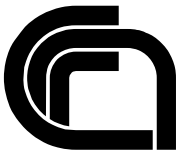

Consiglio Nazionale delle Ricerche

Improved Automatic Maturity Assessment of Wikipedia Medical Articles

E. Marzini, A. Spognardi, I. Matteucci, P. Mori, M. Petrocchi, R. Conti

IIT TR-11/2014

Technical report

Agosto 2014

Istituto di Informatica e Telematica

Improved Automatic Maturity Assessment of Wikipedia Medical Articles

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. We focus on the Wikipedia Medicine Portal, which exposes a large volume of medical articles, representing our dataset. 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 an enhanced instantiation of our methodology for evaluating the quality of such articles. Compared to the previous work, we make use of a lower number of criteria, to reduce both redundant features and those not mentioned by the WikiProject guidelines. Criteria are restricted here only to "computable" features of the article. Intuitively, even if significant, qualitative guidelines may constitute an obstacle when relying on automatic techniques that need quantifiable values as input. We validate the new assessment exploiting the cosine similarity metric. What we obtain is a fine-grained assessment and a better discrimination of the articles' quality, with respect to previous work. In our opinion, the proposed methodology could help automatically evaluating the maturity of Wikipedia medical articles in a easy and 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. As pointed out by a 2013 American survey [14], "one in three American adults have gone online to figure out a medical condition". Further, online seeking goes beyond simple reading: as an example, still in 2013 almost one million US families used video consultations with physicians, mainly through dedicated web portals [1].

Thus, on the one hand, the quest for information is eased by the fact that a myriad of websites containing health-related hypertexts exists on the Internet. For example, the Wikipedia Medicine Portal is a collaboratively edited multitude

of articles with contents often comparable with professionally edited material, whose consultation spans from patients to healthcare professionals, see, *e.g.,* [9]. Indeed, 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 [3, 13]. The portal provides a large number of articles describing diseases and injuries, each of them containing a textual description, pictures, multimedia contents, and links to related content on other webpages.

On the other hand, finding reliable medical articles in an important issue that is worth addressing and successfully solving. For instance, the above cited survey [14] reports that, on the totality of 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.,* [17, 19].

In line with such parallel research, 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 a previous work [6], the authors of this paper proposed a way to extract information from the analysis of the same data-set in order to tag all the Wikipedia medical articles, resulting in associating a newly-defined metric, the *maturity degree*, to each article. This metric summarizes in a series of numerical values the current state of each article, in terms of quality and reliability. The maturity degree was calculated adopting the Analytic Hierarchy Process (AHP) [16], a well-known methodology for multi-criteria decision making. The comparison between our calculated degree and the quality level shown on the portal, for each article, led us to draw some critical points about the automatic assessment of Wikipedia medical pages. In particular, 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 [6]), making use of only quantitative features would greatly ease the process. This paved the way for further investigation, aimed at achieving a new automatic assessment exploiting quantitative measures only.

Starting from these premises, in the current work we enhance our approach for automatic maturity assessment of articles. The contribution is two-fold. First, compared with [6], we prune the list of features considered for the automatic evaluation of the article. We experimentally proved that, using a restricted set of features, we eliminate some extra information not directly leading to a fine quality evaluation. Secondly, we exploit the metric of *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 with respect to our previous instantiation.

The paper is organized as follows: next section discusses related work in the area of automatic quality assessment of online documents. Section 3 recalls our previous approach based on AHP and the definition of maturity degree as a new metric associated to Wikipedia articles. In Section 4, we discuss our results by also comparing them with our previous ones, and Section 5 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.

Several works mainly concern the recognition of featured articles (FA) (they are those articles representing, according to WikiProject, excellent contributions). In [18], the authors show that "edit times" (the number of times an article has been edited) and readability are critical features to discriminate FA from low-quality articles. In [4], the authors consider at about one hundred criteria and conclude that, relying only on the "word count" criterion, it is possible to distinguish more than 97% of FA. Also in [5] the authors propose the word count metrics for measuring the quality of the articles and prove that this metric significantly improves the efficiency with respect to existing approaches dated before 2008. Work in [8] reports similar results on a random dataset of Wikipedia articles. In [23], eight linguistic criteria are taken as evaluation metrics. Relying on a decision tree, FA articles were distinguished from non-FA ones with a precision equal to 83%. In [21], only historical criteria like "number of editors" (anonymous or registered users that write a revision) and "edit times" have been considered to distinguish FA articles from the others.

Works in [20, 22] rely on a different approach. Their aim is 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 to distinguish 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.

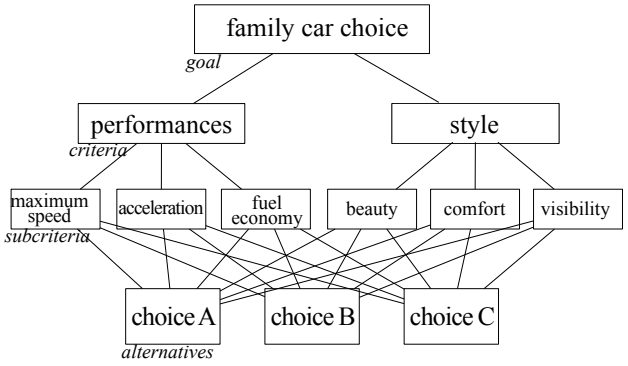

Fig. 1. AHP hierarchy for choosing a family car

In line with the results of [21, 4, 18], which focus on a few number of criteria, in the present work we improve our previous approach in [6], by reducing the number of features the assessment takes into account. As explained in Section 4.3, we extend and improve our proposal by following a more efficient approach, which avoids redundancy and strictly follows the WikiProject guidelines. 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.

Even not specifically focused on Wikipedia articles, for the sake of completeness, we acknowledge here research work on quality assessment of user-generated web content. The literature is quite extensive in this huge area. To cite one for all, work in [2] investigates methods for exploiting feedback by social media communities as a metric to automatically identify high quality content. The proposed methods were successfully tested on the Yahoo! Answers question/answering platform.

3 Settings

We recall our approach of [6], based on an instantiation of the Analytic Hierarchy Process, AHP for short, to assess the maturity degree of the medical articles published on the Wikipedia Medicine Portal. In particular, the considered dataset consists of the whole portal (24,418 medical articles at the time of our study). This dataset is distinctive mainly because it is composed of heterogeneous content, from very short drafts till comprehensive articles with a complex structure and a technical dictionary.

3.1 The Analytic Hierarchy Process

AHP [16], introduced by Saaty in the 70's, is 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. Indeed, AHP divides a complex problem into a hierarchy of sub-problems, based on a set of criteria and subcriteria. In figure 1) we have sketched the problem of choosing a car that best suits the needs of a family, organized in a hierarchy of two criteria (namely, "performances" and "style") and sub-criteria, like "max speed", "fuel economy", and "acceleration" for the performances criterion, and "beauty", "comfort", and "visibility" for the style criterion. To compute the global solution, AHP properly merges the various local solutions for each sub-problem.

In particular, once the hierarchy is built, the method performs pairwise comparisons, from the bottom to the top, in order to compute the relevance, hereafter called *local priority*: i) of each alternatives with respect to each subcriteria (for example, in terms of fuel economy, does choice A consume less than choice B?), ii) of each subcriterion with respect to the relative criterion (for example, in terms of car performance, is fuel economy more relevant than max speed?), and, finally, iii) of each criterion with respect to the goal (in terms of choosing a car, is style more relevant than performance?). Comparisons are expressed in a matricial form, called *pairwise comparison matrix*. A pairwise comparisons matrix A is a square matrix which has positive entries and satisfies the reciprocal property, *i.e.,* $a_{ii} = 1$ and $a_{ij} = \frac{1}{a_{ji}}$. For each level of the hierarchy, the entries of a pairwise comparison matrix represent how much one element at that level is more relevant with respect to another element at the same level. Relevancy is then estimated with respect to elements at the upper level, and according to a pre-defined comparison scale typical of AHP (see Table 1). The scale indicates how many times an element is *more relevant* than another one, with respect to the element at the upper level. The relevance between the criteria also allows to compare subjective characteristics (like car comfort) and objective ones (like max speed). The normalized eigenvector associated with the largest eigenvalue x of each matrix A gives a vector of local priorities, that represents the local solution: $Ax = \lambda x$ [15].

To exemplify, we can imagine that, with respect to the performances criterion, the pairwise comparison matrix of the three sub-criteria is as follows:

$$\begin{array}{c} \\ \textit{max speed} \\ \textit{acceleration} \\ \textit{fuel economy} \end{array} \begin{array}{c} \begin{array}{ccc} \textit{max speed} & \textit{acceleration} & \textit{fuel economy} \end{array} \\ \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{5} \\ 2 & 1 & \frac{1}{3} \\ 5 & 3 & 1 \end{pmatrix} \end{array}$$

where "fuel economy" is the most relevant criterion and "max speed" is the lowest one. In particular, "fuel economy" has been evaluated five times more relevant than "max speed" and three times more relevant than "acceleration", in the above example. This matrix provides a local priority of $(0.122, 0.23, 0.648)$.

Intensity	Description	Explanation
1	Equal	Two elements contribute equally to the objective
3	Moderate	One element is slightly more relevant than another
5	Strong	One element is strongly more relevant than another
7	Very strong	One element is very strongly more relevant than another
9	Extreme	One element is extremely more relevant than another

Table 1. The AHP fundamental scale

To avoid writing inconsistent comparisons and come up with an invalid result, AHP requires a *consistency* check on the matrices. Inconsistency of a reciprocal matrix $n \times n$ can be captured by a *Consistency Index*: $CI = \frac{\lambda_{\max} - n}{n-1}$, where $\lambda_{\max}$ is the maximum eigenvalue of the matrix. *CI* must be less than 0.1. The Consistency Index of the matrix for the above example is 0.04.

Once all local priorities are computed, global priorities $P_g^{a_i}$, given the relevances of the alternatives a_i with respect to the goal g, are computed as a weighted sum. For the sake of simplicity, and without loss of generality, we have in mind a hierarchy tree where the leftmost $n1$ criteria have a set of subcriteria each, while the rightmost $n2$ criteria have no subcriteria below them, and $n1 + n2 = n$ is the number of total criteria. Thus, global priorities are computed as:

$$P_g^{a_i} = \sum_{w=1}^{n1} \sum_{k=1}^{q(w)} p_g^{c_w} \cdot p_{c_w}^{sc_k^w} \cdot p_{sc_k^w}^{a_i} + \sum_{j=1}^{n2} p_g^{c_j} \cdot p_{c_j}^{a_i} \tag{1}$$

where $q(w)$ is the number of subcriteria for criterion c_w, $p_g^{c_w}$ is the local priority of criterion c_w with respect to the goal g, $p_{c_w}^{sc_k^w}$ is the local priority of subcriterion k with respect to criterion c_w, and $p_{sc_k^w}^{a_i}$ is the local priority of alternative a_i with respect to subcriterion k of criterion c_w.

The choice of AHP has been driven by the fact that it is flexible and modular (easily tunable with new criteria), hierarchical (able to group together and weigh different levels of criteria), and it is sound (producing consistent results). We selected AHP instead of mechanisms like simple regression or multi-label classification, since our study is not really a supervised learning problem, but a maturity assessment. In fact, our aim is not to correctly match the WikiProject quality assessment, but, indeed, to provide a measure of the maturity of the medical articles, evaluating how each of them is relevant to each of the WikiProject classes. The vector constituting the AHP output synthesizes the article relevance.

Since AHP best performs with a reduced number of criteria (no more than 9 per level) [11], in Section 4.3 we describe how we reduced to 9 the number of criteria considered in our maturity assessment.

3.2 Maturity assessment

The community leading the Wikipedia Medicine Portal 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 attached to each article are: (1) Stub, (2) Start, (3) class C, (4) class B, (5) Good article, 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 our instantiation (as in [6]), 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 in subsequent Section 4.

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.

Comparison matrices for maturity assessment. Given an article, the relative relevance of two classes (i.e., two alternatives) with respect to a subcriterion in the upper level of the hierarchy depends on the value of that subcriterion for that article. Thus, we have defined several different comparison matrices, depending on the values an article exhibits for a given subcriterion.

The matrices have been defined as follows. Starting from the articles dataset, we build a sample set formed by an equal number of articles belonging to each of the six WikiProject classes (Featured, Good, etc.). For each subcriterion, we sort the values exhibited by our sample set and split such values in intervals with the same number of elements: the values related to the borderline elements are used to define the extremes of the intervals. We decide to use six intervals. In principle, each interval should correspond to each WikiProject class. Since we a priori know the class which each article in our intervals belongs to, we leverage the distribution of the classes in the different intervals to define their relative relevance with respect to each subcriterion. This leads us to build several comparison matrices, reported online (see http://mobicare.iit.cnr.it/wikiassessment/) for the sake of brevity.

For example, let us consider the subcriterion *edit count* (i.e., the number of times an article has been edited) and suppose that its value is 2500 for a given article. From our class-distribution analysis, it results that class Featured Articles is as relevant as class Stub if the value of the subcriterion *edit count* ranges from 84 to 250, ..., and class Featured Articles is extremely more relevant

than Stub if *edit count* is more than 2236. Hence, when *edit count* is 2500, we consider class Featured Articles as extremely more relevant than class Stub.

Finally, the comparison matrices of the upper levels of the hierarchy define the relevance of the subcriteria with respect to each criterion, and of criteria with respect to the final goal (the top element in the hierarchy). To define those matrices, we followed the guidelines proposed by WikiProject and also the main observations suggested by Academia, see, *e.g.,* [22, 18, 23, 4] on the importance of each subcriterion.

Implementation of the assessment. For each article in the dataset, we have run AHP: first, we computed the values of that article for all the subcriteria and we picked up the opportune matrices related to those values, according to the intervals. We computed local priorities as the eigenvectors for each matrix, and we applied Equation 1 to obtain the final vector of global priorities. This final vector represents the maturity degree of the article, namely, a vector with six components, each representing the relevance of the article to the corresponding WikiProject class.

For each vector, we have verified its consistency, as defined in Section 3.2, to check if our assessment produces conflicting results. A conflicting result happens when, for example, an article results to be mature to be a Start as well as a Featured article (indeed, they are not neighboring classes). Finally, we have compared the obtained maturity degrees with the class associated to that article by the WikiProject Medicine. The comparison is useful to check how our quantitative results agree with the quality class. All the results are reported and discussed in the next section.

4 Assessment results

In this paper, we move beyond the work in [6]. Still using AHP for calculating the maturity degree of the all articles in the Portal, we first reduce the number of subcriteria in the AHP instantiation, and we secondly evaluate the goodness of the newly obtained maturity degree by relying on the cosine similarity (*cosSim*). The *cosSim* is a measure commonly used in Information Retrieval and text mining to evaluate the similarity of two multi-dimensional vectors. The *cosSim* between two vectors v_i and v_j 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 the maturity degree vectors of articles belonging to different WikiProject classes have lower

similarity than those of articles of the same class. Moreover, we expect that 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 denoting with v_i the maturity degree of an article, we used:

$$cr\,cosSim(C_1, C_2) = \frac{1}{|C_1| \cdot |C_2|} \sum_{v_i \in C_1} \sum_{v_j \in C_2} cosSim(v_i, v_j) \tag{2}$$

$$cl\,cosSim(C_1) = \binom{|C_1|}{2}^{-1} \sum_{v_i, v_j \in C_1, i \neq j} cosSim(v_i, v_j) \tag{3}$$

To express the above intuition in a more formal way, if an index from 1 to 6 represents the WikiProject class Stub, Start, Class C, Class B, Good article, Featured article, respectively, we should observe that $cr\,cosSim(C_i, C_j) < cr\,cosSim(C_i, C_k)$ when $i < j < k$. Moreover, a too high $cl\,cosSim$ (*i.e.,* higher than 0.95) would mean that our maturity degree simply mimics the WikiProject classification of the articles.

In the following, we report our experimental results, on our database of 24,418 articles taken from the Medicine portal. We recall that, in our previous work, for each article we considered four criteria inspired by [18, 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) anonymous editors ratio; (5) minor edits ratio; (6) article age; (7) edit frequency. Finally, 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 1) the choice of using AHP for assessing the maturity of an article (Section 4.1), 2) the results obtained in [6] (Section 4.2), and 3) the refined version of the approach proposed here, with less criteria (Section 4.4).

4.1 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. In particular, for each article we consider a vector of 25 elements, each with the value of the corresponding normalized feature. The similarity of the vectors are reported in Table 2. 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

	Stub	Start	Class C	Class B	Good Art.	Feat. Art.
Stub	**0.88**	0.89	0.89	0.84	0.83	0.77
Start		**0.91**	0.92	0.88	0.86	0.79
Class C			**0.93**	0.90	0.88	0.83
Class B				**0.90**	0.89	0.86
Good Art.					**0.89**	0.88
Feat. Art.						**0.89**

Table 2. Average similarity of the feature vectors of articles belonging to the same WikiProject class (class *cosSim*, in **bold**) and to distinct classes (cross *cosSim*)

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.

4.2 Similarity of results with 25 subcriteria

The second experiment we perform is the evaluation of the similarity of the maturity degrees obtained in the previous work [6]. The results of the similarity evaluation are reported in the second column of Table 3, where, for completeness, we also include the results of Table 2 in the first column. We have chosen this alternative representation of the results in order to highlight the variation between the different approaches. The third column will be considered in the following sections.

Comparing the similarity of the normalized feature vectors with the similarity of the maturity degree obtained in [6], we can appreciably observe an increase of the class *cosSim* and a decrease of the cross *cosSim*. In particular, we can 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, as a consequence, a low granularity characterization of the maturity of different articles. We can 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 $cr\,cosSim(Stub,\ Feat.\ Art.) = 0.53$.

4.3 Reducing the number of subcriteria

Many studies have shown that AHP performs better with few criteria [12, 16]. In particular, it has been noticed that too many criteria reduce the ability of AHP to make correct decisions (mainly because the judgment capability of the human brain is reduced with too many criteria involved). Here, we reduce the number of criteria to make a more efficient use of the decision process. Reducing

	considered classes	average *cosSim* no AHP	AHP 25 subc. ([6])	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. Article	0.88	0.94	0.91

Table 3. Average similarity for the articles belonging to the same WikiProject class (class *cosSim*) and to distinct classes (cross *cosSim*), with different settings

the criteria, in facts, has 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 start eliminating 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 [7]. 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 [10].

For the 6 subcriteria of 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

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 4. Mutual information of structural and historical features with respect to the WikiProject article class. It captures the dependency of the feature with respect to the class.

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 during the decision process 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 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 with respect to the article class. In particular, the removed features are reported in Table 4, jointly with their mutual information with respect to 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.

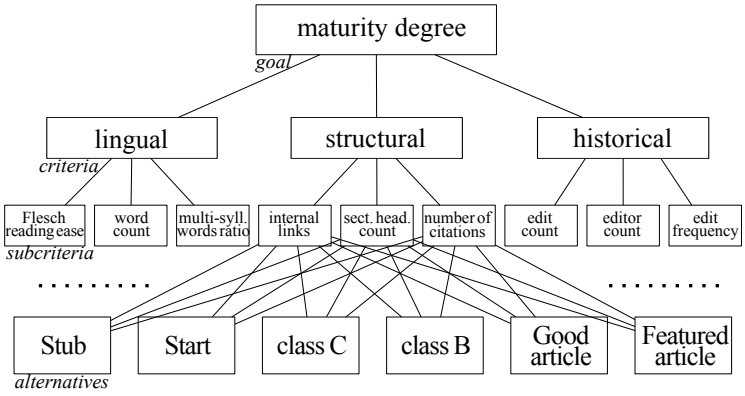

Fig. 2. AHP hierarchy with 9 subcriteria

The reducing 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).

4.4 Maturity assessment with 9 subcriteria

We apply AHP with 9 subcriteria and re-compute the maturity degree of the articles of our dataset. Figure 2 describes the AHP hierarchy we use. Similarly to [6] and as detailed in Section 3.2, we construct the comparison matrices for the subcriteria leveraging the distribution of the articles among the six classes, considering six matrices for each feature. This leads us to build 54 comparison matrices (see http://mobicare.iit.cnr.it/wikiassessment/). The matrices for the criteria and for the final goal are reported in Figure 3. As noteworthy observation, we consider that all the criteria are equally relevant with respect to the goal and that the word count feature is the most relevant among its criterion.

Table 5 compares the consistency (as defined in Section 3.2) of the new results with the old ones obtained with 25 subcriteria. Firstly, we observe that a slightly

	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 5. Percentage of articles that obtained a consistent maturity degree

$$\begin{array}{c} \\ FR \\ MS \\ WC \end{array} \begin{array}{c} \begin{array}{ccc} FR & MS & WC \end{array} \\ \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{8} \\ 2 & 1 & \frac{1}{5} \\ 8 & 5 & 1 \end{pmatrix} \end{array}$$

(a) *Lingual* matrix

$$\begin{array}{c} \\ IL \\ SHC \\ NoC \end{array} \begin{array}{c} \begin{array}{ccc} IL & SHC & NoC \end{array} \\ \begin{pmatrix} 1 & \frac{1}{2} & 3 \\ 2 & 1 & 3 \\ \frac{1}{3} & \frac{1}{3} & 1 \end{pmatrix} \end{array}$$

(b) *Structural* matrix

$$\begin{array}{c} \\ EditorsC \\ EditC \\ EF \end{array} \begin{array}{c} \begin{array}{ccc} EditorsC & EditC & EF \end{array} \\ \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \end{array}$$

(c) *Historical* matrix

$$\begin{array}{c} \\ Ling. \\ Struct. \\ Hist. \end{array} \begin{array}{c} \begin{array}{ccc} Ling. & Struct. & Hist. \end{array} \\ \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \end{array}$$

(d) *Criteria* matrix

Fig. 3. Comparison matrices for criteria and subcriteria

higher consistency ratio holds for the new results (it is almost perfect for Stub articles).

The third column of results in Table 3 reports the class and the cross similarity for the new results and makes possible a comparison with the other 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 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 Section 3.2. 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 contributing factor is the adoption of the intervals: using the same comparison matrices for articles with two different values falling within the same range but near to its two opposite ends, reduces the final maturity degrees of the two articles. 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 4 shows a summary of the 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 (it would be impossible, and, maybe not really meaningful, given the amount of articles), but only to highlight the general agreement of the results for 300 considered articles within the different classes. In particular, for each WikiProject class, we randomly sample

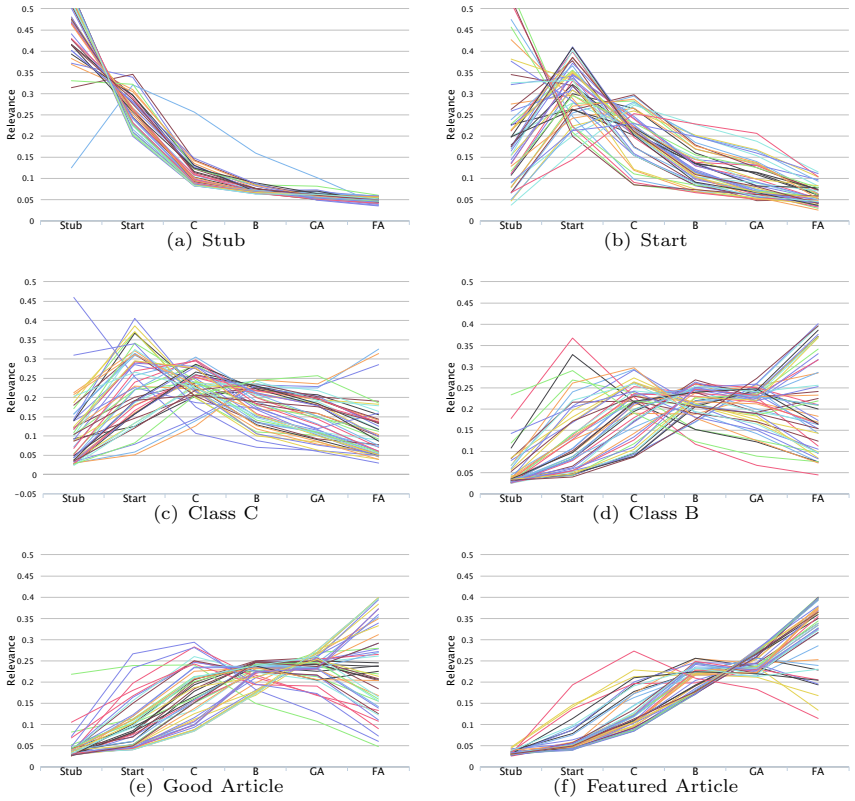

Fig. 4. Maturity degree with respect to the WikiProject assessment, 9 subcriteria. Each line is the maturity degree of an article belonging to a given WikiProject class.

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

Summarizing, reducing the subcriteria set leads us to an efficient application of AHP, as discussed in [12, 16]. 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.

5 Conclusions

This paper enhances our previous automatic assessment of Wikipedia medical articles. We refined our AHP-based approach by identifying and pruning re-

dundant 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 [6], 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. A. Pai. Almost 1M families use video consultations with physicians last year. *Mobihealthnews.com*, Mar '14. http://goo.gl/6sI0tD, Last checked: July 14, 2014.
2. E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 183–194. ACM, 2008.
3. 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: July 14, 2014.
4. J. E. Blumenstock. Automatically assessing the quality of Wikipedia. In *UC Berkeley: School of Information Report 2008-021*, 2008.
5. J. E. 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, pages 1095–1096, 2008.
6. 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.
7. T. M. Cover and J. A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006.
8. 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.
9. C. A. Haigh. Wikipedia as an evidence source for nursing and healthcare students. *Nurse Education Today*, 31(2):135 – 139, 2011.
10. R. W. Hamming. *Coding and Information Theory (2Nd Ed.)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1986.
11. G. A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2), 1956.
12. G. A. Miller. The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. *The Psychological Review*, 63:81–97, 1956.
13. R. Feltman. The #1 doctor in the world is Dr. Wikipedia. *MSN News online*, Jan '14. http://goo.gl/zhTNoa, Last checked: July 14, 2014.
14. S. Fox And M. Duggan. Health Online 2013. *Pewinternet.org*, Jan '13. http://goo.gl/vRBnCA, Last checked: July 14, 2014.
15. T. L. Saaty. A scaling method for priorities in hierarchical structures. *Journal of Mathematical Psychology*, 15(3), 1977.

16. T. L. Saaty. Decision making with the Analytic Hierarchy Process. *Int. Journ. of Services Sciences*, 1(1):83–98, 2008.
17. 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):1853 – 1862, 2007.
18. B. Stvilia, M. B. Twidale, L. Gasser, and L. C. Smith. Information quality discussions in Wikipedia. In S. Hawamdeh, editor, *Intl. Conference on Knowledge Management*. 2005.
19. U.S. Dept. of Health and Human services. Online Health Information: Can You Trust It? . http://goo.gl/G5evGg, Last checked: July 14, 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 *Int. Conf. of Information Science and Management Engineering (ISME)*, 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.