Functional-based Table Category Identification in Digital Library

Seongchan Kim, Ying Liu
Department of Knowledge Service Engineering
Korea Advanced Institute of Science and Technology
291 Daehak-ro, Yuseong-gu, Daejeon, Republic of Korea
{sckim, yingliu}@kaist.ac.kr

Abstract – Better understanding the document logical components is crucial to many applications, e.g., document classification or data integration. As the development of digital libraries, more people realize the importance of the scientific tables, which contain valuable information concisely. Although tons of previous table works focus on table data extraction, few concrete works on understanding and utilizing the extracted table data exist. Based on a large-scaled quantitative study on scientific papers, we believe that identifying the original purpose of the table authors can improve the table data comprehension and facilitate the table data reusability. In this paper, scientific document tables are classified into three topical categories: background, system/method, and experimental, and two functional categories: commentary and comparison. We apply machine learning based methods to implement the table classification task. Our results demonstrate that the proposed features are effective in the classification performance and our proposed method outperforms the rule-based baseline significantly**.**

Keywords- document table; content analysis; table category, function-based classification

I. INTRODUCTION

Table is one of the effective document logical components, which is widely used to compactly present and communicate complex and important information. The concise layout and tabular structure can provide a better information presentation, heavily reduce the time and efforts to read and digest the contents, and improve the overall knowledge capture process. However, with a history that pre-dates that of sentential text, table has not received enough formal investigation and characterization, comparing with other document components, e.g., figures as well as the free-texts. Tables are ubiquitous in many applications for different purposes, for example, introducing the latest experimental results in scientific papers, comparing the product prices in web pages, tracking the stock fluctuation in mobile devices, summarizing the historic data in business reports, showing the financial data in white reports, listing course names in transcripts, or even displaying unrelated information pieces as business advertising strategies to increase the marketing shares and to improve the interface visualization, such as the product categories in online shopping systems.

Understanding table types, functions, and purposes is crucial for a better table understanding, table data sharing and reuse. Moreover, automatic functionality identification of each document table could be useful for many information-processing tasks, including advanced information retrieval, knowledge extraction, article summarization, document classification, mobile access and data integration, etc. While there have been some table analysis works, most of them focus on table boundary identification [1] and table structure decomposition [2]. Even researchers mentioned the table-related applications such as table search [3] or table classification [4], no research investigates table understanding from the functionality perspective. To the best of our knowledge, our paper is the first one to address this issue.

Similar to the table data extraction, table functionality analysis is a document medium-dependent task. Because of the natures and functions of different document media, the functionalities of scientific tables are different from that of web tables. According to WebTable system [5], only 1% web tables contain meaningful information, which is valuable for data sharing and reuse. Wang and Hu [4] classified web tables into genuine tables and non-genuine tables, based on relations among table cells. However, there is no further analysis on the genuine tables. In this paper, we focus on scientific tables in digital libraries. Authors usually adopt tables to display the most important information in a document, to draw more attention from readers. For example, researchers always adopt tables to display their latest experimental results as well as the statistical information. Other researchers, who conduct similar studies in the same topic, can quickly obtain valuable insights by examining these tables. For example, a bio-chemist may want to find tables containing experimental results about "mutant genes" or an economist may look for the tables with "the GDP growth of USA in 2000 -- 2007".

Accurately collecting and identifying such specific table data are our two research problems. In addition to experimental tables, another important function of scientific tables is information comparison. If a researcher wants to understand the state of the art of an unfamiliar topic, all the tables in the survey papers should be the best materials to start with. The existing search engines cannot provide a solution for such function-based table search demands because of the following reasons: 1) Most of existing search engines is document-based instead of table-based. Accurately extracting tables from document repositories is a challenging problem. 2) Even with collected tables, how to filter out the interesting ones is another task. Several table search engines, e.g., *TableSeer* [3] and *Biotext* [6] try to answer these questions; however, the results are still not satisfying because of the low precision. Not all the returned tables satisfy the specific functionality requirement. A table

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.274

with "mutant genes" keywords may only introduce the definition and types instead of the latest experimental results. How to automatically detect the scientific tables from digital libraries and further classify them according to the functionalities are important problems to support knowledge sharing and inter-disciplinary collaboration.

Table categorization has a broad application potential. Categorization-empowered search engine is a typical one to reduce user's cognitive burden. For a search query "mutagen gene", all the matched tables can be automatically organized into three different types according to the functionalities: background, system/method, and experiment. Moreover, user can easily spot the target tables, which simply display the contents, or contain rich comparison and analyzed contents.

Overall, our contribution is considered in two aspects: 1) defining table type categorization in terms of contents and functions, which is a forerunner for table sharing and reuse, triggering related issues about table analysis in different angles; 2) identification and investigation of useful features for function-based table classification.

The paper is organized as follows. Section 2 and 3 present our definition on scientific table type and classification method, result and discussion. We discuss related work in section 4 and close with conclusion in Section 5.

II. Scientific Table Types

There are several types of research publication that appear in journals due to the nature of the science. Astrophysicists, theoretical physicians, mathematicians tend to publish logical argumentation papers that have a general-specific organization while empirical scientists such as chemists, biologists, and computer scientists often follow the standard Introduction-Method-Results-Discussion format (IMRD) [7] or some variants of it in constructing their research papers. Since tables are embedded in almost every research paper, the table essence is seriously affected by the organization of the paper and the intention of authors.

Since there is no previous salient study on defining scientific table types, we made the first attempt based on the affiliated table descriptions. We suggest scientific table category classification from two aspects: table content materials perspective and table function perspective, based on two reasonable and promising considerations: IMRD research paper organization pattern, and table reuse and sharing.

A. Content Material-Based Table Classification

According to the spatial features of tables in digital libraries, we can classify them into three categories: background, method/system, and experiment. Figure 1 shows examples of each type of table.

1) Background Tables

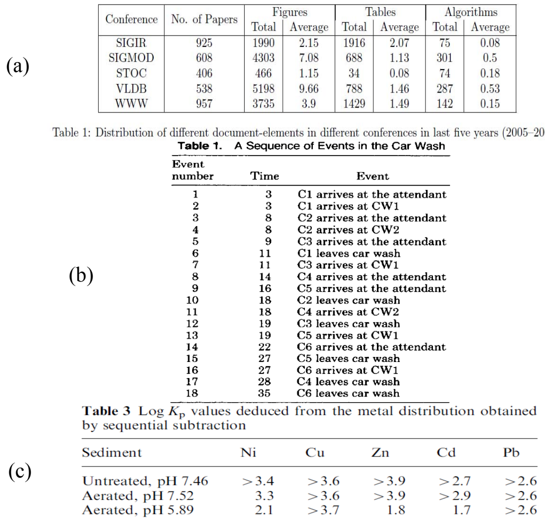
The main aim of the background part of the paper is to supply the particular research question, topic and hypothesis being studied from general discussion to narrow scope. Table in this category is used in supplementing the explanation about background theory, listing and analyzing the related studies, statistics and data, and introducing the paper contribution and implementation agenda to readers.

2) Method/System Tables

Method and system sections address system methodologies, formal procedures, and theoretical materials in various levels. Tables in these sections are always used to discuss the system details, itemize the theoretical steps, and explain the implementation procedures.

3) Experiment Tables

In experiment and discussion sections, results found and what has been learned in the study are offered referring research question, topic, and hypothesis suggested in background part. Tables about those are accompanied for presenting commentary of experimental result and organizing findings, and comparing their results with others.



(a)

Conference	No. of Papers	Figures		Tables		Algorithms	
		Total	Average	Total	Average	Total	Average
SIGIR	925	1990	2.15	1916	2.07	75	0.08
SIGMOD	608	4303	7.08	688	1.13	301	0.5
STOC	406	466	1.15	34	0.08	74	0.18
VLDB	538	5198	9.66	788	1.46	287	0.53
WWW	957	3735	3.9	1429	1.49	142	0.15

Table 1: Distribution of different document-elements in different conferences in last five years (2005–2009).

(b)

Table 1. A Sequence of Events in the Car Wash

Event number	Time	Event
1	3	C1 arrives at the attendant
2	3	C1 arrives at CW1
3	8	C2 arrives at the attendant
4	8	C2 arrives at CW2
5	9	C3 arrives at the attendant
6	11	C1 leaves car wash
7	11	C3 arrives at CW1
8	14	C4 arrives at the attendant
9	16	C5 arrives at the attendant
10	18	C2 leaves car wash
11	18	C4 arrives at CW2
12	19	C3 leaves car wash
13	19	C5 arrives at CW1
14	22	C6 arrives at the attendant
15	27	C5 leaves car wash
16	27	C6 arrives at CW1
17	28	C4 leaves car wash
18	35	C6 leaves car wash

(c)

Table 3 Log K_p values deduced from the metal distribution obtained by sequential subtraction

Sediment	Ni	Cu	Zn	Cd	Pb
Untreated, pH 7.46	>3.4	>3.6	>3.9	>2.7	>2.6
Aerated, pH 7.52	3.3	>3.6	>3.9	>2.9	>2.6
Aerated, pH 5.89	2.1	>3.7	1.8	1.7	>2.6

Figure 1. An example of (a) Background, (b) System/method, and (c) Experimental Table

B. Function-Based Table Classfications

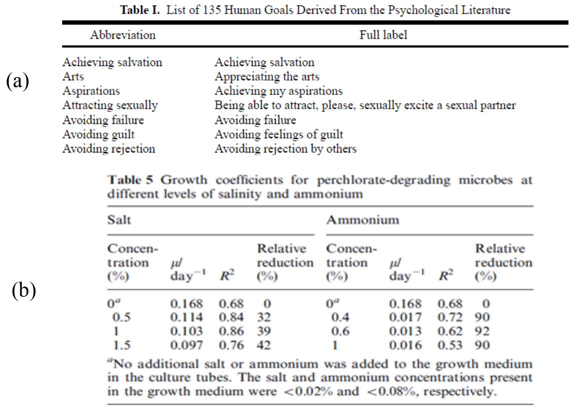
Based on the nature of scientific tables, there are two main functions for its data: commenting facts and drawing a parallel contrast between rows or columns. In other words, authors use the table to deliver the information or make data comparison. Table caption, contents as well as all the associated information are key resources to "reverse engineering" the purposes of table authors, which decide the exclusive table function types. A table that is made for simply displaying and delivering the contents is considered as "commentary type". Contrary, if the author of a table analyze and compare the contents and give emphasis on the differences among the contents, we can consider it as a "comparison" table. Figure 2 shows an example of commentary and comparison tables.

a) Commentary Tables

Commentary tables usually list, describe and comment the contents. These tables are widely used to provide the auxiliary information on objects listed in tables. Usually such tables contain two main parts: the basic items (the column 1 in Figure 1) and the detailed annotation (the column 2 in Figure 1). The basic items are often introduced in the reference texts already. Most of the cases, the structure of such tables is also simple and neat. Attributes in rows and columns are often listed individually instead of grouped together.

b) Comparison Tables

In contrast, comparison table is weighted in comparing, contrasting and balancing the contents of the table by juxtaposing the data. Comparison tables contrast different contents, which are usually represented in quantitative data. Comparison is one of the effective means for researchers to report their latest results and show the contribution by contrasting their works with others. Table reference texts usually give a detailed explanation about the attributes with the comparative degree or the superlative degree (e.g. E2E protocol achieves a throughput of less than 50% of the maximum). Furthermore, attributes in row and columns are often grouped and repeated patterns of attributes in groups can be clearly observed.



(a)

Table I. List of 135 Human Goals Derived From the Psychological Literature

Abbreviation	Full label
Achieving salvation	Achieving salvation
Arts	Appreciating the arts
Aspirations	Achieving my aspirations
Attracting sexually	Being able to attract, please, sexually excite a sexual partner
Avoiding failure	Avoiding failure
Avoiding guilt	Avoiding feelings of guilt
Avoiding rejection	Avoiding rejection by others

(b)

Table 5 Growth coefficients for perchlorate-degrading microbes at different levels of salinity and ammonium

Salt				Ammonium			
Concentration (%)	μ/ day^{-1}	R^2	Relative reduction (%)	Concentration (%)	μ/ day^{-1}	R^2	Relative reduction (%)
0^a	0.168	0.68	0	0^a	0.168	0.68	0
0.5	0.114	0.84	32	0.4	0.017	0.72	90
1	0.103	0.86	39	0.6	0.013	0.62	92
1.5	0.097	0.76	42	1	0.016	0.53	90

^{a}No additional salt or ammonium was added to the growth medium in the culture tubes. The salt and ammonium concentrations present in the growth medium were <0.02% and <0.08%, respectively.

Figure 2. An Example of (a) Commentary and (b) Comparison Table

C. Distribution of Scientific Tables

Based on our definitions, we invited domain experts to label 626 and 626 tables from two datasets of empirical sciences: Computer Science (CS) and Chemistry (CH). Table I and Table II illustrate the geographic distribution of these scientific tables respectively. More than 90% of the tables are experimental tables in CH. On the contrary, 77% of experimental tables are appeared in CS. We can notice that most dominant table type of table is experimental in both empirical sciences and computer scientists use various type of table than chemists. For commentary and comparison, near 60 % and 75 % of comparison table are more obtained from CS and CH respectively.

TABLE I. DISTRIBUTION OF SCIENTIFIC TABLE IN CS

Location	Commentary	Comparison	Total
Background	7.8%	2.7%	10.5%
Method/System	10.7%	1.6%	12.3%
Experimental	22.8%	54.3%	77.2%
Total	41.4%	58.6%	100.0%

TABLE II. DISTRIBUTION OF SCIENTIFIC TABLE IN CH

Location	Commentary	Comparison	Total
Background	0.5%	5.1%	5.8%
Method/System	1.3%	2.1%	3.4%
Experimental	23.3%	67.6%	90.9%
Total	25.5%	74.8%	100.0%

III. TABLE TYPES CLASSIFICATION

A. Experimental Table Detection

An important part of documents related to empirical and experimental science has the experiment section that reports experimental data, settings, results and observations. In order to support researchers, who conduct similar studies in the same topic to quickly obtain valuable insights by examining the experimental tables, we automatically identify, extract and collect experiment-related tables.

Using scientific papers in experiment-intensive areas as examples, authors usually address an experiment in the following manner: a description of instruments used such as models, instrument characteristics, instrument calibration, even how reagents are prepared. Authors then describe experimental procedure and reactions observed during the experiment. Finally, they display results and analyze results in different conditions. The challenging problem is how to automatically detect and extract such information.

Since experimental tables have certain styles, we transform the problem of collecting experimental tables into a binary classification task: experimental tables or non-experimental tables. Considering each table as an instance, which we denote as t_i, each table is either related to experiments or not. There is a set of features $\{f_{ij} \mid j=1\ldots n\}$, which i denotes which table feature belong to and n denotes the total number of features.

We use LibSVM[1], a library of Support Vector Machines (SVMs) [8], which has been widely used for classification tasks. We choose the RBF (Radial Basis Function) kernel, since it shows the best performance in our preliminary experimental results.

1) Dataset and experimental environment

In this paper, we focus on the tables in PDF scientific documents. The document collection comes from two sources: 1) Computer Science (CS) papers in the *CiteSeer*[2] archive and 2) chemistry (CH) papers in *Royal Chemistry Society* [3]. We obtained 626 tables from 244 randomly selected CS papers and 626 tables from 259 randomly-selected chemistry papers in PDF formats. Students from both fields are hired to evaluate the results based on our definition. We had 483 experimental tables and 143 non-experiment tables for CS and 569 experimental tables and 57 non-experiment tables for CH.

For table metadata extraction, we used *TableSeer* [3], which is an automatic table extraction and search engine system. *TableSeer* extracts an extensive set of medium independent metadata for table representation (e.g. caption, reference texts, and contents in the cells). Our classification features are generated based on the metadata.

2) Features

Feature selection plays an important role to performance of classification. Because of the space limitation, we present three typical features in the follow.

[1] http://www.csie.ntu.edu.tw/cjlin/libsvm/

[2] http://citeseer.ist.psu.edu

[3] http://www.rsc.org/

- Keywords: It defines representative words that are frequently appeared in experiment table captions and reference texts. We manually collected these keywords (Table III) appeared in the caption and the reference texts of annotated experimental tables. Different keywords are highlighted in Italics.

TABLE III. EXAMPLE OF EXPERIMENTAL KEYWORDS (STEMMED)

	Keywords
CS	experimen, run, test, perform, prepar, procedur, instrument, measur, evaluat, estimat, result, parameter, variat, execut, improve, effect, siginificant, *precision, cros-valid, overhead, classiflc, compris, subject, negoti, mispredict, assess, outperform, reformul, constrain, useful, finding, dataset*
CH	experimen, run, test, perform, prepar, procedur, instrument, measur, evaluat, estimat, result, parameter, variat, execut, improve, effect, siginificant, *rate, contrast, control, case, expect, agree, analys, exposur, decrease, apparatus, reagent, react, prepar, calibrat*

- Position: This feature represents the relative location of a table within the paper. Experimental tables tend to appear in latter part of the paper. A percentile value from 0 to 1 (0: the beginning, 1: the end of the paper) is used to label the table position.
- Cell-contents type: We are interested in two types of cell data (numerical and textual) because we observed that numerical data is more prevalent than textual data in the experimental tables.

3) Experimental Results

The evaluation was conducted with a 10-fold cross validation. The rule-based method is adopted as the baseline. Our rule-based approach assigns a positive label to any table, which contains at least *k* keywords. The keywords are same as the keyword features adopted in SVM classifier. We present micro-averaged precision (P), recall (R), and F-measure (F) as measurements. Table IV lists the results of both rule-based methods and SVM. The number of keywords *k* varies from 1 to 3.

TABLE IV. EXPERIMENTAL TABLE DETECTION PERFORMANCE

Method		Precision	Recall	F-measure
SVM	**CS**	**0.83**	**0.84**	**0.822**
	CH	**0.889**	**0.98**	**0.94**
Baseline (k=1)	**CS**	**0.713**	**0.777**	**0.740**
	CH	**0.837**	**0.891**	**0.863**
Baseline (k=2)	CS	0.756	0.681	0.717
	CH	0.841	0.845	0.842
Baseline (k=3)	CS	0.783	0.62	0.691
	CH	0.847	0.79	0.815

The results clearly show that our approach with SVM makes the most accurate detection performance: 0.822 for CS and 0.94 for CH in F-measure. Comparing with the baseline with (k=1), our method improved the F-measure by 11.08% for CS and 8.9% for CH respectively.

The main reason of the inconsistent performance between CS and CH fields is the diverse table layouts. Based on our quantity study on the table characterization, we observed an interesting phenomenon about the table formats: tables in CH repository are usually well-defined and standardized, from the caption designing to the measurement of numerical data. However, authors of CS tables are more flexible on the table designing. We could improve the performance by adding more features, such as the section where table belongs to (e.g., background, method/system, and experiment). However, only CS papers follow IMRD pattern for organization. Most CH papers only contain experimental contents or theoretic description. Accurately identifying sections is another challenging problem.

B. Functional-Based Table Type Classification

In this section, we classify tables from the functional perspective: commentary and comparison.

1) Dataset and experimental environment

We used the same datasets and evaluators described above. In result, we identified 259 commentary tables and 367 comparison tables in 626 CS tables and 158 commentary tables and 468 comparison tables in 626 CH tables. *TableSeer* [3] generates a different set of features from table metadata for table type classification.

2) Features

- Keywords: It includes commentary and comparison representative words that are frequently appeared in table caption and reference texts while seldom occurring in other parts. Exemplary keywords for comparison are “compare”, “contrast”, “balance”, “outperform”, and “prefer”, while commentary keywords are “example”, “highlight”, “specify”, “reference”, “belong”, and “pool”. We notice that commentary keywords are focusing on delivering information and comparison keywords are focusing on comparison.
- Cell-content types: How many percentages of numerical and textual of cell data exist in a table? Based on our observation, comparison table has more numerical data than textual data.
- Structure complexity: Comparison table is more complex than commentary table. Comparison table often has nested cells and complicated structures. This feature is determined by the number of cells in the table stub and box.
- Attribute patterns: It is one of the distinctive factors, which is often appeared in comparison tables. Attributes in stub and box usually repeat regularly in comparison tables.
- The length of the table reference texts: Based on our observation, the reference text of comparison tables is longer than that of commentary tables. Especially, comparison table has rich comparing explanation in their reference texts since authors give detailed explanation comparing their research results with others or research results in their works. On the other hand, authors tend to simply mention commentary table with short description.
- The comparative degree and the superlative degree: Comparison-revealing sentences tend to have comparative and superlative expressions. We used the Stanford POS tagger [9] for detecting the

comparative and superlative degree with the related Penn tagsets. (e.g. JJR - Adjective, comparative; JJS - Adjective, superlative; RBR - Adverb, comparative; and RBS - Adverb, superlative).

3) *Results*

The evaluation was conducted with the same method in experimental table detection. Table V shows the results of both rule-based methods and SVM.

TABLE V: TABLE TYPE CLASSIFICATION PERFORMANCE (CS)

Method		Precision	Recall	F-measure
SVM	**CS**	**0.67**	**0.672**	**0.66**
	CH	**0.702**	**0.753**	**0.675**
Baseline (k=1)	**CS**	**0.613**	**0.471**	**0.485**
	CH	**0627**	**0.672**	**0.648**
Baseline (k=2)	CS	0.662	0.315	0.369
	CH	0.652	0.491	0.557
Baseline (k=3)	CS	0.685	0.220	0.292
	CH	0.648	0.385	0.477

For table classification, the best performance, about 0.66 in F-measure score is obtained from SVM for CS. We achieve 0.675 in F-measure from SVM for CH. We obtain 36% of improvement than baseline (k=1) for CS and 4% than baseline (k=1) for CH. Although the results are still far from satisfying, the main reasons come from the diverse table structures, the fuzzy terms, the unclear purposes, etc. For many tables, there are not absolute correct answers. Even our professional evaluators with domain knowledge disagree with each other. However, we are the first to consider the table understanding from this novel perspective. Also we made an important tentative experiment. Our improvement will include optimizing feature selection, extending testing dataset, involving more domain knowledge, etc.

IV. RELATED WORKS

In this paper, we focus on the scientific tables in PDF because of two reasons: First, PDF gains popularity in digital libraries due to the compatibility of output on a variety of devices. Second, PDF documents are overlooked in table analysis field. The limited PDF table analysis works can only extract small-scaled visually-defined tables from PDFs without any further table understanding and classification. After the table identification and data extraction, some researchers try to associate the table extraction with question answering (QA) or information retrieval. The only table classification work is finished based on Web tables. Wang and Hu [4] designed a system that extracts table-related information, stores them in databases, and generates a man-machine dialog to access the table data via a spoken language interface. They finished the first web table classification based on the table content relationships: genuine tables or non-genuine tables. Google claims [5] that only 1.1%--1.6% web table are genuine tables with meaningful relationships, all the others are "non-genuine" (false) tables for displaying. Classification is an important step to identify the genuine ones. However, almost 100% scientific tables are genuine tables. Hu and Bagga's work [10] on functionality based web image classification is relevant. Even though it is on image, it originally proposed classification based on functionality rather than content. There is no similar classification work on tables and our study is the first functional-based classification on the scientific tables.

V. CONCLUSION

To facilitate semantic analysis for a table, we present the definition of table types based on its topic and function. Automatic classification method and feature set are proposed to extract specific type of tables from a set of tables in scientific papers. For extracting experimental tables, a new feature set is proposed combining multiple factors. We also adopt a novel feature combination to distinguish commentary and comparison table. Experimental results demonstrate that classification was very promising in both experimental table detection and type classification. Our next work is to extend table function-based identification and classification on various different datasets (e.g., UW3, UNLV dataset and Web tables) beyond scientific tables.

ACKNOWLEDGMENT

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(2011-0014733)

REFERENCES

[1] Y. Liu, P. Mitra, and C. L. Giles, "Identifying Table Boundaries in Digital Documents via Sparse Line Detection," in Proceedings of CIKM-08, Napa Valley, California, 2008.

[2] T. G. Kieninger, "Table structure recognition based on robust block segmentation," In Proc. Document Recognition V, SPIE, volume 3305, pp. 22-32, January,1998.

[3] Y. Liu, K. Bai, P. Mitra, and C. L. Giles, "TableSeer : Automatic Table Metadata Extraction and Searching in Digital Libraries Categories and Subject Descriptors," In Proceedings of JCDL-07, pp. 91-100, 2007.

[4] Y. Wang and J. Hu, "A machine learning based approach for table detection on the web," In Proceedings of WWW, pp. 242-250, 2002.

[5] M. J. Cafarella, A. Halevy, Z. D. Wang, E. Wu, Y. Zhang, "WebTables: Exploring the Power of Tables on the Web," in Proceddings of Vldb, 2008.

[6] M. A. Hearst, A. Divoli, H. Guturu, A. Ksikes, P. Nakov, M. A. Wooldridge, and J. Ye, "BioText Search Engine," Bioinformatics, vol. 23 n.16, pp.2196-2197, 2007.

[7] L. B. Sollaci and M. G. Pereira, "The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey," J Med Libr Assoc. 92(3): 364–371. 2004

[8] C. J. Burges. "A tutorial on support vector machines for pattern recognition," Data Mining and Knowledge Discovery, 2:121–167, 1998.

[9] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, "Feature-Rich Part of Speech Tagging with a Cyclic Dependency Network," In Proceedings of HLT-NAACL. 2003.

[10] J. Hu and A. Bagga, "Categorizing Images in Web Documents", IEEE Multimedia Special Issue on Content Repurposing , January-March, 2004.