Assessing and Visualising Online Information Quality

Davide Ceolin, Jacco van Ossenbruggen
Information Access Group
CWI
Amsterdam, The Netherlands
davide.ceolin@cwi.nl

Lora Aroyo, Ozkan Sener,
Robin Sharma, Lesia Tkacz
Computer Science Department
Vrije Universiteit Amsterdam
Amsterdam, The Netherlands

Julia Noordegraaf
Media Studies Department
University of Amsterdam
Amsterdam, The Netherlands

***Index Terms*—Information Quality, Online Information**

While the information available on the Web is potentially useful to a variety of users, ranging from laypeople to scholars, it may be provided by incompetent or malicious authors. Additionally, the social media mechanisms incentivise popularity of content, which provides an additional motivation for misuse of the Web democracy to emphasize a particular perspective. Fake news is a noteworthy example of documents containing low-quality (e.g., inaccurate) information, often created with misleading intentions that gain or gained a high popularity. This calls for an increase in the user awareness of the quality of the information they consume. To complicate things further, quality is not a monolithic and binomial concept. The overall quality of a document depends on the user that assesses it and on the intended uses for this document. However, it is possible to decompose quality into dimensions that tackle a specific aspect of quality (e.g., accuracy) and are thus easier to quantify. InfoQ semi-automatically scores online information on a number of these dimensions. These scores can be combined in order to increase the awareness of possible aspects related to information quality [1], [6].

In this abstract, we outline InfoQ - a data assessment tool developed within the context of the QuPiD2 project (Quality and Perspectives in Deep Data). QuPiD2 investigates methods and tools for computational support to capture, model and assess the diversity in quality of online information and the multitude of perspectives (i.e. beliefs, opinions, and world views) being reflected in this online information. Thus, we explore the perception factors and linguistic phenomena reflecting a certain perspective or influencing the perceived quality of text, and develop tools to automatically detect perspectives and assess quality. The quality assessment performed by InfoQ provides a multi-perspective view along multiple quality dimensions (precision, trustworthiness, accuracy, neutrality, readability, relevance with respect to a given topic). In order to address the intrinsic subjectivity of quality assessment, InfoQ is based on a symbiotic pipeline that brings together humans and machines to gather and train information assessments and the factors that impact them. InfoQ machine learning models (multi-label regression and Support Vector Machines) are trained on quality assessments provided by experts and crowds and rely on automatic extraction of document features, such as NLP features, provenance features, web-based technical features (e.g., latency, i.e., Website speed test), and crowdsourced features (e.g., Web of Trust trustworthiness scores) to identify correlations between document features and human assessments, and allow for the assessment of any Web document.

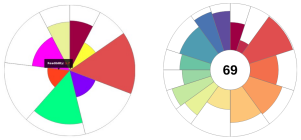

Fig. 1. Examples of visualisations of quality dimensions. Each color stands for a different dimension. The size of the segments indicate the dimension score. While the left diagram allows the user understanding the contribution of each dimension to the overall quality of the document, the right diagram highlights the actual score of each dimension, when hovered with the mouse pointer.

The tool has been tested on a group of 50 documents regarding the vaccination debate (selected in order to represent a small but heterogeneous sample consisting of blog posts, news articles, documents from public authorities, etc.), where it shows a promising performance (up to 90% accuracy) that reflects previous works of ours [3], [6].

The quality assessments produced by InfoQ are presented to the user by employing diverse types of visualisation we are currently exploring. This allows the user to both obtain a summary of the quality of a document (or of a group of documents) without having to introduce artificial aggregations. Figure 1 shows an example of a visualisation that combines insights into each dimension with an overview of the quality of the document assessed: while the overall quality of the document is relatively good, we can see that actually, the readability of the document is low. Such a visualisation also allows the final user to investigate further the quality of a given document, thus increasing user awareness of the quality of documents. These visualisations are a crucial aspect of our tool: on the one hand, they should be as exhaustive as possible, to convey all the possibly useful details regarding the quality of the information assessed; on the other hand, their complexity should be limited by the user understandability.

REFERENCES

[1] Son, Chantal van, Emiel van Miltenburg, and Roser Morante (2016). Building a Dictionary of Affixal Negations. In Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics (ExProM 2016), Osaka, Japan.

[2] Fokkens, Antske, Serge ter Braake, Isa Maks, and Davide Ceolin, On the Semantics of Concept Drift: Towards Formal Definitions of Semantic Change, Drift-a-LOD. Detection, Representation and Management of Concept Drift in Linked Open Data. Workshop at EKAW, Bologna, Italy, 20 November 2016.

[3] Ceolin, Davide, Julia Noordegraaf, Lora Aroyo, Capturing the Ineffable: Collecting, Analysing, and Automating Web Document Quality Assessment, EKAW 2016: Knowledge Engineering and Knowledge Management, 83-97.

[4] Davide Ceolin, Lora Aroyo, Julia Noordegraaf, Identifying and Classifying Uncertainty Layers in Web Document Quality Assessment, International Workshop on Uncertainty Reasoning for the Semantic Web. pp: 61-64

[5] Van der Zwaan, J.M., Maarten van Meersbergen, Antske Fokkens, Serge ter Braake, Inger Leemans, Erika Kuijpers, Piek Vossen and Isa Maks, Storyteller: Visualizing Perspectives in Digital Humanities Projects, in: Bozic, Mendel-Gleason, Debruyne and OSullivan eds., 2nd IFIP Workshop on Computational History and Data-Driven Humanities (2016).

[6] Davide Ceolin, Julia Noordegraaf, Lora Aroyo, and Chantal van Son. 2016. Towards web documents quality assessment for digital humanities scholars. In Proceedings of the 8th ACM Conference on Web Science (WebSci 16). ACM, New York, NY, USA, 315-317.

[7] Davide Ceolin, Chantal van Son, Lora Aroyo, Julia Noordegraaf, Ozkan Sener, Robin Sharma, Piek Vossen, Serge ter Braake, Lesia Tkacz, Inger Leemans, Rens Bod, DHBenelux 2018.