Quality Assessment of Arabic Web Content:

The case of the Arabic Wikipedia

Adnan Yahya
Department of Computer Systems Engineering
Birzeit University, Birzeit, Palestine.
yahya@birzeit.edu

Ali Salhi
Department of Computer Systems Engineering
Birzeit University, Birzeit, Palestine.
eng.salhi.ali@gmail.com

***Abstract*—** With the huge size and large diversity of Arabic web content, machine assessment of document quality acquires added importance. Users are in dire need for quality rating of the material returned in response to their queries. The Wikipedia, with its large metadata, has been a topic of extensive research on document quality assessment. Criteria used include text properties and style parameters, contributor and edit characteristics and multimedia components. In this paper we report on our ongoing work to adapt existing document assessment approaches to Arabic content with concentration on the Arabic Wikipedia and present some of the results. We also try to augment that with features specific to Arabic as well as parameters like author expertise and social media presence. One of our goals is an aggregate measure integrating many of the features into a single document quality index. We plan to use Wikipedia article quality assessment results to train general content assessment methods that can be applied to general content that lacks major Wikipedia features.

***Keywords*—**Document Quality Assessment, Arabic Wikipedia,Web Content Quality, Quality Assessment Parameters

I. INTRODUCTION

Web content is highly diverse, ranging from rigorous online academic publications to quite informal social media posts. In between, we have corporate web sites, online news services and more. We also have the Wikipedia emphasizing collaborative community effort to content generation[4]. Some web content compares well with its traditional counterpart, as seems the case for much of the Wikipedia[4], while other content is of suspect quality[7]. In addition to differences between articles of a single language Wikipedia, differences exist between Wikipedia content in different languages, reflecting factors like different stages of maturity and community engagement in content development, different experiences of authors/editors and available tools for their work [2,6,13,7].

While developing fast, Arabic content seems to have major quality problems[2,6,7]. It is important for consumers to be aware of the quality of the web material they encounter through the various content delivery outlets. While material like electronic journals and books, media sites and even corporate websites have established and strictly enforced rules for quality assurance, the same cannot be said about personal websites, blogs and participatory content sites like the Wikipedia. The quality of material in the latter can vary and may change over time. In such an environment, it is important to relay any information about the quality of content to the user so as to make informed decisions on how to utilize the delivered material. This acquires special importance when the data returned is contradictory/incompatible: quality indicators can be deciding as to which data to accept. The need for quality assessment goes beyond trust, and quality indicators may be utilized to improve content, to rank search results and reward creators of content based on the product quality.

The scientific community devised methods for quality assurance like peer review, editorial oversight, training of content developers or even content "star rating". Much of the web content lacks such mechanisms, and the pressure to post fast doesn't allow for extended review before going online. The vast amounts of online data and its volatility render manual assessment an impossible task. Thus, the need for machine methods for content quality assessment. The term *quality* may not be well defined as it may involve many parameters that relate to form and matter: one can think of well written pieces that are not factually correct, and of material that is factually true and valuable but ill presented (e.g. product of crude machine translation). The topic of this paper is web content quality: how to automatically assess the quality of Arabic web content in a manner that agrees with the generally accepted norms of quality assessment.

We distinguish between the cases when we have sizable metadata about the content as is the case for Wikipedia, where we have detailed information about the contributors (authors/editors), revisions, links, references, categories/tags, quality tags (say through feature/good designations) and even overall, though controversial, quality indicators like *depth*,[1] and much more; and the case of general content where we may have only the text of the document to rely on. The first case is addressed here and we try to utilize the metadata not only for assessing the quality of the material but also to increase the amount of annotated (known quality) data that can be used to train quality estimators for general content. Though not reported here, we also try to use resources external to the items being assessed such as citations, search results, social media citations and web analytics for quality assessment purposes. We are interested in document quality rather than overall site or corpus quality. The ultimate test of success in web content quality assessment will be the agreement of our results with gold standards defined based on human judgments.

The rest of the paper is organized as follows: in the next section we give an introduction and survey the state of the art in assessing web content quality with emphasis on Wikipedia; in section III we discuss parameters for assessing content quality applied to Arabic. We consider quality parameters in the presence of Wikipedia style meta-data and report on our experiments. In Section IV we draw some conclusions and point to possible directions for future research.

[1] http://meta.wikimedia.org/wiki/Wikipedia_article_depth

II. BACKGROUND

Much research has been ongoing into assessing the quality of web documents. The criteria used range from the simple, based on document length and word properties, to the sophisticated, that takes into account authors/editors characteristics and the interaction between them. The latter have focused on the Wikipedia with its extensive metadata. Wikipedia also has a community based quality classification system where articles may be judged *Feature* or *Good* through a nomination/voting process based on criteria relating to being well-written, comprehensive, well-researched, neutral, stable and following style guidelines regarding having a lead section, structure, consistent citations and media content, length and focus.[2] *Good* articles do not qualify for *Feature* status but meet the criteria to a reasonable degree.[3]

One of the simple measures of quality of Wikipedia articles is article word count. As reported in [3] article length can be a good predictor of quality in terms of a Wikipedia being *Feature* or just *Random* (or regular). Under this measure an article is *Feature* (*Good*) if its word count exceeds a predefined, empirically determined, threshold, T_F (T_G). The success rate (accuracy) reported for the *Feature* case in English was 96.31%. However, this measure is problematic if used in isolation, as there is more to quality than word count and given that the percentage of *Feature* articles is quite low for English and Arabic (less than 0.2% in both cases) and that the number of articles with length exceeding T_F is at least 25 times the number of *Feature* articles.

In [9] a machine learning approach to recognize *Feature* articles through their distinct writing style based on character tri-grams distribution is presented. It reports good results and argues that character n-grams relate to sentence transitions, the utilization of stop-words, adverbs, and punctuation, all of which are important authorship/style indicators[9].

In [5] there is another effort at simple evaluation of the quality for two article quality classes, namely *stabilized*: articles without major changes over the recent period and *controversial*: articles undergoing major changes, reverts and vandalism. The assumption is that the users of Wikipedia use varied criteria for quality assessment of articles from different categories. They present a two-tier approach: first find the broad category of the article then use category-specific quality prediction models to compute the quality estimate. Validation is done by comparing the prediction results with assessments made by average Wikipedia visitors. For *Stabilized* articles, features like article length, citation and link, image and section count densities are used. For *Controversial* articles factors relating to revision history such as reverts, edits by different types of users are used. The features were used to train classifiers using WEKA. A weighted average performance of more than 80% was reported.

Article quality is influenced by contributor quality. The quality of the contribution is important for assessing author contributions and estimating the quality of future contributions. [1] talks about edit longevity combining the amount of change performed by an author, with how long the change lasts (survives subsequent edits). The edit quality measures how long the change lasts in the system: it is max for edits that are preserved fully in subsequent revisions, and min for edits that are reverted. The edit longevity of an author is computed as the sum, over all the edits performed by the author, of the edit size multiplied by the edit quality.

In [12] article quality is assessed based on edit longevity of contributions and the contributor authoritativeness metric which is defined through centrality. This is based on the intuition that articles with major contributions by authoritative authors/editors are more likely to be of high quality, and that quality articles generally involve more communication/interaction between authors. While longevity itself is a good measure as evidenced by the experiments on articles of known quality, adding contributors authoritativeness improves that[12].

In [10] an elaborate system for quality evaluation for Wikipedia articles with increasing complexity of the employed parameters is offered. The four models used are: Naïve, Basic, PeerReview and ProbReview. In the Naïve model, as in [3], the quality of an article is directly proportional to the number of words. The Basic model uses the authority of the article authors to define the quality of articles and the authority of authors depends on the articles they coauthored. There is a mutual dependence between these parameters and they enforce each other. The PeerReview model incorporates the authority of editors into the evaluation process on the assumption that a text surviving a peer review by an authoritative reviewer needs to be viewed positively even though it was authored by a less authoritative user. However, given the doubt that an editor approves every word of an edited article, the input is relaxed a little in the ProbReview model to include the probability that a reviewer checked a certain text chunk. Chunks closer to changes have a better chance of having really been reviewed and their weight will be increased accordingly. Reported test results show that ProbReview performs best but that both PeerReview and Basic can have improved performance as hybrid systems augmented by article length. That is not the case for ProbReview. Naïve, based solely on article length is the baseline, and the conclusion is that contributor properties, both authors and reviewers do matter for article quality evaluation. [8] also uses the reputation of the author, defined as the probability to produce good quality content, to assess the added content, and to estimate the percentage of high quality revisions and the proportion of time during which an article is in a high quality state. These measures are tested on *Feature/Random* articles. The shortcomings of basing quality on reputation include anonymity of many authors; sparsity, preventing the assessment of reputation and ignoring expertise in the area in estimating reputation.

In [14] a Fuzzy Logic approach is used to evaluate the quality of Thai Wikipedia articles. The set of features employed is extensive and covers text properties, metadata including contributor parameters such as feature article authoring. The results reported are good. The approach is limited to feature/normal judgments and doesn't discriminate between normal articles: the vast majority of any Wikipedia.

From the reviewed work, and many others, it is clear that web content quality assessment is a complex issue that has the elements of evaluating manuscripts in traditional publishing. Having that done automatically is not straightforward and may

[2] http://en.wikipedia.org/wiki/Wikipedia:Feature_articles
[3] http://en.wikipedia.org/wiki/Wikipedia:Good_articles

involve so many parameters/features of the text itself, the authors, the editors and even the users. Adopting techniques used in guaranteeing quality of scientific works (basically peer review, editorial hierarchy) is out of the question due to the large size and continuously changing nature and the short publication cycle of much of web content and the free community contribution paradigm involved in certain types of content like Wikipedia. It seems that the best one can hope for is an efficient predictive mechanism to assign a quality measure to web content that will be taken as an additional input by the user when presented with links to content. One may also work in a multi-tier system whereby the better quality estimates for the Wikipedia can be used to generate annotated data that can serve to train quality assessment tools that may operate on general, less annotated, content. This is the approach we plan to take regarding the quality of general Arabic web content. We are also working to integrate issues like contributor expertise, social media impact, citations and other factors into the evaluation process. All through, one needs to keep in mind that quality assessment measures/methods may have cultural bias that may render them less applicable in different contexts, despite the substantial cross lingual subject overlap [15]. We also give preference to simple, computationally feasible solutions.

III. Arabic Web Content Quality Assessment Methods

While still far from proportional to the share of Arabic speakers, Arabic web content is large and growing fast. Much of it is spontaneous but several initiatives have been undertaken to encourage the generation of quality Arabic Web content. The Arabic Wikipedia has also been growing fast and is slowly stabilizing. Currently, the article count stands at around 338K articles (October 12, 2014)[4]. As is usual for other languages, the proportion of high quality articles is small: in February 2014 we had only 284 (1 in 998) *Feature* (gold star) articles and another 258 (1 in 1029) *Good* (silver star) articles for a total of 542 labeled *high quality* articles. These designations are done manually and need time to ascertain. The positive side is that these articles can be used as annotated data to automatically learn quality assessment. While the size of the Arabic Wikipedia is small and it constitutes only a small part of general Arabic web content, its role can be substantial due to the availability of metadata and annotations (tags/categories, edit history, links, references, occasional quality designations) and contributors (their classification, contributions, coauthors, ...) . This makes the Wikipedia a good candidate for quality assessment research, both by itself and as an infrastructure for general web content.

Document quality assessment is generally not a trivial task. The concept of quality may have cultural/contextual sensitivities that go beyond the possible shifts in measurement models, vocabulary, metric functions, and measurement representations with regard to scale, precision, and formatting, to differences in value structures, reference sources and more[15]. Our contribution will go beyond the adaptation of the existing tools to Arabic and we investigate the possibility of including factors like behavior of confusion letters and the expertise of contributors by granting an edge to those writing in their area of expertise; by leveraging social media citations and conformance to Arabic writing best practices as potential quality indicators. However, some of that is still work in progress and is not reported here for space considerations.

A. Arabic Wikipedia Quality Assessment:

We start by discussing document quality criteria that are general and do not rely on Wikipedia-style metadata.

1) Naïve approach (word count):

There is strong evidence that size is an important quality indicator and that longer Wikipedia articles have a better chance at being of high quality. Under the Naïve Approach, an article is declared *High Quality* if its length exceeds the threshold for that quality level. The distribution of Arabic Wikipedia articles by length in words is given in Fig. 1.

For English the article length threshold for *Feature* articles, T_F, that gave best results (minimal error rate for classifying articles into *Feature* and *Random*) is 2000 words[3]. Given Arabic writing rules we expected T_F for Arabic to be below 2000 reflecting the fact that equivalent Arabic texts will have smaller word count compared to their English counterparts. To test that we applied the experiment in [3] to the Arabic *Feature* and *Good* articles and an equal number of random articles for each and tested the error rate for different threshold values. Given that *Feature* and *Good* are generally longer articles and tend to overlap, we performed a series of experiments for each class and for the class *High Quality* (*Feature* and *Good* combined). Fig. 2, shows the results.

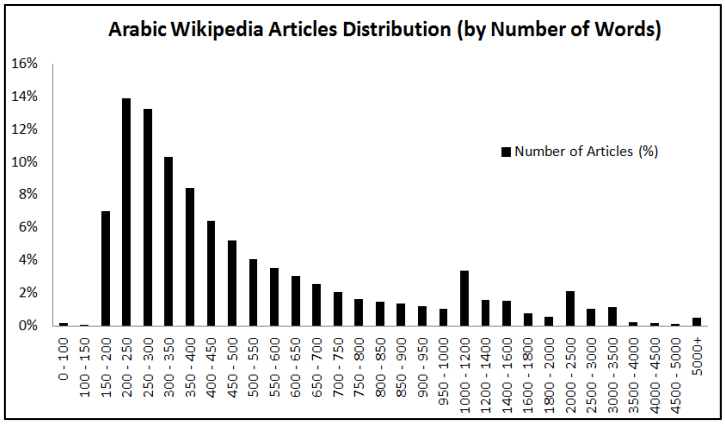

Fig. 1. Arabic Wikipedia Articles Distribution (By Number of words)

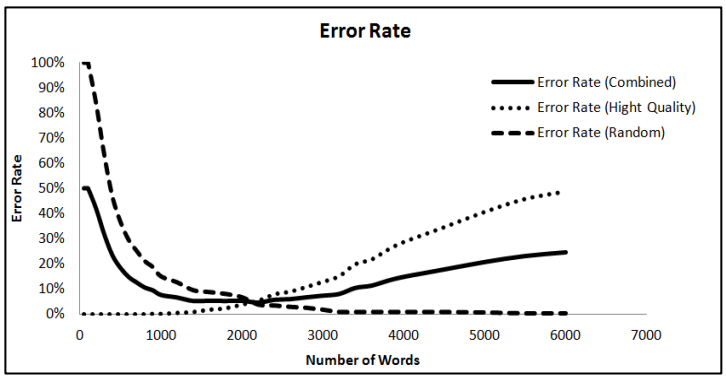

Fig. 2. Error Rate for Article Classification by Article Size in words

[4] http://meta.wikimedia.org/wiki/List_of_Wikipedias

The magic number for *Feature* articles for Arabic (T_F) is 2200 words. For *Good* (T_G) it is 1200. The combined High Quality class (Feature or Good) has the number (T_{HQ}) 2000 which gives a prediction error rate of around 4.4% (2.8% for *Feature* vs *Random*, 4.4% for *Good* vs *Random*). That is, if you classify articles of greater than 2000 words as *High Quality* and others as *Random* we will be right 95.6% of the time. Note that for all our experiments we used balanced pairs of sets of articles.

The number of articles that can be *Feature/Good* (of size>2000 words) is above 8% of the Arabic Wikipedia. This is clearly an indicator that labeling all long articles as *Feature/Good* based on length alone is deceiving as some of these articles didn't make it and others were not nominated to these quality classes. One has to do more to come up with better predictions. We need predictors as to how to find those among the candidate articles that are really Good/Feature.

Fig. 3, gives the distribution of the article length (in words) for *High Quality* and *Random* articles of the Arabic Wikipedia. The average lengths, in words, for each of these classes are 7038 for *High Quality* (9176 for *Feature*, 4694 for *Good*) and 653 for *Random*. The corresponding average sizes in Kilobytes –KB- are 71.5 (93.1, 47.9) and 6.9, respectively.

2) Character N-Grams as quality indicator:

As another measure of quality, we compared between Uni, Bi and Tri grams relative frequencies in the different classes of Wikipedia, We calculated Kullback-Leibler (KL) distance measure[11,13] between HQ (Feature + Good) articles and Random articles. DKL metric is the sum of absolute difference between relative word frequencies in the compared smaller corpora (Equation (1)).

$$D_{KL}(P,Q) = \sum_i P(i).log\frac{P(i)}{Q(i)} \qquad (1)$$

Where *P(i)* is a value of a n-gram *"i"* in *HQ* class and Q(i) is the value of a n-gram "*i"* in *Random* class.

Table I holds the DKL calculations for Uni-, Bi- and Tri-grams in *HQ* vs *Random* classes.

Fig. 4 shows how close the relative frequency of confusion letters ({ة،ه},{ى،ي},{ا،أ،إ}) in both HQ and Random classes.

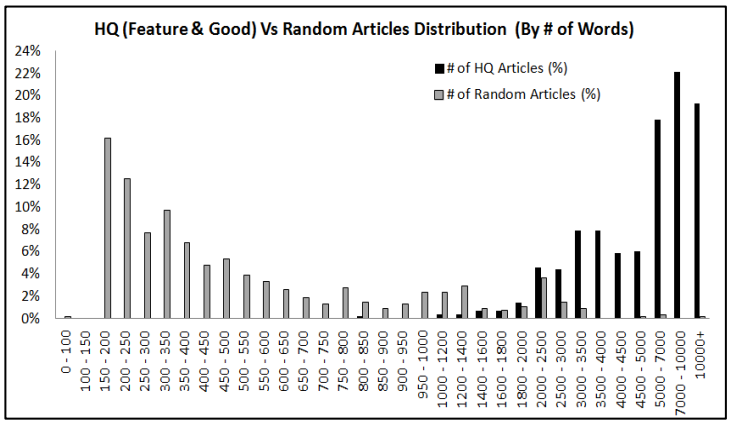

Fig. 3. HQ vs Random Articles Distribution (size in words)

TABLE I. DKL FOR N-GRAMS (HQ VS RANDOM)

Character N-gram	DKL
Uni-grams	0.0061318
Bi-grams	0.0236057
Tri-grams	0.0741143

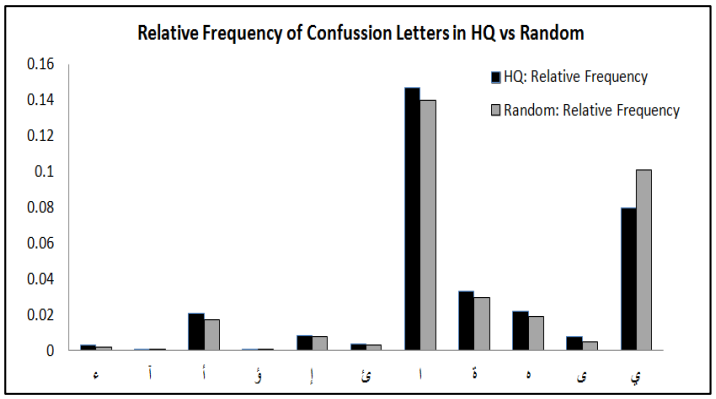

Fig. 4. Relative Frequency of Confusion letters in HQ vs Random

As can be noticed from Table I, the values of DKL distance are small which indicates a similar use of n-grams in both classes, which enables us to conclude that there is no real difference in letter distribution between these quality classes. While this parameter in not discriminating for Wikipedia articles (reflecting a generally good writing style) that is not the case when we considered other web content. Non-Wikipedia material exhibited more writing/spelling errors especially for the confusion letters, indicating worse quality. Meaning character N-grams may still be a quality parameter for general web content.

3) Machine learning based experiments:

We used WEKA machine learning toolkit to classify articles according to quality. We worked with the texts of *Feature* and *Good* articles (together give the *HQ*) and an equal sized *Random_Feature* and *Random_Good* (together give the *Random*) sets. The results are summarized in Table II. For the four way classification with *Feature*, *Good*, *Random_Feature*, *Random_Good* the best precision was around 60%. This is understandable since we don't think the two classes of *Random* are distinguishable. For the three way evaluation: *Feature*, *Good* and *Random* the precision was close to 83%. However, it was 94% for *Random*, 69% for *Feature* and 76% for *Good*, meaning most confusion is between *Good* and *Feature*. For the *Good* and *Feature* classes alone the precision was 72%, with 66% of *Feature* articles predicted correctly and 79% for *Good*. For the two way evaluation: *HQ* and *Random* the precision was close to 95%. The results are summarized in Table II.

We also tested 8K articles (Economics and Politics Wikipedia articles) using the two way approach (*HQ*, *Random*) with our set of *Feature/Good* articles and an equal number of *Random* articles as the training set. Around 14% were judged as *HQ*, a much higher percentage than the general percentage of known *HQ* articles in the Wikipedia

.

TABLE II. WEKA TEXT BASED EXPERIMENTS

Quality Classes	Success Rates (Precision)
Feature,Good, Random_Feature and Random_Good	60%
Good, Feature and Random	83%: 94% (Random) 69% (Feature) 76% (Good)
Feature and Good	72% 66 (Feature) 79% (Good)
HQ and Random	95%

4) Style and punctuation:

We believe that high quality content has better writing style compared to general content. Some of the parameters of good style are 1. Error rate also based on confusion letters which we proved almost neutral in the Arabic Wikipedia, 2. Sentence length and 3. Use of punctuation marks.

We studied the use of punctuation for the *Feature* and *Good* articles and an equal number of *Random* articles. Table III gives a summary of the results. One can observe that average sentence length for *HQ* articles is lower than that of *Random* (24 vs. 28.6) with minor variations within each class. The average distance between commas for HQ articles is close to *Random* (9.7 vs. 11). This means that the writing style doesn't exhibit a real difference between the Wikipedia quality classes. The average number of paragraphs for *HQ* articles is several times that of *Random* articles (62 vs. 7) reflecting longer articles in the *HQ* class. However, average paragraph length in words doesn't reflect a big difference (113 vs. 90) which also doesn't reflect a much different writing style. Here again we are coming to the conclusion that writing style in the case of Wikipedia may not be deciding in defining higher quality articles.

5) Part of Speech (PoS) tags:

Another parameter we tested is the PoS tags of article words. We used Stanford PoS Tagger to label words in the Arabic Wikipedia.[5] We split each article (in *HQ* and *Random*) into sentences, and passed the sentences (each alone) to the tagger. The output is a PoS tag for each word in the article. Table IV shows the results for the various quality classes. There seems to be minor differences between *HQ* and *Random.*

TABLE III. SENTENCES AND PARAGRAPHS STATISTICS HQ VS RANDOM

Content Type	Avg. Sentence Length	Avg. Number of words/ Comma	Avg. Number of Paragraphs	Avg. Paragraph Length (Words)
Feature	24.02	9.6	77.67	118
Good	21.52	10.23	45.93	102
HQ	**23.16**	**9.7**	**62.53**	**113**
Random/F	30.63	10.7	7.78	97
Random/G	26.08	11.45	6.73	81
Random	**28.59**	**11.00**	**7.28**	**90**

[5] http://nlp.stanford.edu/software/tagger.shtml

TABLE IV. POS TAGS AS PERCENTAE FOR THE DIFFERENT CLASSES

Quality Class → / PoS Tag	Feature	Good	HQ	Random
Noun	62.9%	62.9%	62.9%	69.4%
Verb	12.0%	11.20%	11.7%	10.0%
Adjective	10.9%	11.5%	11.1%	10.0%
Adverb	0.43%	0.47%	0.44%	0.27%
Others	13.77%	13.93%	13.86%	10.33%

B. Nontextual Content:

1) Multimedia elements:

Given the prevalence of Multimedia in the Internet, one of the quality measures of Web content is the non-textual component including pictures, info tables, video and sound files. One may also include foreign language content. Table V lists the average of some of these parameters for our classes.

2) Links:

Links are important in web content evaluation. One needs to care about inbound links: from other sources to the document and outbound links from the document to other sources. Even internal links reflecting navigation ease can be indicators of good style and maneuverability. Thus we will make an effort to compute/estimate each of these parameters for our articles. One needs to appreciate that not all links are equal: links to/from good content are better indicators of good quality than links to/from lower quality material. For Wikipedia, we calculated the number of inbound links (called back links) leading from other Wikipedia articles to the article under investigation. We also calculated the number of internal Wikipedia links (how many links from the current article points to other Wikipedia articles); plus external links (how many links in the current Wikipedia article points to links outside the Wikipedia domain), and finally language links (how many links to similar articles in other languages for the current article). Table V lists the average of these parameters for our quality classes.

3) Categoires and author expertise:

Categories are the tags that any Wikipedia article is tagged with. Categories can be a good indicator of how much an article is specialized and detailed. Table V also list the average number of categories found in our quality classes. We plan to use the intersection of article categories and author categories as a measure of author expertise in the article topic.

We can notice the difference between averages for the high quality classes and the random class. It's obvious that the more connected the article the more the quality of the article. This is also reflected in Multimedia and categories parameters, the more multimedia the higher the quality class.

TABLE V. NONETEXTUAL PARAMETERS (AVERAGES)

Quality Classes	Back links	Internal links	External links	Languages Links	Multi Media	Categories
Feature	1127	511	89.5	76	32	14.0
Good	339	291	51	53	15	11.3
HQ	750	406	71	65	24	12.6
Random	141	164	3	16	1	5.9

C. Contributors and Edits:

There is plenty of evidence that contributors and the number of edits are major defining factors in the quality of Wikipedia articles. The quality of contributors can vary and is generally based on the quality of their output: in our case the quality of articles they contribute to. Thus, author "quality" is both defined by and influences the quality of their contributed articles. Another factor in defining contributors' quality is their community: the quality and number of co-contributors. Human contributors are divided into anonymous and registered. The latter can be tracked and their connections studied, thus we give preference to such contributors when measuring author quality.

Table VI lists different article quality metrics based on contributors and seems to suggest that *Feature* and *Good* articles have much more contributors and HQ authors than *Random*. Also it's noticeable that the edit activities (whether in size of edits, number of edits and average times between edits) is higher in HQ content. The more the better.

IV. Conclusions and Future Work

In this paper we addressed the issue of assessing the quality of Arabic Wikipedia articles. Quality assessment is a nontrivial task and it is better done by humans. However, on the web scale human assessment is not practical and machine assessment is an invaluable tool. Machine assessment can be an end by itself but can also be an element of a more elaborate process for further assessment by humans or for efforts at improving article quality. We believe that in the age of the Big Data the user can be greatly helped by making quality as a factor in answering his/her needs, be it through quality based ranking or limiting search to reasonable quality content.

The parameters used in quality assessment are varied and keep evolving. We discussed several of them here and reported on only some for space considerations.

A big challenge is how to combine the various criteria to come up with a single quality indicator[16]. Both machine learning methods and rule based methods (and combinations) can be useful to that end. Some further work will focus on exploring and formalizing additional quality parameters for assessment like Google scholar and social media presence.

We are also interested in cross lingual quality assurance where we care about content in different languages and it quality in each.

TABLE VI. Contributors and Edits: Average per article by Quality Class

Quality Classes → **Parameters (Per Article)**	**Feature**	**Good**	**HQ**
# Registered of Contributors	59	32	46
# of Bot Contributors	31	22.5	27
# of Anonymous Contributors	79	29	55
# of Contributors (Total)	**170**	**83**	**128**
# of Contributors who authored HQ articles	28	17	23
Edit Size (Bytes)	727	935	826
Time between edits (days)	6.3	11.5	8.8
# of edits (by Registered users)	436.5	185	317.5
# of edits (by Bots)	100	64.5	83
# of edits (by Anonymous)	119	45.5	84

For improving Arabic content through consulting, or even translating, good quality material in other languages. Also our current work goes beyond the Wikipedia to cover general Arabic web content with its large diversity, where quality considerations have more value and we believe that work on Wikipedia quality can be of great use.

ACKNOWLEDGMENTS: The authors acknowledge the support of the Palestinian Ministry of Higher Education through grant number 22/1/2013 and thank the 3 referees for their useful comments.

References

[1] B. Adler, L. de Alfaro, I. Pye, and V. Raman. "Measuring Author Contributions to the Wikipedia," in *Proceedings of the 4th International Symposium on Wikis*. ACM, 2008.

[2] M. Al Huziah, M. Al Kahtany, R. Al Ammari, R. Al Faiz. "Assessment of Online Health Information for Arabic Sites", *A Report by King Saud University for Health Science.* Saudi Arabia. November 2009.

[3] J. Blumenstock. "Size matters: word count as a measure of quality on Wikipedia" in *Proceedings of the 17th international conference on World Wide Web (WWW '08)*. ACM, April 2008, pp. 1095-1096.

[4] I. Casebourne, C. Davies, M. Fernandes and N. Norman. "Assessing the accuracy and quality of Wikipedia entries compared to popular onlineencyclopaedias: A comparative preliminary study across disciplines in English, Spanish and Arabic". Epic, Brighton, UK. http://commons.wikimedia.org/wiki/File:EPIC_Oxford_report.pdf

[5] G. De La Calzada and A. Dekhtyar. "On Measuring the Quality of Wikipedia Articles".In *Proceedings of the 4th Workshop on Information Credibilty: WICOW'10.* April 26-30, 2010. Raleigh, NC. PP. 11-18.

[6] ESCWA. "Status of the Digital Arabic Content Industry in the Arab Region", *A Report by United Nations Economic and Social Commission for Western Asia. Saudi Arabia*. November 5, 2012.

[7] M. Gazal. "Most online Arabic Content Lacks Accuracy, Quality" *Jordan Times*. Nov. 8, 2012.

[8] S. Javanmardi and C. Lopes. "Statistical Measure of Quality in Wikipedia". In *Proceedings of the First ACM Conference on Social Media Analytics*. ACM, 2010, PP. 132-138.

[9] N. Lipka and B. Stein. "Identifying Featured Articles in Wikipedia: Writing Style Matters"; *WWW 2010*. Poster April 26-30, 2010. Raleigh, NC. PP. 1147-1148.

[10] H. Liu, E. Lim, H. Lauw, M. Le, A. Sun, J. Srivastava, and Y. Kim. "Predicting trusts among users of online communities: an epinions case study". In *EC'08: Proceedings of the 9th ACM conference on Electronic commerce.* ACM, 2008, pp. 310-319.

[11] D. Pinto, P. Rosso, and H. Jimenez-Salazar, "On the Assessment of Text Corpora". In *Proceedings of the 14th International Conference on Applications of Natural Language to Information Systems NLDB'09*, Saarbruecken, Germany, PP.281—290.Springer-Verlag. 2009.

[12] X. Qin and P. Cunningham. "Assessing the Quality of Wikipedia Pages Using Edit Longivity and Contributor Centrality". arXiv Preprint. June 12, 2012.

[13] P. Rosso, Y. Benajiba, A. Lyhyaoui, "Towards a Measure for Arabic Corpora Quality". In *Proceedings of the 4th Conf. on Scientific Research Outlook and Technology Development in the Arab world, SROIV*, Damascus, Syria. PP11-14. 2006.

[14] K. Saengthongpattana and N. Soonthornphisaj. "Thai Wikipedia Quality Measurement using Fuzzy Logic" . *The 26th Annual Conference of the Japanese Society for Artificial Intelligence*, 2012.

[15] B. Stvilia, A. Al-Faraj and Y. Yi. "Issues of Cross-Contextual Information Quality Evaluation—The Case of Arabic, English, and Korean Wikipedias". *Library and Information Science Research*, 31(4), 232-239. 2009.

[16] M. Warncke-Wang, D. Cosley and J. Riedl. "Tell Me More: An Actionable Quality Model for Wikipedia", In *Proceedings of WikiSym '13*, Aug 05-07 2013, Hong Kong, China. 2013.