Automatic Searching of Tables in Digital Libraries

Ying Liu, Kun Bai, Prasenjit Mitra, C. Lee Giles

The College of Information Sciences and Technology
Pennsylvania State University
University Park, PA 16802
{yliu, kbai, pmitra, giles}@ist.psu.edu

ABSTRACT

Tables are ubiquitous. Unfortunately, no search engine supports table search. In this paper, we propose a novel table specific searching engine, *TableSeer*, to facilitate the table extracting, indexing, searching, and sharing. In addition, we propose an extensive set of medium-independent metadata to precisely present tables. Given a query, *TableSeer* ranks the returned results using an innovative ranking algorithm – *TableRank* with a tailored vector space model and a novel term weighting scheme. Experimental results show that *TableSeer* outperforms existing search engines on table search. In addition, incorporating multiple weighting factors can significantly improve the ranking results.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information Search and Retrieval – search process

General Terms

Algorithms, Experimentation, Documentation, Performance, Design

Keywords

Table search, table crawler, table metadata, table extraction, table indexing, table ranking

1. INTRODUCTION

Tables appear everywhere, from web pages to scientific publications, from financial reports to news papers. Scientists always use tables to display the latest experimental results or statistical data. Tables have gradually accumulated a huge amount of valuable information as the explosive development of the Internet. However, current search engines do not support the table search. When applying a table search query, end-users will receive a flood of unwanted and sometimes unsolicited results from them. Moreover, among the returned documents, the ranking order of the top n results does not precisely reflect the relevance to the queries. Table searching is a challenging problem because of three reasons: the incapability of current search engines to recognize table contents, the impropriate ranking schemes, and the lack of a standard table representation scheme.

WWW 2007, May 8–12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

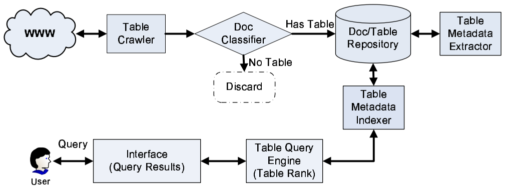


Figure 1: TableRank in the TableSeer System

Our paper has three main contributions: a table search engine *TableSeer*, an innovative table ranking algorithm *TableRank*, and an extensive set of table metadata. Although table-related research received considerable attention, most of them focus on the table extraction from a specific document medium. Although some researchers try to associate the table extraction with question answering (QA) or information retrieval (IR) [5] [6], none of them provides a real table search engine. To the best of our knowledge, *TableSeer* is the first search engine for table search. Empirical results show that *TableSeer* achieves encouraging results. The remainder of the paper is organized as follows. Section 2 presents the architecture of *TableSeer* with explanation for each part. Section 3 discusses the experiment and the result analysis. Section 4 is the conclusion.

2. THE ARCHITECTURE OF TABLESEER

Figure 1 highlights the procedure of the *TableSeer* in handling the table search queries. *TableSeer* crawls documents from the web, classifies them into two groups (document with/without tables) and discards the latter, extracts the metadata [4][3] for each table, and ranks the tables in response to the user query with the *TableRank* algorithm.

2.1 Table Crawler

TableSeer harvests online scientific documents by crawling open-access digital libraries and scientists' web pages. The crawler supports a number of document media, such as PDF, HTML, WORD, PowerPoint, etc. In this paper, we focus on scientific documents in PDF format because it gains popularity in digital libraries and is overlooked in the table extraction and information retrieval fields.

2.2 Table Metadata Extraction and Indexing

We design a universal table metadata representation scheme by classifying the table metadata into six mutually exclusive categories: 1) table environment/geography (document-level), 2) table-frame metadata, 3) table affiliated metadata, 4) table-layout metadata, 5) table cell-content metadata,

6) and table-type metadata. For each identified table, a corresponding table metadata file is created. We design a *page box-cutting* method to detect and extract table metadata (see details in [4]). Table metadata indexer adopts the Lucene Index Toolbox[1] to index and rate the <query, table> pairs instead of the <query, document> pairs. To index a table, a "document" is created where the table metadata fill the "fields".

2.3 Table Ranking

TableSeer search engine adopts an novel table ranking algorithm – *TableRank*. *TableRank* tailors the classical vector space model [1] to calculate the relevance of each <table, query> pair. As shown in *Table 1*, each row represents the vector of a table tb_j or a query q. All the table vectors and query vectors construct a vector matrix. Each table row is compose of k metadata and each metadata is composed of a set of alphabetically ordered terms. $w_{i,j,k}$ is the term weight of the i^{th} term in the k^{th} metadata of the table tb_j and $w_{i,q,k}$ refers to the term weight of the i^{th} query term in the k^{th} metadata. To determine $w_{i,j,k}$, we design an novel term weighting scheme: Table Term Frequency - Inverse Table Term Frequency (TTF-ITTF), a tailored *TF-IDF*[2] weighting scheme. Compared with *TF-IDF, TTF-ITTF* has two major advantages. First, it calculates the term frequency in the table metadata file instead of the document. Second, it calculates the weight of a term with a comprehensive consideration at three levels: the term, the table, and the document level. Cosine measure is used to determine the similarity between the query vectors and the table vectors. The details of the ranking algorithm can be seen in [3].

Table 1: The Vector Space for Tables and Queries

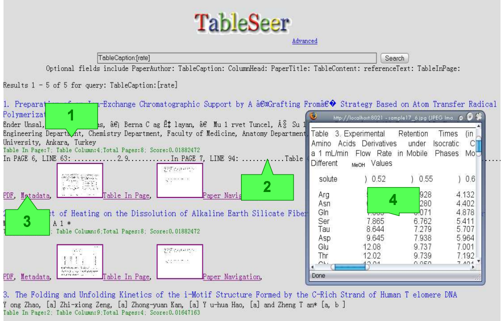
	$m_1(MW_1)$			…	$m_k(MW_k)$		tlb	dlb
	$t_{1,1}$	…	$t_{x,1}$	$t_{1,k}$	…	$t_{z,k}$	…	…
$tb1$	$w_{1,1,1}$	…	$w_{x,1,1}$	$w_{1,1,k}$	…	$w_{z,1,k}$	…	…
$tb2$	$w_{1,2,1}$	…	$w_{x,2,1}$	$w_{1,2,k}$	…	$w_{z,2,k}$	…	…
…	…	…	…	…	…	…	…	…
tb_b	$w_{1,b,1}$	…	$w_{x,b,1}$	$w_{1,b,k}$	…	$w_{z,b,k}$	…	…
q	$w_{1,q,1}$			$w_{1,q,k}$			…	…
$ittf$	…	…	…	…	…	…	…	…

2.4 Query Interface

TableSeer consists of two levels of search: basic search and advanced search. Basic search allows the search with one or more simple search keywords. For the advanced searching, users can set more complex queries. To facilitate the result browsing, *TableSeer* provides a user-friendly interface to present the ranked results (see Figure 2). For each matched table, it not only lists the basic document information (e.g., the document title, the author and the affiliation), highlights the reference texts to the table in the document, but also provides the links for the original PDF document, the table metadata file, and the snapshot of the matched tables.

3. EXPERIMENTAL RESULTS

The total crawled 10000 PDF documents come from three sources: scientific digital libraries (Royal Chemistry Society), the web pages of research scientists in chemistry departments in universities, and the *CiteSeer* archive. We performed a five-user study to evaluate the performance of our *TableSeer*. The evaluation metrics are precision and recall. The experiment on table detection is conducted on a



TableCaption:[rate]

Optional fields include PaperAuthor: TableCaption: ColumnHead: PaperTitle: TableContent: referenceText: TableInPage:

Results 1 - 5 of 5 for query: TableCaption:[rate]

1. Preparat... -Exchange Chromatographic Support by A â€œGrafting Fromâ€� Strategy Based on Atom Transfer Radical Polymerizat...
Ender Unsal, ...as, â€¡ Berna C ag Ë‡ layan, â€ Mu 1 rvet Tuncel, Â§ Su 1...
Engineering Depart...nt, Chemistry Department, Faculty of Medicine, Anatomy Department...
University, Ankara, Turkey
Table In Page:7; Table Column:4;Total Pages:8; Score:0.01882472
In PAGE 6, LINE 63: ....2.9...........In PAGE 7, LINE 94: ....Table
PDF, Metadata, Table In Page, Paper Navi...

2. ...t of Heating on the Dissolution of Alkaline Earth Silicate Fiber...
A 1 *
...: Table Column:6;Total Pages:8; Score:0.01882472
PDF, Metadata, Table In Page, Paper Navigation,

3. The Folding and Unfolding Kinetics of the i-Motif Structure Formed by the C-Rich Strand of Human T elomere DNA
Y ong Zhao, [a] Zhi-xiong Zeng, [a] Zhong-yuan Kan, [a] Y u-hua Hao, [a] and Zheng T an* [a, b]
Table In Page:2; Table Column:9;Total Pages:4; Score:0.01647163

http://localhost:8021 - sample17_6.jpg (JPEG Ima...

Table 3. Experimental Retention Times (in Amino Acids Derivatives under Isocratic C... a 1 mL/min Flow Rate in Mobile Phases Mo... Different MeOH Values

solute	) 0.52	) 0.55	) 0.6
Arg		928	4.132
Asn		280	4.402
Gln		071	4.878
Ser	7.865	6.762	5.411
Tau	8.644	7.279	5.707
Asp	9.645	7.938	5.964
Glu	12.08	9.737	7.001
Thr	12.02	9.739	7.192

Figure 2: An Example of the Query Results by Basic Search

document set with 200 randomly selected PDF documents. Based on testers, the precision and the recall values of table metadata extraction are over 95% respectively. In order to evaluate the *TableRank*, we established a "*golden standard*" to define the "*correct*" ranking based on human judgement and apply *pairwise accuracy* to evaluate the ranking quality. We also set up the common test-bed with the manually "bottom-up" method and the custom search engine method. Experimental results show that *TableSeer* outperforms existing search engines on table search. In addition, incorporating multiple weighting factors can significantly improve the ranking results (See details in [3]).

4. CONCLUSIONS AND FUTURE WORK

In this paper, we present the *TableSeer* system that arms with a novel table ranking algorithm, *TableRank*, to retrieve the tables contained in Web and digital libraries. There are several areas in which we still hope to make progress. First, currently we focus on the scientific documents in PDF format. Next, we will extend to handle other kinds of documents in Web. Second, although we present preliminary results showing the effect of the impact factors proposed, many parameter settings are based on empirical studies. In the future, more extensive experiments are needed to determine more suitable parameter settings.

5. REFERENCES

[1] R. Baeza-Yates and B. Ribeiro-Neto. Modern information retrieval. In *ACM Press/Addison-Wesley*, 1999.

[2] C. B. G. Salton. Term-weighting approaches in automatic text retrieval. In *Information Processing and Management 24(5)*, pages 513–523, 1988.

[3] Y. Liu, K. Bai, P. Mitra, and C. L. Giles. Tableseer: Automatic table metadata extraction and searching in digital libraries. In *Technical Report*, 2006.

[4] Y. Liu, P. Mitra, C. L. Giles, and K. Bai. Automatic extraction of table metadata from digital documents. In *JCDL*, pages 339–340, 2006.

[5] P. Pyreddy and W. Croft. Tintin: A system for retrieval in text tables. In *In Proceedings of the Second International Conference on Digital Libraries*, pages 193–200, 1997.

[6] J. Wang and J. Hu. A machine learning based approach for table detection on the web. In *Proceedings of the 11th Int'l Conf. on World Wide Web (WWW'02)*, pages 242–250, Nov 2002.

[1] http://lucene.apache.org/java/docs/index.html