# Extracting Tables Containing Information on Mutations

**Synopsis**

Handed in by: Sinje Kristin Kühme
Handed in on: 30 January 2018
Supervisor: Prof. Dr. Ulf Leser

## Introduction

The ever-growing amount of biomedical research articles offers a wealth of information on many topics of interest to researchers in the field; however, the more information is available, the more difficult and time-consuming it is to annotate and summarize articles manually. The collective knowledge contained in corpora can, however, be processed by using the methods of computer science. It is the field of information extraction that offers the tools and methods concerned with obtaining specific information from a given knowledge representation (Henrich, 2008).

Academic articles in particular tend to use tables in order to summarize information and present it to the audience in a concise manner. When Kim et al. (2012) investigated the use of tables in research journals, looking at 2500 randomly selected tables from academic journals, they concluded that approximately 75% of tables can be found in the results section of an article. Another 20% can be found in the section on methods used in the research project. One can conclude that a collection of all tables concerning a particular query is a suitable way to summarize a considerable part of the current state of knowledge about that topic (insofar as that state is contained in the corpus).

This work aims to implement and evaluate a system that extracts all those, and only those, tables that contain information on mutations from the corpus provided by BioMed Central[1]. These are particularly important for any research regarding the interaction of mutations and the diseases they may cause (Wei et al., 2013). In the following synopsis, as well as the thesis, the term *relevant* will be used to denote 'containing information on mutations'.

The BioMed Central corpus is available in Extensible Markup Language (XML). As this markup language allows users to define their own tags and the document's structure, it can be customized and used to store any data and can also be extended if necessary. Moreover, as XML files are saved in a plain-text format, they can be stored and shared more easily and are independent of software or hardware[2]. While this necessitates the development of individual parsers, it also allows relevant aspects of a corpus to be tagged and therefore extracted more easily.

To facilitate the extraction of tables that contain information on mutations, various existing procedures of table extraction and the identification of mutation mentions will be reviewed. This will be used to develop and implement four scoring algorithms that can be used individually or in combination. Finally, the system will be evaluated and the effectiveness of the various methods will be compared in order to conclude which methods come closest to the goal of extracting only relevant tables.

## Related Work

Both the identification and extraction of tables from documents as well as the identification of mutation mentions have been considered in previous research projects.

Several corpora have been used for research on tables. Tables became more important as an object of study when they started to occur more frequently in the World Wide Web (WWW), as traditional search engines and their methods of ranking results are not ideally suited to dealing with tabular data (Cafarella et al., 2008b). Even when disregarding all tables that are only used for formatting, there is an enormous amount of information contained in the WWW's tables. Due to their structure, this information is presented in a way that resembles a relational database with a schema. In this case, the table's columns represent a schema's attributes and the rows (excepting any possible header row) correspond to the individual instances (Cafarella et al., 2008a). Almost everyone can contribute to the WWW, which means that the data available is of inconsistent quality; as this work is based on a corpus of academic articles, however, the assumption that any

---

[1] https://old.biomedcentral.com/about/datamining, retrieved on December 18, 2017

[2] https://www.w3schools.com/xml/xml_whatis.asp, retrieved on January 15, 2018

table contains a schema that is maintained throughout the entire table (even if it is implicit) is even more likely to hold true. A seminal work on tables extracted from the web is the WebTables project by Cafarella et al. (2008a, 2008b). On the basis of a Google crawl, a corpus of about 150 million tables was composed. The relations identified in the corpus were then ranked in relation to a given query. In addition to naive ranking algorithms, two algorithms were proposed that take into account structural elements, such as the number of rows or columns, and content-based elements, such as the table's captions or the content of the first column, which is often used as a "semantic key" (Cafarella et al., 2008b). Additionally, the coherence of a table's schema was considered.

Lehmberg and Bizer's work (2017) also uses tables extracted from the web and tries to match them to knowledge bases. In order to improve the performance of algorithms commonly used for this task such as T2K Match or COMA, they aim to combine tables of suitable content to produce larger tables. They propose several matchers that are based e. g. on a table's column headers or the data contained in a column. A combined matcher offers the option to resolve ambiguities (such as the label 'name' that can refer to a work's title or an artist).

Enriching a table's data by extracting information from the surrounding paragraphs is the focus of Braunschweig et al. (2015). This is primarily done by extending column headers by searching the table's context for suitable expansions as well as matching the headers to instances of a knowledge base, which yields additional potential labels for the columns.

Another way to expand tables is introduced by Limaye et al. (2010), who propose a machine learning algorithm in order to annotate a table's columns with a type, adding a relation between two types (i. e. columns) and an entity label to individual cells. This allows for easier recognition of one and the same entity in different tables and therefore easier synthesis of information on entities that occur in more than one table.

Instead of HTML tables, Liu et al. (2007) use a corpus of PDF documents. Their approach to table extraction uses visual and structural elements (such as font size) to identify tables. They also extract metadata and develop a TF-IDF score specifically for tables: the TTF-ITTF score allows ranking tables in relation to a specific query.

The BioMed Central corpus has previously been used in other works. Isberner (2016) examined the possibility of using similarity search on the tables contained in the corpus. After extensive preprocessing, the tables receive scores based on their content and structure and are then ranked according to the final score that measures their similarity to a given query table.

Glushanok (2015) explores the possibility of using clustering algorithms to structure the tables contained in the corpus. After extracting the tables, the k-means algorithm is used to divide them into several clusters.

Identifying mutation mentions in texts is another field of research that has garnered attention in recent years. As there is no standardized nomenclature for mutations, this task is not easy to accomplish. An approach that employs regular expressions in order to recognize mutation names is used in the tool *MutationFinder* developed in Caporaso et al. (2007). An extended version of *MutationFinder* is a part of SETH, which further offers a grammar-based recognition of mutation names, additional regular expressions and also detects dbSNP identifiers (Thomas et al., 2015). Another possible approach is based on machine learning: the tmVar system by Wei et al. (2013) uses conditional random fields to train a system that recognizes mutation mentions.

**Methods**

*Table Similarity Search* by Isberner (2016) has dealt with the extraction of tables form the BioMed Central corpus in an in-depth manner, and the parser and text processor of that work will be adapted in order to be suitable for the purpose of searching for mutation mentions. As this system will use SETH in order to detect mentions of a specific mutation name, it is important that any mutation name remains in its original format and is not changed or deleted during preprocessing;

for example, special characters that occur frequently in mutation names, such as '>', must not be used to separate two tokens and must not be deleted.

In this work, several scoring algorithms will be developed and all extracted tables will then receive a score that should reflect their relevance. There are two events that can signify relevance: Firstly, one or more mutation name(s) may be recognized by SETH. Intuitively, one expects this in a table that compares several mutations on certain attributes, as one column would likely identify the mutations by their name. Another event that can occur and will be used as a signifier for the table's relevance is the occurrence of the term "mutation" or a related concept, e. g. in a column header of the table. As related concepts, this work will use the "associative relationships"[3] that can be found in the Medical Subject Headings' entry for the term "mutation", which are as follows: Antimutagenic Agents; DNA Damage; Mutagenesis; Mutagens; Polymorphism, Restriction Fragment Length; Suppression, Genetic[4].

In the literature related to table extraction and analysis, several criteria stand out to determine a table's content. The table's caption in the `caption` tag, the column headers, the data in individual columns as well as the paragraphs containing further information on the table are often used for this purpose. This work will implement scorers based on these four aspects of any table, which can be used individually or in combination. In the combined configuration, the weighting of the various scores will depend on the evaluation of the individual scorers' performance. Should one scorer not apply (e. g. because no paragraph describing the table can be found), it will not be taken into consideration.

Table caption

In academic articles, table captions follow a certain pattern (as noted by Liu et al., (2007)): in general, they begin with a title such as "Table 3" or "Figure 6", often followed by a closer description in natural language. In the BioMed Central corpus used in this work, the former can be found in the `title` tag, the latter description in the `caption` tag. The scorer will take into account any mention of a specific mutation name as well as the term mutation or any of its synonyms in the `caption` tag. The more mutations names are detected or the more often the term mutation or one of its synonyms is found, the higher the score will be; but even a single event should result in a decisively positive score.

Column headers

As described above, a table that compares several mutations regarding their attributes would treat these individual mutations as instances. In order to identify them, a column header such as "mutation name" or others that include the term "mutation" or a synonym may be found. Any specific mutation name that is detected will also influence the score positively. This scorer will base its assessment of the table's relevance on the percentage of column headers that indicate relevance.

Column data

If several mutations constitute a table's instances, one might also expect to find a column in which the majority of cells contain a mutation name. The relevance score will be assigned based on how many of a column's cells contain relevant data and how many columns in a table can be considered relevant. Moreover, the scorer will try to detect whether one or more of a column's cells contain the term mutation or one of its synonyms, and determine a final score for each table based on the results.

---

[3] https://www.nlm.nih.gov/mesh/meshrels.html, retrieved on December 17, 2017

[4] https://meshb.nlm.nih.gov/record/ui?ui=D009154, retrieved on December 17, 2017

Descriptive paragraphs

Based on the table's title as found in the `title` tag, the system will try to identify a paragraph in the corresponding article which contains additional information on the table. If one or more paragraphs can be found, the scorer will try to identify both mutation names and the term mutation and its synonyms in the text. Its assessment will be based on the number of occurrences of both in relation to the length of the paragraph.

**Evaluation**

The project will result in several documents containing the ranked results of each individual scorer and of the combined scorer. For each of these documents, the results will be divided into groups depending on their scores: top 20, ranks 21 to 50, 51 to 100, 101 to 500 and so on (the final groups will depend on the number of extracted tables). The top 20 results, along with 10 randomly selected tables from each of the other groups, will be manually classified as relevant or not relevant. For each of these groups, the precision of the results will be calculated. The *precision* P indicates how many of the documents in a given set are indeed relevant to the query at hand and is calculated as follows:

$$P = TP / (TP + FP) \quad (1)$$

where TP stands for the true positives (the number of truly relevant tables in the set of results) and FP stands for the false positives (the number of tables that the system considers relevant but that do not contain any information on mutations) (Henrich, 2008).

The system aims to achieve a high precision in the first groups, but should also have a low precision in the groups that contain results ranked as less relevant. Ideally, the average precision mapped to the tables' groups should be steadily decreasing, as a high precision in the lowest groups would indicate that there are many false negatives, i. e. that many relevant tables have been falsely considered irrelevant by the system.

Another metric that is commonly used to evaluate IR systems is *recall*, which specifies how many of the relevant tables have been assigned a score that does indicate relevance. In order to calculate this, however, the absolute number of relevant tables in the corpus needs to be known. Given the size of the BioMed Central corpus and the scope of the work at hand, it will not be possible to calculate the system's recall.

However, another metric that will be taken into account is the number of documents that receive a relevance ranking above a certain threshold and can therefore be considered relevant. This will be determined for each configuration and allow a comparison of their strictness relative to each other.

**References:**


BioMed Central (2017). Using BioMed Central's open access full-text corpus for text mining research. https://old.biomedcentral.com/about/datamining

Braunschweig, K., Thiele, M., Eberius, J., & Lehner, W. (2015). Column-specific context extraction for web tables. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing* (SAC '15), 1072-1077. DOI: http://dx.doi.org/10.1145/2695664.2695794

Cafarella, M., Halevy, A., Wang, Z. D., Wu, E., & Zhang, Y. (2008a). Uncovering the Relational Web. In *Proceedings of the 11th International Workshop on Web and Databases (WebDB 2008)*.

Cafarella, M., Halevy, A., Wang, Z. D., Wu, E., & Zhang, Y. (2008b). WebTables: exploring the power of tables on the web. In *Proceedings of the VLDB Endowment,* 1(1), 538-549. DOI: http://dx.doi.org/10.14778/1453856.1453916

Caporaso, J. G., Baumgartner Jr, W. A., Randolph, D. A., Cohen, K. B., Hunter, L. & Valencia, A. (2007). MutationFinder: a high-performance system for extracting point mutation mentions from text. *Bioinformatics* 23(14), 1862-1865. DOI: 10.1093/bioinformatics/btm235

Glushanok, I. (2015). *Semantische Analyse von Tabellen in Volltexten* (Bachelorarbeit). Henrich, A. (2008). *Information Retrieval. Grundlagen, Modelle und Anwendungen.* Bamberg: Otto-Friedrich-Universität Bamberg, Lehrstuhl für Medieninformatik.

Isberner, A. (2016). *Ähnlichkeitssuche auf Tabellen* (Diplomarbeit).

Kim, S., Han, K., Kim, S. Y., & Liu, Y. 2012. Scientific table type classification in digital library. In *Proceedings of the 2012 ACM symposium on Document engineering* (DocEng '12), 133-136. DOI: http://dx.doi.org/10.1145/2361354.2361384

Lehmberg, O. & Bizer, C. (2017). Stitching Web Tables for Improving Matching Quality. In *Proceedings of the VLDB Endowment*, 10(11), 1502-1513.

Limaye, G., Sarawagi, S., & Chakrabarti, S. (2010). Annotating and Searching Web Tables Using Entities, Types and Relationships. In *Proceedings from the VLDB Endowment*, 3(1), 1338-1347.

Liu, Y., Bai, K., Mitra, P., & Giles, C. L. (2007). TableSeer: Automatic Table Metadata Extraction and Searching in Digital Libraries. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries* (JCDL '07), 91-100. DOI: http://dx.doi.org/10.1145/1255175.1255193

Thomas, P., Rocktäschel, T., Hakenberg, J., Lichtblau, Y., & Leser, U. (2015). SETH detects and normalizes genetic variants in text. *Bioinformatics* 32(18), 2883-2885. DOI: 10.1093/bioinformatics/btw234

U.S. National Library of Medicine. (2001, 20 November). *Chapter 11, Relationships in Medical Subject Headings.* Retrieved from https://www.nlm.nih.gov/mesh/meshrels.html

U.S. National Library of Medicine. (2008, 8 July). *Mutation, MeSH Descriptor Data 2018.* Retrieved from https://meshb.nlm.nih.gov/record/ui?ui=D009154

Wei, C., Harris, B. R., Kao, H., & Lu, Z. (2013). tmVar: a text mining approach for extracting sequence variants in biomedical literature. Bioinformatics 29(11), 1433-1439. DOI: 10.1093/bioinformatics/btt156

w3schools.com (2018). *Introduction to XML.* Retrieved from https://www.w3schools.com/xml/xml_whatis.asp