RESEARCH OF DOCUMENT DATA IDENTITY VERIFICATION SYSTEM USING SEMANTIC INFORMATION

Masanori Ikebe[1], Shigenori Tanaka[2], Hitoshi Furuta[2], and Kenji Nakamura[1]

Abstract: In electronic delivery of engineering works and construction fields, the standard is settled and the base is adequately maintained by the initiative of the Ministry of Land, Infrastructure and Transport. But, problems arise in the actual electronic delivery. These problems are that when order supplier and order demander exchange data, the significantly identity is not assured enough and the coexistence of similar data increase the work cost. Currently, a special application for document processing is being introduced as means to solve these problems. But, it is not a technique for fundamentally solving the problem. In this research, the layout of the document data made by a variety of data forms was analyzed, and its structural information was created. The substantial difference between documents is extracted by analyzing meaning information using the natural-language processing, and the identity is verified. In addition, document management is made more efficient by presenting detailed difference information.

Keywords : Electronic Delivery, Document Processing, Layout Analysis, Natural-Language Processing, Identity Verification

1. INTRODUCTION

In recent years, business transactions using electronic data are becoming socially widespread with the development of information technology. Following this trend, public agencies actively set about coping with electronic delivery and settling the standard. Some of the fruits that electronic delivery is expected to bring are the following changes: Reduction of paper and savings on space; higher efficiency in project enforcement; improvement in quality. From 2001, electronic delivery was implemented in civil engineering and construction field by the construction controlled directly by the Ministry of Land, Infrastructure, and Transport. When electronic delivery was first implemented, there was a restriction on the least estimated sum of money for contract, which was over 300 million yen. Since then, the restriction of estimated sum of money for the contract has been declining year by year, and in 2004, restriction was removed from every construction[1)]. Moreover, according to the survey, 56.0% of municipalities in Japan were using electronic delivery in 2002, and the percentage has increased to 63.8% in 2003, from which we can see that the number of municipalities showing interest in electronic delivery is increasing[2)]. As the statistic implies, public agencies, who are the on the ordering side, are showing active movements toward diffusion of electronic delivery. Businesses in general, the order suppliers, are also steadily preparing to cope with electronic delivery, and according to the results of the survey, 79.4% of general businesses were able to cope with electronic delivery in 2002. The numbers tell us that both order demanders and suppliers are raising their interest in electronic delivery. However, while electronic delivery is diffusing, various problems peculiar to electronic delivery have become apparent in the actual construction sites.

- If the original data is in data form, in which its content can be altered, such as Microsoft Word, there is a risk that its content may be altered easily, and it would be difficult to verify the alteration of data.
- When the data is being revised repeatedly between order demander and supplier, similar data will coexist with each other and data will rewind itself, due to the fact that version management of data is not done adequately[3)].

1) Student Member of JSCE, Ph.D. Candidate, Kansai University Graduate School, Faculty of Informatics (2-1-1 Ryozenji-cho, Takatsuki-shi, OSAKA 569-1095, JAPAN, E-mail: ikebe@kansai-labo.co.jp)

2) Member of JSCE, Dr. Eng., Kansai University, Professor, Faculty of Informatics.

- In case where work is done by several people in the supplier side, many similar data would be created, requiring great deal of labor cost to verify the appropriate data to be presented to the ordering side.

These problems can be categorized into 2 categories; security based problem and data management based problem[4].

The following 5 items can be subdivided under security based problem and are regarded as important items to assure the protection of original electronic document: Assurance of completeness, assurance of confidentiality, assurance of readability, existence verification, and long term safekeeping. The problem of data alteration mentioned earlier can be solved by assuring its completeness and confidentiality. These assurances assure that no changes are made to the content of the data from the time it is created to the time the ordering side receives it. Security problem can be solved by implementing encoded communication or adding electronic signature to data forms. However, due to the fact that data transactions are performed several times between order demander and supplier during electronic delivery, enforcing these security measures in every transaction to deal with security problem is difficult in the aspect of cost and time.

Coexistence of similar data is the principle problem related to data management. In order to solve this problem, when distinguishing appropriate file from inappropriate ones during document management in the actual construction stage, the file content needs to be observed or verified by document comparison tool. If the file content is to be verified by observation, a great deal of human cost will be required at the stage where appropriate data is selected from delivered data or data under construction. If the file content is to be verified using document comparison tool, 2 problems will arise. The first problem is that many document comparison tools being released today are applicable only to the text within the data, therefore distinguishing different layouts and meanings in the sentence is rather difficult. Therefore, when current document comparison tool is applied to a document including layout information, all the subsequent contents are perceived as having been revised, at an instance when the position where comparison is conducted shifts[5]. For this reason, difference information can only be extracted with an extremely low accuracy. The second problem is that when several original data are in different film forms, analysis of content information becomes indispensable in order to compare documents.

Therefore in this research, we will propose new methods on security improvement and a method that can both verify the identities of different data forms and detect their detailed differences.

2. OUTLINE OF THE RESEARCH

(1) PURPOSE OF THE RESEARCH

Previous researches on the method of adding third person confirmation or electronic signature are considered as means of solving security problems. However, these researches deal with detection of alteration in identical binary data or protection of original document, and do not verify documents with the same structural information but in different data forms. Also, even if 2 data have identical extension, they are still in 2 different data forms depending on the type of tool each data was created with or the version of each data. Then in this research, security of the data can be assured by analyzing the documents' content information to verify 2 data with the same structural information but in different forms as 2 identical data.

Problem of coexistence of data are considered solvable by using electronic delivery tool or document management tool. However, these tools need to have information such as explanation of the document added at the time of its creation, and effect of these tools cannot be expected on documents already completed. Furthermore, when documents are exchanged between the ordering side and the supplier side by email, these document data cannot be managed by these tools. As a result, information cannot be added, so information content or revised information needs to be verified from the substance of the document. Accordingly, in this research, detailed difference information between documents will be presented when the difference is detected by identity verification. Consequently, by easily specifying the place of revision or presenting analogical information of revision planning to the user, this research aims to support the verification of appropriate data.

(2) PROBLEMS TO BE SOLVED IN THIS THIESIS

In this research, by sending 2 documents in different types of data form during data exchange, the recipient can verify that the data has not been altered, by

comparing 2 documents. Since this method is in formality different from the previous security measure, it can be used together with the previous method to increase the level of security. Furthermore, since comparison between documents with layout information that was once difficult can now be done easily, not only can we save the labor of document management by users. We can also reduce the labor of electronic work by applying this research when data is inputted in electronic delivery tool at enforcement stage, in which paper medium and electronic medium coexist.

3. SYSTEM IN DETAIL

The system to be developed in this research is composed of the following 3 functions: Document analysis function that analyzes the documents intended to be compared and divides each document into objects, which are units used in document processing; Function that verifies the identical places by extracting correspondence relationship of objects between the 2 documents being compared; Function that outputs the difference by presenting the result of the difference obtained from 2 documents in detail, in object unit.

(1) DOCUMENT ANALYSIS FUNCTION

This function divides the 2 document intended to be compared into object units, the units used in document processing, by reading the documents before and after alteration and by analyzing them. Also, because information such as the layout is added to the object, this information is created with the object as added information.

Dividing the object into large processing units and analyzing the content of processing unit in detail is the general method for extracting layout information[6]. However, since the standard for data used in this research is PDF (Portable Document Format), into which information in maintenance form divided into smallest processing units is saved; we reverse the process of the method by grouping the smallest division units and creating even bigger division units.

a) PROCESS OF DOCUMENT ANALYSIS

Data in PDF document form will be inputted in this system. Data that is not in PDA document form will be converted to PDA form before being inputted into this the system.

If case the document information in PDF document is in text form, text content is divided into letter unit and then saved. For this reason, in this process, text content information of inputted data after division is acquired, before creating objects. Information of vertical and horizontal position when document's top left corner is the origin is acquired. If the analytical content is in text form, text content and information related to the layout such as font, font size, font color, and spacing between lines as well as information of vertical and horizontal position are acquired. As for images, binary data of the actual image and information on the image's horizontal and vertical width are acquired. In this process, PDF document is converted to text form using the tool called "pdftohtml", an improved version of "Xpdf". Text obtained by "pdftohtml" is information of text presented at regular intervals grouped together.

b) PROCESS OF OBJECT CREATION

Although information obtained by document analysis has consistency, it is not in paragraph, chart, or item unit, therefore we need to group it.

First, objects in the document are grouped according to the objects' vertical and horizontal position, and according to the positional relation between objects. Object's top left corner is used as a standard for object's position. Next the group is classified into 3 types; chart, item, and paragraph. Then, text content of the paragraph is united to finish this process. The grouping procedures of objects are explained as of below.

- Objects are sorted in the order according to their vertical position.
- Objects, in which their vertical position is within a certain range, are grouped together.
- Amount of spacing used as a standard of style unit is determined.
- Each object is compared with the object above and below it, and if the spacing between them is within standardized value, these objects are grouped together.

The above-mentioned process is repeated until there are no two objects in the same vertical position. Next, we verify the group type. The procedures for verifying the group type are explained as of below.

- Objects within a group are sorted in order according to their horizontal position.
- If there are more than 2 rows, each with more than 2 objects, it would be considered as a chart.
- As for objects placed after the first object in the same row, if the objects were in the same

horizontal position, it is defined as a chart, and the number of rows and columns will be obtained.

- As for objects not in a chart, they are matched with template with indented lists from the word at the beginning of the row, and if they match, they will be defined as a list.
- As for objects defined as a list, objects having the same relationship and the number of objects are obtained based on information of horizontal position.

Finally, procedures for grouping normal texts as a paragraph are explained as of below.

- Objects not classified in a chart or in a list are defined as being classified in a paragraph.
- Texts grouped according to the vertical position of the objects, are combined together.
- Texts grouped according to the spacing information between objects are combined together.

The above-mentioned process is repeated until the number of unclassified object becomes 0.

(2) FUNCTION OF VERIFYING IDENTICAL AREAS

In this process, data from document before and after alteration, which are divided into object units, are compared, and added objects, corrected objects, and deleted objects are detected by specifying identical areas. To specify identical objects, feature vector is created using TF-IDF (Term Frequency-Inverse Document Frequency) weighting, and the feature vector is expanded in multi-dimensional space. Then, vector space method is used to calculate cosine correlation value. Other than the previous method[7], case-base method[8] is usually used for classification, but since document is classified only by its individual document data in this process, application of case-base method is difficult. For this reason, we have adopted TF-IDF weighting and vector space method in this thesis.

a) PROCESS OF VERIFYING IDENTICAL OBJECTS

In this process, identical objects list is created based on proper matching first, in order to obtain corresponding objects from a document before alteration and from a document after alteration. In proper matching process, each and every object that form the document before alteration are combined with each and every object that form the document after alteration, and if any combination are a perfect match, that combination will be added to the identical objects list.

b) PROCESS OF CREATING FEATURE VECTOR

In this process, objects other than identical objects created in the process of verifying identical objects are considered as altered objects, and the presence of such objects are searched. To explain the process specifically, object units undergo morphological analysis, and nouns are extracted from the analyzed object units. Morphological analysis system, "ChaSen", is used in this process. Finally, TF-IDF value of each noun obtained by morphological analysis is calculated. TF-IDF is defined by **formula (1)**.

$$w(t) = tf(t) \times idf(N,t) = tf(t) \times \log \frac{N}{df(t)} \qquad \textbf{(1)}$$

In this formula, *tf(t)* represents the frequency of word t appearing in the object intended to be analyzed. *N* represents the total number of objects in analyzed document, and *df(t)* represents the number of objects in analyzed document, in which t appears. Then, group of nouns included in object *X* is represented as *Ox={Tx1,Tx2,Tx3,…,Txn}*, and group of TF-IDF values of object *X* are defined as feature vector and is represented as *Vx={Wv(Tx1),Wv(Tx2),Wv(Tx3),…,Wv(Txn)}*.

c) PROCESS OF VERIFYING CORRELATED OBJECTS

In order to verify the objects, in which their content had been revised, feature vectors for all the combinations of objects are calculated, and these feature vectors are expanded in cyberspace of dimension *D*. *D* is defined as the number of object *X* and object *Y* in a set obtained when comparing object *X* and object *Y*, and is represented in **formula (2)**.

$$D(x,y) = (Ox \cup Oy) \qquad \textbf{(2)}$$

$$sim(Vx,Vy) = \frac{Wv(Tx_1)\times Wv(Ty_1) + Wv(Tx_2)\times Wv(Ty_2) + \ldots + Wv(Tx_n)\times Wv(Ty_n)}{\sqrt{Wv(Tx_1)^2 + Wv(Tx_2)^2 + \ldots + Wv(Txn)^2} \times \sqrt{Wv(Ty_1)^2 + Wv(Ty_2)^2 + \ldots + Wv(Tyn)^2}} \quad \textbf{(3)}$$

Therefore, feature vectors expanded in cyberspace need to have the number of components equivalent to that defined in *D(x,y)*. However, feature vectors defined as *Vx* save only n number of values where *D(x,y)>n*. So to sum up the number of dimensions, for each object's feature vector, feature component, in which a value does not exist, will be complemented as 0. Next, cosine correlation value is calculated from feature vector in order to calculate the relevancy between objects. Cosine correlation value *sim(Vx,Vy)* is defined in **Formula (3)**.

In **Formula (3)**, the right side of the numerator represents the inner product of the feature vectors of object *X* and object *Y*, and the denominator represents the product of each feature vector's origin and distance. Cosine correlation values are calculated by comparing the document's feature vectors in terms of their angles, and the values range from 0 to 1. The closer to the value is, the more similar in class the 2 documents are. After cosine correlation values for combinations of all the objects have been calculated, a combination is verified as a revised object if its cosine correlation value is more than a certain threshold. Then, combinations verified as altered objects are added to revised objects list in the order of their cosine correlation value, from the value closest to 1, until the number of combinations with cosine correlation value over the threshold becomes 0. Finally, the remaining objects are added to added/deleted objects list.

(3) FUNCTION THAT OUTPUTS THE DIFFERENCE

This function extracts detailed information of parts where difference exist, from altered objects list obtained by the process of verifying identical objects, and the information is presented to the user. Information to be outputted as the difference not only indicates the addition/revision/deletion of objects, it also presents detailed information so that users can spot every altered part at a glance. In the case where the intension of alteration is not attached, object's meaning information, as information to help the user verify the intension of alteration in the document, are represented by radar chart, and alteration of meaning content is made visible.

a) PROCESS OF EXTRACTING DIFFERENCE INFORMATION

In this process, information on the addition/deletion of sentences is extracted from altered objects list that had been extracted by the function of verifying identical parts. The procedures of the process of extracting difference information within altered objects are listed as of below.

- Types of altered object are distinguished from one another.
- The number of items in the list, or the number of rows and columns in the chart are obtained, as information to be outputted as the difference.
- Sentences within the object are divided by punctuation marks.
- Each and every divided sentences, or divided units undergo identity verification, and if the sentence is identical, it will be added to the revised objects list, and if it is different, it will be added to the added/deleted objects list.
- Divided units within added/deleted list are classified in the following way: Divided units that exist in the list from the document before alteration but do not exist in the list from the document after alteration are classified as deleted units; Divided units that do not exist in the list from the document before alteration but exists in the list from the document after alteration are classified as added units.

The above-mentioned procedures are to be conducted for every revised object in order to extract difference information on all the revised information.

b) PROCESS OF OUTPUTTING RESULTS

In this process, the numbers of added/revised/deleted cases are indicated in object units. Revised objects are in the form, in which its detailed information can be browsed. Detailed information includes the number of added cases and deleted cases, and the content that actually had been added or deleted. This information is specified in the document. In the process of outputting results, interface for presenting this information to users is provided.

Table 1 Internal Composition of Analyzed Document

Document	Total Number of pages	Image	Chart	List	Paragraph
Document A	17	12	1	4	86
Document B	16	8	16	8	101
Document C	10	3	10	2	57
Document D	15	10	2	3	84
Document E	21	14	0	9	127
Total	79	47	29	26	455

Table 2 Number of Alterations of Compared Document

Object type	Added	Altered	Deleted	Identical
Sentence	21	33	6	44
Chart	1	1	0	2
List	3	3	1	2
Total	25	37	7	48

4. EXPERIMENT BY MEANS OF EVIDENCE

(1) ACCURACY OF DOCUMENT ANALYSIS

a) EXPERIMENT METHOD

In this experiment, document data is browsed, and subdivided objects are grouped and transformed into group structure by using document analysis function. Then, objects are classified into two types: Image and text. We verify if the numbers of chart, list, and paragraph in the text are accurately acquired.

b) SUBJECT OF EXPERIMENT

We have prepared 5 document data to measure the accuracy of document analysis function. Since electric delivery in engineering works and construction field is the subject of this research, all the documents used in the experiment are technical documents related to engineering. Documents used in this experiment range from 10 pages to 21 pages, and the number of objects included in each document range from 72 to 150.

c) EXPERIMENT RESULTS

After analyzing 5 documents, we have succeeded in structuring a total of 636 objects by grouping them. By checking the 5 documents used in this experiment with our eyes, we have confirmed the total number of objects to be 636, as indicated in **Table 1**, from which we were able to prove that objects were grouped accurately. Furthermore, of all 636 objects, 47 were classified as images, 29 as charts, 26 as lists, and 455 as paragraphs, by verifying the type of each object. These result also agree in numbers with the numbers of each object type confirmed with our eyes, we can conclude that each object's type was verified accurately.

d) CONSIDERATION

By this research, we have clarified that document analysis function's object creation function is accurate enough to be put into practical use. In this experiment, we have proposed a grouping method for each of the 4 object types frequently used in ordinary documents, which are image, chart, list, and paragraph. We were able to confirm the useful accuracy of each method. Also, we have proposed a method for verifying the 4 object types mentioned above, and we have confirmed that classification was performed accurately. Therefore, we can consider that this research can be applied to the document composed of these object types.

(2) ACCURACY OF VERIFYING IDENTICAL PARTS

a) EXPERIMENT METHOD

In this experiment, 2 documents were read, and added/deleted objects and revised objects were verified using the function of verifying identical parts. Correlation value of 0.7 or more has been set as a standard value for verifying identical parts.

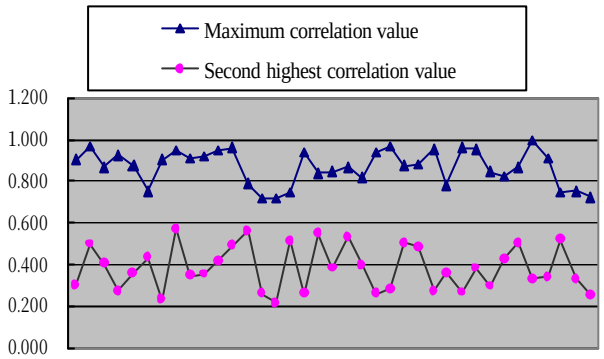

Fig.1: Correlation Value of Revised Objects

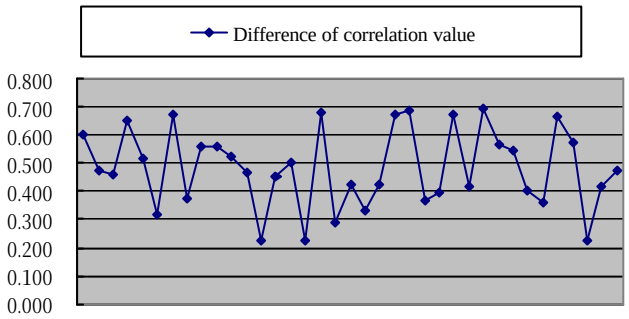

Fig.2: Correlation Value of Added Objects

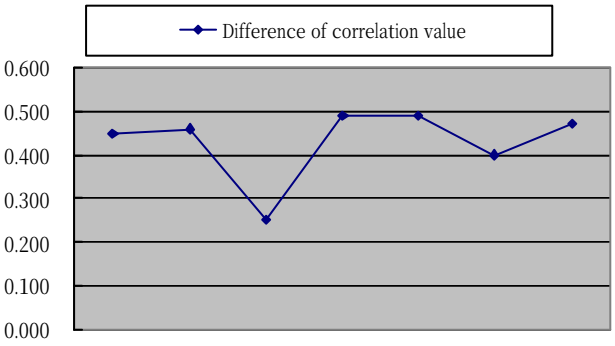

Fig.3: Correlation Value of Deleted Objects

b) SUBJECT OF EXPERIMENT

In order to measure the accuracy of identical parts verification, we have created a document as a subject of comparison by conducting the following procedures on document 1, which had been used in the experiment of measuring the accuracy of document analysis: "Adding/revising/deleting sentences, images, and blocks in lists" and "Changing the numbers of rows and columns in charts". Details on each alteration are shown on **Table 2**.

c) EXPERIMENT RESULTS

As results of the experiment, 48 identical objects, 25 added objects, 37 revised objects, and 7 deleted objects have been confirmed precisely.

d) CONSIDERATION

The correlation value of the combinations of objects added to the revised list range from 0.716 to 0.999. As opposed to this, the maximum value of correlation values of all combinations of objects added to added/deleted list is 0.535. And when we have looked at the distribution of correlation values, they were in the range from 0.5 to 0.7, from which we were able to confirm that rate of over-crowdedness of dots indicating the values was relatively low. Therefore, we were able to confirm that the correlation value of 0.7 was an appropriate value to be set as a standard for the value where division is to be made.

Also, when we have compared the maximum correlation value and the second highest correlation value of each object in the combination of objects in the revised list, the average difference of correlation values came out to be 0.484. The difference between maximum correlation value and the second highest correlation value is illustrated in **Fig.1**. From the above result, we were able to find out that the object verified as an identical object by its meaning information has a big difference with other objects in terms of meaning.

Correlation value of objects in the added list and correlation value of objects in the deleted list are illustrated in **Fig.2** and **Fig.3**, respectively. Objects included in added/deleted lists had correlation values lower than any other objects, with the maximum value of only 0.535. Therefore, excessive recognition, in which added/deleted objects are verified as being revised objects, did not occur, allowing us to be able to confirm the correct number of cases.

5. CONCLUSION

In this research, we have realized the comparison of electronic documents in different data forms, which was previously difficult, by analyzing the meaning of text information within the document. Furthermore, we were able to confirm the usable level of accuracy of the comparison by conducting an experiment by means of evidence. Moreover, by applying this research, we improve security by assuring the identity of documents when these are being exchanged between order supplier and demander during electronic delivery. And what is more, we can make document management more

efficient by presenting the different parts in each documents in detail, enabling revision work to be done smoothly between order supplier and demander.

However, in electronic delivery, its information application has not been examined enough and this remains as a problem. To deal with this problem, by using the fruits of this research, we plan to promote a research and development practical to the phase of information application by constructing a search system attached with meaning information of delivered data.

References

1) Ministry of Land, Infrastructure, and Transport: Guideline on Application of Electronic Delivery (Proposal), 2004.10. (in Japanese)
2) Japan Civil Engineering Contractors' Association: Fact-finding Investigation Report on Diffusion of Information, 2002.12. (in Japanese)
3) Takashi Matsumoto: Present Situation and Problem related to Electronic Delivery of Engineering Products, *Civil Engineering*, Vol.58, No.7, pp.34-41, 2003.7. (in Japanese)
4) Tadashi Okutani, Koichi Aritomi: Research on Business Procedure Reform by Application of Electronic Delivery Information, *Civil Engineering Journal*, Vol.45, No.3, pp.38-43, 2003.5. (in Japanese)
5) Tapas Kanungo, Song Mao: Stochastic Language Models for Style-Directed Layout Analysis of Document Images, *IEEE Transactions on Image Processing*, Vol.12, No.5, pp.583-596, 2003.5.
6) Seong-Whan Lee, Dae-Seok Ryu: Parameter-Free Geometric Document Layout Analysis, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.23, No.11, pp.1240-1256, 2001.11.
7) Yoshihiro Ueda, Naotaka Kato, Katsuaki Hayashi, Hitoshi Narita, Hidetaka Nanbo, Haruhiko Kimura: Automatic Distribution of Email Using Text Mining and Reinforced Study, *Transactions of the Institute of Electronics*, Vol.J87-D-1, No.10, pp.887-898, 2004.10. (in Japanese)
8) Lam Wai: Modeling Textural Document Classification, *Proceedings of 1999 IEEE International Conference on Systems*, Vol.3, pp.946-949, 1999.10.