Recovering the Semantics of Tabular Web Data

Kurzfassung der Dissertation

zur Erlangung des akademischen Grades
Doktoringenieur (Dr.-Ing.)

vorgelegt an der
Technischen Universität Dresden
Fakultät Informatik

eingereicht von
Dipl.Medien-Inf. Katrin Braunschweig
geboren am 28. August 1984 in Suhl

Betreuender Hochschullehrer:
Prof. Dr.-Ing. Wolfgang Lehner

Dresden, im August 2015

1 INTRODUCTION

Data is increasingly becoming a valuable commodity, with personal and business decision making becoming more data-driven. As a result, tools that provide efficient data management, analysis or visualization functionality continue to gain in importance, as they enable users to obtain information from their data. The Web takes on a special role in this data and information oriented society. Described as an *abundance of accessible information* (Kurland, 2006), the Web provides a platform for people to share all of their data as well as to satisfy their diverse information needs. With free access to the data, predominantly through highly efficient Web search engines, it is an important resource for users to get the answers they need.

A special category of Web data, that has only come to the attention of researchers and application developers quite recently, is structured data stored in *tables*. Not natively supported by the text-centric approach of Web search engines, tabular data has received much less consideration compared to unstructured, textual data. Yet, the Web contains a considerable amount of tables, with the same topical diversity. For example, Yakout et al. extracted about 154 million tables containing mostly relational data from Web pages (Yakout et al., 2012). Moreover, the growing Open Data trend, which sees public organizations and government bodies publish their data on designated platforms to increase transparency and accountability, further raises the amount of high-quality factual and relational data that is freely available on the Web.

There is great potential in the content of these tables and a number of applications have emerged that benefit from this rich resource of structured data, including specialized fact search (Balakrishnan et al., 2015), automated knowledge base construction (Wang et al., 2012; Dong et al., 2014) and situational data analysis (Cafarella et al., 2009; Eberius et al., 2012).

Many traditional information extraction and retrieval techniques, including document-centric Web search, are not well suited for Web tables, as they generally do not consider the role of the table structure and layout in reflecting the semantics of the content. As a result, they cannot fully take advantage of this rich data source (Limaye et al., 2010). Therefore, in order to keep the content of Web tables from remaining underutilized, a designated Web table recovery and understanding process is required. The objective of this process is to recover and expose the entities and relations underlying the tables. For humans, understanding the information contained in Web tables is often an easy task, as the majority of these tables are intended for human consumption. However, to enable further utilization of the data and considering the scale of the Web, there is a great need for an automated processing of the tables. Here, we face a wide range of challenges. First of all, the tables are *embedded* in the Web, which means they must be located and extracted (Balakrishnan et al., 2015). To recover the semantics of a table, we often need to consider the interplay between the structured data in the table and the unstructured data of the context (Cafarella et al., 2011). Constraints of the table data are often only mentioned in the caption or surrounding text (Yin et al., 2011). Furthermore, the answers to queries can be distributed among multiple tables from different sources, raising a need for integration of Web tables (Halevy, 2004). However, no uniform schema or controlled vocabulary exists on the Web (Limaye et al., 2010). The tables are very heterogenous, since most of them have been developed individually, leading to different choices in the design of the schema and the selection of attribute labels. And in addition to that, there is no centralized quality control in place, resulting in significant variation in the quality of the data as well as its description.

In recent years, a lot of research effort has been put into addressing these challenges in order to develop a process for Web table understanding. Significant contributions have been made regarding the identification of relational tables on the Web (Cafarella et al., 2008), finding related tables (Das Sarma et al., 2012), or matching attributes in the tables to entries in a knowledge base, in order to understand the content (Wang et al., 2012). However, as pointed out by Yin et al. (2011), the precision and coverage of the data extracted from Web tables is often still quite limited. As Web table recovery and understanding is very complex, with many different challenges, many techniques developed so far make simplifying assumptions about the table layout or content to reduce the complexity. Thanks to these assumptions, many subtasks become manageable. However, they also imply a limited scope or a limited accuracy, if applied to tables that do not conform to these assumptions. In this thesis, our goal is to extend the Web table understanding process with techniques that enable some of these assumptions to be relaxed, thus improving the scope and accuracy of the extracted data. We consider extensions in various aspects of table understanding, including layout classification, context recovery and conceptualization.
This thesis is organized as follows: In Chapter 2, we start with a review of the role and characteristics of tables in general, before outlining the foundations of table understanding. In Chapter 3, we then focus on Web tables in particular. We provide an analysis of Web table characteristics and present key applications to highlight essential requirements that must be met by the table understanding process. Reviewing related work, we then derive a set of limitations and open issues of Web table understanding that we address in this thesis. Chapter 4 focuses on incorporating layout classification into the understanding process. In Chapter 5, we study the importance of contextual information for the understanding and utilization of Web tables. First, we analyze the relevance of various available context resources with respect to the table content. Second, we utilize the context to extract supplementary information that extends the description of attributes in the table. Chapter 6 addresses the conceptual model of Web tables, utilizing semantic normalization to expose the concepts described in the tables. Finally, in Chapter 7, we conclude this thesis with a summary of our findings and directions of future work.

2 FOUNDATIONS OF TABLES AND TABLE UNDERSTANDING

Tables form a key concepts in this thesis. However, there is no single, clear-cut definition, as the notion of a *table* is often associated with different characteristics when used in different application scenarios. Therefore, we analyze and contrast the role and characteristics of tables in the most prominent use cases, in order to derive an understanding of the term *table* that applies to the tables that we encounter on the Web. Based on this definition, we then specify the process of *table understanding* in the larger context of *table recovery*.

Tables in Documents and Databases
Tables are a versatile tool for the representation and communication of relational or similarly structured data (Embley et al., 2006). Through a two-dimensional layout, they provide a compact visualization of the data that particularly facilitates the *search* and *comparison* of values of interest (Zanibbi et al., 2004). As a result, tables are frequently found in printed as well as digital documents. Furthermore, they also represent an important concept in relational databases and spreadsheets. However,

the role of tables in documents differs significantly from the purpose of database tables, which is reflected by substantial differences in the characteristics of these tables.
In relational databases, a table acts as a *logical data structure* designed to hold large amounts of fine-grained data and facilitate efficient *automated* processing and querying. The structure of the table as well as data can be modified by a user or application. As database tables are mainly intended for algorithmic consumption, they generally feature a simple, uniform layout with clear reading paths and require each attribute to be named.
In contrast, document tables, which are primarily intended for human consumption, aim to present a detailed view or highlight certain characteristics, often in the form of summaries or aggregates, of the underlying data. A more complex layout provides the expressiveness required to effectively communicate this information. The layout and content are generally fixed and not meant to be modified. Furthermore, it is also common that some of the information required to understand the semantics is not mentioned in the table directly, but in the surrounding context in the document. Tables on the Web represent a special type of document tables.
To facilitate reuse, automated processing and analysis of the data embedded in document tables, a designated *table recovery* process is required to locate these tables, recognize their structure and semantics and convert them into a format that is better suited for algorithmic consumption.

Table Recovery
The complete table recovery process, depicted in Figure 1, can be divided into four consecutive steps (Göbel et al., 2012). The purpose of the first subprocess, *table detection*, is locating the table within the source document. This involves identifying the outer boundaries of the table. The second subprocess, *table structure recognition*, aims to reconstruct the cellular structure of the table as well as identify merged and nested elements. It requires the identification of cell boundaries as well as the recovery of cell alignments. The output of this step is a description of the *physical structure* of the table. *Functional analysis* then recovers the function or role of each cell in the table, separating label cells from data cells. In addition, each data cell is associated with its corresponding label cells. The result is a description of the table's *logical structure*. In the final subprocess, *semantic interpretation*, the content of the table is analyzed, in order to identify the entities, attributes and relationships presented in the table.
Aspects related to table recovery have been studied from many different angles in various communities. Research initially focused mainly on the detection of tables in the documents as well as the recovery of the tables' physical and logical structures. More recently, research has moved its focus to the semantic interpretation of tables, which is significantly harder to automate, as it requires a more comprehensive analysis of not only the table itself, but also its context. The attention directed at table recovery research has increased considerably in recent years, largely due to general advances in large scale data processing on the Web.

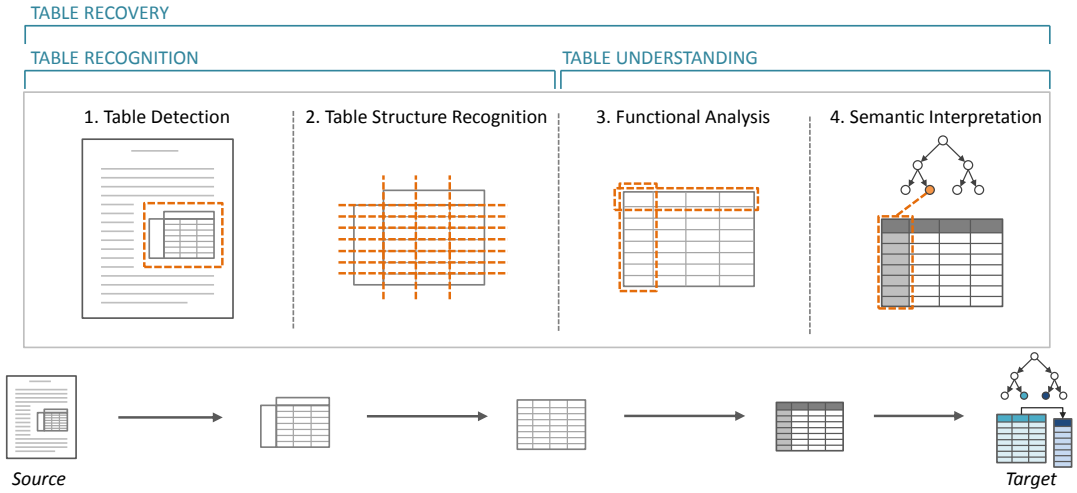


Figure 1: Overview of sub-processes involved in table recovery.

3 WEB TABLE UNDERSTANDING

After describing the general process involved in table understanding, we now focus on tables on the Web in particular. In a detailed study of Web tables and their potential applications, we derive necessary requirements for the table understanding process. Reviewing previous research in the field, we derive a set of limitations and open challenges that we intend to address in this thesis.

Characteristics of Web Tables

On the Web, we can identify two kinds of sources for tabular data: (1) a significant amount of tables is directly *embedded* in Web pages, as plain text tables or HTML tables, with the host page providing the context. (2) Alternatively, tables can reside in *external* documents that are linked to a Web page, which provides access to these documents, but also additional context information. Such external tables are commonly found on designated data portals, such as Open Data platforms, and exist in many different file formats.

The Web features a great variety of table layouts, with indexing schemes varying in semantics, orientation as well as complexity. As most of these tables are created independently by different people, they provide no formal schema that adheres to a controlled vocabulary. As a result, the table content is often ambiguous, requiring additional information to infer its meaning. Without quality control on the Web, the tables are also prone to contain incomplete and inconsistent data. While we can extract millions of distinct tables from the Web, covering a wide range of topics and domains, the size of individual tables is generally rather small compared to tables in enterprise database systems, for instance. Consequently, the sample of attribute values in a single table is often not representative of the attribute domain. These characteristics must be taken into account in order to develop a table understanding process that is suitable for Web tables.

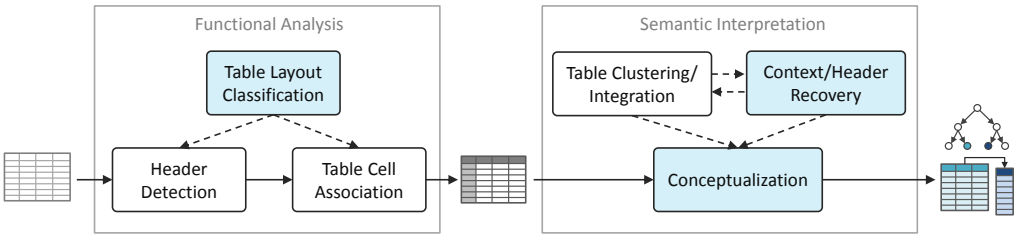


Figure 2: Process involved in Web table understanding.

Application Scenarios

A number of applications benefit from the huge amount of diverse data embedded in Web tables. By analyzing how the tables are utilized in these applications, we can identify important requirements for the table understanding process. In this thesis, we consider three key applications: (1) ontology learning, (2) question answering, and (3) situational data analysis.

The primary requirements include the identification of entities, attributes and relations in the tables. This involves inferring descriptive labels for each attribute, recovering constraints from the context, identifying functional dependencies between the attributes and, finally, matching the tables to entries in a reference knowledge base (a process often referred to as *conceptualization*). Additionally, it is often necessary to identify semantically related tables in order to retrieve all information that is required, due to the small size of individual tables.

Web Table Understanding

With these requirements and the characteristic features of Web tables in mind, we can specify a more detailed outline for the process that is necessary to *understand* these tables, as illustrated in Figure 2. In the literature, a number of important contributions have already been made in order to implement this process, addressing tasks such as table detection or header recovery. However, the precision and coverage of the data extracted from Web tables is often still quite limited. Due to the complexity of Web table understanding, many techniques developed so far make simplifying assumptions about the table layout or content to limit the amount of contributing factors. Thanks to these assumptions, many subtasks become manageable. However, the resulting algorithms and techniques often have a limited scope, leading to imprecise or inaccurate results when applied to tables that do not conform to these assumptions.

In this thesis, we focus on three assumptions in particular: (1) the assumption that all tables feature a single, uniform layout, (2) the assumption that all context is equally relevant for a table, and (3) the assumption that all attributes in a table describe the same semantic concept. Our objective is to extend the table understanding process with techniques that enable these assumptions to be relaxed, thus leading to a higher overall coverage and precision. In the remainder of the thesis, we introduce extension to three aspect of the table understanding process, which are marked in Figure 2.

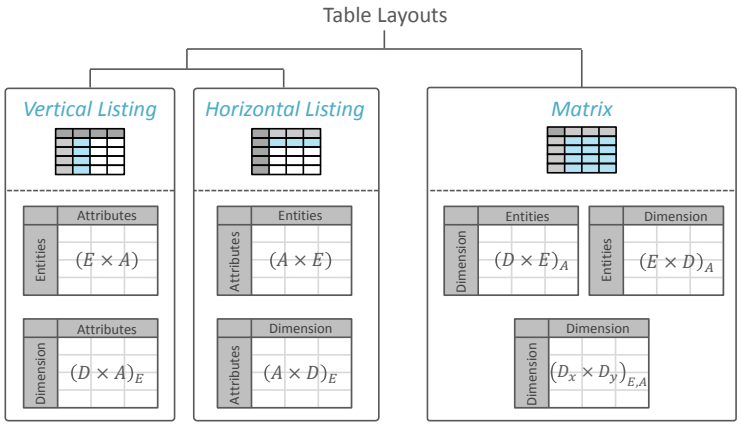


Figure 3: Categorization of Web tables, based on layout characteristics.

4 TABLE EXTRACTION AND CLASSIFICATION

One of the most prominent features of Web tables is the diversity of the table structures and layouts. Some of the information presented by a table is implicitly conveyed through its layout. Thus, correctly recognizing the layout is important in order to understand the role of each table cell and locate attribute and entity names as well as additional categorical dimensions.
As illustrated in Figure 3, we distinguish three main classes of table layouts, based on the alignment of attribute values: (1) *vertical listings*, (2) *horizontal listings*, and (3) *matrix tables*. Each of these classes can be subdivided further, based on additional criteria. Defining a single table model to recognize and process tables with such varied layouts is very complex and challenging. Therefore, in order to reduce the structural diversity and ambiguity, various approaches processing Web tables only focus on tables with a layout similar to relational database tables (Cafarella et al., 2008).
In addition to identifying different table layouts, we face another challenge, especially in the context of tables embedded in Web pages. While HTML documents provide support for tabular structures with designated <TABLE> tags, the same structures are also frequently used for other purposes, such as the spatial arrangement of page content. Instead of relational data, these layout tables contain other document content, such as images or menu items. The vast majority of HTML tables in Web pages are layout tables instead of genuine data tables (Cafarella et al., 2008b).

Web Table Classification
Both tasks can be regarded as table classification problems, although with very different objectives. In this thesis, we analyze and compare two alternative approaches to incorporate layout classification into the table understanding process: (1) combined with table detection as a single classification task, and (2) as a separate classification task performed after table detection.
We consolidate and extend a wide range of features proposed in the literature, including global features describing the entire table as well as local features that only describe selected subsets. We apply feature selection to identify the most relevant features for each classification problem. A thorough evaluation of both approaches is performed, using a large set of tables extracted from the Web and

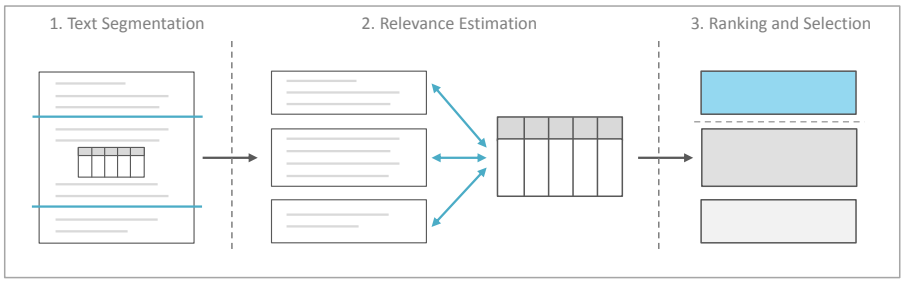


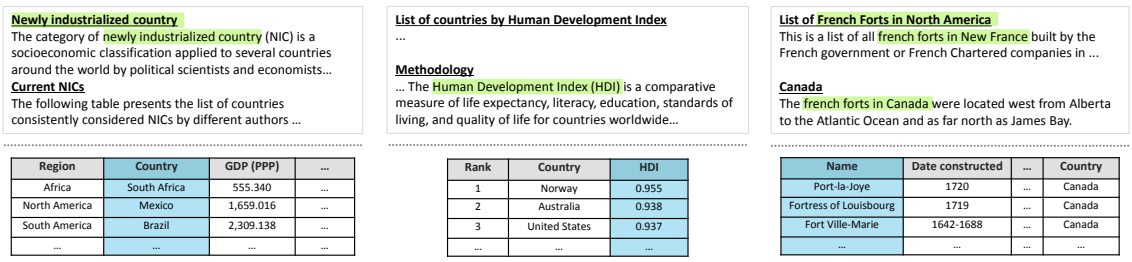
Figure 4: Overview of the paragraph selection task.

various classification algorithms. The results indicate that, in general, both approaches are equally suited to classify Web tables. However, a process that separates the tasks of table detection and layout classification offers more flexibility to tailor the feature selection and training data to each task and, thus, provides more room for further optimization.

5 RECOVERING WEB TABLE CONTEXT

The semantics of Web tables is inherently implicit, and additional context is required to recover the meaning of the table content. Web pages offer various resources for contextual information, including captions, headlines and surrounding text. Context referring to a table can provide a summary of the content or conclusions drawn from it. It also frequently offers a more detailed description of various table entries to clarify terms or indicate restrictions on attributes (Hurst, 2000). Additionally, *hidden* attributes that have been factored out of the table are generally placed in the context. The importance of considering such context in the table recovery process is well established. A wide range of tasks, including table search (Sarawagi et al., 2014) and finding related tables (Yakout et al., 2012), benefits from this supplementary information.
However, not all information mentioned in the context is actually relevant to the table. The verbosity of the context, especially when considering large texts, often introduces noise that leads to incorrect interpretations (Pimplikar et al., 2012). Consequently, evaluating the relevance of potential context segments as well as establishing an explicit link to the table content is essential in order to reduce noise and prevent misinterpretations.

Estimating Context Relevance
The text on a Web page can cover many different aspects of the main topic, while an individual table is often only related to a specific aspect. To identify the text segments that are relevant for the understanding of the table content, we propose a paragraph selection algorithm, which is illustrated in Figure 4. The algorithm encompasses the following subtasks: (1) First, the text is decomposed into topically coherent paragraphs using linear text segmentation. (2) A similarity measure is utilized to match each paragraph to the table in question in order to estimate its relevance. (3) Finally, all paragraphs are ranked according to their relevance and an appropriate threshold is determined to filter out irrelevant, noisy paragraphs. Within this process, we treat a table as a bag of words, assuming independence between individual terms. To estimate the relevance of context segments, we consider



(a) Qualifying Noun Phrase (b) Acronyms (c) Indirect Match

Figure 5: Example tables with corresponding context sections. Marked phrases (green) provide a detailed description of the selected attributes (blue) in each table.

various word-based as well as topic-based similarity measures.
An evaluation on real-world tables extracted from the Web reveals that word-based measures outperform topic-based measures, largely due to the fact that small tables as well as small context segments do not provide sufficient information to infer accurate topic models. However, using word-based measures, we achieve high-quality results that reduce the noise in Web table context effectively.

Attribute-specific Context Information
After identifying relevant paragraphs, we can now utilize the context to extract supplementary information about individual attributes in the tables. Labels in Web tables are often of low quality or too general to provide a suitable description of the attributes. Additionally, many tables provide no attribute labels at all. However, the contextual description of the table content often contains phrases that provide a sufficiently detailed specification of the attributes and can be used as alternative labels, as illustrated in the examples in Figure 5.
Given an attribute A in a table T, with a set of instances $\mathcal{I}_A$ and an original attribute label L_A (if available), our objective is to extract a set of alternative labels $\{L_{C_1}, \ldots, L_{C_n}\}$ from the context of T. Each label L_{C_i} should have the following properties: (1) $L_{C_i} \neq L_A$, and (2) L_{C_i} gives a valid description of attribute A. We refer to these phrases as *attribute-specific context annotations*. In our approach, we consider two types of context annotation, *directly* and *indirectly related* phrases. An extracted label L_{C_i} is *directly related* to original label L_A, if $L_A \subset L_{C_i}$, meaning that the original label is contained in the extracted phrase. On the other hand, the label L_{C_i} is *indirectly related*, if we can derive an intermediate label L_I from attribute A so that $L_I \subset L_{C_i}$. For each annotation type, we consider several extraction and inference techniques. As directly related annotations, we extract qualifying noun phrases as well as expansions of acronyms. To identify indirectly related phrases, we consider two alternative sources in order to infer intermediate labels: a general-purpose taxonomy as well as a linguistic database.
As part of our experimental evaluation, we consider attribute and table search to evaluate the effectiveness of the extracted context annotations. We can show that both search tasks benefit from additional context information that can be linked directly to specific attributes in the table. The supplementary information increases both, precision and recall, as it provides a more detailed and informative description of the table content.

6 SEMANTIC NORMALIZATION

To fully utilize the data stored in tables on the Web, we must be able to *understand* their content. However, without a conceptual model or formal description, we can only attempt to infer the semantics directly from the table, which is a very challenging task. In order to limit the complexity of this inference, the majority of previous research has focused exclusively on tables that only describe a single semantic concept. We refer to these tables as *single-concept tables*. In the context of Web table understanding, single-concept tables are frequently assumed for the extraction of binary relations from tables (Yakout et al., 2012; Zhang et al., 2013), matching tables to concepts in an external knowledge base (Venetis et al., 2011; Wang et al., 2012), as well as finding related tables (Das Sarma et al., 2012). However, the Web also provides access to a substantial amount of larger, more complex tables that describe properties of multiple concepts as well as the relationships between them (i.e. *multi-concept tables*). Figure 6 shows an example of such a table, which mentions the semantic concepts *City*, *Country* and *Mayor*. Treating such a table as a single-concept table, i.e. assuming that all attributes describe the *same* semantic concept, leads to an incorrect interpretation and use of the data. To relax the single-concept assumption, we propose a *semantic normalization* approach, which identifies concept boundaries and enables multi-concept tables to be decomposed into multiple single-concept tables, which can then be processed as before.

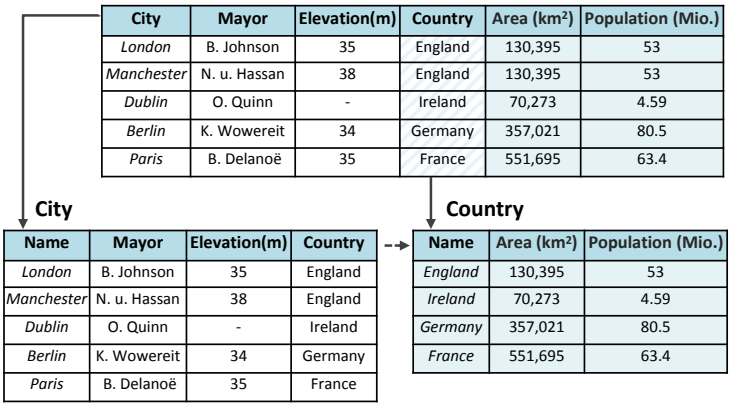


Figure 6: Example of a table describing three separate semantic concepts: the concept *City* with additional property *Elevation*, the concept *Mayor*, and the concept *Country* with properties *Area* and *Population*.

Semantic Normalization

The objective of *semantic normalization* is the detection and separation of individual semantic concepts C_i contained in a relation R. Multi-concept relations are split into multiple single-concept relations $R_{C_i} \subset R$, one for each concept. In this thesis, we focus on *simple concepts*, i.e. concepts that, in relational form, are described by a single key attribute K and a number of additional attributes A_i, which contain property values (Zhang et al., 2013; Yakout et al., 2012). Additionally, we focus on *binary relations* between attributes, distinguishing between *entity-attribute binary (EAB)* relations, which

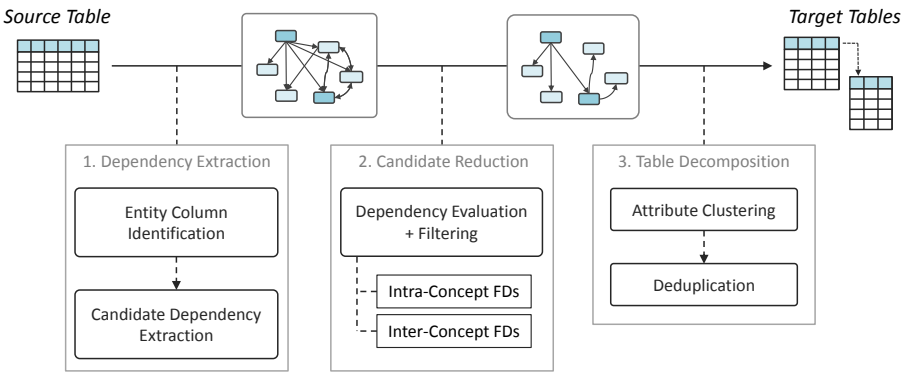


Figure 7: Process overview: Extraction, evaluation and filtering of functional dependencies inferred from table data.

hold between the key and an additional attribute of the same concept, and *entity-entity binary (EEB)* relations that hold between two concepts represented by their keys. The majority of these relations are reflected through *functional dependencies (FDs)* in the data. Relying on expert support and domain knowledge to derive functional dependencies is too extensive and unrealistic considering the scale of a Web table corpus. Instead, we need to infer these dependencies directly from the data.
During the inference of FDs, we face many challenges due to the quality of the data and the small size of Web tables. As a result, a large number of spurious dependencies are inferred from the data, which leads to incorrect normalization results. To address these challenges, we propose a semantic normalization approach, depicted in Figure 7, that systematically extracts, evaluates and filters functional dependencies. Our approach effectively eliminates spurious or transitive dependencies to increase the accuracy of the table decomposition.

Estimating Attribute Relatedness
In order to distinguish between meaningful and spurious functional dependencies inferred from the data, we rely on a set of structural and semantic measures that indicate valid relationships between attributes. The structural measures *Column Distance*, *Value Correlation* and *Null Value Distribution* address the column arrangement as well as the distribution of values in the table. The semantic measures *Common Concept Label* and *Attribute Co-Occurrence* take into account the attribute labels and their semantic relationships.
To score each functional dependency $X \rightarrow A$, we combine all individual measures using a weighted sum. Inter- and intra-concept dependencies feature very different characteristics. As a result, not all indicators are equally useful for both types. We address these differences by using a separate set of weights for each type, which are inferred from pre-labeled training data using linear regression.

Evaluation
We conduct a comprehensive evaluation of the various processing steps involved in our semantic normalization approach, using real-world tables extracted from the Web. The results show that our algorithm successfully handles the challenges introduced by the characteristic features of Web tables.

The systematic approach substantially reduces the number of functional dependencies that need to be considered. Furthermore, the measures we proposed to estimate attribute relatedness effectively identify meaningful functional dependencies for a wide range of tables, which we can then use to decompose the tables.

7 CONCLUSION

The Web offers an abundance of accessible information, including a substantial amount of *tables*, which form a rich resource for factual and relational data. Processing these tables and inferring their meaning algorithmically is a challenging task. The complexity of the recovery process originates from the heterogeneity of table layouts, the semantic ambiguity of the content description, the complex dependencies between table content, structure and context, as well as the general lack of quality control on the Web. In the literature, many important contributions have been made to implement and extend the recovery process for Web tables. However, often various *simplifying assumptions* are made about the characteristics of the tables, in order to reduce the problem space and make the complex sub-tasks involved in Web table recovery and understanding manageable.

In this thesis, we address some of these limitations by proposing various techniques that extend the Web table recovery and, especially, understanding process in order to relax these simplifying assumptions. Our objective is to adjust different aspects of the process to better match the characteristic features of tables on the Web and improve the overall extraction quality.

We relax the assumption of a uniform table layout, by studying different approaches to effectively incorporate table layout classification into the recovery process. Furthermore, we analyze the role of contextual information in Web table understanding. We propose a selection algorithm to identify relevant context segment, which we then utilize to extract attribute-specific information. The extracted context annotations provide a richer description of the table content and improve the accuracy of table relevance decisions, for instance, in search applications. Finally, we address the single-concept assumption by introducing a semantic normalization approach that decomposes multi-concept tables into multiple single-concept tables. This algorithm serves as an essential preprocessing step for various table matching and analysis techniques proposed in the literature.

As part of this thesis, we provide an experimental evaluation of all techniques and algorithms we propose. All experiments are conducted using real-world tables extracted from the Web. The results indicate that our techniques effectively handle the characteristic features of Web tables by achieving high-quality results for a wide range of tables. Therefore, they provide useful extensions to the Web table understanding process in order to improve the scope and quality of the recovered data.

REFERENCES

Balakrishnan, Sreeram et al. (2015). "Applying WebTables in Practice". In: *Seventh Biennial Conference on Innovative Data Systems Research*.

Cafarella, Michael J. and Alon Y. Halevy (2011). "Web Data Management". In: *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, pp. 1199–1200.

Cafarella, Michael J., Alon Y. Halevy, and Nodira Khoussainova (2009). "Data Integration for the Relational Web". In: *Proceedings of the VLDB Endowment* 2 (1), pp. 1090–1101.

Cafarella, Michael J., Alon Y. Halevy, Yang Zhang, Daisy Zhe Wang, and Eugene Wu (2008). "Uncovering the Relational Web". In: *Proceedings of the 11th International Workshop on the Web and Databases: In Conjunction with the 2008 ACM SIGMOD International Conference on Management of Data*.

Das Sarma, Anish, Lujun Fang, Nitin Gupta, Alon Y. Halevy, Hongrae Lee, Fei Wu, Reynold Xin, and Cong Yu (2012). "Finding Related Tables". In: *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*. ACM, pp. 817–828.

Dong, Xin, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang (2014). "Knowledge Vault: A Web-scale Approach to Probabilistic Knowledge Fusion". In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pp. 601–610.

Eberius, Julian, Maik Thiele, Katrin Braunschweig, and Wolfgang Lehner (2012). "DrillBeyond: Enabling Business Analysts to Explore the Web of Open Data". In: *Proceedings of the VLDB Endowment* 5.12, pp. 1978–1981.

Embley, David W., Matthew Hurst, Daniel Lopresti, and George Nagy (2006). "Table-processing Paradigms: a Research Survey". In: *International Journal of Document Analysis and Recognition* 8.2-3, pp. 66–86.

Göbel, Max, Tamir Hassan, Ermelinda Oro, and Giorgio Orsi (2012). "A Methodology for Evaluating Algorithms for Table Understanding in PDF Documents". In: *Proceedings of the 2012 ACM Symposium on Document Engineering*. ACM, pp. 45–48.

Halevy, Alon Y. (2004). "Structures, Semantics and Statistics". In: *Proceedings of the 13th International Conference on Very Large Data Bases*. Vol. 30. VLDB Endowment, pp. 4–6.

Hurst, Matthew (2000). "The Interpretation of Tables in Texts". PhD thesis. University of Edinburgh.

Kurland, Oren (2006). "Inter-document Similarities, Language Models, and Ad Hoc Information Retrieval". PhD thesis. Cornell University.

Limaye, Girija, Sunita Sarawagi, and Soumen Chakrabarti (2010). "Annotating and Searching Web Tables using Entities, Types and Relationships". In: *Proceedings of the VLDB Endowment* 3 (1-2), pp. 1338–1347.

Pimplikar, Rakesh and Sunita Sarawagi (2012). "Answering Table Queries on the Web using Column Keywords". In: *Proceedings of the VLDB Endowment* 5.10, pp. 908–919.

Sarawagi, Sunita and Soumen Chakrabarti (2014). "Open-domain Quantity Queries on Web Tables: Annotation, Response, and Consensus Models". In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 711–720.

Venetis, Petros, Alon Y. Halevy, Jayant Madhavan, Marius Paşca, Warren Shen, Fei Wu, Gengxin Miao, and Chung Wu (2011). "Recovering Semantics of Tables on the Web". In: *Proceedings of the VLDB Endowment* 4.9, pp. 528–538.

Wang, Jingjing, Haixun Wang, Zhongyuan Wang, and Kenny Q. Zhu (2012). "Understanding Tables on the Web". In: *Proceedings of the 31st International Conference on Conceptual Modeling*, pp. 141–155.

Yakout, Mohamed, Kris Ganjam, Kaushik Chakrabarti, and Surajit Chaudhuri (2012). "InfoGather: Entity Augmentation and Attribute Discovery by Holistic Matching with Web Tables". In: *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*. ACM, pp. 97–108.

Yin, Xiaoxin, Wenzhao Tan, and Chao Liu (2011). "FACTO: A Fact Lookup Engine Based on Web Tables". In: *Proceedings of the 20th International Conference on World Wide Web*, pp. 507–516.

Zanibbi, Richard, Dorothea Blostein, and James R. Cordy (2004). "A Survey of Table Recognition: Models, Observations, Transformations, and Inferences". In: *International Journal on Document Analysis and Recognition* 7.1, pp. 1–16.

Zhang, Meihui and Kaushik Chakrabarti (2013). "InfoGather+: Semantic Matching and Annotation of Numeric and Time-varying Attributes in Web Tables". In: *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*. ACM, pp. 145–156.