CWTs: A Public Large-Scale Chinese Web Table Data Set

Changlong Wang¹, Zhilong Ren¹*, Siyun Bi¹, Rong Zhang¹, Qibin Fu¹ and Zhichang Zhang¹

¹ School of Computer Science and Engineering, Northwest Normal University, Lanzhou 730070, China
*Corresponding author’s e-mail: 2019221855@nwnu.edu.cn

Abstract. As the world’s largest information sharing platform, web pages contain a large amount of table data. However, these table data emphasize the presentation of data to users in a more intuitive way, which cannot be understood by computers. How to obtain large-scale web table data and carry out in-depth applications has become the focus of data science. This paper designs an extraction framework for Chinese Web table data. This framework extracts Web table data containing Chinese from the Web and constructs an open Chinese Web table data set CWTs (Chines Web Tables). The data set is quantitatively analyzed in multiple dimensions such as scale and data distribution. The results show that the data set obtained in this paper can provide a preliminary data basis for table data query, question answering system, and knowledge base construction, at the same time provide an optional channel for building a large-scale machine-understandable Chinese knowledge base.

1. Introduction
The development of web technology allows more web users to directly participate in web data editing. Web becomes the world’s largest database sharing and co-construction platform. In addition to unstructured hypertext documents, a large amount of web data also contains a large number of table data represented by HTML table tags (such as <table>, <tr>, <td>). Generally, the use of these table tags is mainly manifested in two aspects: one is to plan the layout of the web page, which is convenient for web front-end developers to control and manage the page module components; the other is to display structured relational data to be more intuitive. The data is presented in a way that is convenient for web users to read and understand. For the tables mentioned below, they often contain structured data used to describe entities and entity relationship events or entities. The tables represent the relationships between entities and the attributes of the entities, as shown in Figure 1.

Figure 1. Web Chinese table.
Figure 1 contains the relationship between the skyscraper entity and the city entity in each place, as well as the related attributes of the skyscraper, such as the height of the building, the number of floors. Each row in the figure represents the relationship between an entity (skyscraper) and other entities (located in the city) and the related attributes of this entity. The table in the figure is called a horizontal table. The first row of this kind of table is the header, which gives the name of each column, and the top of the data table is the title of the table. Another type of table is a vertical table, in which columns are used to represent entities and their attributes [1]. A large amount of web table data emphasizes the display of page content to users in an intuitive table format, which is convenient for people to read and understand. Because the table data lacks semantic information, it cannot be understood by computers, and it is difficult to exert its potential knowledge value [2].

Obtaining tabular data from the web has become a hot spot in fields such as data science and knowledge engineering [3] [4]. For example, web tables are used for the construction of cross-domain knowledge bases [5], the attribute expansion of tables [6] [7] [8], web table knowledge translation [9] and so on. Existing research shows that web table data has potential academic research and practical application value. The web table data sets used by most of the current systems are not the same, and the public cannot access these data.

This article is dedicated to extracting and constructing a Chinese-oriented web table data set in order to provide a test benchmark for the Chinese web table data system. At the same time, it provides a data foundation for the research and application of Chinese web table data. The organization structure of this article is as follows: After introducing the research background; Section 2 reviews existing related work; Section 3 describes the framework of Chinese web table data extraction; Section 4 conducts a quantitative analysis on the data set CWTs; finally (Section 5) summarizes the work of this paper and proposes the next research direction.

2. Related work
In order to facilitate people’s research and use of web data, the Common Crawl project regularly crawls from the web to build an open data set that covers the entire web, Common Crawl [10]. Table 1 lists some existing web table data sets.

Among them, WDC (Web Data Commons, network data sharing) [11] is a research project of the University of Mannheim, Germany under the European Union’s seventh framework, which aims to extract structured data (such as RDF, JSON, and table data) from Common Crawl). The current representative table data sets are WDC2012 and WDC2015. WDC2015 contains about 10.24 billion relational tables, of which relational tables and entity tables account for 0.9% and 1.4% respectively. The number of horizontal tables and vertical tables is 84.78 million and 5.48 million. The average number of columns and rows of the horizontal table are 5.2 and 145.5, respectively.

| Data set          | Table type  | Table capacity | Table source  |
|-------------------|-------------|----------------|---------------|
| WDC 2012 /2015    | Web tables  | 147/233M       | Web crawl     |
| Dresden Web       | Web tables  | 174M           | Web crawl     |
| Tables Corpus     | Web tables  | 154M           | Web crawl     |
| WebTables         | Wikipedia tables | 1.6M   |
| WebTables         | Scientific tables | 0.34M  | arxiv.org     |

WikiTables is 1.6 million high-quality relational tables extracted from Wikipedia by Bhagavatula et al. [12]. Each table is stored as a JSON file, and the data includes table title, page title, table content, column title, number of rows and columns, and connections in the table.

https://github.com/renzhilong52000/CWTs_extract
TableArXiv [13] is a table data set for scientific research. It consists of 341,573 tables, which are taken from the physical electronic version on arxiv.org. It provides 105 information requirements and corresponding correlation judgments for scientific table search tasks.

3. Data extraction framework
The data studied in this paper comes from the data set WDC. This paper uses the regular expression matching method to extend the WDC data extraction framework to identify Chinese web table data. In regular recognition, we make statistics based on the table title, table header, and table data content. The framework uses Amazon’s cloud computing platform, and a master node Master controls each computing module, collects and filters the obtained table information, and its data processing flow is shown in Figure 2.

![Extraction frame](image)

Figure 2. Extraction frame.

1. The master node Master adds pending files to Amazon’s queue server (AWS SQS);
2. In the AWS EC2 environment, the master node launches many instances;
3. The running instance requests tasks from the queue server and starts file processing;
4. Run the instance storage server S3 to download files;
5. The running instance uploads and saves the processed files to the server S3;
6. If the queue is considered empty, the main server starts to collect the extracted data;
7. Categorize web tables;
8. Extract Chinese web table;
9. Divide the tables in CWTs according to different domain names.

This article uses a regular table similar to the following when counting domain names:

\[
\begin{align*}
\text{r1} &= (\text{(?=://)[a-zA-Z0-9]+(?=\\/)}) \\
\text{r2} &= (\text{\.[a-z]{2,5}})
\end{align*}
\]

Among them, r1 is used to extract the entire url, and r2 is used to extract the domain name part of the url for the second time. In doing various statistics on different extraction tables, these formulas are used to analyze table information.

```
{"pageTitle": "2017世界高楼排名，世界上最高的楼在哪个国家"，"title": "2017世界高楼排名前20名"，"url": "https://www.phb123.com/city/jz/18136.html"}
```

Figure 3. JSON format for Web table.
4. Data set analysis
Due to the enlightenment of WDC data representation, this article saves the extracted Chinese web table data in JSON format. For the web page table shown in Figure 1, the JSON part of the format is shown in Figure 3. The article uses the expanded table data framework to construct the Chinese Web table data set CWTs. The data volume in CWTs is 9.0GB, including 1,339,939 relational tables, of which 1,233,751 horizontal tables account for about 92.08%; and 106,188 vertical tables account for about 7.92%.

4.1. Distribution of main data source URLs
The analysis results show that the Chinese web table mainly comes from six top-level domain names: com, org, net, co, edu, and gov. Among them, the number of tables in the com domain name is the largest, and the distribution of table data in each domain name is shown in Figure 4.

4.2. Distribution of main data source URLs
In the analysis of all the table source URLs, Figure 5 shows the top ten websites that get the table rankings. It can be concluded that the first ranking is about the game ranking website; the second ranking is the Google search website; the third ranking is the Wikipedia site wikipedia. In addition, there are websites for ubuntu and debian application software, as well as chinatravel.com for China travel and viaf.org for art.

4.3. Distribution of data columns
Table 2 shows the statistical information of the column horizontal table (CHT), row horizontal table(RHT), column vertical table(CVT), and row vertical table(RVT). The analysis indicators are the min, max, average, and median of the number of rows and columns in the web table. Figure 6 shows the distribution of the number of columns in the web table. The most common level table is a table with 6 columns, with a total of 286,813 tables occupying 22.86% of the total table extracted.

Figure 4. Top-level domain table.
Figure 5. Ranking of source URLs.
Figure 6. Distribution of the number of columns (attributes) in each table.
Table 2. Statistics about columns and rows.

|               | Min | Max | Average | Median |
|---------------|-----|-----|---------|--------|
| CHT (attributes) | 1   | 996 | 4.69    | 4.0    |
| RHT (entities)  | 1   | 14111 | 38.32  | 12.0   |
| CVT (entities)  | 1   | 76  | 4.66    | 4.0    |
| RVT (attributes) | 1   | 4957 | 12.18  | 9.0    |

4.4. Table title analysis
Generally, the summary information of the table data reflected in the table title contains the data about what entity the table is about. Based on the table title extraction, this article obtains the top 6 keywords that appear frequently in the title. The results are shown in Table 3.

Table 3. Word frequency in the table captions.

| Word      | The number of occurrences |
|-----------|--------------------------|
| Google    | 186972                   |
| Patent    | 233680                   |
| Train     | 93493                    |
| Wikipedia | 78023                    |
| Resources | 38831                    |
| Method    | 37760                    |

4.5. Header information
The header refers to the attribute name of the current column in the table. In a horizontal table, the header information is contained in the first element of each column. In a vertical table, the first column contains all the header information. In Table 4, we show the first 10 title names of the horizontal and vertical tables in the CWTs data set.

Table 4. Vertical and horizontal header statistics.

| Vertical table | Horizontal table |
|----------------|------------------|
| Header element | The number of occurrences | Header element | The number of occurrences |
| Hardware architecture | 1911 | Date | 14107 |
| Display options | 620 | Preparation | 12839 |
| Plate | 433 | Time | 12061 |
| Years | 398 | Rank | 9419 |
| Price | 395 | Remarks | 8605 |
| Date | 370 | Character | 8024 |
| Content | 353 | Daily payment | 6336 |
| Actor | 347 | Years | 5880 |
| Book | 328 | Price | 5168 |
| Popularity | 327 | Name | 4999 |
5. Conclusion
This paper extracts page table data from the Web to construct a large-scale Web table data set CWTs. The paper analyzes the data in multiple dimensions such as data scale, title, header. The research in this paper provides a preliminary foundation for the application and development of open structured data on the web. We intend to continue our work in three directions: First, CWTs are only used as preliminary research results, which include redundant and noisy data. How to clean up the data set to construct high-quality data is a key task in the next step. Secondly, because Common Crawl regularly updates the data set, the data in this article comes from Common Crawl. Therefore, studying the update strategy of CWTs to reflect the dynamic evolution of the Web is of great significance to time-dependent data (such as stocks, weather); Finally, we plan to develop a platform based on the data set and provide the necessary application programming interfaces so that users can use the data set conveniently, which can give full play to the potential value of CWTs.

Acknowledgements
We thank the anonymous referees for their critical comments on a previous version of this paper, which encouraged us to significantly improve the paper. This work was supported by the Natural Science Foundation of China (No. 61762081) and the Natural Science Foundation of Northwest Normal University (No. NWNU-LKQN-17-16).

References
[1] Zhang, S., Balog, K. (2020) Web Table Extraction, Retrieval, and Augmentation: A Survey[J]. ACM Transactions on Intelligent Systems and Technology, 11(2): 1-35.
[2] Cafarella, M. J., Halevy, A. Y., Zhang, Y. et al. (2008) Uncovering the Relational Web[C]. 11th International Workshop on the Web and Databases, Vancouver, BC, Canada, June 13.
[3] Cafarella, M. J., Halevy, A. Y., Lee, H. et al. (2018) Ten years of webtables[J]. Proceedings of the VLDB Endowment, 11(12): 2140-2149.
[4] Ritze, D., Lehmburg, O., Bizer, C. (2015) Matching HTML Tables to DBpedia[C]. In Proceedings of the 5th International Conference on Web Intelligence, 10:1–10:6.
[5] Lehmburg, O., Ritze, D., Ristoski, P., et al. (2015) The Mannheim Search Join Engine[J]. Web Semantics: Science, Services and Agents on the World Wide Web, 159-166.
[6] Lehmburg, O., Bizer, C. (2016) Web Table Column Categorisation and Profiling[C]. Proceedings of the 19th International Workshop on Web and Databases, p. 4.
[7] Yakout, M. A., Ganjam, K., Chakrabarti, K., et al. (2012) InfoGather: Entity augmentation and attribute discovery by holistic matching with web tables[C]. SIGMOD. ACM, 97–108.
[8] Eberius, J., Thiele, M., BraunSchweig, K., et al. (2015) Top-k Entity Augmentation Using Consistent Set Covering[C]. In Proceedings of the 27th International Conference on Scientific and Statistical Database Mgmt, p. 8.
[9] Morcos, J., Abedjan, Z., Ilyas, I. F., et al. (2015) DataXFormer: An Interactive Data Transformation Tool[C]. SIGMOD. ACM, 883-888.
[10] http://commoncrawl.org/
[11] http://webdatacommons.org/webtables/index.html
[12] Bhagavatula, C. S., Noraset, T., Downey, D. (2015) TabEL: Entity Linking in Web Tables[C]. In Proceedings of the 14th International Conference on The Semantic Web, 425–441.
[13] http://boston.lti.cs.cmu.edu/eager/table-arxiv/