Guiding Readers to the Focus and Context of Industrial Statistical Reports

Ken Wakita*
Tokyo Institute of Technology

Kohei Arimoto†
Teikoku Databank
The University of Tokyo

ABSTRACT

The poster proposes a bi-modal presentation method of industrial statistical reports to assist readers in comprehending the text by guiding readers attention to the context and the focus on the statiscal data presented in the associated tables and charts. This technique is expected to assist the reader to grasp content of the report through good understanding of the statistics.

Index Terms: Human-centered computing—Visualization—Information Visualization;

1 INTRODUCTION

1.1 Dataset description

Teikoku Databank (TDB) is a Japan-based, business intelligence corporation founded in 1900. It owns Japan's largest corporate database and provides corporate credit research services. This research utilizes TDB's online reports, which are available from the following page: `https://www.tdb.co.jp/report/watching/`. The reports are targeted to industry leaders who can read Japanese, edited by professional editors hired by TDB, and published almost every working day. Each issue of the report covers specialized topics related to Japanese companies, such as "company capital investment incentives", "changes in industrial structure", "latest financial trends", "market surveys", and "employment trends".

Though the reports are written by human editors, they are described in a standardized format. Reports are about eight pages long, single-column, and divided into sections containing paragraph sequences, followed by data illustrations, such as tables or charts (of various kind). The paragraphs refer to illustrated data points and report on facts deduced from the data.

1.2 References and Claims

To investigate the potential of a bimodal presentation of the report by linking the report text and illustration content, we extracted all the *data references* in the text content of the report and categorized them. Table 1 presents two examples of such data references found in a June 24, 2019, report on "Female leaders in Japanese companies". The dataset they refer is a collection of quantitative values, indexed by years (1989, 1999, 2009, 2014, 2018, and 2019) and by seven industry types ("construction", "manufacturing", "wholesale", "retail", "transport and communication", "service industry", "real estate", and "others"). The first example (1) in Table 1 is a simple dataset lookup, $D[2019][\text{RealEstate}]$, where D denotes the dateset. Example (2) indicates that the temporal change of the value for the "real estate" industry can be expressed using simple arithmetic: $(D[2019] - D[2019-30])[\text{RealEstate}]$.

It was also observed that the text contained *claims* based on various types of statistical reasonings. The lower part of Table 1 lists claims from the two paragraphs presented in Fig. 1. Claims (c) and (a) point out the lower extreme value of the column, and an upper extreme value of the temporal changes, respectively. Claims (d) and (e) are regarded as the unique existential and universal properties, respectively. These claims can easily be expressed through statistic and logical functions. The editor's intention behind somewhat unclear phrases such as 'high' and "well below" in claims (b) and (f) is explicitly formulated in the corresponding formulae.

1.3 Focus and Context

The reader must know that these claims are supported by a set of data values under consideration (the *focus*), the value distribution of other data items (the *context*), and the method of statistical reasoning being employed. For example, claim (a) discusses the increase in the proportion of female CEOs over the past 30 years and thus, the data points under consideration are the proportions in years 1989 and 2019 for all industry types, while the focus is the 7.5% increase in "real estate" industry, specifically. We can identify focuses and contexts of all other claims in the table and formulate their intention as described in the **Assertions** column of the table.

The focus and context in our proposal vary from one claim to another. The focus in our case refers to the data points and not variables, while the context corresponds to the scope of comparison under discussion. Latif and Beck [3] introduce a similar notions of *focus variable* and *context variable* in their interactive map report proposal. They assumed that the focus and context variables were embedded in the problem domain and do not change.

2 RESEARCH AIM: CONTEXT AND FOCUS OF STATISTICAL ANALYSIS

The aim of the research is to induce reader attention from report text content to the associated data illustrattions (tables and charts). We hope to evoke a more self-motivated, critical, and explorative reading attitude by encouraging readers to

- locate illustrated data items corresponding to the text;
- verify "claims" described in the text through statistic testing; understand the statistical standards that lead to factual descriptions in the text;
- learn to discover undocumented insights on their own.

To realize this aim, we propose a simple visual interaction mechanism that links text and illustration, that directing reader interest to tables and charts, while highlighting data items shown on the tables and charts to let the reader comprehend the focus and the context of the text.

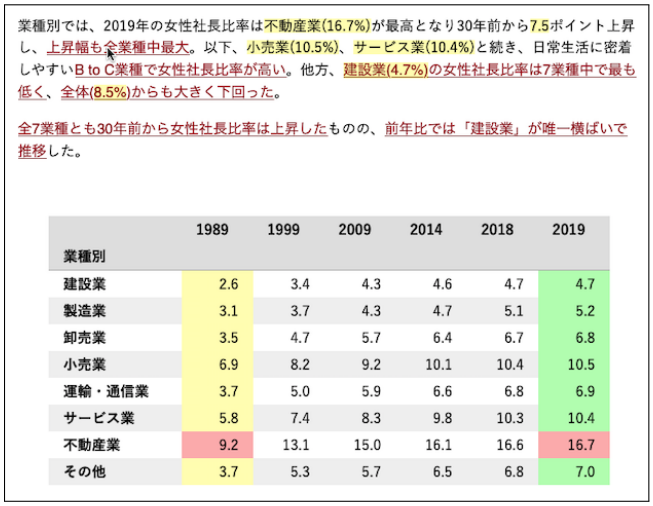
Fig. 1 and Fig. 2 depict our first Web-based prototype of this idea. The Web-based interface carries content indentical to the original PDF-based report, but it also highlights the text that refers to data in yellow, while the claims supported by statistical reasoning are indicated with red underlined font.

The reader can move a pointing device over these highlighted text areas. Hovering over a data reference in the text makes a data item refered to in a table or a chart as the focus of the text and gives it a pink background color. Additionally, when a pointing device hovers over a claim in the text, its focus area is given a pink background color and the context area is given background color according to its degree of importance.

*e-mail: wakita@is.titech.ac.jp
†e-mail: kohei.a.19870908@gmail.com

Table 1: Data references and claims found in the report article with corresponding statistical expressions that reference the database (D) and perform automated fact checks

Data references found in a text	**Statistical formulae**
(1) The ratio of female CEOs in 2019 in the 'Real Estate' (16.7%)	$D[2019][\text{Real Estate}]$
(2) increase of 7.5% over the past 30 years	$(D[2019] - D[2019-30])[\text{Real Estate}]$
Claims found in the text	**Assertions (fact check expressions)**
(a) The "Real estate" industry also had the highest increase of 7.5% over the past 30 years	$(D[2019] - D[2019-30]).\text{idxmax}() = \text{Real Estate}$
(b) Female leadership was also high in the B to C focused Retail and Services industries	$D[2019][\text{B2C}].\text{min}() > D[2019].\text{mean}()$
(c) female leadership was the lowest in the Construction	$D[2019].\text{idxmin}() = \text{Construction}$
(d) Construction was the only industry with no change from 2018	$(D[2019] - D[2018])[\text{Construction}] = 0$ & $(D[2019] - D[2018])[\text{Others}] > 0).\text{all}()$
(e) the share of female leaders has generally increased year-on-year	$(D[2019] > D[2019-30]).\text{all}()$
(f) The share of female CEOs in Construction was well below the all-industry average.	$D[2019][\text{Construction}] < D[2019].\text{mean}() - D[2019].\text{std}()$


業種別では、2019年の女性社長比率は不動産業(16.7%)が最高となり30年前から7.5ポイント上昇し、上昇幅も全業種中最大。以下、小売業(10.5%)、サービス業(10.4%)と続き、日常生活に密着しやすいB to C業種で女性社長比率が高い。他方、建設業(4.7%)の女性社長比率は7業種中で最も低く、全体(8.5%)からも大きく下回った。

全7業種とも30年前から女性社長比率は上昇したものの、前年比では「建設業」が唯一横ばいで推移した。

業種別	1989	1999	2009	2014	2018	2019
建設業	2.6	3.4	4.3	4.6	4.7	4.7
製造業	3.1	3.7	4.3	4.7	5.1	5.2
卸売業	3.5	4.7	5.7	6.4	6.7	6.8
小売業	6.9	8.2	9.2	10.1	10.4	10.5
運輸・通信業	3.7	5.0	5.9	6.6	6.8	6.9
サービス業	5.8	7.4	8.3	9.8	10.3	10.4
不動産業	9.2	13.1	15.0	16.1	16.6	16.7
その他	3.7	5.3	5.7	6.5	6.8	7.0

Figure 1: The table presents the percentages of female CEOs in Japanese industry in the last 30 years and the text above it describe the trends. The mouse pointer hovers over a text that reads: *The Real estate industry had the highest percentage point increase of 7.5% over the past 30 years.* The table columns that correspond to the latest and 30 years back form the context of this statement. The values of the real estate industry for these years form the focus of the statement.

For example, in Fig. 1, a pointing device hovers over text that can be translated as the blue portion of (a) in Table 1. We choose the focus of this claim to be two data points, namely $D[2019][\text{Real Estate}]$ and $D[2019][\text{Real Estate}]$, that contribute as the maximum increase from year 1989 to 2019, and the context to be the two columns (1989 and 2019) being compared under this consideration. Hence, the focus is given a pink background color, and the two columns are highlighted in yellow and green background colors. Another example, which employs a bar chart is shown in Fig. 2.

The prototype system is a simple Web application that is implemented using server-side Python (Pandas, Jinja, and Flask) and front-end JavaScript (D3.js for visualization). The original text is converted into a Jinja template with a set of named "holes" filled in by the Jinja template engine with the evaluation results of the respective statistical formulae expressed as Pandas expressions, similar to the ones given in the second column of Table 1. Given a statistical formulae, we regarded all the statistical variables contributing to the

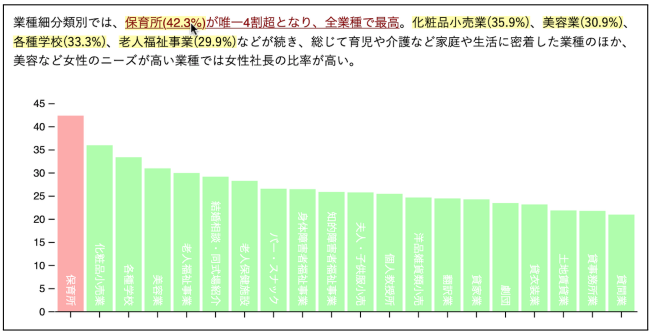


Figure 2: The bar chart depicts the percentages of female CEOs in Japanese industries as of 2018 and the text describes the trend. The mouse pointer hovers over a claim that reads: *Further analysis, by the detailed industry type, shows that the Nursery school sub-industry has the highest share of female CEOs of 42.3%.* The context of this claim is the list of whole sub-industries and the focus is the Nursery school sub-industry.

formula as context and the one(s) most importantly connected to the result as the focus. The focus was assigned manually by the authors when providing the data.

3 Concluding Remarks

The poster discussed the importance of suggesting focuses and context for reported data refrences and claims. We proposed a Web-based non-intrusive interaction technique to show these acpects by highlighting respective table cells and chart shapes.

Future research directions include (1) analyzing more reports to test the generality of the proposed method, (2) determining if reading experiences improve through task-based evaluation as performed by Kumar and others [2], (3) incorporating the automatic text synthesis technique proposed by Latif and Beck [3], (4) incorporating other annotation techniques such as the ones proposed by Brath and others [1], and (5) automating the manual focus specification using the structural analysis of statistical formulae.

Acknowledgments

The authors would like Teikoku Databank, Ltd.'s Center for TDB Advanced Data Analysis and Modeling for providing both the data and financial support.

REFERENCES

[1] R. Brath and M. Matusiak. Automated annotations. In *An IEEE VIS workshop on visualization for communication (VisComm)*, 2018.

[2] A. Kumar, M. Burch, I. van den Brand, L. Castelijns, F. Ritchi, F. Rooks, H. de Smeth, N. Timmermans, and K. Mueller. Eye tracking for exploring visual communication differences. In *An IEEE VIS workshop on visualization for communication (VisComm)*, 2018.

[3] S. Latif and F. Beck. Interactive map reports summarizing bivariate geographic data. *Visual Informatics*, 2019. doi: 10.1016/j.visinf.2019.03.004