Capturing the Ineffable: Collecting, Analysing and Automating Web Document Quality Assessments

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

[1] {d.ceolin,lora.aroyo}@vu.nl
VU University Amsterdam
de Boelelaan 1081a
1081HV
Amsterdam, The Netherlands
[2] j.j.noordegraaf@uva.nl
University of Amsterdam

Abstract. Automatic estimation of the quality of Web documents is a challenging task, especially because the definition of quality heavily depends on the individuals who define it, on the context where it applies, and on the tasks at hand. In this paper, we investigate the quality of Web documents from two perspectives. Firstly, we propose a method for capturing Web quality assessments based on a nichesourcing Web application that we developed. Secondly, we investigate the characteristics of Web documents that hint at their quality levels. We propose a model for automatically estimating such assessments, we analyze users ability to judge quality based on features extracted from these documents and, finally, we decompose overall quality assessments so to identify which quality dimensions (e.g., accuracy, precision) are of higher importance when assessing Web documents. We evaluate our contributions on two use cases involving experts (journalists and media scholars) that make professional use of Web documents. Our results show that: (1) it is possible to automate the process of Web document quality estimation to a level of high accuracy; (2) that document features shown in isolation are poorly informative to users; and (3) that, related to the tasks we propose (i.e., choosing Web documents to use as a source for writing an article on the vaccination debate), the most important quality dimensions are accuracy, trustworthiness, and precision.

1 Introduction

Automatically estimating the quality of Web documents is a compelling, yet intricate issue. It is compelling because the huge amount of Web documents we can access makes their manual evaluation a costly operation. So, to guarantee we access the best documents available on the Web on a given matter (or, at least, that the documents we observe meet some minimum qualitative standards), an automated assessment is needed. However, quality is a rather inflated term, that

assumes different meanings in different contexts and with different subjects. This is the reason why we call them "ineffable".

This paper aims at investigating how it is possible to capture such ineffable judgments and at understanding how they are characterized. In particular, our current focus is on the quality assessment of Web documents to be used for professional use (i.e., by journalists and media scholars). In the first part of the paper, we describe a nichesourcing application for collecting Web document quality assessments (WebQ[3]) that we developed. Then, we describe a thorough set of analyses we performed on the quality assessments gathered through WebQ by means of two user studies. From these analyses, we derive that quality assessments heavily depend on the task at hand. The subjectivity of such assessments is overcome by the fact that users share a common background and assess documents with the same task in mind. Hence, besides the individual peculiarities in the assessments, we can aggregate user contributions and actually automate their learning. This opens up for the possibility of extending the set of assessments to be used for training in the future.

The rest of the paper is structured as follows. Section 2 introduces related work. Section 3 describes the nichesourcing Web application we developed for collecting quality assessments, WebQ. Section 4 describes the two case studies we performed, along with the results collected. These results are discussed in Section 5. Section 6 concludes the paper.

2 Related Work

The problem of assessing the quality of Web documents and, more in general, (Web) data and information, is a compelling one and has been tackled in diverse contexts.

The ISO 25010 Model [9] is a standard model for data quality. From this model, we select those data quality dimensions from this model that apply also to Web documents (e.g., precision, accuracy) and ask the users of WebQ to rate Web documents on them. This set of quality dimensions has been extended to include other measures tailored to Web documents, like neutrality and readability.

The problem of identifying the documents of higher quality for a given purpose is common in information retrieval. Bharat et al [2] copyrighted a method for clustering online news content based on freshness and quality of content. Clearly, their approach differs from ours as they focus on news, and they aim at clustering documents. However, one of the key features for determining the quality of documents is the (estimated) authoritativeness of the source, both in their and in our approach. Kang and Kim [10], instead, find links between specific quality requirements and user queries. We do not make use of queries since we preselect documents and also predefine the task the users are asked to perform. However, we still analyze user assessments to derive their specific definition of

[3] The tool is running at `http://webq3.herokuapp.com`, the code is available at `https://github.com/davideceolin/webq`.

quality, and might consider analyzing user queries in the future, when we will expand the dataset and tasks at hand.

Following up on the use of specific metadata as markers for quality, Amento et al. [1] use link-based metrics to make quality predictions, showing that these perform as good as content-based ones. In our case, we focus mainly on metadata- and context-based features, and will consider link-based ones in the future.

Regarding the use of niche- or crowdsourcing for collecting information and, in particular, quality assessments, Lee et al. [11] provide a framework tailored to organizations. Zhu et al. [14] 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 collecting are based on specific tasks, while they rely on contributions via browsers plugins. Currently, we focus on niches for collecting quality assessments. In the future, we will make use of crowdsourcing as well, embracing methods for extracting ground truth like the CrowdTruth framework [8].

While this paper proposes a framework that aims at generically identifying markers for quality of Web documents, we evaluate such framework with an emphasis on digital humanities applications. Digital Humanities scholars are professionals that are used to critically evaluate the sources they deal with, hence we target this specific class of users to investigate how to extend source criticism practices to cover Web documents as well. Source criticism is the process of evaluating traditional information sources that is common in the (digital) humanities. De Jong and Schellers [5] 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. This paper extends the work we presented at the Web Science conference, where we began the exploration of how it is possible to assess the quality of Web documents, especially for the Digital Humanities [4].

Lastly, one aspect that we partially consider when estimating the quality of Web document is their provenance. Provenance analysis is used to assess the quality of humanities sources, as Howell and Prevenier mention [7]. In Computer Science, the use of provenance information to assess the quality of Web data has been explored by Hartig and Zhao [6], who focus mostly on temporal qualities. More extensively, Zaveri et al. [13] provide a review on quality assessment for Linked Data. We also investigated the assessment of crowdsourced annotations using provenance analysis [3,12].

3 Nichesourcing Web Document Quality Assessments

We developed WebQ, a tool for investigating and collecting judgments about Web documents. WebQ aims at shedding a light on three main aspects:

– Understanding whether (professional) users are able to estimate the quality of Web documents based on limited sets of features of these documents (e.g., the sentiment of these documents, or the list of entities extracted from them);

– Understanding whether user judgments are coherent enough over multiple documents and among diverse assessors, so to allow their automated learning;
– Understanding how can overall quality assessments about Web documents be explained in terms of specific quality dimensions (i.e., understanding what is the weight of precision, accuracy, etc. when users provide overarching quality assessments about Web documents), when focusing on specific tasks.

3.1 Document Features and Document Quality Dimensions

We carachterize documents by means of features we automatically extract about them. In Section 4 we analyze the existence of correlations between these automatically extracted features and the nichesourced qualities.

Document Features These are a series of attributes we automatically extract by means of Web APIs. These features aim at identifying commonalities among documents, opening up for the possibility of predicting their qualities (provided that features and qualities correlate). These features are:

Entities, Sentiment, Emotions We use AlchemyAPI to extract all the features mentioned in the documents, along with their relevance to the document. Also, AlchemyAPI provides us with a quantification of the sentiment expressed by the document (positive or negative, and its strength), and its emotions (joy, fear, sadness, disgust and anger, along with their strength).

Trustworthiness This is an anomalous feature, because trustworthiness belongs also to the document qualities. In this case, we use the Web Of Trust API to obtain crowdsourced trustworthiness assessments about the source publishing the article.

Document Quality Dimensions These are a series of abstractions of these documents qualifying the information therein contained, namely:

Overall Quality This score represents the overall reliability of the document.

Accuracy quantifies the level of truthfulness of the document information.

Precision quantifies whether the document information is precise.

Completeness determines whether the information contained in the document (or facts) are complete.

Neutrality determines whether a particular stance (e.g., pro or con a given topic) is represented in the document.

Readability quantifies whether the document reads well.

Trustworthiness quantifies the perceived level of trustworthiness of the information in the document. Note that the Web Of Trust score refers to the source, while this quality refers to the specific document evaluated.

3.2 Structure of WebQ

Below we describe the structure of WebQ, illustrated in Figure 1.

Architecture The application is developed based on the Flask Python library [4]. As backend storage for Web document assessments, we use MongoDB [5].

Annotations We use AnnotatorJs[6] as a tool for allowing our users to indicate which specific parts of a document mark particular qualities of the whole document. AnnotatorJs is a javascript library ran on the client side. This library records the document annotations by sending HTTP messages to a storage server. We adapted to this purpose the Annotation Store[7], which relies on ElasticSearch for storage and retrieval of annotations.

HTTP Proxy We developed an HTTP proxy to provide to the users with the Web documents to be annotated within WebQ. This proxy allows to present the documents within our application and allowing users to annotate them by enabling AnnotatorJs. In this manner, the users see the exact same document they would see on the Web, but they are able to annotate it, remaining in the context of our application. This proxy is tailored to the documents we identified in our dataset and is able to render them at their best. In particular, it addresses the following issues:

- replace relative paths with absolute ones in image, CSS and link addresses (so that the page can refer to the absolute addresses of the accessory files);
- handle charsets: to properly render the documents, their charset has to be correctly detected and utilized;
- forward the browser headers. Some websites allow being accessed only via (some) browsers, and not being scraped. Our proxy accesses them programmatically, but on behalf of a browser access. Therefore, we forward the browser headers to the URL via the proxy;

In the future, we aim at extending our dataset, so we will work on further extensions of this proxy.

Randomizer WebQ is designed for collecting Web document quality assessments via one or more user studies. In such a scenario, users would access the application more or less simultaneously. We assign to each user a random sequence of documents to assess (we set the length of such sequence to six), but at the same time we also need to guarantee that the dataset is uniformly assessed: documents should get approximatively

$$n_{assessments} = |dataset| \text{ div } |users|$$

assessments, where $|dataset|$ is the cardinality of the document dataset (50), div is the integer division and $|users|$ is the cardinality of the set of users.

[4] `http://flask.pocoo.org/`
[5] `http://mongodb.com`
[6] `http://annotatorjs.org`
[7] `https://github.com/openannotation/annotator-store`

Before running the user studies, we generate $n_{assessments}$ random permutations of documents. We split such sequences in consecutive sequences of 6 documents. Sequences are then uniquely assigned to users as soon as they register in the application.

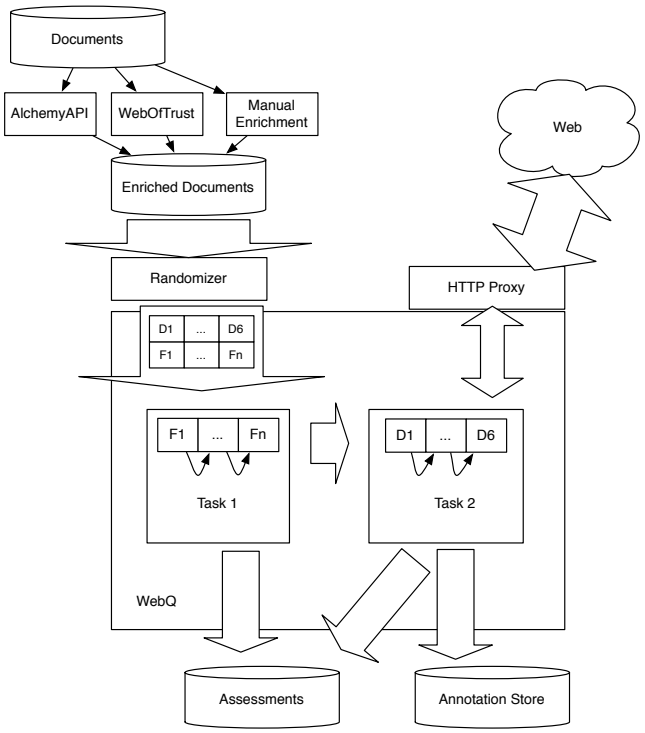

Fig. 1. Overview of the WebQ application. The document set is firstly enriched by using AlchemyApi, Web of Trust and manual enrichment. Then, a random selection of six documents is presented to the users for the first task: selecting the documents with the highest quality on the basis of the value of one document feature. After all the features (sentiment, entities, etc.) have been evaluated, then users are asked to assess each of the six documents assigned (task 2). In this case, documents are rendered through an HTTP proxy, to allow annotating them and visualizing the documents within the app.

3.3 Tasks Description

In WebQ we ask the users to perform two tasks. The first task aims at exploring whether document features could be used as quality indicators alone. The second task aims at collecting assessments about the documents presented. The two tasks are described as follows.

Task 1 The application is divided into two tasks. In the first, task we evaluate the informativeness of specific document features in relation to the quality of the documents. The task is structured as follows:

1. We assign to each user a set of six documents from our overall dataset.
2. We identify six classes of potentially useful features about the documents, namely: the document's sentiment and emotions expressed, its trustworthiness, its title, its source and the list of entities we extract from it.
3. In turn, we show the values for each of these features to the user, i.e., first we present the user with the six lists of entities extracted from the six documents (one list per document), then we present the user with the sentiment and the emotions detected in each document, and so on. Each time only one feature is presented. Note that the selection of documents is fixed per user. However, the user does not know the documents, they only know the values of the features we present. Also, every time we present features to the users we shuffle the document order and we use different identifiers.
4. We ask the user to select which documents among these six she will use as a source for her article, based on the information displayed.
5. Lastly, we ask the user to make the same selection on the basis of all the features presented together.

Task 2 We ask the user to assess the quality of each article in depth. Based on the same selection of six articles the user was assigned to in task 1, they:

1. Read the article
2. Assess the overall quality of the article, as well as the following quality dimensions: accuracy, precision, completeness, readability, neutrality, trustworthiness. Assessments are indicated in a 1 to 5 Likert scale.
3. Highlight in the article the words or sentences that motivate their assessments, tagging each selection with the name of the corresponding quality dimension and with an indication of the fact that the selection represents a positive or negative observation.
4. Revise their quality assessments (step 2.) if they wish so.

4 Case Studies

In this section, we describe the two case studies we run. Both case studies are based on the same set of documents, which we describe as follows.

4.1 Dataset and Scenario

The dataset we base our experiments on is composed by Web documents about the vaccination debate that has been triggered by the measles outbreak happened at Disneyland, California, in 2015[8]. This dataset contains 50 documents, which are diversified in terms of:

[8] The dataset is available at `https://goo.gl/cLDTtS`

Stance Some are pro vaccinations, some con, some neutral.
Type of source We include in our dataset diverse types of sources, comprising official reports from authorities, news articles, editorial articles, blog posts.

The scenario we hypothesize is that users have to write an article about the vaccination debate triggered by such measles outbreak. We propose diverse types of Web documents to the users, and we ask to select those they would use as a source for their article (i.e., those they consider of a higher quality).

4.2 Case Study 1 - Journalism Students

Experimental Setup The first case study involved a class of 20 last-year journalism students from the University of Amsterdam. The students performed both tasks of WebQ in a time frame that lasted between 45 and 60 minutes.

Results We present here a series of analyses on the results collected.

Document Assessments Collected In this first experiment we collected 104 complete assessments (i.e., quality judgments about the diverse quality dimensions of the documents) and 238 annotations.

Comparison of the two document assessments in task 2 In task 2 we ask the users to assess the documents two times: the first time they read the documents, and after having highlighted the motivations for their assessments. We observe no significant difference between these two assessments, using a Wilcoxon Signed-rank test at 95% confidence level.

Document Assessments Predictability The first analysis we perform regards the predictability of Web documents assessments. Only two or three assessments are provided per document, but if users assess the documents coherently enough (i.e., following similar policies), and if the features we extracted (entities, sentiment, emotions, trustworthiness) are considered by the users' policies, then we might be able to automatically learn such predictions. Table 1 shows the results of such predictions using the Support Vector Classification algorithm.

Correlation between quality dimensions and overall quality We checked the correlation between the overall quality score and the scores given to each quality dimensions. Results are reported in Table 2.

Correlation between document selection (task 1) and document assessments (task2) In task 1 we ask the users to select documents they think are of high quality based on diverse features we extracted from these documents. If many users select a document, we might derive that it has high probability to be of high quality. Since each document has been proposed to only either two or three users, we compute such probability using a smoothing factor that allows to account for the uncertainty due to the small samples observed (see Equation (1)). Also,

Table 1. Results of 10-fold cross-validation using Support Vector Classification with different number of features, and predicting either 5 classes (as in the 1-5 Likert scale used in WebQ) or 2 classes (i.e., high- and low-quality documents). We calculated the performance for all possible permutations of the four classes of features. For each cardinality of such permutation (1,2,3,4) we show the best performing combination.

Features used	**SVC 5 classes**	**SVC 2 classes**
trustworthiness	48%	75%
sentiment, trustworthiness	46%	78%
sentiment, emotions, trustworthiness	38%	72%
sentiment, emotions, trustworthiness, entities	39%	72%

Table 2. Correlation between each quality dimension and the overall quality score attributed to the documents.

Quality dimension	**Correlation with Overall Quality**
Accuracy	0.89
Completeness	0.69
Neutrality	0.46
Relevance	0.63
Trustworthiness	0.80
Readability	0.67
Precision	0.77

smoothing allows us to treat differently documents that have been proposed two or three times: if a document has never been selected when it has been proposed two times, its probability to be of high quality is 0.25; if it has been proposed three times, 0.2. This probability is equivalent to the expected value of a Beta probability distribution with a non-informative prior (as we do not know a priori which documents are of higher quality).

$$P = \frac{\#selection + 1}{\#samples + 2} \tag{1}$$

In task 2, users assess these same documents. Table 3 reports the correlation between the probability from task 1 and the overall quality score from task 2.

User Evaluation We asked the users express their opinion about the experiment by means of a questionnaire. The quantitative results of the 13 respondent (52% of the total) are reported in Table 4. To those results, we can add that trustworthiness, accuracy and the source of the document are the features mostly indicated by the users as good quality markers.

Quality Definition and Qualitative Analysis of Annotations and Remarks Lastly, from a qualitative evaluation of the annotations and of the remarks collected, we can derive that users assume that *the documents of higher quality are those showing the following qualities: high trustworthiness, high accuracy and high precision.*

Table 3. Correlation between the probability of documents to be selected in task 1 and their overall quality assessment from task 2.

Feature shown to the users (task 1)	**Correlation (Spearman) with Overall Quality (task 2)**
Entities	-0.07
Sentiment	0.09
Trustworthiness	0.20
Sources	0.29
Title	-0.07
All	0.20

Table 4. Results of the user evaluation questionnaire.

Question	**Result**				
	1 (bad)	**2**	**3**	**4**	**5 (good)**
How was your experience?	**30.8%**	**30.8%**	23.1%	15.4%	0%
Was the experience easy?	15.4%	**53.8%**	23.1%	7.7%	0%
Does this experiment resemble the process that you would follow when you write an article?	**38.5%**	30.8%	23.1%	7.7%	0%
Do you think that the entities extracted from the Web documents are good quality markers?	15.4%	15.4%	**30.8%**	**30.8%**	0%

4.3 Case Study 2 - Media Scholars

Experimental Setup The second case study involves 20 media scholars attending the Research School for Media Studies (RMeS) summer school. We present the users with a setting that closely resembles the one adopted for Case Study 1 but, following the indications received from the user evaluation, we adopt the following improvements:

- we improve the user introduction and add a walk-through session to guide the users in the application;
- task descriptions and user experience are improved (e.g., landing pages).

The users had about 45 minutes at their disposal to complete the two tasks.

Results We present the results obtained and their analyses following the same structure adopted for case study 1.

Document Assessments Collected In this experiment we collected 47 complete assessments about the documents in our dataset and 89 annotations.

Comparison of the two document assessments in task 2 Also in this case we observe no significance difference between the first and the second series of assessments, for any quality dimension.

Document Assessments Predictability Like with the previous case study, we use 10-fold cross-validation to test the predictability performance of Support Vector Classifier on the overall quality assessment. Results are reported in Table 5.

Table 5. Accuracy of the prediction of the overall quality assessments on the second case study. We compute all features permutations and we show the best performing combination per feature set cardinality (1,2,3,4).

Features used	**SVC 5 classes**	**SVC 2 classes**
trustworthiness	63%	89%
sentiment, trustworthiness	53%	86%
sentiment, entities, trustworthiness	34%	85%
sentiment, entities, trustworthiness, emotions	34%	85%

Correlation between quality dimensions and overall quality We checked the correlation between the overall quality score and the scores given to each quality dimensions. Results are reported in Table 6.

Table 6. Correlation between each quality dimension and the overall quality score attributed to the documents.

Quality dimension	**Correlation with Overall Quality**
Accuracy	0.89
Completeness	0.69
Neutrality	0.45
Relevance	0.64
Trustworthiness	0.78
Readability	0.66
Precision	0.76

Correlation between document selection (task 1) and document assessments (task2) We computed the probability of documents to be of high quality based on the number of selections they received in task 1, as indicated by Equation (1). Table 7 reports the correlation between such probability and the overall quality score from task 2.

User Evaluation The results of the user evaluation questionnaire are reported in Table 8. To these quantitative results, we add the fact that users indicate accuracy and also indicators from social media (e.g., discussion on the topic, likes) as possible quality markers. The results described here are based on a very little sample, since only four participants responded to the questionnaire.

Table 7. Correlation between the probability of documents to be selected in task 1 and their overall quality assessment from task 2.

Feature shown to the users (task 1)	**Correlation (Spearman) with Overall Quality (task 2)**
Entities	0.38
Sentiment	0.19
Trustworthiness	0.21
Sources	0.25
Title	0.15
All	0.24

Table 8. Results of the user evaluation questionnaire.

Question	**Result**				
	1 (bad)	**2**	**3**	**4**	**5 (good)**
How was your experience?	**25%**	0%	**75%**	0%	0%
Was the experience easy?	0%	**50%**	**50%**	0%	0%
Does this experiment resemble the process that you would follow when you write an article?	**50%**	0%	25%	25%	0%
Do you think that the entities extracted from the Web documents are good quality markers?	**50%**	25%	0%	25%	0%

Quality Definition and Qualitative Analysis of Annotations and Remarks Lastly, from a qualitative evaluation of the annotations and of the remarks collected, we can derive that users assume that *the documents of higher quality are those showing the following qualities: high trustworthiness (e.g., expressed by the author's authoritativeness), high accuracy and high precision.*

4.4 Comparison between Case Study 1 and 2

We compare the results obtained in case study 1 and 2. First of all, we use a Wilcoxon signed-rank test to compare the performance obtained by support vector machines (Tables 1 and 5). We observe no significant difference neither with 2 nor with 5 classes.

Also comparing the correlations between the quality dimensions and the overall quality (Tables 2 and 6), we observe no significant difference.

Neither the results of Tables 3 and 7, i.e., the correlation between probabilities of a document to be selected and its quality show any significant difference between task 1 and 2.

The second user questionnaire has been fulfilled only by a very limited number of users. A Wilcoxon signed rank test and a χ^2 test both agree that the results from the two case studies are not significantly different, but the sample sizes are so small that we can hardly rely on these results.

5 Discussion

We discuss here the results presented in Section 4. We summarize the discussion by means of a series of statements that emerge from the analysis of the result. Each statement is presented and motivated below.

User assessments are stable and coherent. In both case studies, we observe that the first and the second document assessments are not significantly different. Moreover, in both cases, we can use Support Vector Classifier to automatically learn and predict the quality of documents. This means that, even if users assess different documents (the same document has been assessed by three users at most), assessments are coherent enough to be learned. Also, the features we identified (entities, sentiment, emotions, trustworthiness) correlate with these judgments enough to allow using them as features for prediction.

User assessments are highly related to the task at hand. The extremely high similarity between the results in Tables 2 and 7 shows that, when assessing the quality of documents, the task at hand is the most important factor. In fact, here the users were asked to pretend they were writing an article about the vaccination debate. Consequently, they focused on identifying the most accurate and trustworthy documents. Neutrality is the least significant quality of these documents because, if users want to represent the whole spectrum of the debate, they have to consider also the least neutral documents, provided that they are accurate enough. Different tasks could (and would probably) imply different quality requirements.

This opens up for the possibility to extend our training set with assessments from future experiments. In fact, the fact that the task is the most important aspect in such assessments makes new assessments datasets from future experiments assimilable to the existing one. In this manner, we will be able to scale up our current approach at Web scale. In this light, although in some cases we observe that by considering only a subset of features we obtain a better performance (up to +6% in some cases), we still prefer to consider all the features we collected so far. In fact, we do not know if, by extending the set of documents considered (or by diversifying the tasks at hand), some of the features that are now less significant could become more prominent.

Features in isolation are hardly meaningful (but the user experience plays a role here). Entities, sentiment, and emotions, trustworthiness, title and source are hardly useful to be used to decide if a document is of high quality or not. The fact that these features are profitably used to learn the quality assessments of the documents using SVC means that they are good markers of quality (e.g., the fact that a given document expresses an extremely positive sentiment or show specific entities is correlated with its quality). Nevertheless, users are hardly able to determine the document quality on the basis of a quantification of such features. True is that in the second case study, although the performance is still pretty low, the results are slightly better than those of the

first use case. This might be due to the different user background, but we believe that also the changes we added to WebQ facilitated such improvement.

The application setup should take (also) the user experience into consideration We aim at collecting annotations from users, so we need to balance a couple of trade-offs between the application (and, hence, our) requirements and user-based constraints. First of all, our target users have a professional background that is not necessarily an Information or Computer Science background. So, even if the application is able to capture all the necessary information, the way its functionality is presented and the user is guided plays an important role. In fact, we observed (both via the questionnaire and via a post-study discussion) an improvement in the experience perception from case study 1 to 2. Moreover, our goal is to collect as many assessments as possible, but we have to take into account also that user attention decrease over time. This is the reason why in the second case study users paid more attention to the tasks performed, so provided much fewer assessments than in case study 1.

6 Conclusion

Automatically assessing the quality of Web documents is a crucial task to be able to benefit from the vast amount of information we can access online. In this paper, we investigated how we can collect quality assessments by means of a nichesourcing application we developed. We design and evaluate the application by means of two case studies involving journalists and media scholars. Such application provides all the necessary functionalities to collect such assessments (e.g., the possibility to rate and annotate documents) and the evaluations collected allowed us fine tuning it. Also, by analyzing the assessments collected in detail, we discovered that having a clearly defined task at hand we can overcome the subjectivity in document assessment, thus allowing assessments to be automatically estimated. In the specific task performed (selecting documents to be used as a source for an article on the vaccination debate), the most important quality dimensions considered are accuracy, precision and trustworthiness. We also show that the results collected in the two case studies are assimilable and allow creating a uniform collection of document assessments.

We plan to extend our application in several directions. We aim at considering other typologies of professional users, and at extending the tasks evaluated. Clearly, we intend to extend also the dataset of documents considered, as well as we will incorporate additional features in our models, including link- and network-based features (e.g., based on document interlinking) and social media-based features (e.g., the number of likes a given article received on social media sites, or the number of followers a given blog has). Lastly, as a consequence of such extension of both the number of records and of features, we will have to consider methods for scale our prediction models.

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

ties. We thank the students of the UvA journalism course and the RMeS summer school participants for participating our user studies.

References

1. B. Amento, L. Terveen, and W. Hill. Does “authority” mean quality? predicting expert quality ratings of web documents. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, pages 296–303, New York, NY, USA, 2000. ACM.
2. K. Bharat, M. Curtiss, and M. Schmitt. Method and apparatus for clustering news online content based on content freshness and quality of content source, June 7 2016. US Patent 9,361,369.
3. D. Ceolin, P. Groth, V. Maccatrozzo, W. Fokkink, W. R. van Hage, and A. Nottamkandath. Combining user reputation and provenance analysis for trust assessment.
4. D. Ceolin, J. Noordegraaf, L. Aroyo, and C. van Son. Towards web documents quality assessment for digital humanities scholars. In *Proceedings of the 8th ACM Conference on Web Science*, WebSci '16, pages 315–317, New York, NY, USA, 2016. ACM.
5. 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.
6. 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.
7. M. Howell and W. Prevenier. *From Reliable Sources: An Introduction to Historical Methods*. Cornell University Press, 2001.
8. 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.
9. 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.
10. I.-H. Kang and G. Kim. Query type classification for web document retrieval. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03, pages 64–71, New York, NY, USA, 2003. ACM.
11. 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.
12. 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.
13. A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer. Quality assessment for linked data: A survey. *Semantic Web Journal*, 2015.
14. H. Zhu, Y. Ma, and G. Su. Collaboratively assessing information quality on the web. In *ICIS sigIQ Workshop*, 2011.