Identifying and Classifying Uncertainty Layers in Web Document Quality Assessment

Davide Ceolin[1] and Lora Aroyo[1] and Julia Noordegraaf[2]

[1] {d.ceolin,lora.aroyo}@vu.nl
VU University Amsterdam
Amsterdam, The Netherlands
[2] j.j.noordegraaf@uva.nl
University of Amsterdam
Amsterdam, The Netherlands

Abstract. Assessing the quality of Web documents is crucial, but challenging. In this paper, we outline the different uncertainty bottlenecks that such task implies, and we propose a strategy to tackle them.

1 Introduction

Assessing the quality of Web documents is a necessary, yet challenging issue. For example, if a journalist is writing an article on the vaccination debate, and is looking for Web documents to use as a source. What would her definition of quality encompass? Given that she wants to represent a debate, she needs documents that properly represent each point of view, i.e. they are complete, accurate, precise and reliable documents with a clear provenance. With the proliferation of information on the Web, the potential set of documents she may be confronted with is so vast that it is necessary to make a selection of documents with the highest quality, seen from the perspective of journalistic usage.

As this example shows, the prime source of uncertainty is the fact that the definition of quality depends on the user's perspective on the data. Suppose that this definition comprises the quality dimensions mentioned before: completeness, accuracy, precision, and trustworthiness. On the one hand, we need to understand how these quality dimensions have to be combined together to come up with a final decision about the overall quality of a document (i.e., to decide if the journalist is going to use the document or not). On the other hand, in order to meet the Web scale of the set of documents the user is presented with, we need to understand how to automatically evaluate and quantify these qualities: what is the information we need to extract from the documents to make such quantification? And how can this information be extracted?

Given the complexity of defining Web document quality, it would be useful to accompany estimated quality assessments obtained by automatic predictions with quantification of their confidence. We could always come up with a decision about the quality of a document, but we may be unsure about the accuracy of such decision. To address such a bottleneck, in this paper we propose to identify the possible sources of uncertainty in the process of quality estimation of Web documents, and we discuss an approach to quantify them.

The problem of assessing the quality of Web documents is crucial in information retrieval. Bharat et al [1] copyrighted a method for clustering online news content based on freshness and quality of content, while Kang and Kim [2] find links between specific quality requirements and user queries. We focus on detecting the uncertainty in such clusters and links. Pasi et al. [3] and Floridi and Illari [4] edited two extensive reviews on (Web) information quality and its philosophy. These reviews hint at the uncertainty issues in quality assessment.

The rest of the paper is structured as follows. Sections 2 and 3 introduce a quality assessment pipeline we devised and its sources of uncertainty. Section 4 presents a strategy for uncertainty handling, and Section 5 concludes.

2 Quality Assessment Pipeline

The pipeline for automating the process of quality estimation developed in previous work of ours [5] is depicted in Figure 1 and described below.

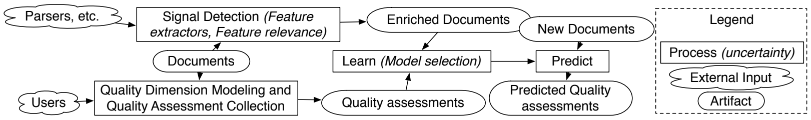
Parsers, etc.
Signal Detection *(Feature extractors, Feature relevance)*
Enriched Documents
New Documents
Legend
Documents
Learn *(Model selection)*
Predict
Process *(uncertainty)*
External Input
Users
Quality Dimension Modeling and Quality Assessment Collection
Quality assessments
Predicted Quality assessments
Artifact

Fig. 1. Overview of the Web Document Assessment pipeline

Signal Detection We automatically extract features from documents to be able to identify similarity and differences among them.

Running example. Consider the example in the previous section. Suppose that the journalist indicated the quality of a set of documents, and we intend to identify more documents she might consider of high quality. Since each document is different from each other, we need to extrapolate information that allows comparing them. If we extract the sentiment and the entities mentioned, we can compare a blog post and a news article on those terms (Do the documents share the same sentiment? Do they mention the same entities?) and, if any correlation with quality is present, these features will be used for quality prediction, as described below.

Quality Dimension Modeling and Assessment Collection Since the definition of quality depends on the specific user, context, and task, we collect assessments to use as a training set.

Running example. Here we record the document quality assessments provided by the journalist (along with her identity, the context and the ask at hand).

Quality Estimation Once that features are extracted and sample quality assessments are collected, we identify correlations and correspondences between these elements. Typically, we employ machine learning algorithms.

Running Example. An automated learning algorithm (e.g., SVM) is used to associate the quality assessments of the journalist in the training set to the document features, to predict the quality of other Web documents.

3 Sources of Uncertainty

Feature Extractors Tools for document feature extraction may produce disagreeing results. This adds additional uncertainty to the process.
Running example. Suppose that we parse the same document with two different NLP parsers, e.g. P_1 and P_2: the resulting sentiment differs of 0.2 on a range from -1 (negative) to 1 (positive), and the sets of entities extracted are different. How shall we handle such discrepancies? How shall we evaluate the tool reliability? Several possibilities apply here.

Feature Relevance These features are collected because they could (jointly) act as quality markers. In principle, the more attributes we collect, the more potential markers we gather. The quality of different types of documents (e.g., newspaper articles, blog posts) could be marked by different features, and a feature that does not mark quality in the documents observed up to a given time could mark quality in the next document collected. However, features could: (1) conflict with each other; and (2) create scalability issues due to dimensionality growth. It is difficult to prune these features, because we do not know which of these might become relevant in the future.
Running example. We collected a sample of assessments, and we use it to make quality predictions. Yet, we do not know if the correspondencies between assessments in the training set and document features we may find are valid also on other documents (and whether those document features that seem useless at the moment might be useful in the future).

Model Selection Correlations and correspondences between features and qualities can be identified by means of diverse algorithms. For similar reasons that hold for the uncertainty linked to the feature relevance, the choice of these algorithms is difficult. They could perform well on a dataset at hand, but not on its extension. Since we aim at allowing quality prediction on large sets of Web documents, we need to carefully choose the learning algorithm.
Running example. Suppose that Support Vector Machines performs well on the training set at hand. We need to seek guarantees on the fact that the performance keeps stable as long as we extend the dataset. E.g., by monitoring the performance and by evaluating alternative approaches in parallel.

4 Uncertainty Handling Strategy

We identify the following strategy based on Semantic Web technologies to address the uncertainty of Web document quality estimations.

Trace the Provenance of Quality Estimates Tracing the provenance of the estimations we make is crucial to investigate the reasons for high or low accuracy, and improve them. We can use PROV [6] to this aim, and by specializing it further, we may be able to better describe the peculiarities of uncertainty bottlenecks we may find.

Reason on and Annotate Provenance Traces Once we identified all the steps that led to a given quality estimate, we can estimate the confidence in the estimate by looking at the provenance. In particular, by collecting a large enough set of provenance traces, and of measurements of the estimation accuracy, we can identify which processes entities used lead to higher uncertainty. To properly trace the quality of these assessments, we can make use of the Data Quality Vocabulary (DQV) [7].

Running example. We extract sentiment and entities from the documents selected for the journalist. We use a Support Vector Machine model to predict the quality of the documents. Once we make the prediction, we can measure its accuracy, and associate it to the current trace. We can then measure the accuracy also with other algorithms (e.g., Bayesian Networks) and input features (e.g., source trustworthiness). By keeping track of the provenance of the estimates, we can infer which parts of the process constitute an uncertainty bottleneck.

5 Discussion

In this position paper, we discuss the possible sources of uncertainty in the process of automated estimation of the quality of Web documents, and we illustrate them by means of a running example. We propose a general strategy for quantifying such uncertainty, so to measure the confidence in quality estimates. This procedure relies on the Semantic Web techniques (in particular, PROV and DQV) to trace all the steps that led to the estimates, and to learn how these correlate with uncertainty, to detect possible bottlenecks in the process.

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

References

1. Bharat, K., Curtiss, M., Schmitt, M.: Method and apparatus for clustering news online content based on content freshness and quality of content source (2016) US Patent 9,361,369.
2. Kang, I.H., Kim, G.: Query type classification for web document retrieval. In: SIGIR '03, ACM (2003) 64–71
3. Pasi, G., Bordogna, G., Jain, L.C., eds.: Quality Issues in the Management of Web Information. Springer (2013)
4. Floridi, L., Illari, P., eds.: The Philosophy of Information Quality. Springer (2014)
5. Ceolin, D., Noordegraaf, J., Aroyo, L., van Son, C.: Towards web documents quality assessment for digital humanities scholars. In: WebSci '16, ACM (2016) 315–317
6. W3C: PROV-O. http://www.w3.org/TR/prov-o/ (2013)
7. W3C: Data quality vocabulary. https://www.w3.org/TR/vocab-dqv/ (2015)