Improved Automatic Maturity Assessment of Wikipedia Medical Articles*
(Short Paper)

Emanuel Marzini, Angelo Spognardi,
Ilaria Matteucci, Paolo Mori, Marinella Petrocchi, Riccardo Conti

IIT-CNR, Pisa, Italy
firstname.lastname@iit.cnr.it

Abstract. The Internet is naturally a simple and immediate mean to retrieve information. However, not everything one can find is equally accurate and reliable. In this paper, we continue our line of research towards effective techniques for assessing the quality of online content. Focusing on the Wikipedia Medicinal Portal, in a previous work we implemented an automatic technique to assess the quality of each article and we compared our results to the classification of the articles given by the portal itself, obtaining quite different outcomes. Here, we present a lightweight instantiation of our methodology that reduces both redundant features and those not mentioned by the WikiProject guidelines. What we obtain is a fine-grained assessment and a better discrimination of the articles' quality, w.r.t. previous work. Our proposal could help to automatically evaluate the maturity of Wikipedia medical articles in an efficient way.

1 Introduction

Recent studies report that Internet users are growingly looking for health information through the Web, by either consulting search engines, social networks, and specialized health portals: in 2013 [14], "one in three American adults have gone online to figure out a medical condition". Further, in 2013 almost one million US families used video consultations with physicians, mainly through dedicated web portals [12]. The quest for medical information is eased by a myriad of Internet websites with health-related hypertexts. For example, the Wikipedia Medicine Portal is a collaboratively edited multitude of articles concernin health whose consultation spans from patients to healthcare professionals [13, 7]. According to a report on online engagement by IMS Health (a world's leading company dedicated to healthcare), 50% of surveyed physicians who use the Internet have consulted Wikipedia for medical information [1].

However, finding reliable medical articles is an important issue that is worth addressing. For instance, the same survey [14] reports that, on the totality of

* Work partially supported by the Tuscany region projects MyChoice and PICs and by the Registro.it project My Information Bubble.

people that searched for health answers on the Web, only 41% say "a medical professional confirmed their diagnosis". Both government departments and scientific reports have recently raised reliability issues of online seeking health information, see, *e.g.,* [16, 19].

This paper proposes an automatic approach for the evaluation of online articles for assessing their quality. Our data-set is the entire collection of articles published on the Wikipedia Medicine Portal. In [4], we proposed a newly-defined metric, the *maturity degree*, to evaluate the quality and reliability of each article. The maturity degree was calculated adopting the Analytic Hierarchy Process (AHP) [15], a well-known methodology for multi-criteria decision making. We showed that the maturity degree is a different metric with respect to the quality level attached to articles by the WikiProject quality assessment. We concluded that a gap exists between the quantitative features that can be computed as metadata of an article and the qualitative features exploited by the quality assessment process of the portal. However, in order to use automatic techniques for article evaluation (like the approach shown in [4]), making use of only quantitative features would greatly ease the process.

Starting from these premises, in the current paper we contribute by 1) pruning the list of features considered for the automatic evaluation of the article (w.r.t. [4]). Then, we experimentally prove thatfeatures, we eliminate some extra information not directly leading to a fine quality evaluation; 2) exploiting the *cosine-similarity* to compare the results obtained with the restricted set of features with respect to the results with the whole set of features. We find out that, besides being more efficient, the new approach also achieves better results in evaluating the maturity of the articles w.r.t. our previous instantiation.

The paper is organized as follows: next section discusses related work. Section 3 briefly recalls the notion of maturity degree and presents our new results, comparing them with previous ones. Section 4 concludes the paper.

2 Related Work

A series of recent work focuses on the assessment of Wikipedia articles, testifying the quest for effective and efficient techniques supporting the community to identify the best-quality material. WikiProject itself has listed a set of criteria to be manually evaluated, useful to determine the quality of an article. In most cases, such criteria express qualitative properties more than quantitative ones, like, for example, comprehensiveness and neutrality. Undoubtedly, such properties are of particular relevance for the evaluation of an article. However, considering them could complicate the process of automatizing the assessment. Instead, this paper focuses on articles features that can be automatically extracted and processed in order to globally evaluate the articles quality level.

Work in [17, 2, 3, 6, 23, 21] mainly concerns the recognition of featured articles (FA) (they are those articles representing, according to WikiProject, excellent contributions). Recognition is mainly based on features like the number of times an article has been edited (edit time), the word count, and the number of editors.

Similarly [18] and [9] propose to evaluate the articles exploiting the relation between editor and text quality.

Works in [20, 22] rely on a different approach to assign an article to one of the existing WikiProject classes. In [22], the authors consider 28 criteria, grouped into four macro-criteria: lingual, structural, historical, and reputational. They use seven different neural networks for the classification of each article. Overall, each criterion is differently weighted according to the considered class, *e.g.,* linguistic criteria are more important than others to recognize articles in the lowest classes, while richness of content and articulated structure are important for articles of the highest classes. In [20], the authors use also the Wikipedia template messages (small notes to inform readers and editors of specific problems within articles or sections) as new features to assess the quality of the articles.

In line with the results of [21, 2, 17], which focus on a few number of criteria, in the present work we improve our previous approach in [4], by reducing the number of features the assessment takes into account. It is also worth noticing that, rather than assigning an article to one of the existing WikiProject classes, our goal is to evaluate the relevance of each article with respect to all the classes.

3 Assessment results

The Wikipedia Medicine Portal community manually assesses the quality level of the published articles, to aid the recognition of excellent contributions and identify topics that instead need further work. The six quality classes are: (1) Stub, (2) Start, (3) class C, (4) class B, (5) Good, and (6) Featured article. The Featured and Good article grades are the highest possible assessments and they require a community consensus and an official review, while all the others can be achieved with a simple review.

In [4], we assessed the maturity degree of all the articles published on the Wikipedia Medicine Portal (24,418 at the time of our study). This dataset is distinctive because it is composed of heterogeneous content, from very short drafts till comprehensive articles with a complex structure and a technical dictionary.

For the assessment, we exploited AHP [15], a multi-criteria decision making technique, which has been largely used in several fields. It helps making decisions when several different *alternatives* can be chosen to reach a *goal*. AHP is able to order the alternatives from the *most relevant* to the *less relevant*, with respect to a set of *criteria* and *subcriteria*, proceeding with a divide and conquer approach.

In our instantiation, the alternatives are the six quality classes and the output of AHP is a vector representing a new metric that we call *maturity degree*. Criteria and subcriteria of the hierarchy are quantitative features of the article and are listed below in this section. Noticeably, we do not use AHP to classify Wikipedia articles as belonging to a single class. Rather, having the maturity degree vector $v = [v_i]$, each v_i represents the relevance of the article to the corresponding WikiProject class i ($i = 1$ is Stub, 2 is Start, and so on). Similarly to the property of unimodality of a function, we say that a maturity degree vector is *consistent* when it has exactly one absolute maximum, i.e., if for some value

m, it is monotonically increasing for $i \leq m$ and monotonically decreasing for $i \geq m$. The property of consistency of the vector ensures that the relevance is maximal either for only one class or for neighboring classes.

Due to page limits constraints, we invite the interested reader to look at the extended version of this paper [10] for a detailed description of the AHP methodology and its instantiation leading to the maturity assessment.

In this paper, we first reduce the number of subcriteria in the AHP instantiation with respect to [4], and we secondly evaluate the goodness of the newly obtained maturity degree by relying on the cosine similarity (*cosSim*). The *cosSim* is commonly used in Information Retrieval and text mining to evaluate the similarity of two multi-dimensional vectors v_i and v_j, and it is defined as:

$$cosSim(v_i, v_j) = \frac{v_i \cdot v_j}{\sqrt{{v_i}^2}\sqrt{{v_j}^2}}$$

Since the maturity degree of an article is always a vector with positive components, the *cosSim* ranges over [0,1]. We called *cross cosine similarity* (cr*cosSim*) and *class cosine similarity* (cl*cosSim*), respectively, the average *cosSim* between all the pairs of vectors of articles that on WikiProject belong to different classes, and to the same class, respectively. Intuitively, we expect that: *i)* the maturity degree vectors of articles belonging to different WikiProject classes have lower similarity than those of articles of the same class; and *ii)* the more the two classes are distant, the lower the similarity will be. Formally, given two WikiProject classes C_1 and C_2 (with $C_i \in$\{*Stub, Start, Class C, Class B, Good article, Featured article*\} and $i \in \{1, 2\}$) and the maturity degree v_i of an article, then:

$$cr\,cosSim(C_1, C_2) = \frac{1}{|C_1| \cdot |C_2|} \sum_{v_i \in C_1} \sum_{v_j \in C_2} cosSim(v_i, v_j) \tag{1}$$

$$cl\,cosSim(C_1) = \binom{|C_1|}{2}^{-1} \sum_{v_i, v_j \in C_1, i \neq j} cosSim(v_i, v_j) \tag{2}$$

In a more formal way, if an index from 1 to 6 represents each WikiProject class, we should observe that $cr\,cosSim(C_i, C_j) < cr\,cosSim(C_i, C_k)$ when $i < j < k$. Moreover, a $cl\,cosSim$ higher than 0.95 would mean that our maturity degree mimics the WikiProject classification.

In the following, we report our experimental results, on the whole database of 24,418 articles. In our previous work we considered four criteria inspired by [17, 22]: *lingual, structural, historical* and *reputational*. Subcriteria belonging to the *lingual* criteria were: (1) Flesch reading ease and (2) Flesch-Kincaid grade level; (3) word count and (4) sentence count; (5) multi-syllable words / words ratio; (6) spell error / words count ratio. *Structural* subcriteria were: (1) number of categories; (2) internal and (3) external links; (4) non-textual resources; (5) further readings; (6) number of symbols in title; (7) section headings count; (8) number of citations. The *historical* criterion was made by sub criteria: (1) edit counts (times that the article has been edited); (2) editor count (number of different users that edited the article); (3) number of devoted editors ratio; (4)

	considered classes	average *cosSim* no AHP	average *cosSim* AHP 25 subc. ([4])	average *cosSim* AHP 9 subc.
class *cosSim*	Stub - Stub	0.88	0.98	0.95
	Start - Start	0.91	0.93	0.87
	Class C - Class C	0.93	0.92	0.85
	Class B - Class B	0.90	0.92	0.87
	Good Art. - Good Art.	0.89	0.94	0.89
	Feat. Art. - Feat. Art.	0.89	0.96	0.95
cross *cosSim*	Stub - Start	0.89	0.87	0.76
	Stub - Class C	0.89	0.75	0.59
	Stub - Class B	0.84	0.63	0.46
	Stub - Good Art.	0.83	0.59	0.40
	Stub - Feat. Art.	0.77	0.53	0.34
	Start - Class C	0.92	0.89	0.81
	Start - Class B	0.88	0.79	0.71
	Start - Good Art.	0.86	0.74	0.64
	Start - Feat. Art.	0.79	0.67	0.55
	Class C - Class B	0.90	0.88	0.83
	Class C - Good Art.	0.88	0.85	0.79
	Class C - Feat. Art.	0.83	0.79	0.73
	Class B - Good Art.	0.89	0.92	0.87
	Class B - Feat. Art.	0.86	0.90	0.86
	Good Art. - Feat. Art.	0.88	0.94	0.91

Table 1. Average similarity for the articles belonging to the same WikiProject class (class *cosSim*) and to distinct classes (cross *cosSim*), with different settings

anonymous editors ratio; (5) minor edits ratio; (6) article age; (7) edit frequency. The *reputational* subcriteria were: (1) average active age of editors; (2) average upload amount of editors; (3) average edit times of editors; (4) average talk times of editors. Hereafter, we validate *i)* the choice of using AHP for assessing the maturity of an article (Section 3.1), *ii)* the results obtained in [4] (Section 3.2), and *iii)* the refined version of the approach proposed here (Section 3.4).

3.1 Similarity of results with 25 subcriteria without AHP

The first experiment is aimed to verify that AHP effectively helps on making decision about the maturity assessment of a medical article. Hence, in this experiment we don't use AHP and we evaluate the *cosSim* of the articles directly on the value of the 25 subcriteria. In particular, to compute the similarity of two articles, we compute the *cosSim* of the two vectors reporting the values of the 25 normalized features corresponding to two articles. The average similarities are reported in the first column of Table 1. It is clear that the vectors of all the classes exhibit a very high similarity. Some classes, indeed, exhibit also higher values of cross *cosSim* than class *cosSim*: Start, for example, has a class *cosSim* of 0.91, but a cross *cosSim* of 0.92 with Class C. This experiment confirms that the straightforward approach of considering the statistic distribution of the 25 features can not produce accurate results. A better approach that takes into account a finer weight of the different features is, then, advisable in order to better deal with the statistical fluctuation of the features among the different classes.

3.2 Similarity of results with 25 subcriteria

The second experiment evaluates the similarity of the maturity degrees obtained in [4]. The results are shown in the second column of Table 1. Comparing the similarity of the normalized feature vectors with the similarity of the maturity degree obtained in [4], we can appreciably observe an increase of the class *cosSim* and a decrease of the cross *cosSim*. In particular, we notice that the class *cosSim* always has values above 0.92, meaning that the maturity degrees of the articles within the same WikiProject class are very similar among them. However, this may produce a low granularity characterization of the maturity of different articles. We also observe that, as expected differently from the previous experiment, the cross cosine similarity decreases as the two considered classes are distant among them. For example, the average *cosSim* of articles in class Stub decreases as the other class is more distant: $cr\,cosSim(Stub,\ Class\ C) = 0.75$, while $cr\,cosSim(Stub,\ Feat.\ Art.) = 0.53$.

3.3 Reducing the number of subcriteria

Many studies have shown that AHP performs better with few criteria [11, 15]. Here, we reduce the number of criteria having a twofold beneficial effect: it speeds up the decision making process and improves its results in terms of quality.

To reduce subcriteria, we consider several aspects: adhesion to the Wikipedia guidelines and reduction of redundancy. We eliminate the whole reputational criteria, not considered by the guidelines. Then, we adopted the *mutual information* measure (or *information gain*) to evaluate whether a feature brings more information. This measure evaluates the dependency of two random variables: the mutual information $I(X;Y)$ represents the reduction in the uncertainty of X due to the knowledge of Y [5]. It is defined as $I(X;Y) = H(X) - H(X|Y)$ where $H(X) = -\sum_x p(x) \log p(x)$ is the usual definition of the *entropy* of a random variable X and $H(X|Y)$ is the *conditional entropy* of X given Y [8].

For the lingual criterion, we started removing spell error, because of its bias against complex and composite words, typical of the medical terminology. Then, we noticed that some of the lingual subcriteria actually consider the same property of an article. In particular, word count and sentence count are tightly related one with each other, considering the length of the article. Similarly, Flesch reading ease and Flesch-Kincaid grade level both consider the complexity of the articles. Considering both sentence count and word count has the effect of doubling the influence of the article length property on the final outcome. The same happens for the two features that consider the article complexity. Then, we removed the sentence count and the Flesch-Kincaid grade level, since their mutual information with word count and Flesch reading ease (namely the amount of information gained about Y after observing X) is very high, 0.83 and 1.39, respectively. This would lead to overestimate the same property.

To decide which of the subcriteria belonging to the structural and historical criteria should be removed, for each article we computed the mutual information

feature	mutual information	removed
structural features		
section headings count	0.59	no
internal links	0.52	no
number of citations	0.43	no
non-textual resources	0.12	yes
spell error / words count ratio	0.10	yes
external links	0.10	yes
further readings	0.02	yes
number of categories	0.01	yes
number of symbols in title	0.00	yes
historical features		
edit count	0.44	no
edit frequency	0.43	no
editors count	0.36	no
anonymous editors ratio	0.13	yes
article age	0.13	yes
number of devoted editors ratio	0.03	yes
minor edits ratio	0.05	yes

Table 2. Mutual information of structural and historical features w.r.t. the WikiProject article class. It captures the dependency of the feature w.r.t. the class.

of each criterion w.r.t. the article class. In this way, we identified a set of subcriteria that do not provide any help to the decision process, namely those that exhibit a uniform distribution of the articles among all the WikiProject classes, with the lowest mutual information w.r.t. the article class. In particular, the removed features are reported in Table 2, jointly with their mutual information w.r.t. the article class: the more the value is close to 0, the more the feature is unrelated with the class assigned by WikiProject. Such subcriteria, actually, only introduce random noise and, consequently, complicate the decision process.

The procedure leads us to only 9 subcriteria: 1) Flesch reading ease (FR), 2) word count (WC), 3) multi-syllable words / words ratio (MS), 4) internal links (IL), 5) section headings count (SHC), 6) number of citations (NoC), 7) edit count (EditC), 8) editors count (EditorsC), and 9) edit frequency (EF).

3.4 Maturity assessment with 9 subcriteria

We apply AHP with 9 subcriteria and re-compute the maturity degree of the articles of our dataset (see Figure 1). We remind the reader that the full AHP instantiation leading to the maturity degree is described in [10]. Technically, the steps leading to the results are available online at http://goo.gl/li6CpI.

Table 3 compares the consistency (as defined at the beginning of this section) of the new results with the old ones. Firstly, we observe that a slightly higher consistency ratio holds for the new results (it is almost perfect for Stub articles). The third column of Table 1 reports the class and the cross similarity for the new results. It is evident the reduction of cross similarity between the different classes, but also a reduction of the class similarity. We can consider those two phenomena as two different beneficial effects. Firstly, the cross similarity reduction happens because the new assessment it is able to better discriminate between articles belonging to different WikiProject classes, better capturing the

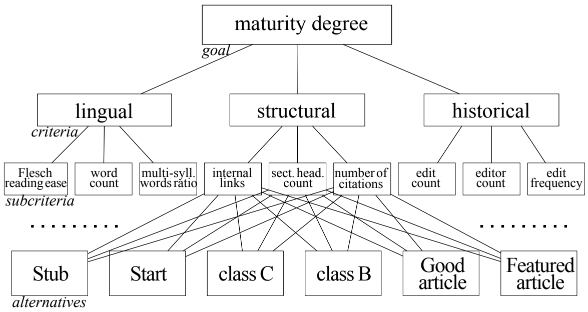


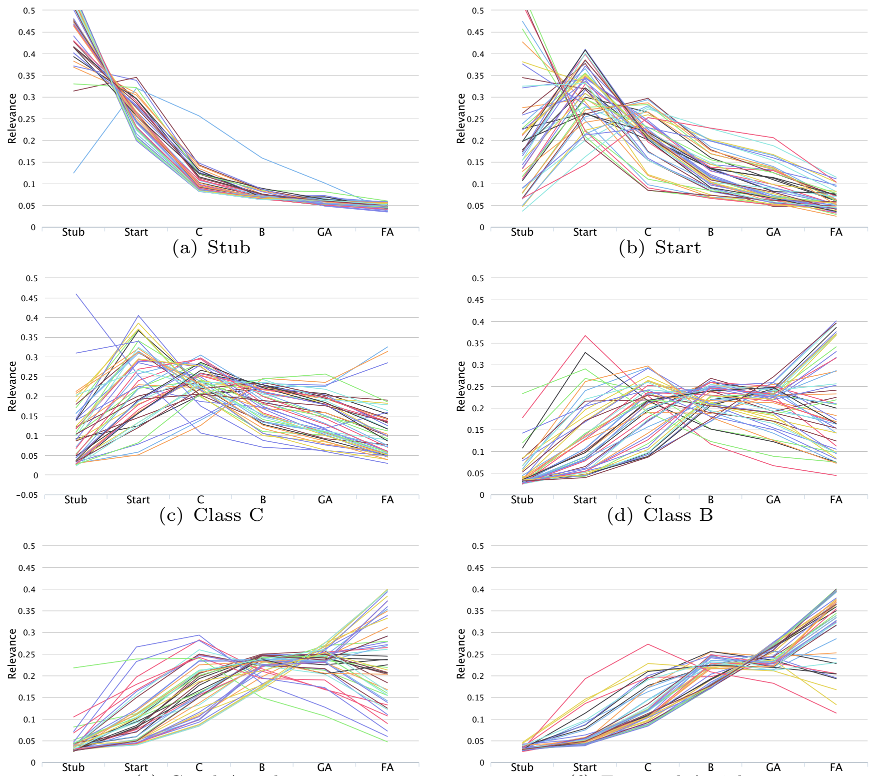
Fig. 1. AHP hierarchy with 9 subcriteria

	consistent maturity degrees	
WikiProject Class	25 subcriteria	9 subcriteria
Stub	95%	99%
Start	98%	98%
Class C	94%	93%
Class B	88%	87%
Good Article	94%	96%
Featured Article	86%	88%

Table 3. Percentage of articles that obtained a consistent maturity degree

varied maturity degrees within the classes. Secondly, the sensible increase of the class *cosSim* means the ability to more precisely characterize the articles belonging to the same class, providing finer and more specific levels of maturity. This can be ascribed to the reduction of redundancy and to the adoption of the intervals, as described in details in [10]. With a high redundancy, features that capture the same aspects (like the article length in case of words and sentences count) overemphasize them, vanishing the smaller differences introduced by the features that consider other aspects. When the overemphasized aspects are also considered more relevant during the AHP process, this effect of flattening is further stressed. Another inherent contributing factor is the adoption of the intervals: articles with two subcriterion values falling within the same interval but near to its two opposite ends, have more similar final maturity degrees. This flattening effect is reduced with the new assessment, since the results for the articles within the same class are slightly different among them, leading to a finer granularity of the maturity degrees.

Figure 2 shows a summary of some results of the new assessment, in order to give a glance of the obtained maturity degrees. The figure does not intend to detail the results for each of the analyzed articles, but only to highlight the general agreement of the results for 300 considered articles within some of the different classes. In particular, for each WikiProject class, we randomly sample 50 articles belonging to that class and draw their resulting maturity degree as a line following the relevance of each class. As for the original work, we have some articles with a maturity degree significantly different from the WikiProject class they belong to. For example, in Figure 2(f) that shows the results for



(a) Stub (b) Start (c) Class C (d) Class B (e) Good Article (f) Featured Article

Fig. 2. Maturity degree with respect to the WikiProject assessment, 9 subcriteria. Each line is the maturity degree of an article belonging to a given WikiProject class.

Featured Articles, we can notice a couple of articles that have their maximum relevance on the C quality class: this reflects the fact that the manual assessment by WikiProject considers also qualitative guidelines, as the neutrality and the comprehensiveness, hard to compute in a quantitative way.

Summarizing, reducing the subcriteria set leads us to an efficient application of AHP, as discussed in [11, 15]. Further, the new set of criteria yields a finer assessment of the articles belonging to the same WikiProject classes and a more evident separation between the articles belonging to different classes.

4 Conclusions

This paper enhances our previous automatic assessment of Wikipedia medical articles. We refined our AHP-based approach by identifying and pruning redundant features. In this way, we obtained fine-grained results to evaluate the relevance of each article with respect to the WikiProject classes. To validate our results, we computed the similarity of each pair of articles exploiting cosine similarity. We observed that, with a reduced set of features with respect to the one in [4], the average cross-similarity of articles (those belonging to two distinct classes) is lower, leading to a more evident separation into classes. The average class-similarity for articles of the same classes is also lower, possibly yielding a finer intra-class assessment. This led us to conclude that the new assessment with the reduced set of features better discriminates the articles, since their evaluation is, at the same time, closer to the one given by WikiProject and fine-grained, with the added-value of an automatic process behind the evaluation outcome.

References

1. M. Aitken, T. Altmann, and D. Rosen. Engaging patients through social media. Technical report, IMS Institute for healthcare informatics, Jan 2014. http://goo.gl/BoFJA8, Last checked: August 27, 2014.
2. J. E. Blumenstock. Automatically assessing the quality of Wikipedia. In *UC Berkeley: School of Information Report 2008-021*, 2008.
3. J. E. Blumenstock. Size matters: Word Count as a measure of quality on Wikipedia. In *17th International Conference on World Wide Web*, 2008.
4. R. Conti, E. Marzini, A. Spognardi, I. Matteucci, P. Mori, and M. Petrocchi. Maturity assessment of Wikipedia medical articles. In *Computer-based Medical Systems*, CBMS '14. IEEE, 2014.
5. T. M. Cover and J. A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006.
6. H. Dalip et al. Automatic quality assessment of content created collaboratively by web communities: A case study of Wikipedia. In *9th ACM/IEEE-CS Joint Conference on Digital Libraries*. ACM, 2009.
7. C. A. Haigh. Wikipedia as an evidence source for nursing and healthcare students. *Nurse Education Today*, 31(2), 2011.
8. R. W. Hamming. *Coding and Information Theory (2Nd Ed.)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1986.
9. M. Hu, E.-P. Lim, A. Sun, H. W. Lauw, and B.-Q. Vuong. Measuring article quality in wikipedia: Models and evaluation. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, pages 243–252, New York, NY, USA, 2007. ACM.
10. E. Marzini, A. Spognardi, I. Matteucci, P. Mori, M. Petrocchi, and R. Conti. Improved automatic assessment of Wikipedia medical articles. Technical report, IIT-CNR, TR-IIT-11-2014, Aug. 2014.
11. G. A. Miller. The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. *The Psychological Review*, 63, 1956.
12. Mobihealthnews.com. Almost 1M families use video consultations with physicians last year. Mar '14. http://goo.gl/6sI0tD, Last checked: August 27, 2014.
13. R. Feltman. The #1 doctor in the world is Dr. Wikipedia. *MSN News online*, Jan '14. http://goo.gl/zhTNoa, Last checked: August 27, 2014.
14. S. Fox And M. Duggan. Health Online 2013. *Pewinternet.org*, Jan '13. http://goo.gl/vRBnCA, Last checked: August 27, 2014.
15. T. L. Saaty. Decision making with the Analytic Hierarchy Process. *Int. Journ. of Services Sciences*, 1(1), 2008.
16. E. Sillence, P. Briggs, P. R. Harris, and L. Fishwick. How do patients evaluate and make use of online health information? *Social Science and Medicine*, 64(9), 2007.
17. B. Stvilia, M. B. Twidale, L. Gasser, and L. C. Smith. Information quality discussions in Wikipedia. In *Knowledge Management*. 2005.

18. Y. Suzuki and M. Yoshikawa. Assessing quality score of wikipedia article using mutual evaluation of editors and texts. In *22Nd Conference on Information & Knowledge Management*, CIKM '13, 2013.
19. U.S. Dept. of Health and Human services. Online Health Information: Can You Trust It? . http://goo.gl/G5evGg, Last checked: August 27, 2014.
20. M. Warncke-Wang, D. Cosley, and J. Riedl. Tell Me More: An actionable quality model for Wikipedia. In *9th Intl. Symposium on Open Collaboration*. ACM, 2013.
21. D. M. Wilkinson and B. A. Huberman. Cooperation and Quality in Wikipedia. In *Intl. Symposium on Wikis*. ACM, 2007.
22. K. Wu, Q. Zhu, Y. Zhao, and H. Zheng. Mining the factors affecting the quality of Wikipedia articles. In *Inform. Science and Management Eng.*, volume 1, 2010.
23. Y. Xu and T. Luo. Measuring article quality in Wikipedia: Lexical clue model. In *Web Society (SWS), 2011 3rd Symposium on*, 2011.