ORIGINAL RESEARCH

Technique for searching tabular form documents using metadata harvested by table structure analysis

Isaac Okada *[1], Minoru Saito[1], Yoshiaki Oida[1], Hiroyuki Yamato[2], Kazuo Hiekata[2], Satoru Nakamura[2], Naoto Fukada[2]

[1] *Systems Integration Technology Unit, Fujitsu Limited, Japan.*
[2] *Graduate School of Frontier Sciences, The University of Tokyo, Japan*

Received: August 2, 2013 **Accepted:** October 17, 2013 **Online Published:** January 14, 2014
DOI: 10.5430/air.v3n1p46 **URL:** http://dx.doi.org/10.5430/air.v3n1p46

Abstract

Conducting full-text searches on collections of tabular files, in which a single sheet corresponds to a single document and each file consists of multiple sheets, typically involves retrieving many candidate files that include the search terms. Opening each of these tabular files to determine whether it is the desired sheet is labor-intensive. Searching with high precision thus requires expert intuition born of operational experience. Therefore, it would be advantageous to enable the pinpointing of desired documents with greater accuracy regardless of the operator's level of experience.

In the present study, we propose a method in which operational classifications are assigned as metadata on the basis of the table structure of a sheet. We obtain the table structure of the sheet and assign metadata based on a set of rules established individually for each pattern in the structure. We propose two methods for representing the table structures obtained: a method using node property matrix, and a method in which positional data regarding cells containing specific operation-description data are indexed. Comparing the results of searches that use assigned metadata to the results of traditional full-text searches reveals that our method has greater search accuracy.

Key Words: Table structure analysis, Tabular form document search, Operation-description metadata, Keyword metadata, Cosine similarity, Metadata, Document management, Knowledge

1 Background

System engineering documents, such as Fujitsu's system design documentation, are stored as tabular files. Each document corresponds to a single sheet, and these are stored in databases (DBs) as tabular files containing multiple such sheets.

Fujitsu's system development is standardized as SDEM, and the sheets used at each process stage are standardized as well. These sheets are templatized on the basis of a specific style, such as that shown in Figure 1; the top of the sheet includes operation-description data, such as the work process stage and the system name, and one or several of these sheets together constitute a single tabular file that is stored in a DB.

To perform tasks such as creating a system design document, one must search through existing system design documentation of similar content in order to reuse information and verify changes. However, a full-text search through a collection of tabular files, each consisting of multiple sheets, using a particular set of operation-description data as reference typically results in multiple matching files. The user must then painstakingly open each of these files to determine if the desired sheet was found. Searching with high precision requires making good initial guesses, a task which demands extensive operational knowledge and experience. The development of new methods for performing higher-precision searches and successfully reaching the desired documents regardless of operational experience would reduce search costs and improve the quality of search results. Moreover, such methods would reduce personnel training costs.

*Correspondence: Isaac Okada; Email: isaac-okada@jp.fujitsu.com; Address: Systems Integration Technology Unit, FUJITSU LIMITED, Japan.

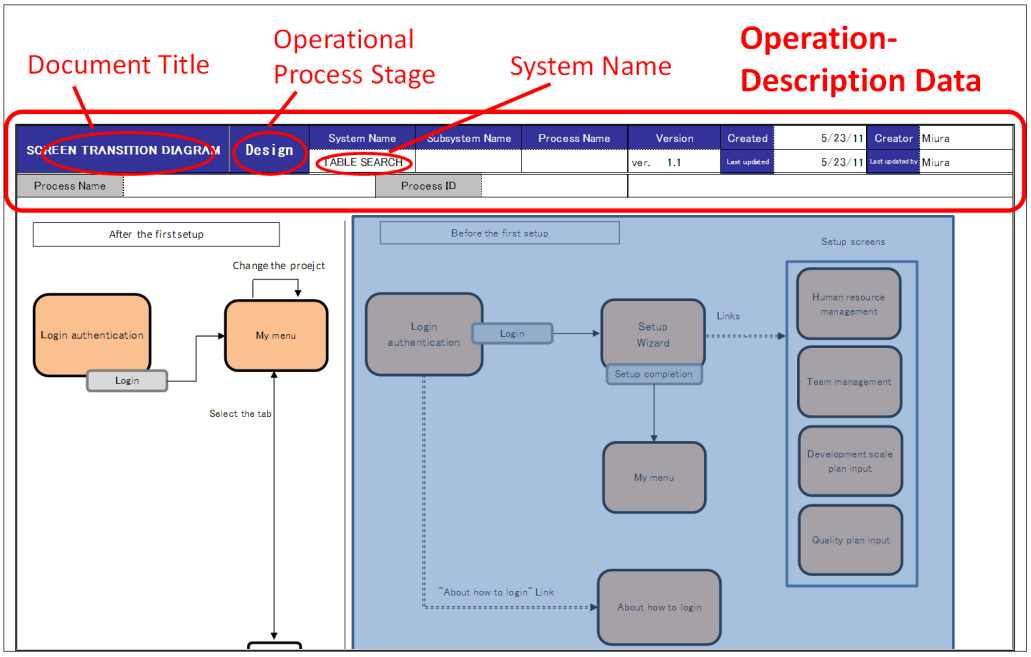

Figure 1: Example of a tabular data sheet

2 Objective

In this study, we propose a technique for searching tabular data sheets in Fujitsu system designs. Our technique is capable of identifying the desired documents with high precision regardless of the user's length of operational experience. More specifically, we obtain the characteristic table structure exhibited by each tabular sheet in a target dataset; then, using rules established individually for each pattern in the table structure, we assign operational classifications as metadata. In this study, we propose two methods for describing table structures: a method in which using node property matrix, and an alternative method in which positional data regarding cells containing specific operation-description data fields are indexed.

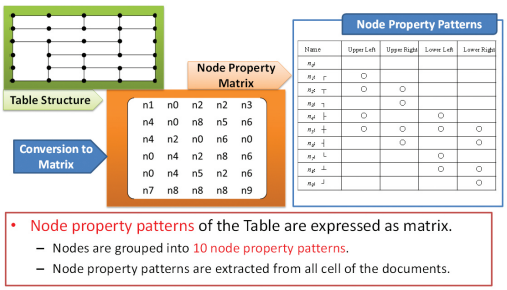

Figure 2: Method for identifying table structures

In the present study, we consider tabular files representing design documents arising in system design projects; a primary point of distinction with previous studies is our use of table-structure identification techniques to assign metadata describing operational classifications.

3 Related work

Several previous studies have consider tabular form documents within corporations.[1,2,4] In particular, Tanaka et al.[3] proposed an automatic classification system for tabular form documents based on a technique for identifying the table structures of tabular data sheets; as shown in Figure 2, lines within the table correspond to connections, and their proposed method makes use of nodes, which are points at which these lines intersect.

However, these studies primarily worked with images of documents with table structures. We are unaware of any previous study that addresses specific examples of application to tabular sheets and practical search problems.

4 Proposed method

Figure 3 presents an outline of the proposed method. In this method, operation-description data within existing tabular data sheets are assigned as metadata (hereinafter operation-description metadata). Searching for operation-description metadata retrieves existing sheets that contains them.

The proposed method analyzes the table structure incorporated within a sheet and assigns metadata by applying rules determined individually for each pattern in the table structure.

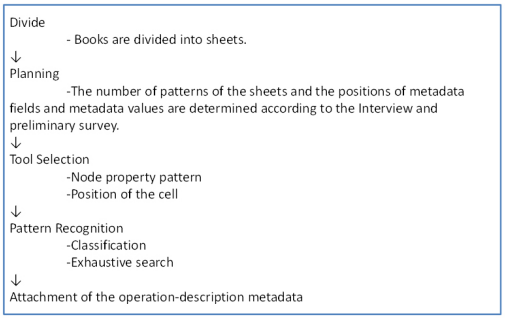

Figure 3: Outline of the proposed method

In what follows, we discuss our methods for expressing the table structure analyzed, for identifying metadata by applying rules determined individually for each pattern, and for quantifying the degree of similarity in each of the two possible expressions of structure that we consider.

4.1 Method for assigning metadata according to node property matrix

In this method, we convert node property on cells into a matrix (as discussed in Section 3) and adopt an expression technique that preserves the table structure. We refer to this method as the node property matrix method.

For patterns of the type shown in Figure 4, operation-description metadata, which includes the document title, operational process stage, system name, and subsystem name, are extracted on the basis of a set of identification rules illustrated as arrows in the figure.

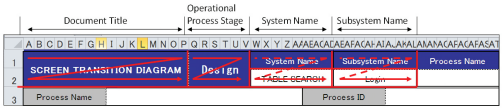

Figure 4: Illustration of the node property method of metadata identification

The degree of similarity between nodal-data matrices is computed as follows. Letting Tq denote the node property matrix describing the table structure of the sheet targeted for metadata identification and Tt denote the node property matrix describing the table structure of the sheets for the various patterns, we compute the Hamming distance between these two matrices, $HD_{Tq,Tt}$, normalize by Nq (the number of nodes in Tq), and define the degree of similarity as

$$Sim_{Tq,Tt} = 1 - \frac{HD_{Tq,Tt}}{Nq} \tag{1}$$

4.2 Method for assigning metadata according to the positions of cells containing specific metadata fields

In this method, we use the positional data regarding the cells containing specific operation-description metadata fields as indices to preserve the expression of the table structure. We refer to this method as the position-data method. Each index consists of a pair of numbers (x,y), where x is the line number (vertical position) and y is the column number (horizontal position) of the cell. Referring to the sheet shown in Figure 5, if the metadata fields from which we create our indices are "System name," "Operation," and "Operation type," then we would identify three indices: (1) system name: (1,2), (2) operation: (1,3), and (3) operation type: (1,4).

The metadata-identification rules for individual patterns are based on rules for specifying the positional relationships between the metadata field and the metadata value. These rules are used to identify the operation-description metadata, of system name, operation, and operation type.

Letting Iq denote the index data describing the table structure of the sheet targeted for metadata identification and It denote the index data describing the table structure of the sheets for the various patterns, we define the degree of similarity between the positional data indices for specific cells as follows:

$$Sim_{Iq,It} = 1 - \frac{Number\ of\ metadata\ fields\ matching\ indices}{Number\ of\ fields\ in\ the\ specified\ funcitional-description\ metadata} \tag{2}$$

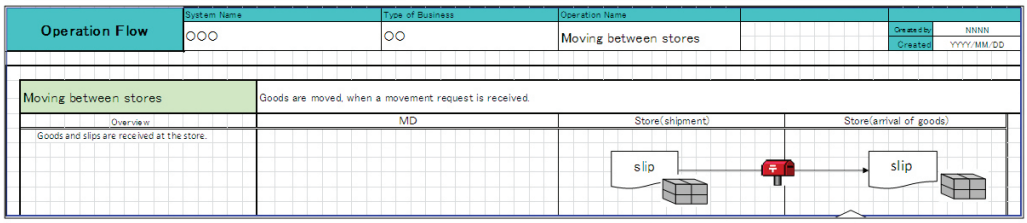

Figure 5: Illustration of the position-data method of metadata identification

4.3 Gudelines for choosing the node property method or the position-data method

As discussed in Sections 4.1 and 4.2, in this study we propose the node property method and the position-data method as two distinct techniques for expressing the table structure of a sheet. To offer guidelines for choosing which method to use, we now discuss the characteristics of each technique. The node property method identifies and preserves all node property in the lines of the table structure to be analyzed. This increases the amount of computation that must be performed (e.g., in the process of computing the degree of similarity) but has the advantage of not taking the contents of the cells into account, which makes it applicable to sheets with arbitrary table structures. In contrast, the position-data method preserves as indices only information regarding the positions of the cells in which specific metadata fields appear. Moreover, in the position-data method, the rules for identifying metadata are based on the positional relationships between metadata fields and their values, which has the potential to reduce the amount of computation and allows metadata to be identified with high precision. However, the position-data method has the drawback that it cannot be applied to sheets in which metadata fields (e.g., the document title and work process stage shown in Figure 4) are not explicitly stated, because in such cases the differences between metadata values would increase the number of patterns.

Thus, when choosing between the two proposed methods for expressing table structure, one must first conduct interviews, or carry out a preliminary survey of the target data set, to ascertain the characteristics of the target sheets.

5 Experiments

We next describe experimental assessments performed on collections of real-world internal corporate tabular data sheets. We assessed the performance of our methods in classifying patterns according to the table structure and the performance of searches of identified metadata.

5.1 Testing the accuracy of pattern classification using the node property method

5.1.1 Summary of the experiment

On the basis of a preliminary survey of the target data and interviews at system design sites, we determined that the target data for this experiment contained three distinct types of pattern sets, as shown in Figures 6 through 8. Because, as is clear from Figures 6 and 7, the metadata fields and metadata values were not clearly stated in the sheets, we use this data set to test the accuracy of the node property method in categorizing patterns. As detailed in Table 1, the data used in this experiment consisted of approximately 2,500 tabular data sheet files from two internal corporate development projects.

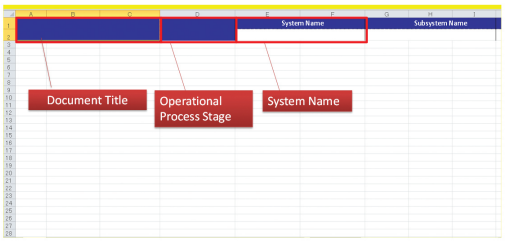

Figure 6: Type 1 table

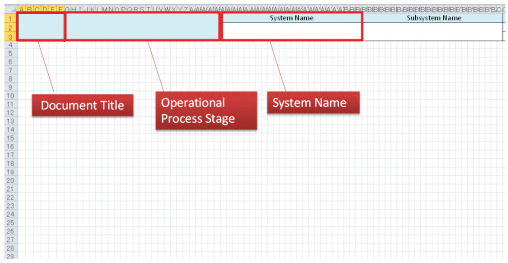

Figure 7: Type 2 table

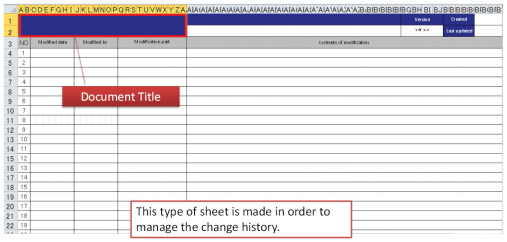

Figure 8: Type 3 table

Table 1: Data used in the experiment.

	Number of tabular documents (book)	Number of tabular documents (sheet)
Project1	694	1,732
Project2	318	857

5.1.2 Summary of the experiment

Our experimental results are displayed in Table 2.

We obtained match rates and recall rates of approximately 100The interviews conducted in advance at design sites and the advance survey of the target data indicated that three distinct types of table structure existed within the target data. However, the actual data additionally contained a total of 531 tabular data sheets exhibiting minor formatting modifications, as illustrated in Figure 9. We grouped these sheets as “others” into Type 3, which then resulted in a relatively low match rate of 65.0 percent for Type 3 sheets.

Table 2: Accuracy of classification for each type of pattern.

Type	Project1 (Number of match)	Project2 (Number of match)	Total (number of match)	Precision	Recall
1	1148(1143)	423(421)	1571(1564)	99.6%	84.2%
2	11(11)	11(11)	22(22)	100%	95.7%
3	305(165)	226(180)	531(345)	65.0%	99.7%
Average precision			88.2%		
Average Recall			93.2%		

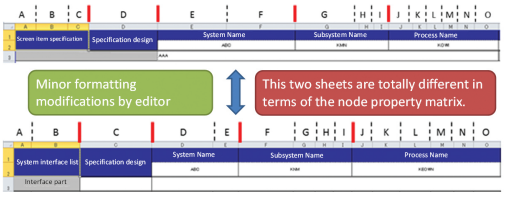

Figure 9: Slightly modified formatting

System design sites clearly do not recognize such minor formatting modifications.

5.2 Assessing the accuracy of searches that use operation-description metadata

5.2.1 Summary of the experiment

This experiment was designed to compare the accuracy of searches that use assigned operation-description metadata to the accuracy of full-text searches in a real-world usage scenario. In this experiment, we assumed that the number of distinct patterns contained in the target data is unknown. Because preliminary surveys confirmed that metadata fields and metadata values were clearly separated in the target sheets, we used the position-data method to assign operation-description metadata. Thus we assumed that the assignment of operation-description metadata correctly assigns both metadata fields and metadata values to the full document. The data used in this experiment were as follows.
Target documents: System design documentation for "Industry-specific sales management software"
Total number of files: 1,989 spreadsheet files
Number of tabular data sheets after separating compound sheets: 6,594 spreadsheets.

5.2.2 Experimental results

As a specific experimental case study, we tested "Referencing the operation flow diagrams for moving merchandise between stores."
<Experimental results> Results obtained by searching via the usual method of "full-text search" (Windows searches using the index):
Searching tabular files with the string "Moving between stores" retrieved 40 files, each including tabular multiple sheets. These 40 tabular files included several files whose main text included the character string "Moving between stores," but which were not files of the type sought in this particular search; for example, some files were "Screen/Ledger view" files or "Code system" files.
Searching tabular data sheets with the string "Moving between stores" retrieved 69 sheets. These 69 sheets included several sheets whose main text included the character string "Moving between stores," but which were not sheets of the type sought in this particular search; for example, some sheets were "Supplemental materials" sheets or "Detailed definitions of program function" sheets.
Results obtained by searching for operation-description metadata:
Metadata searching of tabular data sheets with the string "Moving between stores" retrieved 7 sheets, of which 6 were the operation flow sheets describing the movement of merchandise between stores.

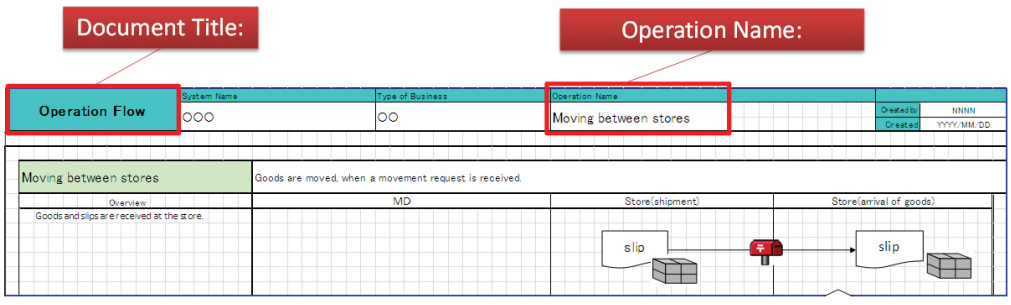

Figure 10: Operation flow for moving merchandise between stores

These results demonstrate that the use of operation-description metadata can assist in further homing in on desired documents among a collection of multiple candidate documents with similar operation attributes (process stages, categories, etc.) more precisely than traditional search techniques. Thus, the ability to narrow down and reduce a selection of candidates (an area in which previous methods were lacking, having required a human operator to open the file and select candidates by eye) has been verified, along with the efficacy of our proposed technique.

6 Conclusions

In this study, we proposed two methods for identifying operation-description metadata by using information on the table structure of tabular files. The proposed methods are the node property method and the position-data method. Experimental assessments of the accuracy with which the node property method classifies sheets found an average match rate of 88.2 percent and an average recall rate of 93.2 percent. In addition, the results of a comparison of search accuracy using operation-description metadata with that of full-text searches, by considering a specific real-world usage scenario, suggest that the proposed method provides greater accuracy.

Future work will consider how our two proposed techniques for assigning operation-description metadata can be combined in order to improve the generality of the approach. In addition, tests of the efficacy of our proposed methods in improving actual operational processes will be conducted.

References

[1] Ando Satoshi, Sawabe Kazuhide, Matsuoka Makoto, Ueda Yumiko, Shigenaga Shin-ichi, "A Method of Detecting Errors on Business Letter Composition", IPSJ SIG Notes, vol.95, no.27, pp.31-36, 1995 (in Japanese)

[2] Doi Miwako, Fukui Mika, Yamaguchi Kouji, Takebayashi Youichi, Iwai Isamu, "Development of Document Architecture Extraction", The transactions of the Institute of Electronics, Information and Communication Engineers, vol.76, no.9, pp.2042-2052, 1993 (in Japanese)

[3] McCullagh, P. (1980). Regression models for ordinal data, Journal of the Royal Statistical Society, 42(2), 109-142. http://dx.doi.org/10.1109/ICDAR.2001.953882

[4] Watanabe T., Luo Q. and Sugie N., "Layout Recognition of Multi-Kinds of Table-Form Documents", IEEE Transactions of Pattern Analysis and Machine Intelligence, Vol.17, no.4, pp.432-445, 1995. http://dx.doi.org/10.1109/34.385976