ExQuisiTe: Explaining Quantities in Text

Yusra Ibrahim
Max Planck Institute for Informatics
Saarbrücken, Germany
yibrahim@mpi-inf.mpg.de

Gerhard Weikum
Max Planck Institute for Informatics
Saarbrücken, Germany
weikum@mpi-inf.mpg.de

ABSTRACT

Web pages and other documents often contain tables to provide numerical details in a structured manner. Typically, the text explains and highlights important quantities, often using approximate numbers and aggregates such as totals, averages or ratios. For a human reader, it is crucial to navigate between text and tables to understand the key information in its context, drill down into tables when details are needed, and obtain explanations on specific figures from the accompanying text.

In this demo, we present ExQuisiTe: a system to align quantity mentions in the text with related quantity mentions in tables, and enable extractive summarization that considers table contents. ExQuisiTe links quantity mentions in the text with quantities in tables, to support user-friendly exploration. ExQuisiTe handles exact single-cell references as well as rounded or truncated numbers and aggregations such as row or column totals.

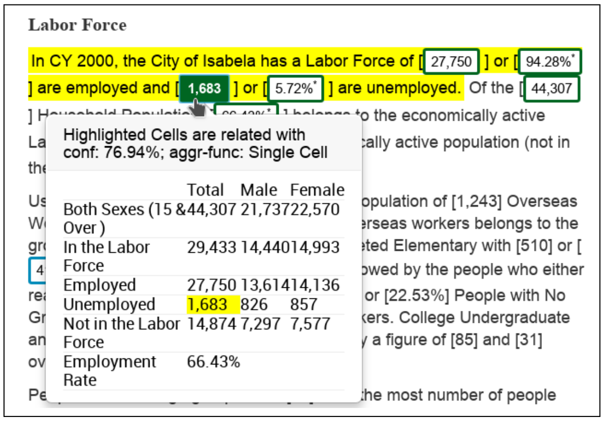
CCS CONCEPTS

• **Applied computing** → **Document management and text processing**; • **Information systems** → *World Wide Web*.

KEYWORDS

Information Extraction, Quantities in Text, Web Tables

ACM Reference Format:
Yusra Ibrahim and Gerhard Weikum. 2019. ExQuisiTe: Explaining Quantities in Text. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3308558.3314134



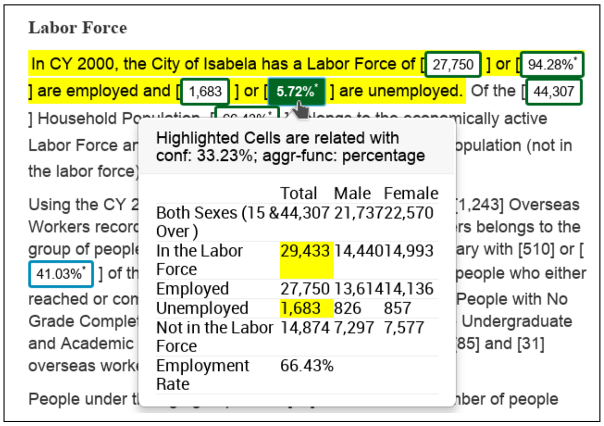
Labor Force

In CY 2000, the City of Isabela has a Labor Force of [27,750] or [94.28%*] are employed and [1,683] or [5.72%*] are unemployed. Of the [44,307
] H ... cally active population (not in
the
Us ... opulation of [1,243] Overseas
W ... erseas workers belongs to the
gr ... eted Elementary with [510] or [
4 ... owed by the people who either
re ... or [22.53%] People with No
Gr ... kers. College Undergraduate
an ... y a figure of [85] and [31]
ov
Pe ... the most number of people

Highlighted Cells are related with
conf: 76.94%; aggr-func: Single Cell

	Total	Male	Female
Both Sexes (15 & Over)	44,307	21,737	22,570
In the Labor Force	29,433	14,440	14,993
Employed	27,750	13,614	14,136
Unemployed	1,683	826	857
Not in the Labor Force	14,874	7,297	7,577
Employment Rate	66.43%		

Figure (1) Simple quantity reference.



Labor Force

In CY 2000, the City of Isabela has a Labor Force of [27,750] or [94.28%*] are employed and [1,683] or [5.72%*] are unemployed. Of the [44,307
] Household Po ... mically active
Labor Force ar ... opulation (not in
the labor force)
Using the CY 2 ... [1,243] Overseas
Workers recor ... rs belongs to the
group of peopl ... ary with [510] or [
41.03%*] of th ... people who either
reached or con ... People with No
Grade Comple ... Undergraduate
and Academic ... 85] and [31]
overseas work
People under t ... nber of people

Highlighted Cells are related with
conf: 33.23%; aggr-func: percentage

	Total	Male	Female
Both Sexes (15 & Over)	44,307	21,737	22,570
In the Labor Force	29,433	14,440	14,993
Employed	27,750	13,614	14,136
Unemployed	1,683	826	857
Not in the Labor Force	14,874	7,297	7,577
Employment Rate	66.43%		

Figure (2) Aggregate quantity reference.

1 INTRODUCTION

Motivation. The Web contains a wealth of pages with embedded tables, and reports with spreadsheets are abundant in enterprises. Such documents, with financial or statistical data, are challenging to read, as they are often packed with numbers and tables. For example, in a financial report, a reader can stumble upon a statement like "..overall revenues were up 21 percent year-over-year... ", giving rise to questions such as: "What was the revenue of the previous year?" or "Which particular product or sector contributed to this increase?". In such cases, the table(s) accompanying the text can provide answers. However, long documents contain several tables, and table cells are referenced at many spots throughout the report. Moreover, many textual references round or truncate numbers, or refer to aggregates such as row or column totals, which are not explicitly given in the table(s). Therefore, it is tedious work to navigate between text and tables to answer the reader's questions.

Generally, what a reader would desire is an easy and seamless way of drilling down from text passages to the relevant table cells for additional detail, and zooming out from tables to the relevant sentences that explain the numbers.

Contribution. To address the above desiderata, we propose *ExQuisiTe*, a system that identifies relations between quantities in text and tables. ExQuisiTe automatically detects these relations and generates an easy-to-read document where numbers in text are linked to their source tables and respective cells. It identifies simple mentions of single-cell table quantities as well as mentions of aggregate quantities. For example, in Figure 1 the mention "1,683" in the text refers to a simple quantity in the table; and in Figure 2

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3314134

the mention “5.72%” refers to an aggregate quantity (percentage) in the same table.

Furthermore, ExQuisiTe can guide *Extractive Text Summarization (ETS)* systems by emphasizing sentences with aggregate quantities. Current summarization systems [2, 9] do not include table data, and ExQuisiTe opens the opportunity for them to harness table data. Once ExQuisiTe identifies references of simple and aggregate table quantities in the text, it can suggest sentences with aggregations to be included in the summary generated by the ETS algorithm.

For example, in Figure 2 the highlighted sentence covers more cells in the table than the other sentences. It contains more aggregate mentions, and hence it provides a better summary, with judicious consideration of the numbers in the tables. ExQuisiTe is base on the BriQ algorithm [6], and consists of four configurable stages: (i) Document Extraction, (ii) Local Resolution, (iii) Global Resolution, and (iv) Markup and Summary Generation. The first stage extracts text segments and their possible related tables using string similarity measures. The second stage identifies potential alignments between quantity mentions in text and tables based on local features. Then, the third stage collectively aligns quantities in the text to their relevant quantities in tables. Finally, the system generates markup for the document with the inferred alignments and selects important sentences for summarization.

The code of ExQuisiTe as well as all the annotated data used for training is available on the project web page [1]. Our main contributions are:

- an end-to-end system for quantity alignment,
- a system that generates salient suggestions for a downstream ETS method,
- an open-source efficient pipeline that can be flexibly configured on a Spark cluster for online document processing.

2 COMPUTATIONAL MODEL

Our algorithm handles the following inputs:

- a piece of text, like a (part of a) web page, with a set of m text mentions of quantities $X = \{x_i : i = 1, \ldots, m\}$,
- a table q with r rows and c columns and a set of n mentions of quantities $T = \{t_j : j = 1, \ldots, n\}$.

Text mentions include terms containing numbers or numerals such as “123 patients”, “37K EUR”, “1.5%” or “twenty pounds”.
Table mentions include two types of quantities. The first are *simple mentions*, such as ‘1,683’ in Figure 1. Given a table with r rows and c columns we have at most $r \cdot c$ single-cell quantity mentions. The second type is *aggregate mentions*, computed as an aggregation of one or more table cells, such as ‘5.72%’ in Figure 2.

In this demo we consider the following *aggregate functions*: average, sum, difference, percentage, and change ratio.

For aligning quantity mentions between text and tables, we aim to compute as output a subset of mention pairs $\langle x_i, t_j \rangle$ where $x_i \in X$ is a text mention and $t_j \in T$ is a table mention, including aggregate quantities. These pairs should denote the same quantity with high confidence.

3 SYSTEM COMPONENTS

3.1 Document Extraction

Long web pages can cover a variety of thematic aspects, along with several tables. Therefore, we decompose the input web page into coherent segments which we refer to as *documents*. We define a coherent document to be a paragraph together with all “related” tables from the same web page. Each document can be processed independently from the other documents. Hence, we can leverage a distributed computing framework, Spark, for online page processing.

This module first decomposes the input web page into paragraphs, then recognizes related tables for each paragraph using pairwise similarities between all paragraphs and all tables in the web page.

3.1.1 Quantity Extraction. For each document, quantity mentions are extracted from the text and the tables, using regular expressions. Our method pays particular attention to the challenges of aggregated quantities (e.g., column totals). Therefore, we generate table candidates as combinations of cells with an associated aggregate function. For example, we generate aggregate mention for a column total even if the table does not explicitly show the total. Aggregate quantities are automatically generated by considering (i) all rows and columns for totals and averages and (ii) all pairs of cells in the same row or column for difference, percentage, and change ratio. We prune the aggregate quantity candidate, to ensure computational tractability and to control spurious matches.

3.2 Local Resolution

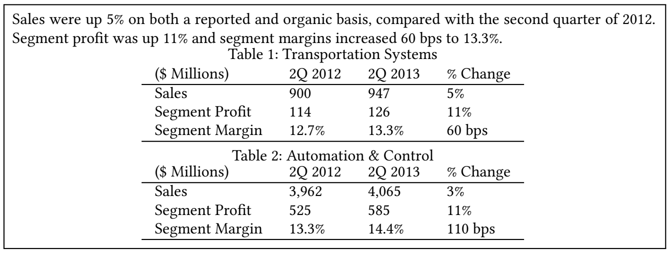
This module first computes features for each text mention and each table mention by analyzing the surrounding context. Then, it computes similarity-based features for each pair of text mention and table mention including aggregate quantities. After that, it uses a binary classifier that accepts or rejects candidate mention-pairs. This binary classifier assigns a confidence score to each mention-pair, and we use this score in the following steps. At the end of this module, we filter the candidate mention-pairs according to their confidence score and other measures which we will explain later.

3.2.1 Mention-Pair Classification. We use manually annotated web pages with ground-truth alignment to train a *Random Forest (RF)* classifier. The classifier operates *locally* in the sense that it predicts the alignment confidence for each mention-cell pair in isolation, It serves two purposes: First, it enables the subsequent filtering step, which significantly reduces the number of candidate pairs for achieving an acceptable running time in the global resolution step. Second, it provides a prior for that global resolution step.

3.2.2 Classifier Feature. For the mention-pair classifier, we designed a variety of features that capture information a human reader would use in order to determine if text mention x and table cell t denote the same quantity. This includes *surface form similarities*, *context features*, and *quantity features*. For more details refer to Ibrahim et al. [6].

3.2.3 Adaptive Filtering. This stage reduces the number of mention-pair candidates from 1000s of candidates to 100s for tractability of global inference algorithms. We design the *adaptive filtering*

[1] https://www.mpi-inf.mpg.de/briq/



Sales were up 5% on both a reported and organic basis, compared with the second quarter of 2012. Segment profit was up 11% and segment margins increased 60 bps to 13.3%.

Table 1: Transportation Systems

($ Millions)	2Q 2012	2Q 2013	% Change
Sales	900	947	5%
Segment Profit	114	126	11%
Segment Margin	12.7%	13.3%	60 bps

Table 2: Automation & Control

($ Millions)	2Q 2012	2Q 2013	% Change
Sales	3,962	4,065	3%
Segment Profit	525	585	11%
Segment Margin	13.3%	14.4%	110 bps

Figure (3) Example with Coupled Quantities

algorithm to work in two stages. In the first stage, we develop a *text mention tagger* to predict the aggregation function for each text mention or tag the mention as a single-cell match. Then, we prune mention-pairs based on this tagger's outcome. In the second stage, we prune mention-pairs based on *value difference* and *unit mismatch*. Finally, we sort mention-pairs according to classifier scores and select top-k mention-pairs for each quantity mention based on *mention type* and *score distribution*. For more details refer to Ibrahim et al. [6].

3.3 Global Resolution

This module takes as input the candidate mention-pairs from the classifier and outputs the alignment of quantity mentions. We harness dependencies among mentions to resolve ambiguities. Consider the example in Figure 3. The text mentions "11%" and "13.3%" have exact matches in both of the shown tables, and local-resolution algorithms cannot infer the proper alignment. However, when considering these two mentions jointly with "60 bps" and "5%", it becomes clear that all of these refer to the first table.

We devised an unsupervised algorithm for this kind of global resolution. The algorithm encodes dependencies among mentions into a graph and uses *Random Walks with Restarts(RWRs)* to infer the best joint alignment.

3.3.1 Graph Construction. We construct an undirected weighted graph $G = (V, E)$ for each document:

- The node set V consists of all quantity mentions in the document's text and tables.
- The edge set E consists of three kinds of edges connecting related nodes: text-text edges, table-table edges, and text-table edges as explained below.

(i) Text-text edges: connects each pair of text quantity mentions that are within a certain proximity or have similar surface forms. Edge weights are computed based on a linear combination of proximity and string similarity. *(ii) Table-table edges:* connects each pair of table quantity mentions in the same row or the same column of the same table, and edge weights are set uniformly. *(iii) Text-table edges:* connects each pair of text and table mention that is kept by the adaptive filtering stage, and edge weights are set to the confidence scores returned by the classifier. After this initial graph construction, all edge weights are normalized to obtain a stochastic graph, via dividing each node's outgoing weights by the total weight of these edges.

3.3.2 Graph Algorithm. In our setting, we employ random walks with restart: starting from a text mention, the graph is stochastically traversed, with a certain probability of jumping back to the initial node. This technique is also known as topic-specific or personalized PageRank [4]. Our implementation iterates RWRs for each text mention until the estimated visiting probabilities of the candidate table mentions change by less than a specified convergence bound. This way we obtain a ranked list of table mentions for each text mention x.

Alignment decisions: The RWR from text mention x computes the stationary probability $\pi(t|x)$ for each table mention t. Pair $\langle x, t^* \rangle$ forms an alignment if and only if (i) t^* is the table mention with the highest overall score, and (ii) its overall score exceeds the defined confidence threshold. We then exploit the alignment decisions to update the graph, such that after identifying an alignment $\langle x, t^* \rangle$, x we modify the graph by removing all edges (x, t) for any $t \neq t^*$.(If no alignment is found for x, then all text-table edges adjacent to x are removed.)

3.4 Markup and Summarization

This module integrates the output of our system with the content of the web page and displays the results to the user in the form of an HTML page. This module is also responsible for analyzing the aligned quantities and highlighting the important sentences for the summarization engine. It estimates the importance of a sentence based on its coverage of table cells.

We define the *coverage of a quantity mention* in the text to be the number of individual single-cells it refers to. For mentions referring to an aggregate table quantity such as the sum of a column, we include all the individual single-cells in this column. Then, we compute the *coverage of a sentence* as the sum of the coverage of its mentions.

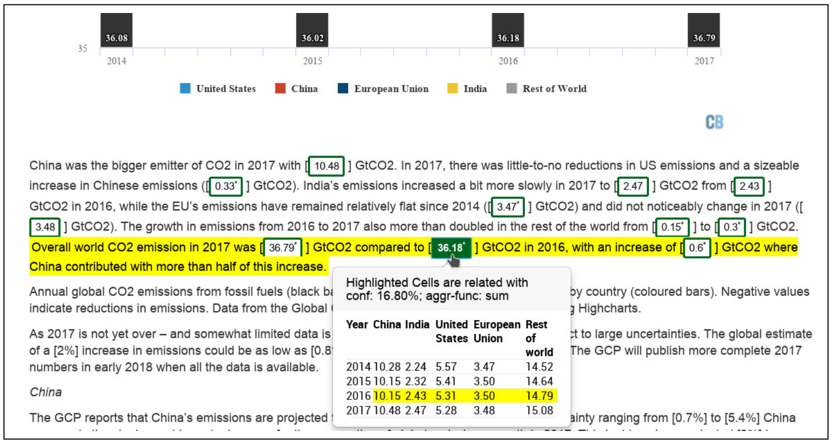
For each table, we extract all the sentences that reference it. Then, we compute a score for each sentence based on its coverage. After that, we highlight the sentence with the maximum score in the generated HTML. For example in Figure 4 even though the second sentence has more simple quantity references to the table than the last sentence, the latter provides a better summary of the table. The last sentence discusses the overall CO2 emission in the world and has the highest coverage of table cells, while the second sentence only discusses the emission of India and the EU.

4 EXPERIMENTAL RESULTS

We trained and evaluated our system on a manually annotated corpus of 495 web pages. The F1 score of our system is 79% for the simple quantity mentions, and 40% for the aggregate mentions. We carried out a run time analysis on a Spark cluster with 10 executors, each with 6 cores and 30GB of memory, and with 50GB of driver memory. The throughput of our system is 2.5K documents per minute. For full evaluation results refer to Ibrahim et al. [6].

5 DEMO OVERVIEW

In this demonstration, we will show how ExQuisiTe aligns quantity mentions in web pages. ExQuisiTe is web-based therefore it only requires a modern web browser. It has two main views: the first view is for configuring the system and selecting the type of input document, and the second view is for displaying the alignment results.



China was the bigger emitter of CO2 in 2017 with [10.48] GtCO2. In 2017, there was little-to-no reductions in US emissions and a sizeable increase in Chinese emissions ([0.33*] GtCO2). India's emissions increased a bit more slowly in 2017 to [2.47] GtCO2 from [2.43] GtCO2 in 2016, while the EU's emissions have remained relatively flat since 2014 ([3.47*] GtCO2) and did not noticeably change in 2017 ([3.48] GtCO2). The growth in emissions from 2016 to 2017 also more than doubled in the rest of the world from [0.15*] to [0.3*] GtCO2. Overall world CO2 emission in 2017 was [36.79*] GtCO2 compared to [36.18*] GtCO2 in 2016, with an increase of [0.6*] GtCO2 where China contributed with more than half of this increase.

Annual global CO2 emissions from fossil fuels (black ba by country (coloured bars). Negative values indicate reductions in emissions. Data from the Global g Highcharts.

As 2017 is not yet over – and somewhat limited data is ct to large uncertainties. The global estimate of a [2%] increase in emissions could be as low as [0.8 The GCP will publish more complete 2017 numbers in early 2018 when all the data is available.

China

The GCP reports that China's emissions are projected ainty ranging from [0.7%] to [5.4%] China

Highlighted Cells are related with conf: 16.80%; aggr-func: sum

Year	China	India	United States	European Union	Rest of world
2014	10.28	2.24	5.57	3.47	14.52
2015	10.15	2.32	5.41	3.50	14.64
2016	10.15	2.43	5.31	3.50	14.79
2017	10.48	2.47	5.28	3.48	15.08

Figure (4) The figure shows the results of the system as described in Section 5

Configuration and Input. ExQuisiTe gives the user control over the settings of the different components. The user can select the type of input she wants to process. Currently, we support HTML input given as a valid page URL or a file containing the HTML content.

For the global resolution the user can choose either (i) Random Walk With Restart (RWR) or (ii) No Global Resolution. RWR is the default option. The second option deactivates the global resolution module and uses only the outcome of the classifier. It uses the confidence value given by the classifier to select the highest-confidence table mention for each text mention.

The user can adjust the threshold for the final confidence score: mapping the text mention to a table mention or to a NIL. In the case of local resolution only, this threshold is applied to the classifier's confidence score. In the case of RWR global resolution, the threshold is applied to the RWR outcome. Finally, there is an option to turn adaptive filtering on or off.

Results Display and Interactive Exploration. The system processes the document according to the user's configuration. Then, it displays the results embedded in the original HTML document as shown in Figure 4.

The text quantity mentions appear between square bracket. Each text mention is hyperlinked to its aligned table mention using a colored button with the mention's surface form as the hypertext. The color of the button is determined by the table, such that each table is assigned a color and all its related mentions in the text are assigned the same color.

The system marks the aggregate quantity mentions with a '*' superscript. When a text mention is clicked, the system displays a pop-up with the table. This pop-up includes—in addition to the table—the confidence of the alignment, the type of alignment, and the related cells in the table. The type of alignment can take one of the following values: single-cell, average, sum, difference, percentage, and change ratio.

The system marks the salient sentences for downstream summarization with yellow background. For each table, the system marks the sentence with the highest coverage. In Figure 4, the last sentence has the highest coverage of table cells.

6 RELATED WORK

Although information extraction research has targeted Web tables, no prior work has examined the relation between mentions in the text and tables within a document. Quantity annotation has been addressed in [5, 12], but these methods rely on external knowledge bases, linking table cells to entries in the knowledge base. Further methods focused on named entities, by annotating table cells with entities and classes from the knowledge base [1, 3, 8, 10, 11]. Table data fusion for search and schema inference was studied in [7, 13–15].

Our work differs from all these prior works in two main aspects: (i) we do not rely on any external knowledge base, and (ii) we handle approximate and aggregated quantities mentions which do not have exact matches.

REFERENCES

[1] C. S. Bhagavatula, T. Noraset, and D. Downey. 2015. TabEL: Entity Linking in Web Tables. *ISWC* (2015).
[2] M. Gambhir and V. Gupta. 2017. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review* 47, 1 (2017), 1–66.
[3] A L. Gentile, P. Ristoski, S. Eckel, D. Ritze, and H. Paulheim. 2017. Entity Matching on Web Tables: a Table Embeddings approach for Blocking. *EDBT* (2017).
[4] T. H. Haveliwala. 2003. Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search. *TKDE* (2003).
[5] Y. Ibrahim, M. Riedewald, and G. Weikum. 2016. Making Sense of Entities and Quantities in Web Tables. *CIKM* (2016).
[6] Y. Ibrahim, M. Riedewald, G. Weikum, and D. Zeinalipour-Yazti. 2019. Bridging Quantities in Tables and Text. In *ICDE 2019*. IEEE.
[7] O. Lehmberg and C. Bizer. 2017. Stitching Web Tables for Improving Matching Quality. *PVLDB* (2017).
[8] G. Limaye, S. Sarawagi, and S. Chakrabarti. 2010. Annotating and Searching Web Tables Using Entities, Types and Relationships. *VLDB* (2010).
[9] A. Nenkova and K. McKeown. 2012. A survey of text summarization techniques. *Mining text data* (2012).
[10] D. Ritze and C. Bizer. 2017. Matching Web Tables To DBpedia - A Feature Utility Study. *EDBT* (2017).
[11] D. Ritze, O. Lehmberg, and C. Bizer. 2015. Matching HTML Tables to DBpedia. *WIMS* (2015).
[12] S. Sarawagi and S. Chakrabarti. 2014. Open-domain quantity queries on web tables: annotation, response, and consensus models. *SIGKDD* (2014).
[13] P. Venetis, A. Y. Halevy, J. Madhavan, M. Pasca, W. Shen, F. Wu, G. Miao, and C. Wu. 2011. Recovering Semantics of Tables on the Web. *PVLDB* (2011).
[14] M. Yakout et al. 2012. Infogather: entity augmentation and attribute discovery by holistic matching with web tables. *SIGMOD 2012* (2012).
[15] M. Zhang and K. Chakrabarti. 2013. Infogather+: Semantic matching and annotation of numeric and time-varying attributes in web tables. *SIGMOD* (2013).