InfoQ: Computational Assessment of Information Quality on the Web[1]

Davide Ceolin[a]*, Chantal van Son[b], Lora Aroyo[b], Julia Noordegraaf[c], Ozkan Sener[b], Robin Sharma[b], Piek Vossen[b], Serge ter Braake[c], Lesia Tkacz[b], Inger Leemans[b], Rens Bod[c]

[a] Centrum voor Wiskunde en Informatica (CWI)
[b] Vrije Universiteit Amsterdam
[c]Universiteit van Amsterdam

Keywords: information quality, trust, nichesourcing, machine learning.

As a result of the democratic nature of the Web, people can contribute different types of information by means of blog posts, articles, tweets - a huge amount of information is available on the Web and this information can be potentially useful to a variety of users, ranging from laymen to scholars. However, this results in a *vast variety of quality* of the published documents expressing a *multitude of perspectives* on different topics. Online information could be provided by incompetent authors or by malicious ones. Additionally, the social media mechanisms for sharing and liking introduce the *popularity currency* (number of retweets, reposts, likes, etc.), which provides an additional motivation for 'misuse' of the Web democracy to generate artificial popularity of a particular perspective, or to disguise incompetency behind 'professionally looking' websites (e.g. low-quality documents can easily be crafted to appear credible and to gain popularity). 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 adds another layer of uncertainty regarding information quality. To complicate things further, quality is not a monolithic and binomial thing: it is hardly possible to judge documents as 'good' or 'bad' in absolute terms. The overall quality of a document depends both on the topics, the user that assesses it, and on the intended uses for this document. However, it is possible to decompose quality into objective 'dimensions' or 'aspects' that can be combined in order to increase the awareness of possible aspects related to information quality, and also to increase the awareness in terms of the relation between quality and perspectives [1-6].

In this demo, we introduce InfoQ - a data assessment tool developed within the context of the QuPiD2 project (Quality and Perspectives in Deep Data)[2]. 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. InfoQ offers the following functionalities: (1) **document-centric assessment:** for a given URL InfoQ provides a detailed analysis of its information quality and (2) **topic-centric assessment:** for a given topic InfoQ provides in-depth comparative analysis of the quality of all the documents related to this topic. The quality assessment performed by InfoQ provides a comprehensive, exhaustive and 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 are trained on (1) quality assessments provided by experts and crowds and rely on (2) automatic extraction of document features, such as NLP features (e.g., sentiment, named entity recognition), provenance features (source type, etc.), web-based technical features (e.g., latency, i.e., Website speed test), and crowdsourced features (e.g., Web of Trust[3] trustworthiness scores) to identify correlations between document features and human assessments, and allow for the assessment of any Web document. As machine learning algorithm, we employ multi-label regression and Support Vector Machines. Figure 1 gives an overview of InfoQ. Ultimately, the quality assessments produced by InfoQ are presented to the user by employing a radar chart visualization. On the one hand, 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 2 shows a comparison of five documents of different quality. This overview does not provide details about the exact meaning of each dimension of the radar graph but allows a quick comparison among the assessed documents (the small the colored area in the graph the lower the quality of the document, and vice versa).

On the other hand, such a visualisation, allows the final user to investigate further the quality of a given document. When restricting the focus on a lower number of documents, the user can understand precisely how the document scores in each

[1] This demo paper is a revised and shortened version of the honomymus paper accepted at DHBenelux 2018.
[2] https://qupid-project.net/
[3] http://mywot.com

dimension, or how two documents compare to each other (see Figure 2). This allows increasing the user awareness regarding the quality of documents. In turn, allowing users deciding whether documents meet their contextual and subjective requirements, without, for instance, having to decide which (combination of) quality dimension determines the overall quality of a document.

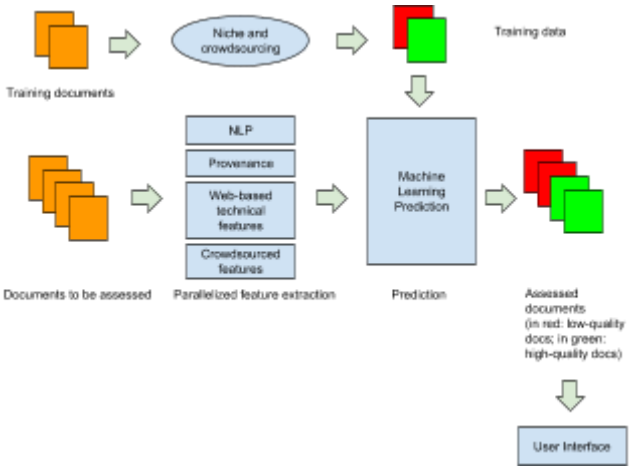

Figure 1. Overview of InfoQ, our online quality assessment tool.

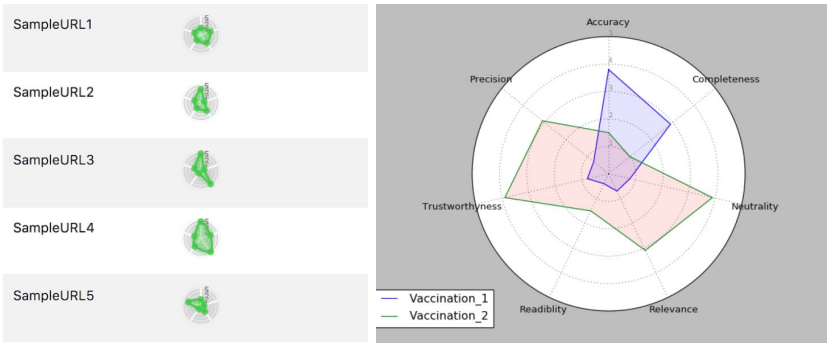

Figure 2. Example of a high-level view of a list of assessed documents and of the comparison between two documents.

InfoQ is a tool for computationally assessing the quality of Web information. The tool is currently running on local instances, and soon to be deployed online. 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]. We envision three main future developments for InfoQ. First, the extension of the domains and topics covered (currently the tool allows any document to be assessed, but the accuracy of such assessments is under evaluation). Second, the improvement of the computation speed. Lastly, the personalisation of the quality assessments, such that different users can obtain assessments matching their specific needs and requirements.

References

[1] Son, C. van, van Miltenburg E., and Morante R. (2016). Building a Dictionary of Affixal Negations. In ExProM 2016.

[2] Fokkens, A., ter Braake S., Maks I., and Ceolin D., 'On the Semantics of Concept Drift: Towards Formal Definitions of Semantic Change', Drift-a-LOD. Detection 2016.

[3] Ceolin D., Noordegraaf J., Aroyo L., 'Capturing the Ineffable: Collecting, Analysing, and Automating Web Document Quality Assessment', EKAW 2016, 83-97.

[4] Ceolin D., Aroyo L., Noordegraaf J., Identifying and Classifying Uncertainty Layers in Web Document Quality Assessment, URSW 2016. pp: 61-64

[5] Van der Zwaan, J.M., van Meersbergen M., Fokkens A., ter Braake S., Leemans I., Kuijpers E., Vossen P. and Maks I., 'Storyteller: Visualizing Perspectives in Digital Humanities Projects', in 2nd IFIP Workshop on Computational History and Data-Driven Humanities (2016).

[6] Ceolin D., Noordegraaf J., Aroyo L., and van Son C.. 2016. Towards web documents quality assessment for digital humanities scholars. In WebSci '16. ACM, 315-317.