155.

Uchida, M., N. Shibata, Y. Kajikawa, Y. Takeda, S. Shirayama and K. Matsushima (2009). Identifying the large-scale structure of the blogosphere. *Advances in Complex Systems,* vol.12, pp. 207-219.

http://mecab.sourceforge.net/

http://www.med-safe.jp/contents/english/index.html

http://www.who.int/patientsafety/implementation/taxonomy/en/

http://www.who.int/patientsafety/implementation/taxonomy/icps_statement_of_purpose.pdf

International Association for Management of Technology
IAMOT 2011 Proceedings

Information Science Linkage of Service Innovation

ICHIRO SAKATA

Policy Alternatives Research Institutes, The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan,

isakata@ipr-ctr.t.u-tokyo.ac.jp

JUNICHIRO MORI

Innovation Policy Research Center, The University of Tokyo, 2-11-16, Yayoi, Bunkyo-ku, Tokyo, 113-8656, Japan

jmori@ipr-ctr.t.u-tokyo.ac.jp

NAOKI SHIBATA

Innovation Policy Research Center, The University of Tokyo, 2-11-16, Yayoi, Bunkyo-ku, Tokyo, 113-8656, Japan

shibata@ipr-ctr.t.u-tokyo.ac.jp

MASANORI AKIYAMA

Policy Alternatives Research Institutes, The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan,

makiyama@pp.u-tokyo.ac.jp

YUYA KAJIKAWA

Innovation Policy Research Center, The University of Tokyo, 2-11-16, Yayoi, Bunkyo-ku, Tokyo, 113-8656, Japan

kajikawa@ipr-ctr.t.u-tokyo.ac.jp

Abstract

It is with this recognition that policies for strengthening international competitiveness regarding service innovation are being adopted by many countries. While planning and implementing these policies, what is required in essence is an objective analysis regarding the current status of knowledge related to this field and the linkage between science and innovation. However, the knowledge infrastructure of this kind is inadequate. The aim of this paper is to develop the way to identify the meta structure of knowledge and measure "information science linkage of service innovation". We use bibliometrics, co-citation network analysis and co-occurrence analysis to objectively identify them.

In the field of service innovation, our results show that there are mainly two groups of elements related to service innovation: applications of service innovation and basic theories for service innovation. In the field of service science, we also identified major knowledge groups such as machine learning, and information retrieval. Then we calculated the co-occurrence of characteristic terms in journal paper abstracts belonging to the SSME sub-clusters and author keywords for papers belonging to the sub-sub-clusters related to information science. We clarified which information sciences are used heavily in service innovation. We also determined which areas of innovation make heavy use of information science and which do not. In the field of medical care, the high value we obtained demonstrates that digital health and EHR research is being vigorously conducted worldwide. By contrast, in other fields related to medical treatment, such as mental healthcare and patient satisfaction, we found a big room to promote the spread of information science.

Overall, we have demonstrated the possibility of using bibliometrics to objectively identify the meta structure of knowledge and measure the semantic relationships between science and technology.

Keywords: SSME, Information Science, Science Linkage, Technology Roadmap

Introduction

It is widely recognized that the concept of service innovation is significant for innovation strategy and economic growth. It is with this recognition that policies for strengthening international competitiveness regarding service innovation are being adopted by many countries. While planning and implementing these policies, what is required in essence is an objective analysis regarding the current status of knowledge related to this field and the technological linkage between science and innovation. However, knowledge infrastructure of this kind is inadequate. The sense of concept SSME is so broad that there is not the common understanding about what is service innovation even among experts. Another reason is the lack of suitable indicator or the way of measurement. Even though, the indicators of science linkage which uses citation data of patents or corporate studies have been developed (Motohashi and Yun, 2007, Bonaccorsi and Thoma, 2007 as examples), they provide us limited information of semantic relationships of technologies. Therefore, using bibliometric methods, we shed light on the structure of knowledge pertaining to service innovation and the linkage or semantic similarity between service science and innovation.

The first aim of this paper is to create a knowledge landscape of service innovation and service science from a number of academic publications. In this paper, of the many definitions of service innovation, we focus on the concept "Service Science, Management, and Engineering (SSME)" proposed by IBM. This paper uses same data with our previous study of SSME (Shibata et al., 2009). As an indicator of service science, we use the data in the fields of information science, library science, and computer science, which are the most important basic knowledge for service innovation. We collect data of academic papers, create citation networks regarding papers as nodes and citations as links, categorize papers into sub-clusters or sub-sub-clusters, extract topics of each sub-cluster, and finally discuss the results with experts. The second aim is to calculate the degree of distance in the semantic connections between service innovation sub-clusters on the one hand and information science, library science, and computer science sub-clusters on the other. The degree of distance in these semantic connections may be employed as an indicator of the depth of the relationship between innovation and individual scientific technologies. This is referred to as "Information Science Linkage of Service Innovation." In addition, the relationship between

information science and service innovation as viewed from the perspective of information science is referred to as "forward linkage," and the same relationship as viewed from the perspective of service innovation is defined as "backward linkage." In concrete terms, based on the likelihood that author keywords for journal papers within the clusters of information science, library science, and computer science will appear in abstracts of papers in the clusters within service innovation, we calculate the depth of the relationships between the various clusters that belong to both fields. Finally, we discuss the appropriate policy for promoting service innovation based on the knowledge infrastructure.

Methodology

First of all, the methodology for creating academic landscape is shown. Analyzing schema is depicted in Fig. 1. The step (1) is to collect the data of the knowledge domain. We collect citation data from the Science Citation Index Expanded (SCI-EXPANDED), the Social Sciences Citation Index (SSCI), and the Arts & Humanities Citation Index (A&HCI) compiled by the Institute for Scientific Information (ISI), which maintains citation databases covering thousands of academic journals and offers bibliographic database services, because these are three of the best sources for citation data. The problem, how we should define a research domain, is difficult to solve. One solution is to use a keyword that seems to represent the research domain. When we collect papers retrieved by the keyword, we can make the corpus for the research domain.

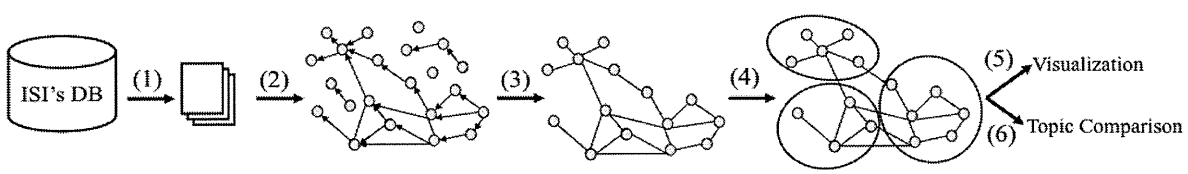

Figure 1. Methodology proposed in this paper.

The step (2) is to make citation networks for each year. We construct citation networks by regarding papers as nodes and intercitations as links. The network created for each year facilitates a chronological analysis of citation networks. According to a previous study, intercitation, which is also sometimes known as direct-citation, is the best way to detect emerging trends (Shibata et al., 2009, Shibata et al., 2010, Sakata et al., 2010). In network analysis, only the data of the largest component on the graph was

used, because our study focuses on the relationships among documents, and we therefore want to eliminate from our study those not linked with any others in step (3).

After extracting the largest connected component, in step (4), the network is divided into clusters using the topological clustering method (Newman, 2004), which does not need the number of clusters by users. Newman's algorithm discovers tightly knit clusters with a high density of links within cluster. In step (5, 6), experts in the research domain assign a name to each cluster manually after they had seen titles and abstracts of the papers in each cluster, supported by the methodology of visualization developed by Adai et al. (2004).

Second, we calculate "Information Science Linkage of Service Innovation." The linkage is defined as the degree of distance in the semantic connections between service innovation clusters on the one hand and information science clusters on the other. The degree of distance in these semantic connections may be employed as an indicator of the depth of the relationship among different research fields.

For respective clusters of information science, we calculate the semantic connection to each cluster of service innovation. We define this linkage, as viewed from the perspective of information science to service innovation, as "forward linkage". And the inverse linkage, as viewed from the perspective of service innovation to information science, is called as "backward linkage".

More formally, let *linkage*(I,S) be the linkage between a cluster I of information science cluster and a cluster S of service innovation:

$$linkage(I,S) = \begin{cases} \sum_{w} \dfrac{freq_s(w_I)}{|A_s|} & forward\ linkage \\ \sum_{w} \dfrac{freq_s(w_I)}{|W_I|} & backward\ linkage \end{cases},$$

where W_I is a set of author keywords, which are assigned to a paper by its author(s), of papers in I, w_I is an individual keyword in W_I, A_s is a set of abstracts of papers in S, and $freq_s(w_I)$ denotes a term frequency of w_I in A_s. The forward linkage is normalized with the size of A_s so that linkages from one of information science clusters to service science clusters are uniformly evaluated. And the backward linkage is normalized with the size of W_I.

The number of abstracts in each cluster of service science is shown in Table 1 and the number of author keywords in each cluster of information science is shown in Table 2.

In our experiment, we exclude an author keyword that its frequency is less than a certain threshold. We set the threshold as 100 based on our preliminary experiment.

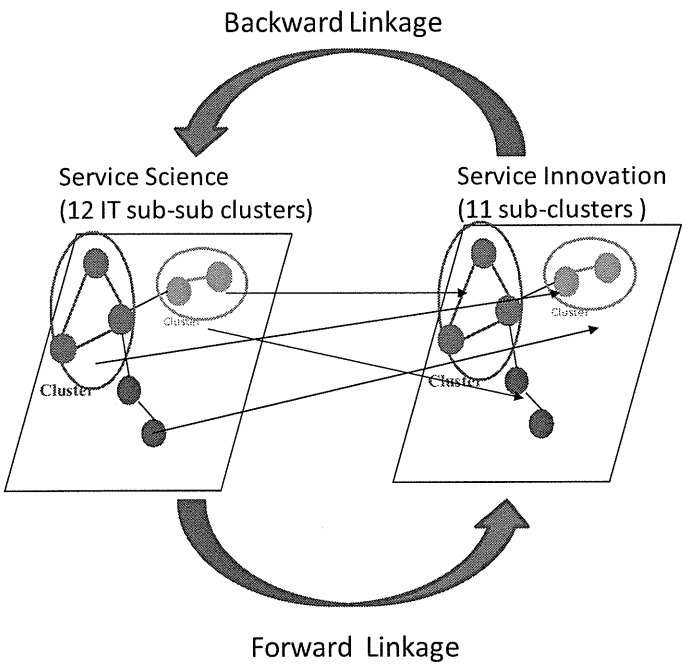

Figure 2.Concept of information science linkage

Results

With respect to service innovation, the number of academic papers that form the target of this analysis stands at 54,928. Our results shows that there are mainly two groups of elements related to service innovation: applications of service innovation such as health and medical care, IT, and public service, and basic theories for service innovation such as management, ecosystem, and QoS (Table 1).

Table1.Major sub-clusters of service innovation (SSME)

Service Cluster	Label	# of Abstracts
1	Management	1783
2	Medical Care	1660
3	Mental Health Care	1276
4	Ecosystem	901
5	Quality of Service (QoS)	903
6	Public Service	837
7	Public Medical Care	627
8	IT and Web	454
9	Patient Satisfaction	351
10	Clinical Pharmacy	305
11	Telemedicine	308

The number of academic papers related to information science, library science, and computer science stands at 314,806. Major sub-clusters include (1) machine learning, neural network, computer vision and computer graphics, (2) artificial intelligence, network, information retrieval, information theory and database, (3)distributed or parallel computing, computer architecture and information system, (4)fuzzy, (5)bioinformatics, (6)security and cryptography, (7)library and information science, (8)computer physics, (9)math application, (10)telecommunication, (11) reliability, (12)graphics, display and color, (13)computer and geo science, (14)health information, (15)circuit device. Our analysis targets the top three sub-clusters which include more than 60,000 papers. Major sub-sub-clusters of the top three sub-clusters include machine learning, computer vision, neuroinformatics, computer graphics, artificial intelligence, telecommunication network, information retrieval, information theory,