Towards Web Documents Quality Assessment for Digital Humanities Scholars

Davide Ceolin
VU University Amsterdam
The Netherlands
d.ceolin@vu.nl

Julia Noordegraaf
University of Amsterdam
The Netherlands
j.j.noordegraaf@uva.nl

Lora Aroyo
VU University Amsterdam
The Netherlands
lora.aroyo@vu.nl

Chantal van Son
VU University Amsterdam
The Netherlands
c.m.vanson@vu.nl

CCS Concepts

•**Information systems** → *Data cleaning;* Answer ranking; Personalization;

Keywords

Quality of Web documents; Digital Humanities

1. INTRODUCTION

We present a framework for assessing the quality of Web documents, and a baseline of three quality dimensions: trustworthiness, objectivity and basic scholarly quality. Assessing Web document quality is a "deep data" problem necessitating approaches to handle both data size and complexity.

Traditional quality assessment methodologies are tailored to physical documents such as books, and qualitatively evaluate their authors and other metadata. These practices need to be extended to respond to the specific nature of online sources. We propose a framework for identifying which characteristics of online sources can be used to signal their quality in different contexts, to benefit both scholars and laymen. This framework enriches the documents analyzed with features from NLP and provenance. In this feature space, quality indicators are sought, to estimate quality assessments that different typologies of users provide in different contexts. We show a preliminary evaluation of employing three techniques for targeting three quality dimensions: crowdsourcing (trustworthiness), manual assessments (objectivity and basic scholarly quality), and best practices (basic scholarly quality). Estimation accuracy reaches 90%, but some quality dimensions are more difficult to estimate.

The paper develops as follows. Section 2 introduces related work. Section 3 presents the framework. Section 4 provides a preliminary evaluation, and Section 5 concludes.

WebSci '16 May 22-25, 2016, Hannover, Germany

ACM ISBN 978-1-4503-4208-7/16/05.

DOI: http://dx.doi.org/10.1145/2908131.2908198

2. RELATED WORK

The ISO 25010 Model [15] is a standard model for data quality. From this, we will select the data quality dimensions that are mostly relevant to assess the quality of the information contained in the data (e.g., credibility), and extend them to cover other aspects that are relevant to users.

Lee et al. [16] provide a framework for assessing the quality of information tailored to organizations. Zhu et al. [20] propose a method for collaboratively assessing the quality of Web documents that shows some similarity with ours (e.g., we both collect collaborative quality assessments), but the assessments we aim at gathering are based on specific tasks, while they rely on contributions via browsers plugins.

Several works have been proposed on the assessment of the quality of Wikipedia articles, as exemplified by the works of Dalip et al. [8] and of Anderka et al. [3]. These works provide a useful basis for our investigations, although our analyses involve a broader set of information sources than Wikipedia.

Source criticism is the process of evaluating traditional information sources that is common in the (digital) humanities. De Jong and Schellers [10] provide an overview of source criticism methods, evaluated in terms of predictive and congruent validity. We will advance such evaluations to identify which document features determine their quality.

Provenance analysis is used to assess the quality of humanities sources, as Howell and Prevenier mention [13]. In Computer Science, the use of provenance information to assess quality of Web data has been explored by Hartig and Zhao [12], who focus mostly on temporal qualities. More extensively, Zaveri et al. [19] provide a review on quality assessment for Linked Data. We also investigated the assessment of crowdsourced annotations using provenance analysis [6, 17]. These methods will be adopted also in this context.

Quality signals will be evaluated utilizing the CrowdTruth framework [14], and referring to the work of Bessi et al. [4], that shows how communities act as echo chambers.

3. FRAMEWORK OVERVIEW

The framework that we introduce aims at identifying document features that hint at quality estimation ("signals"), given the context and purpose with which the document is utilized by scholars. The framework is structured as follows.

Document Enrichment. Signals are searched within

document characteristics and their abstractions. We identify the following categories of signals: (a) *Natural Language Processing Signals*, i.e., the presence of entities and other NLP-based features; and (b) *Provenance Signals*, i.e., information about how data and enrichment came to be.

Quality Dimension Modeling. Given the informal definition of quality as "fitness for purpose", various purposes are possible in different contexts, as well as diverse subjects will measure fitness differently. We will make use of crowdsourcing, nichesourcing [9] and machine learning to identify quality dimensions that are relevant in different contexts.

Signal Detection. Once we identify a set of candidate features and a set of quality dimensions, we build machine learning models to identify the features that hint at quality.

Quality Estimation. Once we identify quality signals, we aim at automating this process using machine learning.

Currently, the framework makes use of NLP analyses from AlchemyAPI [1] and targets the set of quality dimensions described below. The framework code is available online[1].

4. PRELIMINARY EVALUATION

We ran a preliminary analysis on a corpus of 47 Web documents about vaccinations,[2] selected to guarantee diversity in sources (blogs, news, etc.) and manually assessed stance (pro, con, neutral). The documents were enriched with the following information categories from AlchemyAPI: Taxonomy, Relation, Entity, Concept, Sentiment. Every element of each category is a column in the feature matrix. The relevance of the item in the document is a number between zero and one. Using these features, documents are classified according to the qualities described below using Support Vector Machine [7], preceded by Stochastic Gradient Descent [5] ("SGD-SVM"), to handle the large number of features used. On each setting, we run 10-fold cross validation (i.e., the dataset is split in 10 subsets; in round, 9 subsets are used as training set, and 1 as test set; the accuracy is averaged).

Trustworthiness Estimation. We use Web Of Trust (WOT) [18] website trustworthiness as a proxy for document trustworthiness. WOT collects website trust assessments from their users, along with motivations for distrust. This information is made available via an API that exposes aggregations of the collected judgments per category (e.g., trustworthiness, child safety), together with an estimated confidence score. We are interested in the trustworthiness score of the websites (for simplicity, we do not consider their confidence level). This score ranges from 0 to 100. To make the number of classes manageable, we aggregated them into 10 superclasses (corresponding to the websites which trustworthiness is between 0 and 10, etc.), and we obtain an accuracy on 66.25%. With 3 superclasses (trustworthiness between 0 and 33, etc.) accuracy is 70.33%.

Stance and Objectivity Estimation. Stance (e.g., neutral, pro, con a given topic) allows discriminating the most neutral Web documents from the most biased ones.

A Spearman rank-correlation test shows a weak correlation (0.4) between sentiment and stance. In fact almost all articles related to vaccinations show a negative sentiment, although for different reasons. Hence, we need to expand the search for features, and for algorithms to handle them.

We get an accuracy of 63.33% when estimating stance based on sentiment. We also use stance as a proxy for objectivity: "pro" or "con" (vaccinations) documents are considered less objective than the "neutral" documents. Accuracy is 90.67% when predicting objectivity based on sentiment. When considering all the features resulting from enrichment, the accuracy is 70% for stance and 90.67% for objectivity.

Basic Scholarly Quality Estimation. We created a basic (simplified) version of the scholarly quality assessment checklist by the American Library Association (ALA) [2]: (1) Is the author qualified for the topic discussed? (2) Is the reputation of the publisher good? (3) Is the source unbiased? (4) Does the document show a bibliography? (5) Is it peer reviewed or edited? The resulting score ranges between 0 (low) and 5 (high). This questionnaire focuses on quantifiable items from the ALA checklist. Provenance-related items require further research to provide a quantitative result. Prediction accuracy for this score is 35%. Scores are treated as categories, so accuracy does not account for the prediction error. Improvements are discussed below.

5. DISCUSSION

We introduce a framework and a baseline for the evaluation of quality assessment of Web documents, consisting of trustworthiness, objectivity, and basic scholarly quality. Documents enrichment is based on NLP analyses from AlchemyAPI. The highest performance is achieved when estimating manually assessed objectivity. We will address possible inter-rater disagreement in the future.

Accuracy of trustworthiness prediction is about 70%. This result provides a decent starting point for future explorations. We use only the WOT trustworthiness scores, but we plan to take into consideration also their confidence in the future. We will also extend these assessments with others from similar services (e.g., Google Safe Browsing [11]).

The lowest performance is achieved when trying to predict the basic quality score. Hence, we will expand the feature set, and evaluate different automated reasoning methods. The checklist that we describe in Section 4 is meant to provide a quantitative quality score, but different users could provide different answers to those questions. For example, is the extensive use of links in a Web page to be considered equivalent to the use of a traditional bibliography? We will gather assessments from a number of scholars to help clarifying these issues. Following the CrowdTruth practices [14], we will identify and decompose controversial features. For example, if scholars disagree on the "bibliography" feature, we might evaluate also its type and length.

In the basic scholarly quality, all the questions weigh the same, but some could be more important than others. We aim at learning these weights from a set of assessments provided by scholars in different fields. This requires treating the scores as ordinal values (and no more as categories), and employing adequate machine learning algorithms.

In the future, we will extend the feature set by including, for instance, other NLP analyses and author reputation. Also, a series of crowdsourcing tasks will tell us which quality dimensions are more relevant in different contexts.

6. ACKNOWLEDGMENTS

This work was supported by the Amsterdam Academic Alliance Data Science (AAA-DS) Program Award to the UvA and VU Universities.

[1]The code is retrievable at https://goo.gl/GrXpWj.
[2]Available at https://goo.gl/sbVJQq.

7. REFERENCES

[1] AlchemyAPI, Inc. Alchemyapi. http://www.alchemyapi.com, 2015.

[2] American Library Association. Evaluating information: A basic checklist. Technical report, American Library Association, 1994.

[3] M. Anderka, B. Stein, and N. Lipka. Predicting quality flaws in user-generated content: The case of wikipedia. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 981–990, New York, NY, USA, 2012. ACM.

[4] A. Bessi, M. Coletto, G. Davidescu, A. Scala, G. Caldarelli, and W. Quattrociocchi. Science vs conspiracy: Collective narratives in the age of misinformation. volume 2, 2015.

[5] L. Bottou. Stochastic learning. In O. Bousquet and U. von Luxburg, editors, *Advanced Lectures on Machine Learning*, Lecture Notes in Artificial Intelligence, LNAI 3176, pages 146–168. Springer Verlag, Berlin, 2004.

[6] D. Ceolin, P. Groth, V. Maccatrozzo, W. Fokkink, W. R. van Hage, and A. Nottamkandath. Combining user reputation and provenance analysis for trust assessment.

[7] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297.

[8] D. H. Dalip, M. A. Gonçalves, M. Cristo, and P. Calado. Automatic assessment of document quality in web collaborative digital libraries. *Journal Data and Information Quality*, 2(3):14:1–14:30, Dec. 2011.

[9] V. de Boer, M. Hildebrand, L. Aroyo, P. Leenheer, C. Dijkshoorn, B. Tesfa, and G. Schreiber. *Knowledge Engineering and Knowledge Management: 18th International Conference, EKAW 2012, Galway City, Ireland, October 8-12, 2012. Proceedings*, chapter Nichesourcing: Harnessing the Power of Crowds of Experts, pages 16–20. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[10] M. De Jong and P. Schellens. Toward a document evaluation methodology: What does research tell us about the validity and reliability of evaluation methods? 2000.

[11] Google, Inc. Safe Browsing – Transparency Report. https://www.google.com/transparencyreport/safebrowsing/.

[12] O. Hartig and J. Zhao. Using web data provenance for quality assessment. In *Proceedings of the Intenational Workshop on Semantic Web and Provenance Management*, 2009.

[13] M. Howell and W. Prevenier. *From Reliable Sources: An Introduction to Historical Methods*. Cornell University Press, 2001.

[14] O. Inel, K. Khamkham, T. Cristea, A. Dumitrache, A. Rutjes, J. Ploeg, L. Romaszko, L. Aroyo, and R.-J. Sips. *The Semantic Web – ISWC 2014: 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part II*, chapter CrowdTruth: Machine-Human Computation Framework for Harnessing Disagreement in Gathering Annotated Data, pages 486–504. Springer International Publishing, Cham, 2014.

[15] International Organization for Standardization. ISO/IEC 25012:2008 Software engineering – Software product Quality Requirements and Evaluation (SQuaRE) – Data quality model. Technical report, International Organization for Standardization, 2008.

[16] Y. W. Lee, D. M. Strong, B. K. Kahn, and R. Y. Wang. Aimq: A methodology for information quality assessment. *Inf. Manage.*, 40(2):133–146, Dec. 2002.

[17] A. Nottamkandath, J. Oosterman, D. Ceolin, G. K. D. de Vries, and W. Fokkink. Predicting quality of crowdsourced annotations using graph kernels. In *Trust Management IX*, pages 134–148. Springer International Publishing.

[18] WOT Services, Ltd. http://www.mywot.com, 2006.

[19] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer. Quality assessment for linked data: A survey. *Semantic Web Journal*, 2015.

[20] H. Zhu, Y. Ma, and G. Su. Collaboratively assessing information quality on the web. In *ICIS sigIQ Workshop*, 2011.