DRAFT: Predict the Information Quality of Web documents

Ozkan Sener
Vrije Universiteit Amsterdam
ozkansener@gmail.com

Davide Ceolin
Vrije Universiteit Amsterdam
d.ceolin@vu.nl

ABSTRACT

Web users are overloaded with information and Web users often do not know how good the quality of a Web document is by just looking to the URL. This impacts the way a Web user perceives information. Compared to the model of Ceolin et al. we did three things differently: 1) We make the model of Ceolin publicly available by using open-sourced libraries. 2) We add features that that represent categories like readability, numbers of years of existence ,Website performance and social media analysis of a Web document 3) Used the crowd to collect more participants to assess new 4) Used an multiple output regression instead of singular regression. 5) Our model updates over time and tries to be time aware. We believe that IQ of Web documents can be scored, but that user understanding and context understanding is required.

1 INTRODUCTION

Web user are overwhelmed by the quantity of information that is available on the Web [14]. This makes it difficult for a Web user difficult to receive information of high quality. Web user are also subject to human bias, which can be influenced by varying backgrounds and expertise [11]. Information overload is occurs when the brain of the Web user process more information then it can handle. When Web user gets confused they will no longer be able assess the quality of the Web documents [13].

We propose an information retrieval framework that can be used to Semi-Automatically Assess the Quality of Websites (SAATQOW). We ATQOW of a new Web document by collecting historical labeled data (quality assessment scores) and features that are (partially) representing the Quality labels. We then training our model to learn a function that can be used to predict the assessment scores (the labels) for a new instance by automatically collecting the features of the Web document. This function is of the form: $\{\phi\}\chi\{\psi\}$ Where $\{\phi\}$ is the set of features and ψ is the IQ score and χ is the function that our model beliefs can be used to predict ψ. When users of our framework provide feedback about a prediction we update the belief of χ.

Ceolin et al. [4] created already a framework where they collected the features sentiment, emotions and trustworthiness. They then performed an user study with media experts and journalism where they asked them to annotate the scores of the Web documents. They collected 138 assessment scores by showing two or three Web document to the each annotator. In total 51 unique documents have been assessed. They created their model by training their function on a Support Vector Machine algorithm.

Compared to previous research: 1) We make our framework publicly available by using open-sourced libraries. 2) We add features that that represent categories like readability, numbers of years of existence ,Website performance and social media analysis of a Web document 3) Used the crowd to collect more participants to assess new 4) Used an multiple output regression instead of singular regression. 5) Our model updates over time and tries to be time aware.

When our framework is used for a new website, our framework will predict an assessment score for this website. If this Website is not relevant (because the low quality of a Website) to the user, then Web users can decide to not visit the Website.

This paper continues as follows: Section 2 introduces related work; Section 3 describes the methods adopted; Section 4 presents the results collected, that are analyzed in Section 5. Section 6 concludes the paper.

2 RELATED WORK

Several authors [1, 5, 17, 27] report that there is need for a framework that is able to assess the quality of Websites automatically. In America [2, 15] fake news was spread during political election. Facebook and Google was not able to prevent fake news occurring on their platform. Especially younger Web users [6] are vulnerable for fake news.

Ceolin et al. [5] created an model that semi automatically automatically estimates the assessment quality of Websites. Their model does this by performing computations that creates features that functions as input for their model by assuming that the features that they use would represent the quality scores of Web sites. The second component are the assessment scores. The third elements are the Machine Learning algorithms that automatically estimates the IQ scores of Websites.

We found that some of the features that are used in the model of Ceolin et al. create too much variance (noise). A model with high variance is very likely to over fit [26]. We also found that features like the trustworthiness are interpreted differently among different Web users and that sometimes it is not possible to collect the trustworthiness scores.

2.1 Frameworks

Ding et al. [7] report that analysis like sentiment and emotion analysis (granular form of sentiment analysis) can be used to discovery opinions and feeling or moods of Web writers. This technique is called Natural Language Processing. The NLP algorithm does this by analyzing the words that are bad or good words. Then the algorithm searches for modifiers that tells something about bad/good words. With emotion analysis we are interest to automatically assess the feeling of our website that a Web user has toward the quality of the Website. With emotion analysis we analyze the sensitive

aspects that are on the Website. The other future that is been used by Ceolin et al. is Web of Trust (WOT). WOT is a Crowdsourcing platform which we use to gather ratings of Web users that scored the quality of Web pages. This tell us whether Web users trust the content of the website. The WOT rating are domain based. This limits to see the rate scores for a specific URL. Another limitation of the WOT score is that it is based on crowd sourced ratings. This means that Web users can directly affect the results. This means that the WOT ratings can have other scores in the future. WOT scores are from 0 (very untruthful) till 100 (very trustful). Polarity scores are from -1(very negative) till 1 (very positive). The model that was trained by Ceolin et al. had not the opportunity to learn what each features represents and how this influences the assessment scores. For example: For example the feature sentiment: the feature can have positive and negative values. But the model of Ceolin et al. is only trained for values inside the range: 0 till -0,6. We believe that especially the extreme values of the features outliers carry a lot meaning about the meaning of the features. When the interpretation of feature and how the feature influences the assessment scores is known by the Machine Learning Algorithm (MLA), we will be more accurate at predicting the assessment scores of a Website. Ceolin et al. performed their assessments with media experts. We believe that annotators with a heterogeneous background is needed in order to generalize this Framework for individual Web users.

Pinto et al. [22] report that there are many Natural Language Processing (NLP) tools. They describe that it is challenging to select one of the many NLP tools. The performance of a NLP tool depends on the type of source of text it is used. APIs like: Calais, Google Natural Language, Havenondemand, Aylien, TheySay PreCeive, Qemotion and Monkey learn are some examples of NLP Apis that provide the same NLP features as IBM Watson API. Ceolin et al. used IBM Watson in order to collect their NLP features. IBM Watson is not able to correctly analyze large Web documents. IBM only analyze the first 50,000 character of these Websites. There are also libraries for programming languages available that can collect these features NLTK, Textblob, tm and Stanford CoreNLP are some examples of open libraries that provide NLP features.

2.1.1 Textual Statistics. Most of the scientific publishers are limiting researchers in the amount pages that a conference report can have in order to be published. The reason for this is that readers of research papers should only present the essentials/the most important/relevant details of their research so that readers easily can understand what the paper is about instead of getting confused (not understandable). Therefore we believe that having features that describes the amount of text on a Web document can be used to assess the quality of the Web. The writing style of a Web documents can be used to semi-automatically assessing the quality of Wikipedia documents [11, 16]. Dalip et al. [11] report that textual features related to length, structure and style of a Wikipedia document are the most relevant important features in order to assess the quality of a Web document. We believe that Web users are more interested in Web documents that are simple and clear.

Si and Callan [24] developed a model where they used textual statistics in order to predict the readability of the Website. Formulas like the, The Flesch Reading Score abd The Dale-Chall can be used to assess the readibility Si and Callan showed that the readability of Web pages are influenced by the writing style. Flesch reading Score (FRS) is very often used by researchers in order to score the readability of Web document about health related information [21]. With the formula bellow we show how the FRS can be calculated: $0.39\left(\frac{\text{total words}}{\text{total sentences}}\right) + 11.8\left(\frac{\text{total syllables}}{\text{total words}}\right) - 15.59$

The higher the output of the score is the easier the Web document is to read. The Lower scores indicates that the Web document is not easy to understand for the average Web user, but for an university graduated person it is understandable Fitzsimmons et al [9].

2.1.2 Source reputation. Alexa Rank ranks websites based on the popularity of the Website over a period of 3 months. It does this by analyzing the number of times a Web domain that has visited by Web users that use the Alexa tool-bar over a period of three months. The higher the ranking number is the less often the Web domain has been visited. When a Web domain is scored N/A this means that the Web source has no ranking. Google PageRank (PR) is a metric that is used by Google in order to determine which Web pages will appears first in their search engine based on the importance, reliability and authority of a Web page. However Google hasnt updated page ranks since 2013 and it is not open for public anymore. We believe that Websites that have in general a better reputation/popularity are more likely to have a higher IQ. The reason why we expect this is because users in general understand Web document better if they already can make some inferences on how the web document works technically before carry out there task (reading information).

2.1.3 Response time. Nielsen [20] reports that user-perceived IQ is highly influenced by the response times and the variance of the response times of a Web document. When Websites are taking too long to respond to the Web user request the Web user is less likely to be satisfied about the IQ of a Web document. Web users also think that Websites with a high response time are seen by Web users as incompetent. We believe that the response time also provides information about the maturity of the Web source and that these mature Web sources are more likely to have Web documents with a higher IQ.

2.2 Assessments

There are different standards in the field of information retrieval that describe how the quality of documents can be assessed. The DAMA [19] has created a data quality framework with the dimensions: Accuracy, Consistency, Integrity, Timeliness, Completeness and Validity. Ceolin et al.[3], Wang[25],

Held and Lenz [12] provides us the definition of the quality dimensions that we use in this paper: The overall quality score summarizes the opinion about the quality of the website. Accuracy explains how correct the website is. If the accuracy of the Website is high this means that there are not too many wrong (generalized) statements on the website. Completeness means that the information that is available on the website fulfills our needs. Neutrality means that the author of the website is not pro or con towards the subjects. Relevance is the closeness between the information we need and the information that is provided. Trustworthiness means that we are able to see the protocols and procedures that are used to provide the information that is provided to us. The website is considered readable if we can read the content on the website and that the information on the website is clear to us. Precision means that the website has sufficient detailed information that is required to fulfill our task. In order to compare our framework with the framework of Ceolin et al. we are choosing the same quality dimensions. However our assessment is performed on the crowd. We are not able to clarify or help the assessors on the crowd. Therefore we need to clarify our goals, expectations so that the assessors are to perform the task as we wish. This is one of the reasons why we try to keep our assessments as simple as it can so that there is no noise caused by us. Crowd sourcing makes it for us possible to acquire a huge amount of occupants in a small amount of time. Crowd sourcing makes it also possible for us to reach Web users all over the world and acquiring their assessment. *Nonfunctional qualities* Raiber and Kurtland [23] report that the content of the website is not the only key factor that affects the quality scores of our quality dimension attributes. Raiber and Kurtland report that factors like the usability, scalability, originality of the content, personalization of the content, the performance of the Website and the design of the Website are also factors that influence the quality of the website.

2.3 Machine Learning

The choice of choosing an algorithm should be based on the properties of the data set that we want to analyze because different Machine Learning Algorithms (MLA) have different characteristics [18, 26]. Domingos [8] reports that the success of the Machine Learning approach is the combination of highly sophisticated algorithms and large amounts of data of high quality. When assessment scores are not reliable our model will be less accurate at predicting the assessment scores.

Algorithms. Deciding which type of algorithms we apply on our model can be done based on functional and non-functional criteria [10]. Parametric Machine Learning Algorithms are more reliable when the quality of the data is not of high quality compared to non-Parametric algorithms. Parametric algorithms are also easier to understand and do not require complex computation (saves time and computation performance). Parametric algorithms When the sample size of the dataset is small a parametric algorithm is preferred over non-parametric dataset and in scenarios where the fit of a model to the data is not perfect. However these parametric algorithms can not learn the complex patterns from a dataset. For example non-parametric algorithms are capable of interpreting the interactions between our set of features and how these set of features and their interactions represent our IQ scores . However in uncertain scenarios or unseen learning the general patterns of a Machine Learning algorithm is much more important.

Research Question

In this paper we answer the question: "How can we estimate the quality of Websites with the use of machine learning algorithms?"

3 METHODS

We collect Web documents via Search engines, recursively collecting and all the URLs that are mentioned inside a URL for a specified URL. We then collect the features of the URLs that are in accessible and where the content of the Web document is in English. From the collected set of Websites with features we create a sample via diversified sampling.

We will collect the URL of Web documents where the IQ of a Web document has been scored. We then checked whether the Web document are accessible. We check whether we have access to the Web document (the Web publisher responds to our HTTP request and we are not redirected to another page) and we measure how long it takes the Web publisher to responds to our request message. Capturing the response time of Web document is done by sending ten times an HTTP request message and taking the average of these scores, but we also remove the cache after each request. We use the Python library request for this. When the URL is accessible for us we will collect the other features.

We extract the the content of each Web document with the python library goose, we collected several readability metrics (Flesch Kincaid) of a Web document via the Python library textatistic, we collected the Subjectivity and Polarity via the Library Textblob. Via Web of Trust we extract the website reputation rating. By using the Internet Archive check we check the first date that a owner of a Website has extracted. We 4 times open an URL in order to detect the latency (application latency) time of the Website and compute the average latency time. In order to do this at large scales we implemented threads in order to scale up the performance of our architecture. Via the crowd we collect our labels. We did this by asking the same question as Ceolin et al. did in their research.

We perform an test of association with the test of Kendall coefficient in order to test if there is a significance relation between the features that we collected and how our crowd-workers scored these URLs.

Via a multi label regression we trained our framework to learn the function χ

4 RESULTS

We crawled, archived and scraped and collected the features of more than 24000 Webdocuments. We then created an sample of 1000 URLs. On the crowd we asked crowdworkers to label these URLs.

4.1 Statistical Analysis

We performed an test of association with the test of Kendall coefficient in order to test if there is a significance relation between the features that we collected and how our crowdworkers scored these URLs.

5 DISCUSSION

We found that the features readability, trustworthiness and Alexa ranking are stronger correlated with the IQ scores compared to the Natural language Processing scores. We found that the feature Trustworthiness has the strongest correlation with the IQ scores in the study of Ceolin et al. However we also discovered that this is not generalizable for everybody. Scientist for example prefer to have access to unknown information. We also observed during the feature capturing process that some Websites don't have a trustworthiness scores. We found that there is an significance difference between how different NLP tools compute the sentiment scores.

We do believe that our set of features can be used in order to predict the Quality of a Web document. However we believe that the lack of consensus between annotators can cause disagreements. This study showed that different annotators have sometimes different preferences for the same type of Web user have a different view on the features represent the IQ of Websites. However we believe that our framework should not behave statically but also dynamically. Ideally we would like to receive real-time feedback so that the execution of our framework will provide representative assessment scores that is personalized, based on the feedback that the Website user provides us. In order to validate this we would like to perform an experiment where we gather feedback from Web users about our predictions about the Quality We demonstrated that the IQ quality of Web documents can be scored. However our study also reveals that different annotators, annotated the quality scores differently, but we do believe that our framework could be used in order to show weather Web sources have a good IQ or not. We also believe that it is important to identify the cause of errors.

Potential limitations of our study are: Only English language sites were evaluated, and therefore the findings may not be generalizable to those websites written in other languages. The subjective nature of the IQ annotators and the definitions of the IQ criteria can be the cause human bias. It can be the cases that our features over or underestimated the their actual scores of the features. We also believe that differences of methodology, IQ criteria, the capability of assessing the IQ criteria, demographical factors, and the content and context in which the IQ scores was scored could have influence our results. In order to compare IQ scores among different Web users we need to be sure that they performed the same assessment. A cross sectional study could help us. Our study showed that different Readability metrics are highly related with the IQ scores of a Web document. These readability metrics are for each type of Web document accessible and there. We believe that the reading level and topic distributions provide and important new representation of Web content and user interests, Therefore we suggest this as future work. We also investigated whether there are differences of user preferences could be caused by our set of features could identify and predict disagreements of assessment scores based on the different user preferences. We tested whether obtaining user feedback explicitly (or implicitly in a future study) and recording the feedback and the set of features that is related with the Website and the actual prediction could increase the adaptability the estimations of the IQ scores based on the personal preferences of Web users.

6 CONCLUSION

This paper answered the question: "Can we Semi-Automatically Assess the Quality of the Web with the use Machine Learning?" We tested the model: $\{\phi\}\chi\{\psi\}$ where $\{\phi\}$ is named the set of features and ψ is named the IQ assessment scores (the labels) and χ is the function that our model learned in order to Semi-Automatically Assess the Quality of the Web (SAATQOTW).

Our tool improved the previous work about the IQ of the Web by:

- Our framework can be publicly available
- We tested our framework with an huge number of participants on the crowd
- We have an more complete set of features
- Also real people using the application

Web users can use our framework in order to select Websites that are scored with high quality. Search engines can use our framework to provide Web users a list of URLs of Web documents with high quality content. Website masters can use our framework to assess the quality of their Website. However we only tested our framework on English Web documents and it might be the case that this does not apply for Websites that are not English. We also believe that there are features which might be better at representing these labels. We also want to emphasize that assessments are vulnerable for human bias. With this we mean .However this does not imply that these apply for every use Web users since the definition of the quality should be studied in the context of a Web user. Therefore we would like to see However evaluation of the dataset revealed that the assessment scores are highly personal and that we need to predict assessments based on the experience feedback that we receive from our framework. We showed that disagreements of scores can, be explained by our tool in cases that multiple users assessed the URLs.

We also found that our framework can be used in order to predict the amount of disagreement of the IQ scores. This would help us to better understand the cause of it and fix this type of causes in cases a Website its content is unclear.

7 FUTURE WORK

- Following annotators over a long period of time and performing different types of users studies.
- Recommending Web documents of high quality
- Study real people using and their behavior by gathering implicit feedback
- Adding more features like numbers of Links, Alexa Ranking.
- Investigating the effects of the Visual Design and ads on the Quality perception of the Web users.
- User Personalization and Profiling user
- Studying the Content of the Web documents and try to understand their effect on the Quality scores caused by the stance of the source.
- Our features are ignoring video's and pictures and the Design of the Website. We believe that this does effect the Information Quality of the Websites.
- During the ETL procedure we already implemented Multi-Threads to collect the features of multiple Web documents at the same time. We believe that this can be handy for Web users in case we want to provide some suggestions for the Web users.

REFERENCES

[1] Aladwani, A. M., and Palvia, P. C. Developing and validating an instrument for measuring user-perceived web quality. 467 – 476.

[2] Allcott, H., and Gentzkow, M. Social media and fake news in the 2016 election. Working Paper 23089, National Bureau of Economic Research, January 2017.

[3] Ceolin, D., Noordegraaf, J., and Aroyo, L. Capturing the ineffable: Collecting, analysing, and automating web document quality assessments. In *20th International Conference on Knowledge Engineering and Knowledge Management - Volume 10024* (New York, NY, USA, 2016), EKAW 2016, Springer-Verlag New York, Inc., pp. 83–97.

[4] Ceolin, D., Noordegraaf, J., and Aroyo, L. Web data quality assessment.

[5] Ceolin, D., Noordegraaf, J., Aroyo, L., and van Son, C. Towards web documents quality assessment for digital humanities scholars. In *Proceedings of the 8th ACM Conference on Web Science* (New York, NY, USA, 2016), WebSci '16, ACM, pp. 315–317.

[6] Clark, L., and Marchi, R. *Young People and the Future of News: Social Media and the Rise of Connective Journalism*. Communication, Society and Politics. Cambridge University Press, 2017.

[7] Ding, X., Liu, B., and Zhang, L. Entity discovery and assignment for opinion mining applications. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2009), KDD '09, ACM, pp. 1125–1134.

[8] Domingos, P. A few useful things to know about machine learning. *Commun. ACM 55*, 10 (Oct. 2012), 78–87.

[9] Fitzsimmons, P., Michael, B., Hulley, J., and Scott, G. A readability assessment of online parkinson's disease information. *The journal of the Royal College of Physicians of Edinburgh 40*, 4 (December 2010), 292296.

[10] Graczyk, M., Lasota, T., Telec, Z., and Trawiski, B. *Nonparametric Statistical Analysis of Machine Learning Algorithms for Regression Problems*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010, pp. 111–120.

[11] Hasan Dalip, D., André Gonçalves, M., Cristo, M., and Calado, P. Automatic quality assessment of content created collaboratively by web communities: A case study of wikipedia. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries* (New York, NY, USA, 2009), JCDL '09, ACM, pp. 295–304.

[12] Held, J., and Lenz, R. Towards measuring test data quality. In *Proceedings of the 2012 Joint EDBT/ICDT Workshops* (New York, NY, USA, 2012), EDBT-ICDT '12, ACM, pp. 233–238.

[13] Ho, J., and Tang, R. Towards an optimal resolution to information overload: An infomediary approach. In *Proceedings of the 2001 International ACM SIGGROUP Conference on Supporting Group Work* (New York, NY, USA, 2001), GROUP '01, ACM, pp. 91–96.

[14] Kapyla, T., Niemi, I., and Lehtola, A. Towards an accessible web by applying push technology. In *Fourth ERCIM Workshop on "User Interfaces for All"* (Stockholm, Sweden, 1998).

[15] Kwak, H., Lee, C., Park, H., and Moon, S. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web* (New York, NY, USA, 2010), WWW '10, ACM, pp. 591–600.

[16] Lipka, N., and Stein, B. Identifying featured articles in wikipedia: Writing style matters. In *Proceedings of the 19th International Conference on World Wide Web* (New York, NY, USA, 2010), WWW '10, ACM, pp. 1147–1148.

[17] Loiacono, E. T., Watson, R. T., and Goodhue, D. L. Webqual: A measure of website quality. *Marketing theory and applications 13*, 3 (2002), 432–438.

[18] Mitchell, T. M. *Machine Learning*. McGraw-Hill, 1997.

[19] Nicola Askham, Ulrich Landbeck, J. S. The six primary dimensions for data quality assessment, defining data quality dimensions. DAMA UK Working Group.

[20] Nielsen, J. *Designing Web Usability: The Practice of Simplicity*. New Riders Publishing, Thousand Oaks, CA, USA, 1999.

[21] Paasche-Orlow, M. K., Taylor, H. A., and Brancati, F. L. Readability standards for informed-consent forms as compared with actual readability. *New England Journal of Medicine 348*, 8 (2003), 721–726. PMID: 12594317.

[22] Pinto, A. M., Oliveira, H. G., and Alves, A. O. Comparing the performance of different nlp toolkits in formal and social media text. In *SLATE* (2016).

[23] Raiber, F., and Kurland, O. Using document-quality measures to predict web-search effectiveness. In *Proceedings of the 35th European Conference on Advances in Information Retrieval* (Berlin, Heidelberg, 2013), ECIR'13, Springer-Verlag, pp. 134–145.

[24] Si, L., and Callan, J. A statistical model for scientific readability. In *Proceedings of the Tenth International Conference on Information and Knowledge Management* (New York, NY, USA, 2001), CIKM '01, ACM, pp. 574–576.

[25] Wand, Y., and Wang, R. Y. Anchoring data quality dimensions in ontological foundations. *Commun. ACM 39*, 11 (Nov. 1996), 86–95.

[26] Witten, I. HFrank, E. *Data mining*. Morgan Kaufmann, 2000.

[27] Zhung, Y., and Mecer, R. A machine learning approach for rating the quality of depression treatment web pages.