Measuring Article Quality in Wikipedia using the Collaboration Network

Baptiste de La Robertie
Université de Toulouse
IRIT UMR5505
F-31071, France
baptiste.delarobertie@irit.fr

Yoann Pitarch
Université de Toulouse
IRIT UMR5505
F-31071, France
yoann.pitarch@irit.fr

Olivier Teste
Université de Toulouse
IRIT UMR5505
F-31071, France
olivier.teste@irit.fr

***Abstract*—Collaboratively edited articles such as in Wikipedia suffer from well-identified problems regarding their quality, e.g., information accuracy, reputability of third-party sources, vandalism. Due to the huge number of articles and the intensive edit rate, the manual evaluation of article content quality is inconceivable. In this paper, we tackle the problem of automatically establishing the quality of Wikipedia articles. Evidences are shown to consider the interactions between authors and articles to assess the quality score. Collaborations between authors and reviewers are also considered to reinforce the discriminative process. This work gives a generic formulation of the Mutual Reinforcement principle held between articles quality and authors authority and take explicitly advantage of the co-edits graph generated by individuals. Experiments conducted on a set of representative data from Wikipedia show the effectiveness of our approach.**

I. Introduction

Context. The free encyclopedia Wikipedia probably constitutes the most well-known collaborative system where any user can create and edit articles. Recent statistics[1] report almost 35 millions of articles in more than 280 languages, among which close to 2 and 5 millions of French and English articles respectively. This collaborative process, involved by more than 55 millions of contributors, generates 10 millions of edits each month, that is approximately 10 edits per second. This continuously increasing production of text data leads to various scientific challenges, among which the automatic quality assessment of the generated content.

The main strength of Wikipedia is to allow anyone to contribute to its content. Potential pitfall of such an open collaborative editing process is the emergence of doubtful or even radically poor quality contents, e.g., hoaxes, publicity, disinformation or acts of vandalism, that can be available for consultation for several weeks before being detected and corrected. A well-known example has concerned the involvement of the journalist John Seigenthaler in the Kennedy assassination, i.e., a fake content appeared in the wikipedia biography page of the former in 2005. The biography had also spread to the websites Answers.com and Reference.com and the erroneous content remained online for more than five months [1].

To overcome these limitations, various works tackle the problem of automatically assessing the quality of articles. Our work falls into this problem category. However, while substantial efforts have been made using explicit features such as length of the articles [2] or implicit ones such as life cycles of texts [3][4][5], little attention [6] has been paid to indicators in the co-edit graph of authors. As motivated in the next section, we do think that *considering some structural properties of the co-edit graph can significantly help in assessing the quality of Wikipedia articles*.

Motivations. Besides the features considered by the state-of-the-art approaches, we motivate the need for considering the co-edit graph in the quality score calculation by answering the following questions:

Are the co-edit graphs of top quality articles denser than of poor quality articles? To answer this question, we built the co-edit graph for each of these 2 categories[2] according to the following principles: editors having authored some contents in at least one article of the desired quality represent the set of vertices. It exists an undirected edge between two editors if they have co-edited at least an article. Edges are weighted by the number of articles the pair of author has co-edited normalized by the number of documents in the category. The answer to this question as well as others basic statistics on the graphs are shown in Table I: the graph associated with high quality articles is 12 times denser than the one associated with poor quality articles.

TABLE I. Some statistics about the co-edit graphs per article quality level

Type of articles	Poor quality	High quality
Articles	18,823	245
Editors	36,973	9,110
Collaborations	369,844	546,273
Density (10^{-5})	0.53	6.8
Avg. Degree	15	26
Med. Degree	5	10

Because comparing the density of very different size graphs might be legitimaly discussed, the average and the median degree are also reported: nodes associated with users collaborating in high quality articles have twice as many neighboors than those associated with poor articles. Based on these observations, we can suspect that high quality articles

[1] http://en.wikipedia.org/wiki/Wikipedia:Statistics

[2] As stated in Section IV-A the Wikipedia Editorial Team Assessment has manually labelled 30K articles. This enables the possibility to build some statistics on these categories.

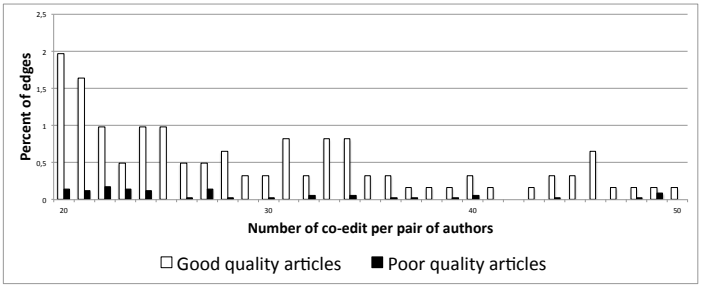

Fig. 1. The histogram of edges per weight on high quality articles (white) and poor quality articles (black).

involve much more collaboration than poor quality articles. Thus, considering the collaborative process between users looks to be a promising track to assess the quality of an article.

Are the co-editors used to work together on top quality articles? Besides the density of the co-edit graphs, we would like to analyze the nature of the relations between the editors in order to determine in what extent authors collaborate together. Our intuition is that the higher the quality of an article is, the more frequent the interactions between co-editors are. The percentage of edges per weight per category is shown in Fig. 1. For readability purpose and because we aim at emphasizing on the impact of high weights, weights lower than 20 where omitted. Our intuition is obviously confirmed since high weights significantly appear more in the graph associated with high quality articles. Such an observation may indicate that good articles are authored by editors who should rely upon themselves since they are used to work together. Thus, considering the strength of the relation between co-editors looks to be a promising track to assess the quality of an article.

Contributions. We base our approach on the Mutual Reinforcement principle [7][8][9] which assumes the interconnection between the author authority and the article quality and propose the following contributions:

- We propose a generic formulation of the Mutual Reinforcement principle. In this way, state-of-the-art approaches can be seen as instances of this general model. This makes the comparison much easier at a theoretical point of view;
- We instantiate our model by incorporating features extracted from the co-edit graph;
- We evaluated our approach and empirically demonstrate that it significantly perform better than the state-of-the-art approaches.

Paper Organization. Related work is discussed in section II. Our proposed model is presented in section III. Section IV presents our experimental results. Finally, we conclude and give some perspectives in section V.

II. Related work

One of the earliest work [2] has shown that the number of words in a document is a good predictor for article quality. More particulary, a specific class of articles in Wikipedia, known as *Features articles*, is easily predictable using the number of words in the article. Others correlations between simple features such as the number of edits and unique editors are empirically demonstrated [10]. While these results are persuasive, both previous works consider the task as a binary classification problem where all non *Features articles* are considered as negative examples. Various works make use of the notion of *lifespan* to infer the quality or thruthness of elements in the documents. Adler et al. [3] introduce an approach that consists in measuring the reputation of the editors. The more an edit is preserved by subsequent reviewers, the more its author gains in reputation, and conversely, authors can lose reputation when their edits are reverted or deleted. The intuitive heuristic is that high-quality contributions survive longer throw the edit process because all subsquents editors implicitly approved the contribution simply by leaving it [11]. Hu et al. [7] also model the authority of the authors and the reviewers to compute the quality of each word in the article. Three models based on the mutual dependency between articles quality and contributors authority are proposed such as the *Peer Review* model which takes advantage of the implicit approbation of the reviewers, and the *Prob Review* model, which applies a decaying function over the authority of the reviewers around each approved word. Experiments show good results but are conducted over solely 200 documents. Moreover the structure between editors is not explicitly used. Adler et al. [12] also propose to credit each word with the reputation of the reviewers in proximity of the word and observe that words with low-trust assignment have high probability of being edited. In [13], the quality of an article is modeled as a time-dependent function, allowing quality of articles to evolve during time. In their work, only two states are considered making the study specific to the binary classification problem. Wohner et al. [4] propose to use the editing intensity during a period of time but once gain, experiments are conducted over 200 articles and consider only two classes : *Features articles* i.e. good articles and *articles for deletion* i.e low quality articles. A more recent work [8] reuses the mutual dependency concept between editors and texts and integrates the concept of lifespan as well as an adjustment of the authority of the reviewers in order to reduce the impact on text quality by vandal edits. A very recent work [9] explicitely formulates the mutual dependency between authors and articles with an Article-Editor network. The proposed model computes the quality of a document according to the editing relationships between article nodes and editor nodes using a variant of the PageRank algorithm [9]. Experiments are performed over only two classes of articles and none consider the relations between users. Finally, Suzuki [6] implicitly uses the structure of the co-edit network via the *h-index* measure in order to calculate the authority of the editors, but the quality of an article is directly computed using solely the derived authorities. In the conclusion, the author emits some doubts about the uniqueness of a quality function to distinguish good from bad articles.

TABLE II. Related approaches

References	Considered relations		
	Authors/Articles	Reviewers/Articles	Authors/Reviewers
[3][8][9]	✓		
[5][7][12]	✓	✓	
[6]	✓	✓	✓

These approaches are summed up in Table II according

to the type of relations they consider to address the article quality evaluation problem. Most of the works consider the task of assessing the quality of Wikipedia article as a binary classification problem whereas six different grades have been proposed by the Editorial Team of Wikipedia. Our work can interestingly deal with a non-predetermined number of classes by formulating the problem as a ranking problem. Moreover, none of the previous work has explicitly exploited the author interactions. However, as pointed out in Section I, some evidences exist to integrate the author interactions in the process. In this work we properly make use of these interactions in our model as described in the next section.

III. Proposed models

Our work is grounded on 2 intuitions :

(1) Good articles are likely to be written by good editors, and conversely, good editors are more authorative if they participate, by writing and reviewing, good articles [7][12]. In our model, we capture this intuition postulating that the more authoritative users participate to the ellaboration of the article, the more the article is likely to be of good quality. The intrinsic dependency between quality and authority inevitably leads to an interdepedant pair of equations, where the quality of an article is defined over the qualities of its individuals piece of contents, and the authority of a user over the individual piece of contents he/she authored and *approved*. With our formulation, the amount of contribution of each editor in each article is catched in order to quantify the notion of *implicit approvement*, saying that if many authoritatives users who widely participate in an article leave previous edit on place it is because they judge it as good quality.

(2) Good articles are the result of a grouping of expert users. This second intuition seems to be particulary true for specialized topics. The proposed model quantifies in what extent these experts are used to collaborate together. Reviews between these authors are very expressive because probably much more *trustworthy*. Hence, we postulate that an article has a much more quality efficiency when these experts work together, and express that intuition using the co-edit graph between the authors. More formally, one can express this intuition using the probability $P_X(k)$ that two editors have together edited at least k articles belonging to a given class X of articles. To illustrate the intuition, let suppose 3 classes of articles A, B and C such as articles of class A are of better quality than those of type B and those belonging to class B of better quality than articles in class C. If editors of class A collaborate more than editors of class B and C, we should have $\int_k P_A(k) \geq \int_k P_B(k) \geq \int_k P_C(k)$. It should be noted that this unequality might not hold for close classes on real data for all k. However, as illustrated in Fig. 1, it does hold when considering the two classes "good" and "poor" quality articles. The cumulative sum of edges weigth of good quality articles is indisputably greater than those of poor quality articles. This will be sufficient to push upward the scores of good quality articles.

A. Notations

Let $\mathcal{X} = \{1, 2, ..., N\}$ be the set of N articles and $\mathcal{U} = \{1, 2, ..., M\}$ the set of M users. Each article i is modeled by a sequence of char sequences $\langle x_i^k \rangle_{1 \leq k \leq n_i}$, where n_i is the number of sequences in the article. We denote by $y_i \in \mathcal{Y}$ the quality of an article i. The author j of a sequence x_i^k is a_i^k, i.e., $j = a_i^k$. The set of sequences that a user j has authored is $S(j)$.

Example III.1. Let $\mathcal{X}_{toy} = \{1, 2, 3, 4\}$ be a set of 4 articles authored by $\mathcal{U}_{toy} = \{1, 2, 3, 4\}$ a set of 4 editors. Corresponding char sequences are illustrated on Fig.2. For instance, the article 1 is composed by 4 sequences $\langle x_1^1, x_1^2, x_1^3, x_1^4 \rangle$. The first one x_1^1, i.e., the beginning of the article, has been authored by the user 1, i.e., $a_1^1 = 1$. The set of sequences that have been authored by the user 1 (in gray in the picture) is $S(1) = \{x_1^1, x_1^3, x_2^2, x_2^6, x_3^2\}$.

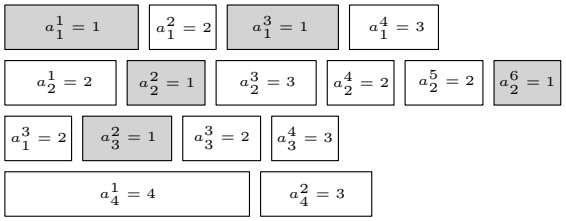

Fig. 2. Four articles and their respective char sequences. For instance, article 4 (bottom) is composed of 2 sequences authored by users 4 and 3 respectively.

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be the co-edit graph defined by the set of nodes $\mathcal{V} = \mathcal{U}$ and the edges set $\mathcal{E}$, where edges $e_{ij} \in \mathbb{R}^+$ between each pair of users $(i, j) \in \mathcal{U}^2$ measure the number of articles both users i and j have co-edited, formally :

$$e_{ij} = |\{k \in \mathcal{X} : \exists\,(u, v) : a_k^u = i \wedge a_k^v = j\}| \qquad (1)$$

Example III.2. Let $\mathcal{G}_{toy} = (\mathcal{U}_{toy}, \mathcal{E}_{toy})$ be the co-edit graph constructed using the previous example (see Fig. 3). The weight of the edge between 1 and 2 is $e_{1,2} = 3$ because they have been co-authors in the articles 1, 2 and 3.

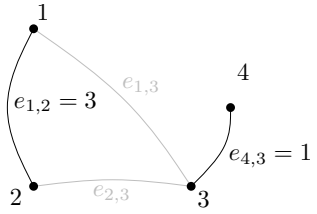

Fig. 3. Representation of the co-edit graph $\mathcal{G}_{toy}$ associated with $\mathcal{X}_{toy}$ and $\mathcal{U}_{toy}$. For instance, $e_{4,3}$ is 1 because users 4 and 3 have co-edited only one article (article 4).

Finally, the quality of a sequence x_i^k is Q_i^k and the authority of a user j is A_j. Score calculations are discussed in the next sections.

B. Definitions

We first introduce the notion of *lifespan* of a sequence with is necessary to define the *anterior* and *posterior* neighborhood of a sequence and the notion of *approvement*.

Definition III.1. (*Lifespan* of a sequence) The *lifespan* l_i^k of a sequence x_i^k is the number of revisions it survives until the latest version of the article.

Definition III.2. (*Anterior neighborhood*) The anterior neighborhood of a given sequence x_i^k, denoted by $\mathcal{N}^-(x_i^k)$, is the set of sequences in the article i such that $\forall x_i^{k'} \in \mathcal{N}^-(x_i^k)$, $l_i^{k'} > l_i^k$.

Definition III.3. (*Posterior neighborhood*) The posterior neighborhood of a given sequence x_i^k, denoted by $\mathcal{N}^+(x_i^k)$, is the set of sequences in the article i such that $\forall x_i^{k'} \in \mathcal{N}^+(x_i^k)$, $l_i^{k'} < l_i^k$.

Definition III.4. (*Approved sequence*) The sequence x_i^k is an approved sequence w.r.t. the user j (the user j *has approved* the sequence x_i^k) if the user j has authored at least one sequence in the posterior neighborhood of x_i^k, i.e, if $\exists x_i^{k'} \in S(j) \cap \mathcal{N}^+(x_i^k)$.

Example III.3. The above defined concepts are illustrated on Fig. 4. It illustrates the edit process of the article 1 from the initial commit of the user 1 (top of the figure) to the latest revision (bottom of the figure). Let us consider the sequence x_1^2 in the latest version of the article. Since it appears three revisions before the latest version of the article (the sequence x_1^2 has been authored by the user 2 in the second revision), its lifespan is 3, i.e., $l_1^2 = 3$. The anterior neighborhood of the sequence x_1^2 is $\mathcal{N}^-(x_1^2) = \{x_1^1, x_1^3\}$ and its posterior neighborhood is $\mathcal{N}^+(x_1^2) = \{x_1^4\}$. Finally, the sequence x_1^2 has been *approved* by the author 1.

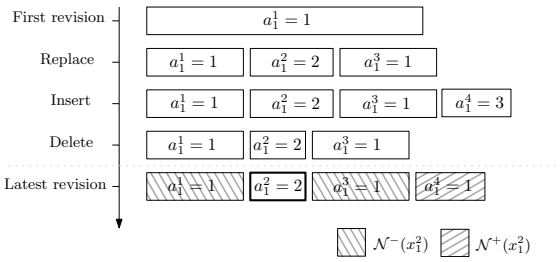

Fig. 4. Illustration of an edit history of an article. Neighborhoods of sequence x_1^2 are computed considering the latest version of the article.

C. Model

When an authoritative user writes a content which is successively reviewed and approved by authoritative users, it is likely to be of good quality. This is even reinforced when many reviewers approved it and even more *when these reviewers have widely participated to the article*, e.g., the user 1 in Fig.4. The more a reviewer throws himself/herself in an article, the more he/she is susceptible to see new edits and perform modifications if he/she judges the quality insufficient. This assumption becomes stronger when the number of authoritative reviewers increase. In other words, the amount of contributions of each reviewer seems to be fundamental to measure the concept of *approvement*. To formalize this intuition, the final quality Q_i^k of an individual sequence x_i^k is expressed by an *approvement function* which generically reflects the weighting schemes (or relations) between authors and reviewers. The quality of the sequence x_i^k is formally defined as:

$$Q_i^k = \sum_{x_i^{k'} \in \mathcal{N}^+(x_i^k)} \underset{a_i^{k'} \to a_i^k}{\mathcal{K}} \left(A_{a_i^k}, A_{a_i^{k'}} \right) \tag{2}$$

where $\underset{j \to i}{\mathcal{K}} : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}^+$ is the generic *approvement* function quantifying the implicit approvement by the user j of the sequence authored by the user i. Two forms of functions will be discussed in Section III-D. In a symmetric way, the authority A_j of the author j is based on the quality of the sequences he/she authored and on those he/she approved and is defined as:

$$A_j = \sum_{x_i^k \in \mathcal{S}(j)} \sum_{x_i^{k'} \in \mathcal{N}^-(x_i^k)} \underset{j \to a_i^{k'}}{\mathcal{K}} \left(Q_i^k, Q_i^{k'} \right) \tag{3}$$

Again, $\mathcal{K}$ must be instantiated in order to capture the approvement of a sequence by its reviewer. Finally, the global quality Q_i of article i is simply defined over the quality of the sequences it contains in the latest revision, formally :

$$Q_i = \sum_{k \leq n_i} Q_i^k \tag{4}$$

The computation of the model is described in Section III-E. It calculates for each article i a score $Q_i \in \mathbb{R}^+$. The obtained list should be ranked to obtain the documents in decreasing (predicted) level of quality.

D. Approvement functions

The core of our model is the definition of the approvement act of a sequence by a reviewer. In this section, two approvement functions between an author i and a reviewer j are introduced and discussed:

$$\underset{j \to i}{\mathcal{K}}{}^1(a, b) = a + b \tag{5}$$

$$\underset{j \to i}{\mathcal{K}}{}^2(a, b) = (ab)^{\lambda(1 - e_{ij})} \tag{6}$$

Equation (5) (in short $\mathcal{K}^1$) is a slight but crucial variation of the *Peer Review* model. In the original formulation [3] the quantity Q_i^k is computed as the sum over the authority of the author and the authorities of each *distinct* reviewer who approved x_i^k. More formally, if $\{z_1, z_2, ..., z_p\}$ is the set of p distinct users (including the author herself/himself) who approved the sequence x_i^k, then the computed quality of the sequence is indeed :

$$Q_i^k = A_{z_1} + A_{z_2} + ... + A_{z_p} \tag{7}$$

One can demonstrate that this expression is a result of a particular instantiation of our model.

Lemma 1. Let z_1 be the author of the sequence x_i^k and $\{z_2, ..., z_p\}$ be the set of $p - 1$ distinct reviewers who have approved x_i^k (including eventually the author himself/herself). Under our model, the use of the approvement function $\mathcal{K}^1$ leads to a linear combination of the authority of the reviewers:

$$Q_i^k = \theta_1 A_{z_1} + \theta_2 A_{z_2} + ... + \theta_p A_{z_p} \tag{8}$$

where $\forall j > 1$, $\theta_j = |\{x_i^{k'} \in \mathcal{N}^+(x_i^k) \cap S(j)\}|$ i.e. number of sequences of user j bellonging to the posterior neighborhood $\mathcal{N}^+(x_i^k)$, and $\theta_1 = |\mathcal{S}^+(x_i^k)| + |\{x_i^{k'} \in \mathcal{N}^+(x_i^k) \cap S(1)\}|$.

Proof: Let $\mathcal{N}^+(x_i^k) = \{x_i^{k_1}, x_i^{k_2}, ..., x_i^{k_s}\}$ be the set of s sequences authored by a set $\{z_1, z_2, ..., z_p\}$ of p distinct

reviewers who approved sequence x_i^k. With $\mathcal{K}^1$ to compute the quality Q_i^k, by definition :

$$Q_i^k = \sum_{x_i^{k'} \in \mathcal{N}^+(x_i^k)} \underset{a_i^{k'} \to a_i^k}{\mathcal{K}} (A_{a_i^k}, A_{a_i^{k'}}) \tag{9}$$

$$= \sum_{x_i^{k'} \in \mathcal{N}^+(x_i^k)} A_{a_i^k} + A_{a_i^{k'}} \tag{10}$$

$$= s \cdot A_{a_i^k} + A_{a_i^{k_1}} + A_{a_i^{k_2}} + A_{a_i^{k_3}} + ... + A_{a_i^{k_s}} \tag{11}$$

By regrouping the identical reviewers, i.e., by identical $a_i^{k_j}$

$$Q_i^k = \underbrace{s \cdot A_{a_i^k} + A_{a_i^{k_1}}}_{\theta_1 A_{z_1}} + \underbrace{A_{a_i^{k_2}} + A_{a_i^{k_3}}}_{\theta_2 A_{z_2}} + ... + \underbrace{A_{a_i^{k_s}}}_{\theta_p A_{z_p}} \tag{12}$$

we finally obtained :

$$Q_i^k = \theta_1 A_{z_1} + \theta_2 A_{z_2} + ... + \theta_p A_{z_p} \tag{13}$$

■

Hence, the *Peer Review* model is a particular case of our model assuming all θ_i equals 1 (no weighting scheme is considered between the reviewers).

Our model enables the quality of a sequence to increase with the number of authoritative reviewers who have approved the sequence and also with the amount of contributions θ they have authored in the article. The desired intuition is captured: if many authoritative users who have predominantly participated to the article have approved a given sequence, its quality will fairly increase.

The co-edits relations between users are captured by equation (6) ($\mathcal{K}^2$ in short). When e_{ij} is close to 0 (none or very few relations between author j and reviewer i), the quality of a sequence increases only when both authors and reviewers are authoritative. Hence, unlike function $\mathcal{K}^1$, $\mathcal{K}^2$ causes the approvement of an unauthoritative user to be almost unconsidered (final quantity is bounded by $\max(u_i, u_j)$ because of the normalization of the authority score). Strong relations notably enable new registered users to rapidly gain in authority, and even more if they collaborate with authoritative users. To control the strength of the co-edit weights and consequently the quality scores, the user-parameter $\lambda \in [0, 1]$ is introduced. When λ is close to 1, the function is elitist and tends to disfavor isolated users. Conversely, when λ is close to 0, the function is permissive and tends to encourage unauthoritative users. It should be noted that λ, being a hyperparameter, is not aimed to be learned but to be fixed before the execution of the algorithm in order to build a permissive or elitist function.

E. Calculation

The system formulated by the interdependent pair of equations 2 and 3 is solved by an iterative process that consists in alternatively computing authorities A and qualities Q. More details about the theoretical computation of the associated eigen values problem can be found in [14]. The following generic process is used:

1) Initialize randomly both authorities and qualities scores
2) Compute quality scores with equation 2
3) Compute authority scores with equation 3
4) Normalize scores

The last three steps are repeated until convergence. Let $Q^t \in \mathbb{R}^N$, resp. $A^t \in \mathbb{R}^M$, be the vector of quality scores, resp. authority scores, at the t^{th} iteration of the algorithm. The convergence is reached when $d(Q^t, Q^{t-1}) + d(A^t, A^{t-1}) < \epsilon$, where d is a distance function and ϵ is aimed to control the convergence. In the experimentations, the L_2 norm was used as distance measure. For $\epsilon = 10^{-3}$, the convergence is rapidly reached (less than 10 iterations). Details about the computation are given by the following algorithms.

Algorithm 1 Quality($\mathcal{X}$)

```
for all i ∈ X do
    Q_i ← 0
    for all x_i^k do
        Q_i^k ← Σ_{x_i^{k'} ∈ N^+(x_i^k)} K_{a_i^{k'} → a_i^k} (A_{a_i^k}, A_{a_i^{k'}})
        Q_i ← Q_i + Q_i^k
    end for
end for
```

Algorithm 2 Authority($\mathcal{U}$)

```
for all u_j ∈ U do
    A_j ← 0
    for all x_i^k ∈ S(j) do
        A_j ← Σ_{x_i^{k'} ∈ N^-(x_i^k)} K_{j → a_i^{k'}} (Q_i^k, Q_i^{k'})
    end for
end for
```

With a basic implementation of the anterior and posterior neighborhood search, the quality computation is performed in $\mathcal{O}(Nn^2)$, where $n = \max_{i \leq N}\{n_i\}$ and the Authority computation is performed in $\mathcal{O}(Msn)$ where $s = \max_{j \leq M}\{|S(j)|\}$.

IV. Experiments

This section is dedicated to the presentation of our result. We first properly introduced the methodology used for these experimentations. Quantitative results are then presented followed by a qualitative interpretation of two representative co-edit graphs.

A. Protocol

Dataset description. We used a set of English documents from Wikipedia articles that have been reviewed by the Editorial Team Assessment of the WikiProject. Each article has been labelled according to the WikiProject quality grading scheme, and belongs to one of the following class $\mathcal{Y} = \{S, C, B, GA, A, FA\}$. The user preferences are defined over $\mathcal{Y}$ as follows :

$$FA \succ A \succ GA \succ B \succ C \succ S$$

The label S stands for *Stub Articles* (very bad quality articles with no meaningful content) while FA stands for *Featured Articles* (complete and professional articles). This scale is used as a ground truth for evaluation. It should be specified that

labels were given according to the latest version of an article. We developed a crawler in Java to parse grades, topics and articles (from its first revision to the latest version). The history of edits are stored in a relational database. The raw dataset represents almost 130 Gb of text data. Statistics over these data are summarized in Table III.

TABLE III. RAW DATASET STATISTICS

Grade	y_i	Articles	Revisions	Authors	Gb
FA	5	611	765,917	174,178	29
A	4	67	32,700	7,492	2
GA	3	462	398,757	112,266	15
B	2	1.459	1,283,225	330,010	44
C	1	3.382	1,521,416	423,021	32
S	0	26.736	1,164,376	236,715	7

Preprocessing. A diff tool was developed in Python to extract the sets of sequences that survive until the last revision. During this preprocessing step, the *lifespan* of each sequence is updated. For each pair of consecutive revisions, the sequences are propagated, split and/or removed according to the possible sequence operations an editor can perform (replace, insert and delete characters). This preprocessing step is applied over a stratified random collection of the raw data of nearly 23,000 articles (a fix number of articles per category is randomly selected). More than 110,000 distinct users have produced around 2.8 million sequences. The co-edit graph built over this dataset is composed by more than 111,000 nodes and 5 millions of edges. Statistics about our dataset and resulted co-edit graphs are synthesized in Table IV.

TABLE IV. DATASET

Grade	FA	A	GA	B	C	S
Articles	245	51	346	1,012	1,946	18,823
Authors per article (mean)	61	37	37	46	41	7
Lifespan per article (mean)	275	166	154	141	126	12
Sequences (mean)	1,114	963	809	695	439	40
Sequences length (mean)	78	44	86	85	88	73
Nodes (10^3)	9.11	1.41	8.54	28.3	48.1	36.9
Edges (10^3)	546	62	406	1,505	2,279	369

Evaluation Metrics. Performances are evaluated using both standard ranking and classification evaluation metrics. To evaluate the ranking, the *Normalized Discount Cumulative Gain at k* (NDCG@k) was used [15]. It computes a normalized score based on the degree of relevance of each document and a decaying function of their rank. In our case, the degree of relevance of a document is directly associated with its label : from 0 for documents belonging to class S (poor quality articles), to 5 for documents in class FA (very good articles). A score equals to 1 indicating a perfect ranking. Once the permutation is computed by the model and documents ranked in decreasing order of (predicted) quality, one can split the list of documents in 6 categories according to the repartition of articles per grade in the ground truth. Hence, the number of positive examples per grade is evaluated using the Recall metric. A recall of 1 indicating a perfect classification.

Competitors. We compare our model to the following competitors [3] :

Naive model. Final ranking is obtained by sorting articles by length.
Basic model. The effect of the reviewers is not taken in consideration.
Peer model. Reviewer effects are considered as important as the author ones. Final authority of an author is the sum of the authority of the distinct reviewers.
Prob model. Authority of the reviewer is slightly decreasing with the distance between author and reviewer words. In the following experiments, the best decaying function $f(d) = \frac{1}{\max(0,d-\beta)+1}$ in [3] was used. In the formulation, d is the distance between the words of the author and the closest word of the reviewer and β is a user-parameter to control the maximum distance over which authorities of reviewers are not fully considered. In the experiments, parameter β was set to 1000 (distance in characters), corresponding to the best run among different values of the parameter.

B. Quantitative experiments

The four competitors are compared with the two instantiations of our model, i.e., with approvement functions $\mathcal{K}^1$ and $\mathcal{K}^2$. First, NDCG@k metric is used to compare the rankings over the articles at different levels, i.e., values of k are directly derived from the repartition of the article quality in the dataset (cumulative sum beyond the number of articles per class). Second, the recall for each of the 6 classes is used to compare to the competitors. The first class (FA) is evaluated using the first 245 documents, the second class (A) is evaluated using the next 51, and so on. Because $\mathcal{K}^2$ depends on the parameter λ, only the best run (at least for $k < 642$) is used for the comparison ($\lambda = 0, 7$).

Experiments are conclusive: both proposed functions outperformed competitors for $k \leq 642$. Both the amount of the contributions and the co-edits weights of the reviewers seem to be interested indicators to discriminate good to very good articles. Interestingly, the performances of $\mathcal{K}_1$ using the NDCG@k (see Fig. 5) increase faster with k than $\mathcal{K}_2$, making the former globally more persuasive for $k \geq 642$. The discrimination of articles of mid-quality using the co-edit relations seems to be more challenging. Nonetheless, when we examine the metrics considering all the articles ($k = 22.423$), $\mathcal{K}_2$ again seems to benefit from the co-edit weights. The very few numbers of co-edit relations betweens authors of poor quality articles might be one factor which explains this sudden increase in the NDCG value.

This behavior can also be noted on Fig. 6. We plotted the evolution of the NDCG@k for the $\mathcal{K}_2$ model in function of λ. We choosed to display the evolution for the top of the list documents ($k = 245$) and for the queue of the list ($k = 3.600$). The evolution clearly indicates that $\mathcal{K}_2$ is much more competitive to discriminate very good articles for high values of λ (best NDCK@245 for $\lambda = 0, 7$), while the performances for the queue of the list ($k = 3.600$) remains globally constant.

TABLE V. RECALL PER CLASS

Model	FA	A	GA	B	C	S
Naive	25.30	3.92	11.56	32.21	59.66	98.68
Basic	0.4	0	0.86	13.93	14.33	90.99
Peer	28.16	1.96	3.17	28.45	55.3	98.95
Prob	20.81	1.96	4.33	27.76	39.77	96.99
K1	61.22	3.92	10.98	34.68	53.85	98.07
K2	**75.91**	3.92	**67.91**	**90.21**	**95.94**	**99.82**

Finally, Table V summarizes the capacity of the models to discriminate the six class of articles. Even if the ranking is not

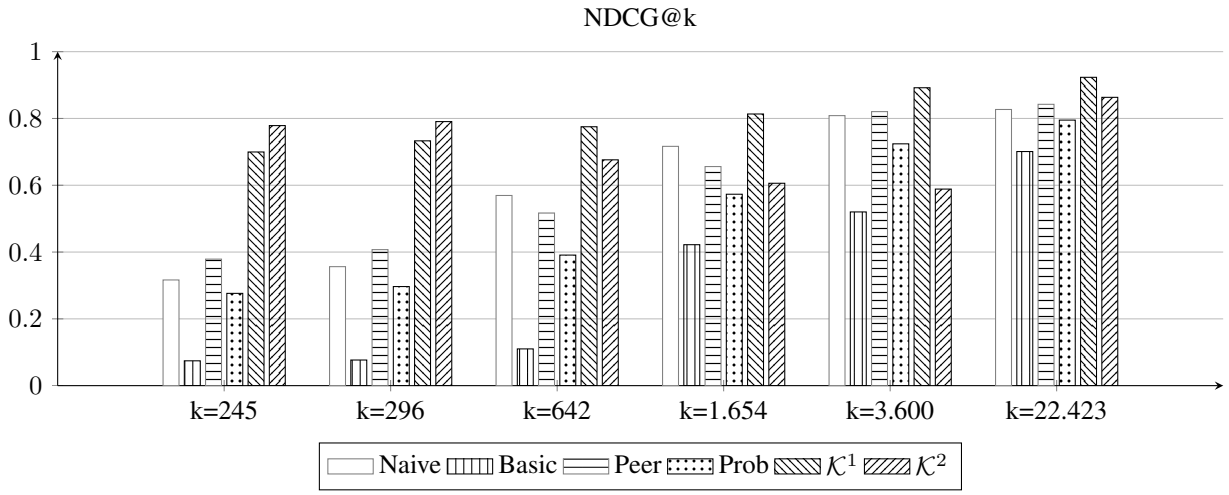

Fig. 5. Evaluation with the NCDG@k metric for $k \in \{245, 296, 642, 1.654, 3.600, 22.423\}$ of the competitors and the proposed model with 2 approvement functions $\mathcal{K}_1$ and $\mathcal{K}_2$.

optimal as we just already said it, the separation induced by $\mathcal{K}_2$ is incomparable with the others competitors.

It is interesting to see that the articles belonging to class A are very badly discriminated by every competitors (see Table V). Even if they are under-represented in the dataset (there are very few A articles in the English version of Wikipedia), they are even though considered in a drastically different way by $\mathcal{K}_1$ and $\mathcal{K}_2$. By analyzing the ranks of each A article in the returned list by both models, we found that mean rank produced by $\mathcal{K}_1$ is 1.194 (it roughly corresponds to B articles) while it is 382 for $\mathcal{K}_2$ (closer to GA articles). Moreover, the dispersion of the ranks of A articles is 1.000 times smaller for the latter than for the former: considering co-editions seems to improve the precision.

Results confirm that combining different structural properties of the Wikipedia co-edit graph is beneficial and make the solution closer to the optimal solution : the co-edit graph is clearly discriminating to capture authoritative users and thus, good articles.

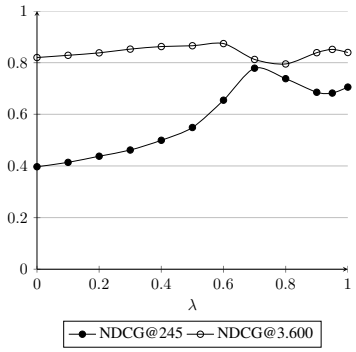

Fig. 6. Evaluation of the proposed model using approvement function $\mathcal{K}^2$ for different values of λ.

C. Qualitative interpretation

We now present the two co-edit graphs associated with a representative poor quality article (see Fig. 7 (left)) and a top quality article (see Fig. 7 (right)) according to our proposed metric. The former is dedicated to *A. Hillgruber*[3], a conservative German historian, and has been labelled as a C-class article by the Editorial Team Assessment because of its partisanship. The latter is dedicated to *Kaga*[4], a Japanese aircraft career, and has been labelled as a FA-class article. Both graphs have been obtained using the following methodology: vertices are the union of the authors and their respective co-editors (in all the dataset). For ease of reading and because very weak co-edit relations are not of interest in this study, edges with weight equals 1 have been filtered out as well as subsequent isolated vertices. Table VI sums up some statistics on these two graphs. In the following we will refer to the non-filtered graph related to *A. Hillgruber*, resp. *Kaga*, as $\mathcal{G}_C$, resp. $\mathcal{G}_{FA}$ and to its filtered version as $\mathcal{G}'_C$, resp. $\mathcal{G}'_{FA}$.

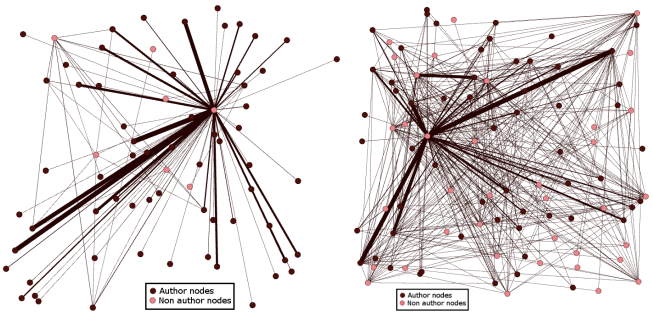

Fig. 7. The filtered co-edit graphs of the articles related to *A. Hillgruber* and *Kaga*

Since less than 1% of the edges have weights greater than 1, authors most often collaborate one time only whatever the quality of the article. This observation can though be sharpened by carefully analyzing the two graphs. On the one hand $\mathcal{G}'_C$ shows the following characteristics: (1) it is very sparse, (2) very few edges connect two author nodes, and (3) very fews non-author nodes remain in $\mathcal{G}'_C$ after the filtering step. On the

[3]https://en.wikipedia.org/wiki/Andreas_Hillgruber
[4]https://en.wikipedia.org/wiki/Japanese_aircraft_carrier_Kaga

TABLE VI. SOME STATISTICS ABOUT THE CO-EDIT GRAPHS OF THE ARTICLES RELATED TO *A. Hillgruber* AND *Kaga*

	A. Hillgruber	*Kaga*
#vertices	42572	14023
#edges	128322	55368
#authors	63	58
#vertices (filtered)	70 (0.17%)	93 (0.66%)
#edges (filtered)	87 (0.07%)	420 (0.76%)
Class repartition (%)	FA: 4.3 / A: 0 / GA: 5.6 B: 1.4 / C: 85 / S: 3.7	FA: 55.7 / A: 0.8 / GA: 3.6 B: 25.3 / C: 8 / S: 6.6

other hand $\mathcal{G}'_{FA}$ shows some opposite characteristics: (1) it is denser than $\mathcal{G}'_{C}$, (2) much more edges connect two author nodes, and (3) more non-author nodes remain in $\mathcal{G}'_{FA}$ after the filtering step. These observations confirm the soundness of our approach. Indeed, in poor quality article, the collaboration is punctual only; the added value of reviews is thus lower than for top quality articles where authors are more used to collaborate. Additionally, we compute the class of articles in which authors mostly participate. The repartition per class is given in the last line of Table VI. This shows that top quality articles, resp. low quality articles, are mostly written by authors who are used to write such good quality articles, resp. low quality articles. This is a clear evidence of the pertinence of the Mutual Reinforcement Principle.

V. CONCLUSION

Crowdsourcing platforms provide the possibility for anyone to freely contribute to their publicly available content. One inherent drawback of this collaborative process is the emergence of poor quality content. In this paper, we tackled the problem of automatically assessing articles quality in the particular case of Wikipedia. We proposed a generic formulation of the quality assessment problem based on the Mutual Reinforcement principle. We generalized previous works by introducing the notion of *approvement functions* which can take advantage of the relations between the editors. Such a formulation facilitates the theoretical comparison with state-of-the art approaches which can naturally be expressed as instances of our model. Motivated by some strong hints that legitimate the importance of considering the co-edit graph, two novel approvement functions were designed. The first function reinforces the quality of a content as a function of both the authority of the reviewers and the amount of their contributions in the article. The second function aims at capturing the relation between the authors and the reviewers since we have considered that the reviews of editors who are used to work together are more trustworthy. For this purpose, the co-edit network between editors was constructed and has appeared to have very interesting features. Experiments conducted on real Wikipedia articles are very conclusive. The proposed model, by improving state-of-the-art methods, empirically confirmed our two intuitions and open several perspectives.

In future work, we first plan to extend our model by generalizing the notion of neighborhood. Indeed, we think it would be beneficial to consider both horizontal (time) and vertical (documents) aspects of the neighborhood of a sequence. Notably, such an operator would enable to even more generalize our model and to reformulate other state-of-the-art works, e.g., the *Prob Review* model. Scalability is a major concern. We plan to study the possibility to adapt our model to a Big Data environment using some parallelization strategies. Finally, because of the intensive edit rate of Wikipedia articles, adapting our model to a streaming environment would enable the quality calculation on the fly.

REFERENCES

[1] L. P. Cox, "Truth in crowdsourcing," *IEEE Security Privacy*, vol. 9, no. 5, pp. 74–76, 2011.

[2] 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*, ser. WWW '08. New York, NY, USA: ACM, 2008, pp. 1095–1096. [Online]. Available: http://doi.acm.org/10.1145/1367497.1367673

[3] B. T. Adler and L. de Alfaro, "A content-driven reputation system for the wikipedia," in *Proceedings of the 16th International Conference on World Wide Web*, ser. WWW '07. New York, NY, USA: ACM, 2007, pp. 261–270. [Online]. Available: http://doi.acm.org/10.1145/1242572.1242608

[4] T. Wöhner and R. Peters, "Assessing the quality of wikipedia articles with lifecycle based metrics," in *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*, ser. WikiSym '09. New York, NY, USA: ACM, 2009, pp. 16:1–16:10. [Online]. Available: http://doi.acm.org/10.1145/1641309.1641333

[5] R. F. H. Zeng, M. Alhossaini and L. McGuinness, "Mining revision history to assess trustworthiness of article fragments," in *Proceedings of the 2nd International Conference on Collaborative Computing: Networking, Applications and Worksharing*, 2006.

[6] Y. Suzuki, "Quality assessment of wikipedia articles using ¡i¿h¡/i¿-index," *Journal of Information Processing*, vol. 23, no. 1, pp. 22–30, 2015.

[7] 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*, ser. CIKM '07. New York, NY, USA: ACM, 2007, pp. 243–252. [Online]. Available: http://doi.acm.org/10.1145/1321440.1321476

[8] Y. Suzuki and M. Yoshikawa, "Assessing quality score of wikipedia article using mutual evaluation of editors and texts," in *Proceedings of the 22Nd ACM International Conference on Conference on Information & Knowledge Management*, ser. CIKM '13. New York, NY, USA: ACM, 2013, pp. 1727–1732. [Online]. Available: http://doi.acm.org/10.1145/2505515.2505610

[9] X. Li, J. Tang, T. Wang, Z. Luo, and M. de Rijke, "Automatically assessing wikipedia article quality by exploiting article-editor networks," in *ECIR 2015: 37th European Conference on Information Retrieval*, Springer. Springer, 03/2015 2015.

[10] D. M. Wilkinson and B. A. Huberman, "Cooperation and quality in wikipedia," in *Proceedings of the 2007 International Symposium on Wikis*, ser. WikiSym '07. New York, NY, USA: ACM, 2007, pp. 157–164. [Online]. Available: http://doi.acm.org/10.1145/1296951.1296968

[11] S. Biancani, "Measuring the quality of edits to wikipedia," in *Proceedings of The International Symposium on Open Collaboration*, ser. OpenSym '14. New York, NY, USA: ACM, 2014, pp. 33:1–33:3. [Online]. Available: http://doi.acm.org/10.1145/2641580.2641621

[12] B. T. Adler, K. Chatterjee, L. de Alfaro, M. Faella, I. Pye, and V. Raman, "Assigning trust to wikipedia content," in *Proceedings of the 4th International Symposium on Wikis*, ser. WikiSym '08. New York, NY, USA: ACM, 2008, pp. 26:1–26:12. [Online]. Available: http://doi.acm.org/10.1145/1822258.1822293

[13] S. Javanmardi and C. Lopes, "Statistical measure of quality in wikipedia," in *Proceedings of the First Workshop on Social Media Analytics*, ser. SOMA '10. New York, NY, USA: ACM, 2010, pp. 132–138. [Online]. Available: http://doi.acm.org/10.1145/1964858.1964876

[14] G. H. Golub and C. F. Van Loan, *Matrix computations*. JHU Press, 2012, vol. 3.

[15] W. Yining, W. Liwei, L. Yuanzhi, H. Di, C. Wei, and L. Tie-Yan, "A theoretical analysis of ndcg ranking measures," in *Proceedings of the 26th Annual Conference on Learning Theory*, 2013.