Scientific Table Type Classification in Digital Library

Seongchan Kim, Keejun Han
Dept. of Knowledge Service Engineering
KAIST
Daejeon, Korea
sckim, keejun.han@kaist.ac.kr

Soon Young Kim
Dept. of Overseas Information
KISTI
Daejeon, Korea
maya@kisti.re.kr

Ying Liu
Dept. of Knowledge Service Engineering
KAIST
Daejeon, Korea
yingliu@kaist.edu

ABSTRACT
Tables are ubiquitous in digital libraries and on the Web, utilized to satisfy various types of data delivery and document formatting goals. For example, tables are widely used to present experimental results or statistical data in a condensed fashion in scientific documents. Identifying and organizing tables of different types is an absolutely necessary task for better table understanding, and data sharing and reusing. This paper has a three-fold contribution: 1) We propose Introduction, Methods, Results, and Discussion (IMRAD)-based table functional classification for scientific documents; 2) A fine-grained table taxonomy is introduced based on an extensive observation and investigation of tables in digital libraries; and 3) We investigate table characteristics and classify tables automatically based on the defined taxonomy. The preliminary experimental results show that our table taxonomy with salient features can significantly improve scientific table classification performance.

Categories and Subject Descriptors
H.3.m [**Information Storage and Retrieval**]: Miscellaneous; I.2.6 [**Artificial Intelligence**]: Learning – knowledge acquisition.

General Terms
Algorithms, Measurement, Experimentation

Keywords
Scientific tables, IMRAD, fine-grained, taxonomy, classification

1. INTRODUCTION
Tables are ubiquitous in all types of documents such as scientific publications, web pages, financial reports, newspapers, magazine articles, etc. Understanding table type, function, and purpose is crucial for better table understanding and for more accurate table data sharing and reuse. Moreover, automatic identification of the functionality of each document table could be useful for many information-processing tasks, including advanced information retrieval, knowledge extraction [2], mobile access, and data integration.

Tables are differently used to present different types of information with various purposes. Scientists use various types of tables to display such things as experimental results or statistical data for multiple purposes. Scholars can easily obtain valuable insight by examining such type-specified tables. For example, a medical scientist may want to search for tables containing information about "cancer." He or she may want only tables that contain definitions, experimental results, or medical interview questions. However, none of the currently available table search engines (e.g., BioText Search Engine [5], *Tableseer* [7]) support table categorization by type. When issuing a query to these table-specialized search engines, end-users will get only a list of keyword-relevant tables, regardless of their types. The purpose of search by table type is, therefore, to help users to easily recognize the relevance of results by referring to the same type of tables.

Table type related research has recently been receiving considerable attention. Crestan and Pantel [2] report on a census of the types of HTML tables on the Web and propose a fine-grained classification taxonomy. However, their taxonomy is too limited to apply to scientific tables, since table types are heavily dependent on the nature of documents. Kim and Liu [6] first suggest functional-based table types for scientific tables; however, they do not provide fine-grained taxonomy but only two types of table: commentary and comparison. To the best of our knowledge, this is the first study of fine-grained table taxonomy for scientific tables in digital libraries.

In this paper, we observe tables in scientific papers, abstract the underlying table functional-based types, investigate the table characteristics, and demonstrate the distribution of different table types. We focus on scientific papers, since they are one of the most important media in digital libraries and contain many tables (1.28 tables per paper in our dataset), which are all genuine, unlike Web tables [2]. The preliminary experimental results show the effective performance of our system of automatic table type classification. The contributions of this paper are as follows: 1) We propose the finest (IMRAD-based) table functional classification system for scientific documents by considering the structural position of tables within a document; 2) A fine-grained table taxonomy is introduced first, based on an extensive period of table observation and investigation in digital libraries; 3) We investigate the scientific table characteristics and automatically classify tables based on the defined taxonomy; and 4) The whole system and methodology can be easily applied to tables in any fields and formats without much modification in order to achieve a fully automatic table classification and understanding.

The paper is organized as follows. In section 2 we present our table type taxonomy including definitions and descriptions for each category. Section 3 reports the design of our experiment and our results. Finally, we conclude in section 4.

2. Table Type Taxonomy
We propose a brand-new table type taxonomy that is the results of our extensive observation and investigation of 2,500 tables that were randomly collected from 25 randomly selected scientific journals [1] published by Springer from 2006 to 2010 in five

DocEng'12, September 4–7, 2012, Paris, France.

[1] All lists are available at http://issl.kaist.ac.kr/table

domains (Biomedical and Life Science, Chemistry and Materials Science, Computer Science, Electrical Engineering, and Medicine). We manually examined a large sample of tables and propose two table taxonomies in different perspectives: 1) IMRAD-based table taxonomy, which considers the structural position of tables within a document, and 2) fine-grained table taxonomy, which looks further and analyzes table content.

2.1 IMRAD-Based Table Taxonomy

The IMRAD[2] structure is currently the most prominent norm to represent the document structure of a scientific paper. Many scientific journals now prefer this structure and have adopted the IMRAD, which is an acronym for Introduction, Methods, Results, and Discussion, as an instructional device for their authors, recommending the usage of the four terms as main section headings. We define IMRAD-based table taxonomy by borrowing the IMRAD paper structure. In this taxonomy, table type is simply decided by the location of the table; in other words, in which section the table appears. For example, if a table is in the introduction part of the paper, the type of that table is introduction. Though this method is coarse-grained, it can help us to understand the basic purpose of the authors who designed each table.

Introduction Tables The introduction section of a paper is designed to inform readers of the background of the research. It usually includes a short preface or relevant background that leads to a statement of the problem that is being addressed. Tables in these sections are usually used to supplement the explanation of the background theory, to analyze the related studies, and to list statistical data.

Methods Tables The methods section of the paper usually addresses contents in various degrees of detail, methodologies, materials (or subjects), and procedures. Tables in these sections are mainly used to present the system details, itemize the theoretical steps, and explain the implementation procedures.

Results Tables The results section of the paper contains the major findings of the experiment, which were performed to approve the research question, topic, and hypothesis suggested in background part. Tables in the results section are widely used for describing results found and what has been learned in the study. Results tables are often accompanied by others in order to allow a comparing of results as well as a presentation of commentary about experimental results and an organization of findings.

Discussion Tables The discussion section of the paper usually offers interpretations and conclusions about the findings. Tables are used to support interpretation, conclusion, and discussion.

2.2 Fine-Grained Table Taxonomy

In this section, we present a fine-grained table taxonomy, organized according to what the tables contain and by the purposes for which they are used. Since there has been no previous comprehensive study that has attempted to identify and classify tables found in scientific articles in any detailed degree, we both define the table types and describe them. Our analysis yields seven functional types for tables[1].

Definition Tables These consist of defining terms and their explanations. It is very nice to provide readers with definitions of the terms that for the protocol of the study by using a compact table. This type of table usually appears before the experiment and results sections. When the experiment section begins, the number of occurrences of this type of table becomes significantly smaller.

Statistics/distribution tables Statistics/distribution tables are often used to support the main topic of the paper by citing common statistical or distribution data that has been used in the other work or was presented by others. We define statistics/distribution tables as limited to tables whose contents are not related with the current experiment being carried out in the paper. Statistics/distribution tables are usually confused with experimental results tables because experimental results tables also frequently use statistical results to show the outcome of experiments. In this case, the classification between the two types of tables becomes clear when we consider the location of the table in the paper by again using the IMRAD approach.

Survey question/result tables Survey question/result tables contain questionnaires of the survey and the results of those questionnaires. They list a series of questions that have been used to find out information about interviewee's opinions or behavior, usually by asking the questions to the interviewees; or, these tables present detailed output of certain examinations.

Example tables Example tables show instances that introduce and emphasize something that needs to be explained clearly. These tables consist of a target and its examples and instances. They focus on one of the possible consequences of experiments rather than explaining all of the outcomes.

Procedure tables Procedure tables describe the sequence, step, flow, or schedule of the methods. This type of table describes steps, processes, or sequences of a task with a timeline. Sometimes, they are drawn in the form of a pseudo code. This type of table can be easily defined due to its clear contents, which are organized in a logical flow according to the introduced algorithms or methods in the paper.

Experimental setting tables Experimental setting tables can be described as having items required for the experiment including configurations, parameters, data, apparatus, etc. Such tables contain necessary arrangements for the experiment that was performed in the study.

Experimental result tables Experimental result tables are accompanied with a summary describing the output of the experiment. The results are usually shown using specific measures to evaluate the performance of the methods. Some are shown comparing the results with results from other well-known methods; others include statistics and distribution.

2.3 Table Type Distribution

In order to estimate the proportion of table types with our taxonomies, we manually annotated 2,500 tables according to these categories by hiring 15 domain specialists in five domains. Three annotators were assigned per domain; as a result, each table was judged by three annotators. Annotators were instructed to classify a given table according to the IMRAD-based and the fine-grained taxonomy. The inter-annotator agreement among the three annotators was $k = 0.64$ for IMRAD annotation, considered to indicate substantial agreement; the agreement was $k = 0.53$ for fine-grained annotation, considered to indicate moderate agreement according to Fleiss kappa [3]. Finally, we obtained only 2,380 and 2,324 tables that had agreed on labeling from more than two annotators in IMRAD and fine-grained annotation, respectively; we considered these tables to be our Gold Standard for IMRAD and fine-grained classification. We also obtain 2,293

[2] http://en.wikipedia.org/wiki/IMRAD

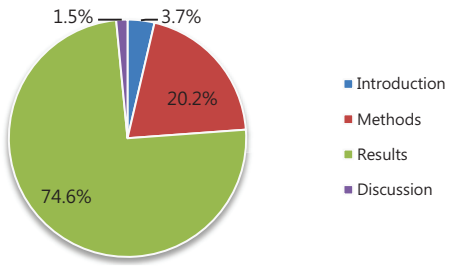

Figure 1. Distribution of Tables with IMRAD Taxonomy

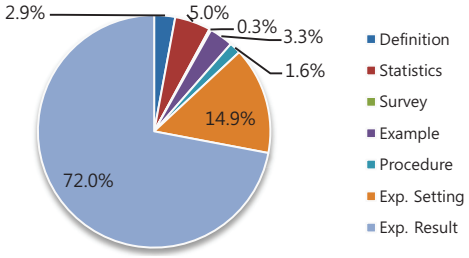

Figure 2. Distribution of Tables with Fine-grained Taxonomy

tables that were agreed on by more than two annotators using both IMRAD and fine-grained annotation at a same time.

Figures 1 and 2 show the distribution of the tables by each taxonomy with 2,293 tables. According to IMRAD classification, the dominant majority of tables were the results tables (74.6%); the second highest number of tables was methods tables (20.2%). As we expected, scientists use tables in the results and methods section, more than in other sections of their papers (94.8%). In fine-grained classification, 1,651 (72.0%) of the tables we analyzed dealt with experimental results, 342 (14.9%) dealt with experimental setting, and, surprisingly, only 6 (0.3%) were in the survey question/result. This last value is very small and different from what we had predicted. However, the percentage of this type of table is expected to heavily increase if we include other fields (such as the field of the humanities). We summarize that tables in scientific papers are highly skewed toward experimental results; further, we assert that scientists usually employ tables to describe their experiment, rather than for other purposes.

In Figure 3, we show a further estimation of the proportion of fine-grained type tables based on the IMRAD taxonomy. In other words, we can find scientific authors' table usage in papers with Figure 3. In the introduction, example, definition, and statistics/distribution tables (87%) were mainly found to appear. The methods section mainly consists of experimental setting tables (60%) and a small portion of other types of tables. In results and discussion, chiefly experimental result tables are found (92%, 83%). With the results, we discover that authors move from general discussion of the topic using various tables (example, definition, and statistics/distribution tables) and details settings in the methods section, to reporting on and discussing the topic with results tables in the latter part of the paper. This is parallel to the general IMRAD writing style.

3. EXPERIMENTS

We automatically conducted a preliminary classification for the IMRAD and the fine-grained taxonomy by using textual analysis of the tables. In this study, we only focus on table classification with complete tables which were extracted by *TableSeer*. Table extraction is not scope of this study.

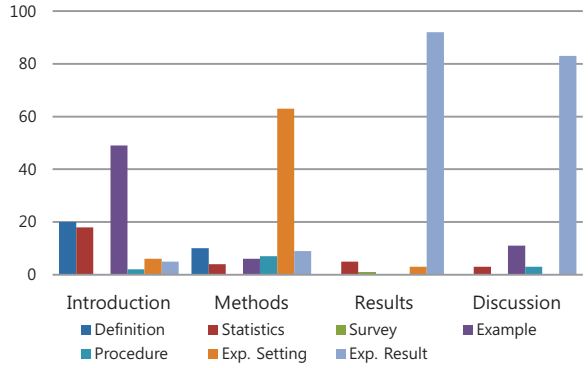

Figure 3 Fine-grained Types based on IMRAD taxonomy

3.1 Experimental Settings

We extracted 2,500 tables using *TableSeer* [7] and used 2,380 tables for IMRAD and 2,324 tables for fine-grained classification. First, we considered table captions and reference texts, which were obtained from table metadata extractor of *TableSeer*. Table captions are caption (sentence(s)) that appear along with the tables; table reference text is the text in the document body that refers to the table and discusses the content of the table. Only unigram tokens were extracted from the texts, tokens were converted to lowercase, and stemmed; however, stopwords were removed.

3.2 Textual Features

In this experiment, we present only the textual features necessary for our results to be discriminative enough to determine table type. Given a table caption and reference text T, we compute the following features and feed them into the classification algorithms. Feature selection was performed by chi-square to determine the correlation of each term with a desired class. The top 300 terms were used as features; all other terms were discarded. Once a decision was made about what to consider as a feature term, the meaning of the numerical feature had to be determined. Several different techniques can be used to generate feature values: binary, Term Frequency (TF), or Term Frequency-Inverse Document Frequency (TF-IDF). However, we designed an innovative term weighting scheme: Table Term Frequency-Inverse Category Frequency (TTF-ICF), which is a tailored Table Term Frequency-Inverse Table Term Frequency (TTF-ITTF) [7]. Because Liu et al. [7] showed the advantages of TTF-ITTF in table search by calculating the term frequency in the table metadata instead of in the whole document, and because Cho and Kim [1] showed the effectiveness of using TF-ICF over TF-IDF for text categorization, we combined these two ideas for our table classification work. Let's assume that category C_1 has tables T_1 and T_2, and that category C_2 has tables T_3 and T_4. A table term W_1, which is a word in the table metadata, appears in T_1 and T_2; W_2 appears in T_1 and T_3. In this case, W_1 is more powerful than W_2 to determine the category; however, two words have the same weight by TTF-ITTF while W_1 is given more weight than W_2 by TTF-ICF.

C_1: T_1, T_2 $\quad$ W_1 : appears T_1 and T_2

C_2: T_3, T_4 $\quad$ W_2 : appears T_1 and T_3

TTF-ICF estimates the term weighting as follows:

$$W_i = freq_i * \log(M) - \log(CF_i) + 1$$

where $freq_i$ is the term frequency of the table term in the table captions, reference text, or both, M is the total number of categories, and CF_i is the number of categories that contain the term W_i.

Table 1. Performance of IMRAD Classification by Features

Features	SVM			Decision Tree		
	P	R	F	P	R	F
Cap.(Baseline)	0.836	0.506	0.543	0.947	0.550	0.792
Ref.	0.875	0.705	0.761	0.930	0.639	0.73
Cap.+Ref.	**0.967**	**0.784**	**0.866**	0.938	0.746	0.831

Table 2. Performance of Fine-grained Classification by Features

Features	SVM			Decision Tree		
	P	R	F	P	R	F
Cap.(Baseline)	0.627	0.333	0.397	0.522	0.271	0.302
Ref.	0.707	0.615	0.649	**0.790**	**0.673**	**0.716**
Cap.+Ref.	0.701	0.657	0.668	0.764	0.62	0.671

Table 3. Performance of IMRAD Classification by Types

Type	SVM			Decision Tree		
	P	R	F	P	R	F
Introduction	0.968	0.6	0.741	0.907	0.68	0.777
Methods	0.901	0.996	0.943	0.912	0.992	0.95
Results	1	1	1	1	1	1
Discussion	1	0.543	0.704	0.933	0.314	0.44
Macro Avg.	**0.967**	**0.784**	**0.866**	0.938	0.746	0.831
Micro Avg.	**0.977**	**0.976**	**0.973**	0.973	0.975	0.972

Table 4. Performance of Fine-grained Classification by Types

Type	SVM			Decision Tree		
	P	R	F	P	R	F
Definition	0.689	0.609	0.646	0.837	0.522	0.643
Statistics	0.699	0.879	0.779	0.898	0.681	0.775
Survey	0	0	0	0	0	0
Example	0.716	0.725	0.72	0.875	0.525	0.656
Procedure	0.9	0.486	0.632	1	0.649	0.787
Exp. Setting	0.905	0.9	0.902	0.74	0.963	0.837
Exp. Result	1	1	1	1	1	1
Macro Avg.	0.701	0.657	0.668	**0.764**	**0.62**	**0.671**
Micro Avg.	0.947	0.947	0.946	**0.94**	**0.936**	**0.987**

3.3 Experimental Results

The classification was conducted with a 10-fold cross validation. We used *SVM* and *Decision Tree*, which have been widely adopted for classification, in Weka tookit [4]. We used default parameters in Weka for classifiers. We report 3 measures: *precision (P)*, *recall (R)*, and *F-measure (F)*. Tables 1 and 2 show the results (macro average of each table type class) of IMRAD and fine-grained classification using different features. The results clearly reveal the effectiveness of the lexical feature. SVM with both captions and reference text gives the best result in IMRAD while Decision Tree with only reference text does in fine-grained. Performance is increased by 59.5% (F-measure) in IMRAD and by 137% in fine-grained when comparing with only captions (baseline). Using both captions and reference text shows better performances without the case of Decision Tree in fine-grained.

In order to analyze the performance by table type, the IMRAD and fine-grained classification results using both captions and reference text for each table type with the macro and micro average are reported in Tables 3 and 4. Macro average of 0.866 (F-measure) with SVM in IMRAD and 0.671 with Decision Tree in fine-grained were best achieved. One of the remarkable things is the performance of fine-grained. In Figure 2, experimental result type is 72.0% of all 7 types in fine-grained. This implies that, hypothetically speaking, for any given table, if we declare it as experimental result type in fine-grained, then our expected accuracy would be 72.0%. However, the results reported in Table 4 do not assure that our classifier is really good. For example, the macro average F-measure of our fine-grained is low (0.67), because the performance of the survey question/result class is 0, whereas the experimental result class is 1. This discrepancy is thought to originate from our data set's being very skewed; the number of samples for the survey question/result type is definitely small (i.e. only 0.6% of the data is survey question/result while 74 % is experimental results). Though we have the results, we still propose these seven types based on a consideration of all the fields in scientific digital libraries. Since our experimental data were limitedly selected from only five domains, it is natural to have skewed results. Once we change the table source and enlarge the data scalability, the performance will be greatly improved.

4. CONCLUSIONS

In this paper, we introduce our study of table types and classifications in scientific papers; we report our experimental results, which not only delineate the underlying table types, but also demonstrate the distribution according to those types. With preliminary experiments, we show the effective performance of automatic type classification, which can be used to better understand table contents and author motivation. For future work, we will develop salient features from the table layout and content to improve the overall performance, not only in digital libraries, but also in any other fields with tables.

5. ACKNOWLEDGEMENTS

We thank Jinhyuk Choi and Jinsup Shin for helpful discussions. This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2012-0004316)

6. REFERENCES

[1] Cho, K. and Kim, J. 1997. Automatic Text Categorization on Hierarchical Category Structure by using ICF(Inverted Category Frequency) Weighting KOREA INFORMATION SCIENCE SOCIETY, 507-510.

[2] Crestan, E. and Pantel, P. 2011. Web-scale table census and classification. In *Proceedings of the fourth ACM international conference on Web search and data mining* (Hong Kong, China2011), ACM, 1935904, 545-554.

[3] Fleiss, J.L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin 76*, 5, 378-382.

[4] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I.H. 2009. The WEKA data mining software: an update. *SIGKDD Explor. Newsl. 11*, 1, 10-18.

[5] Hearst, M.A., Divoli, A., Guturu, H., Ksikes, A., Nakov, P., Wooldridge, M.A., and Ye, J. 2007. BioText Search Engine. *Bioinformatics 23*, 16, 2196-2197.

[6] Kim, S. and Liu, Y. 2011. Functional-Based Table Category Identification in Digital Library. In *Proceedings of the 11th International Conference on Document Analysis and Recognition* (Beijing, China2011), 1364-1368.

[7] Liu, Y., Bai, K., Mitra, P., and Giles, C.L. 2007. TableSeer: automatic table metadata extraction and searching in digital libraries. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries* (Vancouver, BC, Canada2007), ACM, 1255193, 91-100.