Automatic Assessment of Document Quality in Web Collaborative Digital Libraries

DANIEL HASAN DALIP and MARCOS ANDRÉ GONÇALVES,
Universidade Federal de Minas Gerais
MARCO CRISTO, Federal University of Amazonas
PÁVEL CALADO, Instituto Superior Técnico/INESC-ID

14

The old dream of a universal repository containing all of human knowledge and culture is becoming possible through the Internet and the Web. Moreover, this is happening with the direct collaborative participation of people. Wikipedia is a great example. It is an enormous repository of information with free access and open edition, created by the community in a collaborative manner. However, this large amount of information, made available democratically and virtually without any control, raises questions about its quality. In this work, we explore a significant number of quality indicators and study their capability to assess the quality of articles from three Web collaborative digital libraries. Furthermore, we explore machine learning techniques to combine these quality indicators into one single assessment. Through experiments, we show that the most important quality indicators are those which are also the easiest to extract, namely, the textual features related to the structure of the article. Moreover, to the best of our knowledge, this work is the first that shows an empirical comparison between Web collaborative digital libraries regarding the task of assessing article quality.

Categories and Subject Descriptors: H.3.7 [**Information Storage and Retrieval**]: Digital Libraries, User Issues

General Terms: Human Factors, Measurement, Experimentation

Additional Key Words and Phrases: Quality assessment, quality features, wiki, machine learning, SVM

ACM Reference Format:
Dalip, D. H., Gonçalves, M. A., Cristo, M., and Calado, P. 2011. Automatic assessment of document quality in web collaborative digital libraries. ACM J. Data Inform. Quality 2, 3, Article 14 (December 2011), 30 pages. DOI = 10.1145/2063504.2063507 http://doi.acm.org/10.1145/2063504.2063507

1. INTRODUCTION

The Web 2.0 phenomenon and its highly collaborative nature is currently giving rise to a new type of repository of human knowledge. Such repositories exist in the form of blogs, forums, or collaborative digital libraries, whose collection of documents is maintained by the Web community itself [Dondio et al. 2006; Krowne 2003]. More importantly, they are characterized as being freely accessible, both for reading and for writing.

This work is partially supported by INWeb (MCT/CNPq grant 57.3871/2008-6) and by the authors' individual grants and scholarships from CNPq, CAPES, and FAPEMIG.
Authors' addresses: D. H. Dalip and M. A. Gonçalves, Computer Science Department, Universidade Federal de Minas Gerais; email: {hasan, mgoncalv}@dcc.ufmg.br; M. Cristo, Federal University of Amazonas; email: marco.cristo@dcc.ufam.edu.br; P. Calado, Instituto Superior Técnico/INESC-ID; email: pavel.calado@tagus.ist.utl.pt.
Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from the Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701, USA, fax +1 (212) 869-0481, or permissions@acm.org.
© 2011 ACM 1936-1955/2011/12-ART14 $10.00
DOI 10.1145/2063504.2063507 http://doi.acm.org/10.1145/2063504.2063507

The proliferation of collaboratively generated content leads us to think that, in the near future, this content will be predominant on the Web. Currently, there is a very large number of hosting services that allow the free editing of its content by the end users [Fogg et al. 2003; Rubio et al. 2010]. Each of these services hosts numerous collections dedicated to specific communities and subjects, such as geographic information, sports, technology, science, TV shows, science fiction, books, general knowledge, among others.

The most popular of such services, *Wikia*[1], has grown from one hundred to several thousands of collections in just a few years, containing more than four million pages of rich content. Another example of how communities can produce collaborative content on a large scale is that of Wikipedia[2]. This online encyclopedia took only two years to reach as many articles as the Encyclopedia Britannica. It currently contains more than seventeen million articles, written in dozens of different languages [Wikipedia 2010a]. There are also many blogs and micro-blogs, such as *Twitter*[3], where users can exchange opinions about diverse subjects, such as politics, daily life, culture, among others [Maged et al. 2006].

However, such freedom also creates an important issue: given the rhetoric of democratic access to everything, by everyone, at any time, how can a user determine the quality of the information provided? Currently, content generated in a more traditional, centralized manner and published using physical media, such as books or journals, is still naturally seen as higher quality and more trustworthy [Dondio et al. 2006]. Nevertheless, the growth and dissemination of collaboratively created content is such that mechanisms to assess the quality and trust of this type of material should be provided. Thus, an answer to this question is needed.

A possible solution to this problem would be to automatically estimate the quality of the documents in the digital library. Such estimates could be used as an indicator of which documents need revision, to detect vandalism or inadequate revision methods, or even to recommend articles based on their estimated quality and reliability level. In this work, we study a method for automatically estimating the quality of collaborative content. The method was first proposed in Dalip et al. [2009] where it was used to estimate the quality of Wikipedia articles. Here, we further improve the method and apply it to other online collaborative digital libraries, presenting a detailed study of different available pieces of evidence in the estimation of content quality. Also, by testing our method in three different online Web collaborative digital libraries, we provide an assessment of their main differences and similarities.

From our experiments, we observed that the most useful features were those associated with the text structure, which were in fact the easiest to extract from freely available collaborative encyclopedias. We observed that the best results are achieved when Structure features are combined with Network and Revision features and that the articles whose quality level is harder to assess are those of higher quality. For Wikia collections, which use a star-based quality taxonomy, we found that the articles most difficult to classify were the extreme cases, that is those with the highest or lowest quality. We also noted that collections using the star-based taxonomy were, in general, harder to assess than those using the Wikipedia-based taxonomy, probably due to the lack of a standard quality criteria.

In sum, the main contribution of this article is a thorough analysis of the capability of an automatic method to estimate content quality. The studied method learns content quality from articles described using a very large set of features (some proposed

[1]http://www.wikia.org
[2]http://www.wikipedia.org
[3]http://www.twitter.com

by us), using a state-of-the-art learning strategy, which was chosen based on our explicit definition of the problem. The articles are derived from three disparate online collaborative digital libraries using different quality rating systems. Our consistent results throughout a large body of experiments, and analysis allow us to make more generalizable conclusions than any previous work.

This article is organized as follows. Section 2 covers related work. Section 3 discusses our proposed methods for automatic article quality assessment. Section 4 presents the features we explore to represent an article and its relative quality in a Web collaborative digital library. Section 5 describes our datasets, methodology, experiments, and discusses the results. Section 6 concludes the article, including possible future work.

2. RELATED WORK

Some previous work has focused on the quality of collaborative digital libraries on the Web, many of which use Wikipedia as a study case. For example, Brown [2009] justifies the good quality of Wikipedia articles by comparing their evolution with that of biological species. According to the author, revisions are similar to mutations, and articles resulting from bad revisions are not well adapted, thus not surviving for long, since a *natural selection* process would favor articles resulting from good revisions. Brown cites Giles [2005] where a comparison is performed showing that the quality of Wikipedia is similar to that of the Encyclopedia Britannica. Nevertheless, the weak editorial control of Wikipedia has allowed major errors to escape this selection process. An example is the case of a false biography of a well-known North American reporter, which was online for four months before it was detected. This case stirred much controversy and led to the revision of editorial policies of Wikipedia [P. Dondio and Weber 2006].

The need to assess the quality of the content available on the Web has motivated several efforts reported in the literature. In Veltman [2005], the author suggests that, in the future, the Internet should provide mechanisms to deal with multiple variants of content, as well as the degree of certainty and importance of its claims. An example of such a mechanism is proposed by Chu [1997], who proposes it is possible to compute the credibility of a claim based on its sources and editors. These solutions, however, assume that all the necessary information will be provided by authors and/or users. If we consider the free nature of the Internet, such a requirement may be very hard to achieve. Furthermore, this may not even be desirable, due to privacy and security concerns.

All these challenges have stimulated the development of solutions that attempt to estimate content quality and credibility in more realistic scenarios, that is, not expecting any extra information from the authors and users and using only the sources of evidence available. Such sources of evidence previously explored in the literature, some in different contexts, include, for example, hyperlinks [Alexander and Tate 1999], trustworthiness [Fogg et al. 2001], writing style [Argamon et al. 2003; Zheng et al. 2006b], network features [Korfiatis et al. 2006], intensity of cooperative behavior [Wilkinson and Huberman 2007], history of reviews [Adler and de Alfaro 2007; Cusinato et al. 2009; Hu et al. 2007; Wöhner and Peters 2009; Zeng et al. 2006a]. Particularly, Hu et al. [2007] were the first to propose a metric in which the quality of an article is based on the quality of its reviewers and, recursively, the quality of the reviewers is based on the quality of the articles they reviewed. We use this metric as a feature and as a baseline in our experiments [4].

[4]A few other simple extensions of this original scheme do exist (e.g., Cusinato et al. [2009], Wöhner and Peters [2009]) but are only preliminary work which has been tested in restricted scenarios and has not been demonstrated to be superior to the original proposal.

Based on these previous studies, several authors have proposed combining sources of evidence into a unique value to represent quality. For instance, P. Dondio and Weber [2006] and Dondio et al. [2006] suggested a methodology for estimating the quality and credibility of articles in Wikipedia. They realized that specific features of this collection, such as the continuous updating of its content and the relative independence between article versions, make it difficult to apply any of the previously proposed methods. Thus, they combine several pieces of evidence to build an article ranking that tries to jointly capture certain aspects of quality, such as stability, editing quality, and importance. These pieces of evidence are extracted from the article revision history, textual content, and hyperlink structure and combined into a unique final ranking.

Differently from the approaches proposed by P. Dondio and Weber [2006] and Dondio et al. [2006], which use simple linear combination methods, a few other efforts have been proposed for combining the available evidence using machine learning techniques. One such case is that of Rassbach et al. [2007], which suggest the use of a Maximum Entropy Model [Borthwick et al. 1998] to estimate the quality of the articles, according to quality classes previously assigned by human evaluators. In addition, the authors also propose some new text-based sources of evidence for the problem like the number of phrases, auxiliary verbs, and the Kincaid readability index [Ressler 1993]. Similarly, Bethard et al. [2009] have proposed to deal with the issue as a classification problem, presenting a method to estimate document quality in educational digital libraries. Since quality can change according to the user's perspective, they define different dimensions of quality and create an indicator for each dimension. For instance, for the dimension "appropriate pedagogical guide", the indicators used were "contains instructions?" and "identifies the learning objects?". Once these indicators were defined, a Support Vector Machine was trained to classify the library article. Further, De la Calzada and Dekhtyar [2010] propose a machine learning approach to estimate the quality of articles regarding two categories: stabilized articles and controversial articles.

In Dalip et al. [2009], we have proposed treating quality estimation as a regression problem. In other words, we estimate the quality of articles in Wikipedia as a grade in a continuous quality scale. To accomplish this, we use a Support Vector Regression method [Drucker et al. 1996; Vapnik 1995]. Our main contribution in this work is a detailed study of the various sources of evidence and their impact on the prediction of the quality of a Wikipedia article. Furthermore, the proposed method was shown to achieve better results overall than the best approaches previously proposed in the literature.

This work greatly extends the work of Dalip et al. [2009] by experimenting with two other collaborative digital libraries, with some interesting properties regarding estimation of quality (e.g., different criteria for quality evaluation) in addition to a different much larger sample of Wikipedia than what was used before. Furthermore, a detailed study of the different sources of evidence used to evaluate the quality of the articles is performed, with interesting conclusions about their usefulness in each scenario.

3. ASSESSING ARTICLE QUALITY

In Wikipedia, the quality of an article is assigned as a value on a discrete scale. Articles are classified (from the lowest to the highest quality) as "stub", "start", "BC", "GA", "AC", and "featured". We note, however, that in general quality can be seen as a value on a continuous scale. In fact, this is the most natural interpretation for the problem, if we consider that there are better or worst articles even inside the same discrete

category. For instance, in Wikipedia, we have class AC articles that (a) have recently been promoted and await expert evaluation; (b) have been evaluated by experts and await corrections; and (c) have been corrected and await promotion to featured article. In the case of other Wikis, a continuous scale is commonly used where users score each article with a value from 1 to 5, and the final quality value is the average of all scores.

Considering quality as a continuous value is also appropriate for applications in which we need to evaluate articles within the same class. This would be the case for instance of a system that would suggest which articles in a class should be reviewed for promotion or demotion. Those that are closer to a certain level of quality should have priority. Continuous scales are also easier to interpret in a probabilistic model. Finally, for a human user, and has to choose between alternative versions of an article (e.g., two different translations), a numerical value will allow a finer distinction even if the articles are in the same class.

For these reasons, in this work, we consider quality on a continuous scale. Consequently, the problem of learning to evaluate quality will be modeled as a numerical regression task. More specifically, we will apply a state-of-the-art method for regression learning, Support Vector Regression (SVR) [Drucker et al. 1996]. This method is described in the following section.

3.1 Support Vector Regression

To apply SVR to the quality estimation task, we represent the articles to be classified as follows. Let $A = \{a_1, a_2, ..., a_n\}$ be a set of articles. Each article a_i is represented by a set of m features $F = \{F_1, F_2, ..., F_m\}$, such that $\mathbf{a_i} = (f_{i1}, f_{i2}, ..., f_{im})$ is a vector representing article a_i, where each f_{ij} is the value of feature F_j in a_i. In this work, the term *feature* describes a statistic value that represents a measurement of some quality indicator associated with an article. For instance, f_{ij} could represent the length of article a_i.

In our proposal, we assume that we have access to some *training data* of the form $\{(a_1, q_1), (a_2, q_2), ..., (a_n, q_n)\} \subset A \times \mathbb{R}$, where each pair (a_i, q_i) represents an article and its corresponding quality assessment value, such that if $q_1 > q_2$, then the quality of article a_1 as perceived by the user is greater than the quality of article a_2. The solution we propose to this problem consists of (1) determining the set of features $\{F_1, F_2, ..., F_m\}$ used to represent the articles in A; and (2) applying a regression method to find the best combination of features to predict the quality value q_i for any given article a_i.

The problem of regression is to find a function f which approximates the mapping between an input domain and real numbers based on a training sample. In our case, the input domain is given by the set of articles, A, and the real numbers correspond to quality assessments q. We refer to the difference between the hypothesis (i.e., the prediction) and the true value $(f(a) - q), q \in \mathbb{R}, a \in A$, as the *error*. The importance of the error is measured by a loss function. The main idea behind SVR is to use a loss function (called ϵ-intensive) that does not consider error values situated within a certain distance of the true value. One way of visualizing this method is to consider a region of size $\pm\epsilon$ around the hypothesis function, where ϵ denotes a margin. Any training point lying outside this region (i.e., beyond the margin) is considered an example of an error as illustrated by Figure 1(a). In the figure, f_1 and f_3 represent the margins around hypothesis function f_2. Thus, in this work, our goal is to find a function $f : A \rightarrow \mathbb{R}$ that has at most ϵ deviation from the actual targets $q \in \mathbb{R}$ for all the training data.

In SVR, the input article a is first mapped onto an m-dimensional feature space using some nonlinear mapping Φ. Then, a linear model is constructed in this feature space. More formally, the linear model $f(\mathbf{a}, \mathbf{w})$ is given by $f(\mathbf{a}, \mathbf{w}) = \langle \mathbf{w}, \Phi(\mathbf{a}) \rangle + b$, where $\mathbf{w}$ is a weight vector of m feature values, b is the bias term, and $\langle \mathbf{w}, \Phi(\mathbf{a}) \rangle$

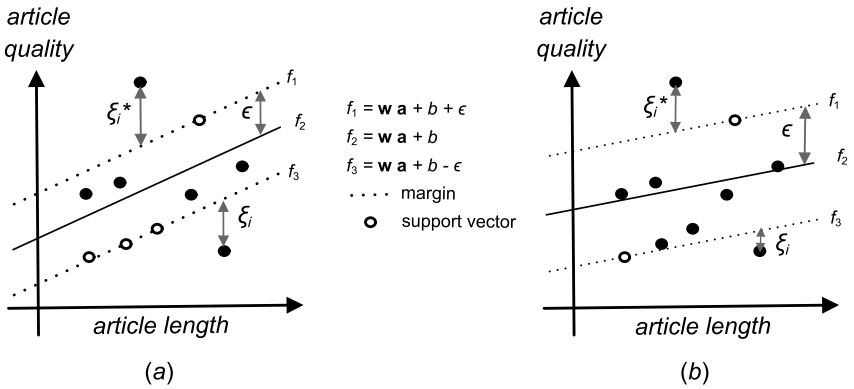

Fig. 1. Regression problem with one numeric target (article quality) and ten articles (points) represented by their lengths. Note that two articles in both graphics are considered examples of an error because they lie outside the area delimited by the margins. Their distances to the margins are given by ξ_i^* and ξ_i, respectively. Figures (a) and (b) represent regressions performed using two different ϵ-values.

denotes the inner product between $\mathbf{w}$ and $\Phi(\mathbf{a})$. The quality of estimation is measured by the ϵ-intensive loss function $L^\epsilon(q,\, f(\mathbf{a}, \mathbf{w}))$ defined in Eq. (1):

$$L^\epsilon(q,\, f(\mathbf{a}, \mathbf{w})) = \begin{cases} 0 & \text{if } |q - f(\mathbf{a}, \mathbf{w})| \leq \epsilon \\ |q - f(\mathbf{a}, \mathbf{w})| - \epsilon & \text{otherwise.} \end{cases} \tag{1}$$

SVR performs a linear regression in the high-dimension feature space using the ϵ-insensitive loss function while it tries at the same time to reduce the model complexity by minimizing $\|\mathbf{w}\|^2$. The linear regression of the loss function is performed by minimizing error estimates $(q_i - f(\mathbf{a_i}, \mathbf{w}))$ and $(f(\mathbf{a_i}, \mathbf{w}) - q_i)$, measured, respectively, by nonnegative slack variables ξ_i^* and ξ_i. If we consider f_1 the margin above f and f_3 the margin below f, ξ_i^* measures deviations above f_1, whereas ξ_i measures deviations below f_3, as shown in Figure 1. Thus, SVR can be formulated as the convex optimization problem of minimizing:

$$\frac{1}{2}\|\mathbf{w}\,\|^2 + C \sum_{i=1}^{n} \left(\xi_i + \xi_i^*\right), \tag{2}$$

subject to:

$$\begin{aligned} |q_i - f\,(\mathbf{a_i}, \mathbf{w})| &\leq \epsilon + \xi_i^* \\ |\,f\,(\mathbf{a_i}, \mathbf{w}) - q_i| &\leq \epsilon + \xi_i \\ \xi_i, \xi_i^* > 0, 0 < i &\leq n, \end{aligned}$$

where $C > 0$ is a constant parameter. This optimization problem can be transformed into the dual problem and its solution is given by Eq. (3):

$$f(\mathbf{a}) = \sum_{i=1}^{n_{\mathrm{SV}}} \left(\alpha_i + \alpha_i^*\right) \kappa\,(\mathbf{a_i}, \mathbf{a}),\ \text{subject to } 0 < \alpha_i, \alpha_i^* \leq C, \tag{3}$$

where n_{SV} is the number of support vectors (vectors lying on the margins, depicted as white circles in Figure 1), and κ is an inner product function (*kernel function*) in a given vector space, given by $\kappa(\mathbf{a_i}, \mathbf{a}) = \sum_{j=1}^{m}(\Phi_j(\mathbf{a_i})\Phi_j(\mathbf{a}))$.

Note that the SVR estimation accuracy depends on a good setting for C, ϵ, and the kernel parameters. C determines the trade-off between the model complexity (flatness)

and the degree to which deviations larger than ϵ are tolerated. If C is too large, the objective becomes simply to minimize $\frac{1}{n}\sum_{i=1}^{n} L^{\epsilon}(q_i, f(\mathbf{a_i}, \mathbf{w}))$. Parameter ϵ controls the width of the ϵ-insensitive zone, which is used to fit the training data. Bigger ϵ-values use fewer support vectors at the expense of providing more "flat" estimates as we can see in Figure 1(a) and Figure 1(b).

We have chosen SVR because of its advantages over other methods, such as the presence of a global minimum solution resulting from the minimization of a convex programming problem, its relatively fast training speed, and its capability of dealing with sparseness in the representation of the solution [Chu et al. 2001]. In this work, we solve the quadratic optimization problem given by Eq. (3) using the SVMLIB software package [Chang and Lin 2001]. In our experiments we have used a radial basis function (RBF) as κ. Other parameters were chosen using cross-validation [Mitchell 1997] within the training set with the data scaling and parameter selection tool provided by the SVMLIB package [Hsu et al. 2000]. In the next section, we discuss in detail the features used to represent the articles.

4. ARTICLE REPRESENTATION

Determining which features should be used to represent an article is a key decision in a regression-based quality assessment. Here we focus on features we believe to be discriminative enough to determine content quality. We based our selection on the quality criteria defined for Wiki articles [Wikipedia 2010e; Wookieepedia 2010b, 2010c] such as completeness, neutral point-of-view, good organization, factual accuracy, and provision of references to relevant sources of information. More specifically, we have implemented all of the features found in recent previous work (e.g., P. Dondio and Weber [2006], Dondio et al. [2006], Hu et al. [2007], Rassbach et al. [2007]) that try to capture these quality aspects or to combine them, disregarding only a few features that were too specific to a particular library so that our methods could be easily generalized for different libraries[5]. These were also combined with other features, such as those related to the connectivity of the articles [Benevenuto et al. 2009] not exploited in previous work. Additionally, we also propose new features indicated in the following. In the process of creating these new features, we tested some variations but here we only report those that produced the best results in our experiments[6]. In the next sections, we present in detail the quality features adopted and how to compute their values.

4.1 Review Features

Review features are those extracted from the review history of each article. These features are useful for estimating the maturity and stability of an article. It can be expected that good quality articles have reached a maturity level at which extensive corrections are unnecessary. Furthermore, the lack of stability could indicate controversial content and a non-neutral point-of-view, aspects to be avoided in good articles. The features we study here are the following.

—*Age (r-age).* Age (in days) of the article. Very recent articles are not normally considered of very high quality since they usually have not had time to go through a refinement process.

[5]There were in fact only two features related to some specific Wikipedia tagging regarding blocked and controversial articles.

[6]In fact, the only new features not reported here were those that were clearly redundant and did not produce good results.

— *Discussion count (r-dc).* Number of discussions posted by the users about the article. This is useful to infer conflict resolution, coordination of reviews, and teamwork dynamics. The authors in Kittur and Kraut [2008], for example, demonstrate that an article gains an advantage from a larger number of reviewers only when their reviews are conducted in a coordinated manner.
— *IP and user count (r-ipc,r-uc).* Number of reviews made by anonymous (*r-ipc*) and registered (*r-uc*) users.
— *Review count (histNumRevisoes).* Total number of reviews.
— *Reviews per user (r-rpu).*

$$r\text{-}rpu = \frac{r\text{-}rc}{usrCount}, \tag{4}$$

where *usrCount* is the number of users who reviewed the article. This feature is useful to infer how much an article is reviewed when contrasted with the number of reviewers. The idea behind this feature is that, if an article has many revisions but made only by few users, the chance that the article is expressing a specific point-of-view of the reviewers is higher.
— *Standard deviation of the reviews per user (r-rsdpu).* This feature is useful to infer how balanced the reviewing process is among the reviewers.
— *Modified lines rate (r-mlr).* Number of lines modified when comparing the current version of an article to a version three-months old. This is a good indicator of how stable an article is.
— *Last three-months review rate (r-lurr).*

$$r\text{-}lurr = \frac{threeMounthRevCount}{r\text{-}rc}, \tag{5}$$

where *threeRevMounthCount* is the number of revisions in the last 3 months. As in the following two features, if an article is less than 3 months old, the review rate will be 100%.
— *Most active users review rate (r-maurr).* Percentage of reviews made by the top 5% most active reviewers.
— *Occasional users review rate (r-ourr).* Percentage of reviews made by users who edited the article less than four times.
— *ProbReview (r-pr).* Proposed by Hu et al. [2007], this measure tries to assess the quality of a Wiki article based on the quality of its reviewers. Recursively, the quality of the reviewers is based on the quality of the articles they reviewed. Formally, the quality of article i, Q_i, is calculated as

$$Q_i = \sum_k q_{ik}, \tag{6}$$

where q_{ik} is the quality of the k^{th} term of article i. The quality of a term is given by the probability of a user having modified this term multiplied by the quality of the user, as shown in Eq. (7).

$$q_{ik} = \sum_j prob(w_{ik}, u_j) U_j. \tag{7}$$

The quality of the user is obtained by the summation of the probabilities of the user having reviewed each word multiplied by the quality of each word according to Eq. (8):

$$U_j = \sum_{i,k} prob(w_{ik}, u_j) q_{ik}. \tag{8}$$

The probability of a user u editing the term p is defined as 1 if u has created p in the article. Otherwise, the probability is reduced according to the distance between p and the closest term created by u, using the formula:

$$prob(p,u) = \frac{1}{\sqrt{max(|d_{pl}| - 7, 0) + 1}}, \tag{9}$$

where d_{pl} is the distance of p to the closest term l also created by u.

Feature r-age was proposed by Rassbach et al. [2007] and Mingus [2008]. The remaining features, except r-pr, were proposed by P. Dondio and Weber [2006]. The following features we propose in this work.

— *Age per review (r-ar).* Ratio between age and number of reviews on the last 30 revisions. Used to verify the average length of time an article remains without revision. If the article has less then 30 reviews, the total number of reviews is used.
— *Reviews per day (r-pd).* Percentage of reviews per day, to verify how frequently the article has been reviewed. Reviews per day is defined as:

$$r\text{-}pd = \frac{r\text{-}rc}{r\text{-}age}. \tag{10}$$

4.2 Network Features

Network features are those extracted from the connectivity network inherent to the collection. In this case, we see the collection as a graph where nodes are articles and edges are the citations between them. The main motivation for using these features is that citations between articles can provide evidence about their importance. For instance, as pointed out by P. Dondio and Weber [2006], a good and mature article should be stable. However, it may be stable because no one is interested in it due to its poor quality or lack of information. Therefore, it is important to take into account the popularity of the article, something that can be estimated by measures taken from its connectivity network. Network features we study here are as follows.

— *PageRank (n-pr).* Used in Rassbach et al. [2007], this is the PageRank value of an article, calculated according to Brin and Page [1998]. This feature captures the popularity of an article among the editors of the collection. The general idea of Pagerank is that a link from page p_1 to page p_2 indicates that p_1 considers the content of p_2 important. Thus, the importance of a page p is proportional to the importance and quantity of pages that point to p. Since page importance could be correlated to content quality, this metric can be viewed as a quality metric. Given its widespread use in search systems as a quality indicator, we will also use it as a baseline in this work.
— *In degree (n-id).* Number of citations of an article from other articles. Like PageRank, this feature reflects the popularity of an article.
— *Out-degree (n-od).* Proposed in Rassbach et al. [2007], Mingus [2008], and P. Dondio and Weber [2006], this feature counts the number of links to other articles.
— *Link count (n-lc).* Number of links to other articles, used in Rassbach et al. [2007], Mingus [2008], and P. Dondio and Weber [2006]. This feature is different from out-degree, since it also counts links to articles yet to be written.
— *Translation count (n-tc).* Proposed by Rassbach et al. [2007] and Mingus [2008], this feature computes the number of versions of the article in other languages. It captures the universality, popularity, and importance of an article.

The next features were previously used to detect spam in Web pages and online video systems, such as YouTube [Benevenuto et al. 2009]. However, to the best of our knowledge, they have not been previously used for assessing the quality of articles:

— *Assortativity in-in, in-out, out-in, and out-out* (*n-aii*, *n-aio*, *n-aoi, and n-aoo*). The *assortativity* of a node is defined in Eqs. (11), (12), (13), and (14).

$$n\text{-}aii = \frac{n\text{-}id}{avgInDegree}, \tag{11}$$

$$n\text{-}aio = \frac{n\text{-}id}{avgOutDegree}, \tag{12}$$

$$n\text{-}aoi = \frac{n\text{-}od}{avgInDegree}, \tag{13}$$

$$n\text{-}aoo = \frac{n\text{-}od}{avgInDegree}, \tag{14}$$

where *avgInDegree* and *avgOutDegree* are, respectively, the in-degree and out-degree average of the node graph neighbors. This feature tries to estimate the similarities between the node and its neighbors. For example, if an article has a high in-degree and $n - aii = 1$, it means that this node is as popular as its neighbors.

— *Clustering coefficient (n-cc)*.

$$n\text{-}cc = \frac{edgesCount(k)}{maxEdges(k)}. \tag{15}$$

Here $edgesCount(k)$ is the number of the k-nearest-neighbors (i.e., number of nodes which have a path to n whose length is, at most, k edges), and $maxEdges(k)$ is the number of possible edges between a node and its k-nearest-neighbors. First used in Benevenuto et al. [2009] and Dorogovtsev and Mendes [2003], this feature is used to indicate if an article belongs to a group of articles related to each other.

— *Reciprocity (n-rcy)*. Ratio between the number of articles that cite the current article and those that are cited by it. This feature tries to estimate the quality of the pointers from an article to others. For example, if one node has a high reciprocity, it might mean that the article points to very related topics and documents not just to completely random or uncorrelated ones given that most of the referenced articles point back to this article.

4.3 Text Features

Text features are those extracted from the textual content of the articles. Since half of the features we study are derived from the text, we further divided them into four subgroups: length, style, structure, and readability.

The major advantage of using text features is that these are less demanding to obtain, in terms of the prepossessing required when compared to other features, such as the review history or those based on link analysis. Moreover, these features are simpler to extract from articles in any collaborative encyclopedia or digital library, making the method more easily applicable in different contexts.

4.3.1 Length Features. Length features, used by Hu et al. [2007], P. Dondio and Weber [2006], and Rassbach et al. [2007], are indicators of the article size. The general intuition behind them is that a mature and good quality text is probably neither too short,

which could indicate an incomplete topic coverage, nor excessively long, which could indicate verbose content. Further, in Wikis, *stub articles* (draft quality) are expected to be short, which reinforces the correlation between length and quality. In this work, we use the following features.

— *Character count (tl-c).* Number of characters in the text, including spaces.
— *Word and Phrase count (t-lw,tl-pc).* These two features correspond to the number of words and phrases in the text, respectively.

4.3.2 Structure Features. These features are indicators of how well the article is organized. According to Wiki quality standards [Wikipedia 2010e; Wookieepedia 2010b, 2010c] a good article must be organized so that it is clear, visually adequate, and provides the necessary references and pointers to additional material. Thus, we use features derived from the article structure in an attempt to describe its section organization and its use of images, links, and citations. These features are the following.

— *Section count (ts-sc).* The number of sections in the article. The intuition behind this feature is that a good article is organized in sections. In particular, in Wikipedia, a good article is expected to include sections such as introduction, summary, list of references, and external links.
— *Standard deviation of the section size (ts-3sd).* This feature represents how balanced the sections are. The intuition behind it is that the content of a good article is not concentrated in a few sections.
— *Subsection count (ts-sbc).* The intuition behind this feature is that large sections of a good article are further organized into subsections.
— *Citation count (ts-ctc).* The number of citations in the article. The intuition behind this feature is that a good article supports claims with references.
— *External link count (ts-xlc).* Number of external links. A good article should provide additional information available on the Web.
— *Image count (ts-ic).* The number of images in the article. Pictures should contribute to make content clearer and visually pleasant.

The features *ts-sc, ts-3sd, ts-ctc, ts-xlc, ts-ic* were proposed by P. Dondio and Weber [2006] to predict the article's editing quality, while *ts-sc* and *ts-sbc* were proposed by Rassbach et al. [2007] and Mingus [2008]. The following features are original proposals of this work.

— *Mean section size (ts-mss).* This feature is useful to verify how well organized an article is, since draft articles usually have only one big section or sections with short length.

$$ts\text{-}mss = \frac{tl\text{-}c}{ts\text{-}sc} \tag{16}$$

— *Mean paragraph size (ts-mps).* This feature is useful to verify the length distribution of the paragraph, since a good article is separated in paragraphs, which are not too long which could indicated a lack of structure in the text.

$$ts\text{-}mps = \frac{tl\text{-}c}{paragraphCount}, \tag{17}$$

where *paragraphCount* is the number of paragraphs on the text.
— *Size of the largest and the shortest section (ts-sls, ts-3s).* These features are useful to detect unusual section organization or articles with empty or very small sections, which could indicate incomplete content and drafts.

— *Mean number of subsections per section (ts-msbs)*. Like the features *ts-mss*,*ts-mss*, and *ts-mss* this feature verifies how well the article is organized.

$$ts\text{-}msbs = \frac{ts\text{-}sbc}{ts\text{-}sc}. \tag{18}$$

— *Abstract size (ts-as)*. The size in characters of the introduction section. Mature articles are expected to have an introductory section summarizing its content.

— *Citation count per text length (ts-ctcl)*. This is a relative version of *ts-ctc*, which takes into account the length of the article.

$$ts\text{-}ctcl = \frac{ts\text{-}ctc}{tl\text{-}c}. \tag{19}$$

— *Citation count per section (ts-ctcs)*. Relative version of ts-ctc which takes into account the number of sections in the article. The intuition behind this feature is that a good article provides a balanced distribution of citations.

$$ts\text{-}ctcs = \frac{ts\text{-}ctc}{ts\text{-}sc}. \tag{20}$$

— *Links per text length (ts-ltl)*. This feature is used to see how well the links are distributed throughout the text.

$$ts\text{-}ltl = \frac{n\text{-}lc}{tl\text{-}c}. \tag{21}$$

— *External links per section (ts-xls)*. Number of external links divided by the number of sections.

$$ts\text{-}xls = \frac{ts\text{-}xlc}{ts\text{-}sc}. \tag{22}$$

— *Images per section (ts-is)*. We use this feature as an indicator of the distribution of images throughout the text.

$$ts\text{-}is = \frac{ts\text{-}ic}{ts\text{-}sc}. \tag{23}$$

4.3.3 Style Features. These features are intended to capture the way the authors write the articles through their word usage. The intuition behind them is that good articles should present some distinguishable characteristics related to word usage, such as short sentences. To compute them and the Readability indicators described in the next Section, we use the Style and Diction[7] software program. Table I shows the terms considered to compute each feature. For example, for feature *ty-scc*, the terms used were “the”, “a”, and “an”.

— *Size of the largest phrase (ty-slp)*. Number of words of the largest phrase.

— *Large phrase rate (ty-lpr)*. Percentage of phrases whose length is ten words greater than the article’s average phrase length. The value of ten was empirically obtained from the training set, using the Style and Diction software program.

— *Short phrase rate (ty-spr)*. Percentage of phrases whose length is five words less than the article’s average phrase length. The value of five was obtained as for the previous feature.

— *Auxiliary verb, question, pronoun and passive voice count (ty-avc, ty-qc, ty-pc, ty-pvc).* Number of auxiliary verbs, questions, pronoun, and passive voice sentences in the text.

[7]http://www.gnu.org/software/diction/

Table I. Terms to Compute Each Style Feature

Feature	Utilized Terms
ty-avc	will, shall, cannot, may, need to, would, should, could, might, must, ought, ought to, can't, can
ty-pc	I, me, we, us, you, he, him, she, her, it, they, them, thou, thee, ye, myself, yourself, himself, herself, itself, ourselves, yourselves, themselves, oneself, my, mine, his, hers, yours, ours, theirs, its, our, that, their, these, this, those, your
ty-cjr,ty-scc	and, but, or, yet, nor
ty-nr (suffixes)	tion, ment, ence, ance
ty-pr,ty-spc	aboard, about, above, according to, across from, after, against, alongside, alongside of, along with, amid, among, apart from, around, aside from, at, away from, back of, because of, before, behind, below, beneath, beside, besides, between, beyond, but, by means of, concerning, considering, despite, down, down from, during, except, except for, excepting for, from among, from between, from under, in addition to, in behalf of, in front of, in place of, in regard to, inside of, inside, in spite of, instead of, into, like, near to, off, on account of, on behalf of, onto, on top of, on, opposite, out of, out, outside, outside of, over to, over, owing to, past, prior to, regarding, round about, round, since, subsequent to, together, with, throughout, through, till, toward, under, underneath, until, unto, up, up to, upon, with, within, without, across, along, by, of, in, to, near, of, from
ty-ber	be, being, was, were, been, are, is
ty-scc	the, a, an
ty-sscc	after, because, lest, till, 'til, although, before, now that, unless, as, even if, provided that, provided, until, as if, even though, since, as long as, so that, whenever, as much as, if, than, as soon as, inasmuch, in order that, though, while
ty-sipc	why, who, what, whom, when, where, how

— *Conjunction rate (ty-cjr)*.

$$ty\text{-}cjr = \frac{cjCount}{t\text{-}lw}, \tag{24}$$

where $cjCount$ is the total number of conjunctions in the text.

— *Nominalization rate (ty-nr)*.

$$ty\text{-}nr = \frac{sbCount}{t\text{-}lw}, \tag{25}$$

where $sbCount$ is the number of nominalizations on the text.

— *Preposition rate (ty-pr)*.

$$ty\text{-}pr = \frac{prepCount}{t\text{-}lw}, \tag{26}$$

where *prepCount* is the number of prepositions on the text.

— *"To be" verb rate (ty-ber)*.

$$ty\text{-}ber = \frac{toBeCount}{t\text{-}lw}, \tag{27}$$

where $toBeCount$ is the number of "To be" verb on the text.

— *Beginning of sentence features*. Number of phrases that start with a pronoun (*ty-sac*), article (*ty-scc*), conjunction (*ty-scc*), subordinating conjunction (*ty-sscc*), interrogative pronoun (*ty-sipc*) and preposition (*ty-spc*).

The use of these features was first proposed by Rassbach et al. [2007] and Mingus [2008].

4.3.4 Readability Features. These features, first used in Rassbach et al. [2007], are intended to estimate the age or US grade level necessary to comprehend a text. They comprise several metrics combining counts of words, sentences, and syllables. The intuition behind these features is that good articles should be well written, understandable, and free of unnecessary complexity. The features are as follows.

— *Automated Readability Index (tr-ari)*. This metric was proposed in Smith and Senter [1967] and consists of using the average number of words per sentence and the average number of characters per word to estimate the readability. To accomplish this, the authors analyzed samples of the textbooks used in the Cincinnati Public School System from which they derived the multiple regression formula given by Eq. (28).

$$tr\text{-}ari = 4.71\frac{characters}{words} + 0.5\frac{words}{sentence} - 21.43. \tag{28}$$

— *Coleman-Liau (tr-cl)*. This metric was proposed in Coleman and Liau [1975] and consists of the average number of characters per word and the number of sentences in a fragment of 100 words (wf). The use of fragments was suggested to make it feasible to use an optical scanning device. The resulting regression equation, derived from 36 previously ranked 150-word passages, is given by Eq. (29).

$$tr\text{-}cl = 5.89\frac{characters}{words} - 0.3wf - 15.48. \tag{29}$$

— *Flesch reading ease (tr-fre) and Flesch-Kincaid (tr-fk)*. These metrics were presented in Flesch [1948] and Ressler [1993], respectively, to infer the reading comprehension difficulty of passages of contemporary English. The first formula, given by Eq. (30), was built to predict the average grade level of a child who could correctly answer three-quarters of a standard reading test (McCall-Crabbs' Standard test) about a given passage. To accomplish this, the authors used the average sentence length and the average number of syllables per word. This metric is then normalized so that it computes a value between 0 and 100, where 0 indicates a text which is hard to understand.

$$tr\text{-}fre = 206.835 - 1.015\frac{words}{sentences} - 84.6\frac{syllables}{words}. \tag{30}$$

The second metric provides US grade levels instead of values between 0 and 100, making it easier for teachers, parents, librarians, and others to judge the readability level of textual content.

$$tr\text{-}fk = 0.39\frac{words}{sentences} + 11.8\frac{syllables}{words} - 15.59. \tag{31}$$

— *Gunning Fog Index (tr-fog)*. This metric was proposed in Gunning [1952] and uses the average number of words per sentence and the average number of complex words in the text, such that short sentences in plain English achieve a better score than long sentences written in a complicated language. A complex word is defined as a word with three or more syllables. The resulting formula, derived by regression analysis, is given by Eq. (32).

$$tr\text{-}fog = 0.4(\frac{words}{sentences} + 100\frac{complexwords}{words}). \tag{32}$$

— *Läsbarhets index (tr-lix)*. This metric was proposed in Björnsson [1968] to assess text difficulty in several languages. It is similar to *tr-fog*, but complex words are simply defined as words with more than six characters. The result is a positive

number, where the higher its value, the more difficult the text is to read. The regression formula is given by Eq. (33).

$$tr\text{-}lix = \frac{words}{sentences} + 100\frac{complexwords}{words} \tag{33}$$

— *Smog-Grading (tr-sg)*. This metric was proposed in McLaughlin [1969] and uses the average number of polysyllabic words, excluding proper names, taken from a sample of 30 sentences. It was devised to eliminate multipliers and use only a simple constant, so that it would be easy to apply. As in previous works, the formula Eq. (34) was derived by regression analysis of text fragments used in McCall-Crabbs' Standard test.

$$tr\text{-}sg = 3 + \sqrt{polysyllables}. \tag{34}$$

In the following Section, we apply the described features, along with the SVR method, to automatically assess the quality of articles in three different collections and evaluate the results.

5. EXPERIMENTS

Using the features described in Section 4, we performed a set of experiments using three different test collections. We now describe our experimental design, the collections, and the results.

5.1 Datasets

In our experiments, we used a sample of Wikipedia in English and samples of two other collaborative encyclopedias provided by the Wikia service. For all these collections, the articles can be freely read and edited, except for those that were vandalized, which can be edited only by registered users. Moreover, the articles have a discussion page, where users can discuss their editions and resolve possible conflicts. All three collections also provide human judgments about the content quality of many of their articles. The contents of these datasets are freely available for download and use the same repository format, which allows us to use the same programs to extract features.

We chose the English Wikipedia because it is a large collaborative encyclopedia with more than three million articles, where more than half have their quality evaluated by users [Wikipedia 2010d]. As previously mentioned, the full content of Wikipedia is freely available for download [Wikipedia 2010f]. From now on, we refer to this encyclopedia as WIKIPEDIA.

Any user can evaluate a Wikipedia article, according to the following quality taxonomy [Wikipedia 2010d][8].

— Featured Article (*FA*). These are, according to their evaluators, the best Wikipedia articles.
— A-Class (*AC*). These are articles considered complete, but with a few pending issues that need to be solved in order to be promoted to Featured Articles.
— Good Article (*GA*). Articles without problems of gaps or excessive content. These are good sources of information, although other encyclopedias could provide better content.
— B-Class (*BC*). Articles that are useful for most users, but researchers may have difficulties in obtaining more precise information.

[8]Note that, currently, there is also an intermediate class between ST and BC, the *C-Class*. We do not use this class because it did not exist at the time we performed our crawling.

— Start-Class (*ST*). Articles still incomplete although containing references and pointers for more complete information.
— Stub-Class (*SB*). These are draft articles with very few paragraphs. They also have few or no citations.

Articles from Wikipedia generally are evaluated by groups of users who belong to different projects. A project is a group of articles sharing a common topic (e.g., Evolution, Biographies, Places). Projects are maintained by users who can help to organize, manage, evaluate, and, in some cases, define a standard editing process [Wikipedia 2010g]. This is the case of the article about Charles Darwin, which is associated with projects Biographies and Evolution. Since an article can belong to multiple projects, it can receive multiple and distinct quality evaluations. For instance, the article about Darwin can be considered *FA* by the Evolution project but only *GA* by the Biographies project, since they have different requirements. As we decided not to deal with multi-classification, we only use articles which were classified into one category by all their projects.

Also note that the process of quality assignment is not the same for all categories. In particular, only the classification of an article as *FA* or *GA* requires a prior indication by a registered user. After the indication, other registered users can vote for the article recommending or not recommending its promotion. Through this voting, they judge if the article meets the quality requirements for the target class [Wikipedia 2010b, 2010c].

From the Wikia service, we selected the encyclopedias Wookieepedia[9], about the Starwars universe, and Muppet[10], about the TV series "The Muppet Show". These are the two Wikia encyclopedias with the largest number of articles evaluated by users regarding their quality[11]. Their repositories are freely available for download [Muppet 2010; Wookieepedia 2010a].

The Wookieepedia collection provides two distinct quality taxonomies. The first is a subset of the Wikipedia quality taxonomy. It comprises the classes *FA*, *GA*, and *SB*. The second is based on the taxonomy commonly provided with Wikia datasets, that is, a star-based taxonomy where the worst articles receive one star and the best articles receive five stars. Unlike Wikipedia, the final rating of a Wookieepedia article is obtained as the average of the ratings provided by all the users that evaluated it. As a consequence, Wookieepedia articles can have a fractional rating value, such as 2.7 stars. Since these taxonomies are not compatible with each other, we extracted two different samples of Wookieepedia. The first sample was built according to the Wikipedia-based taxonomy and, from now on, we refer to it as STWR_3CLASS. The second sample was derived according to the star-based taxonomy, and we refer to it as STWR_5CLASS. Finally, the Muppet collection provides only a star-based taxonomy, which we refer to as MUPPET.

To create our sample for each Wiki collection, we first extracted all the articles from the smallest quality class and then randomly drew the same number of articles from the remaining classes. Tables II and III present the class distribution of the datasets, at the time we collected them. Note that, for the star-based taxonomies in Table III, we considered the rounded ratings. We chose to use balanced samples since SVR can be biased towards the class that has more elements, which could harm our analysis regarding the relative difficulty of classification in each class. Moreover, recent research

[9]http://starwars.wikia.com/
[10]http://muppet.wikia.com/
[11]To obtain the article evaluations we used the APIs provided at http://starwars.wikia.com/api.php and http://muppet.wikia.com/api.php.

Table II. Article Distribution Per Class, for Samples WIKIPEDIA and STWR_3CLASS

Class	# Articles in sample/# Articles in the Collection					
	FA	AC	GA	BC	ST	SB
WIKIPEDIA	549/1.868	549/549	549/2.935	549/47.827	549/272.552	549/721.995
STWR_3CLASS	482/699	–	482/543	–	–	482/482

Table III. Article Distribution Per Class, for Samples STWR_5CLASS and MUPPET

Class	# Articles in sample/# Articles in collection				
	5	4	3	2	1
MUPPET	310/1382	310/1053	310/1101	310/384	310/310
STWR_5CLASS	1836/6773	1836/6376	1836/6045	1836/2320	1836/1836

Table IV. Sample Sizes

Dataset	# Articles	# Revisions	# Edges	# Nodes	Version date
WIKIPEDIA	3.294	1.992.463	86.077.675	3.185.457	jan/2008
MUPPET	1.550	38.291	282.568	29.868	sep/2009
STWR_3CLASS	1.446	127.551	1.017.241	106.434	oct/2009
STWR_5CLASS	9.180	369.785	1.017.241	106.434	oct/2009

has shown that training samples do not need to follow the original class distribution in order to obtain good results and that a balanced distribution is a good alternative [Weiss and Provost 2003].

For all datasets, we also collected the links between the articles, in order to extract network attributes. Table IV presents information about the total number of articles and revisions of each sample and the network graphs derived from the datasets. In this table, the edges correspond to links between pages, and the nodes correspond to the article pages of the complete collection and article-redirecting pages. We used the program Web Graph [Boldi and Vigna 2004] to create the graph and extract all the Network attributes.

5.2 Evaluation Methodology

We performed experiments with two main goals. The first was to analyze the impact of each group of features in the quality assessment of online encyclopedia articles. The second was to analyze the differences and relative difficulties for assessing quality in distinct online Web collaborative digital libraries.

To evaluate the impact of the selected features, we used the *information gain* measure (infogain, for short) [Mitchell 1997]. Infogain is a statistical measure of how much a given feature contributes to discriminate the class to which any given article belongs. Although it is normally used for feature selection, since it provides a ranking of features based on their discriminative power, here we use it simply to study the analyzed features. Infogain was, therefore, computed for all features and the results are reported in Section 5.3.3. We also experimented with other feature selection methods, such as Chi-Square with very similar results, hence we only report the use of infogain.

The effectiveness of the proposed classification method was evaluated using the *mean squared error* measure (MSE). MSE is defined as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n} e^2, \tag{35}$$

where e is the error value and n is the number of articles. We compute error e as the absolute difference between the quality value predicted and the true quality value extracted from the database. In our experiments, we used quality values from 0 (Stub article) through 5 (Featured Article) for WIKIPEDIA, 0 (Stub article) through 3 (Featured Article) for STWR_3CLASS, and 1 (one star) through 5 (five stars) for STWR_5CLASS and MUPPET.

To compare our approach with previously proposed methods, we also used a ranking comparison metric called *Normalized Discounted Cumulative Gain at top k* (NDCG@k, for short). This metric, first proposed in Järvelin and Kekäläinen [2000], allows us to measure how close the predicted quality ranking of articles is to their true quality ranking. More formally, NDCG@k is defined as:

$$NDCG@k = \frac{1}{N} \sum_{i=1}^{k} \left(\frac{r_i}{\log_2(i+1)}\right), \tag{36}$$

where r_i is the true quality assessment for the article at position i in the ranking, and N is a normalization factor. The factor N is equal to the *discounted cumulative gain* (the sum part in Eq. (36)) of an *ideal ranking*, that is, a ranking where all FA Class articles come first, followed by the AC articles, and so on. Thus, the higher the high quality documents are placed in the ranking, the higher the value for NDCG@k. This NDCG@k formulation was also used in Hu et al. [2007].

To perform the comparative experiments, we used a 10-fold cross-validation method [Mitchell 1997]. Each dataset was randomly split in 10 parts, such that, in each run, a different part was used as a test set, while the remaining parts used as the training set. The split on training and test sets was the same in all experiments. The final results of each experiment represent the average of the 10 runs.

For all comparisons reported in this work, we used the signed-rank test of Wilcoxon [1945] to determine if the difference in effectiveness was statistically significant. This is a nonparametric paired test that does not assume any particular distribution on the tested values. In all cases, we only drew conclusions from results that were considered statistically significant with a 95% confidence level. To make sure that the results were not biased by an inappropriate choice of parameters, several experiments were performed and, in all cases, we report only the best results obtained.

5.3 Results

In this section, we present our experimental results. First, we performed a thorough analysis of the utilized features. Since an analysis considering all possible feature combinations would be prohibitive due to the high number of features, we conducted our study by initially evaluating groups of features organized according to the description in Section 4 (Section 5.3.1 and 5.3.2), and next we investigated the sets of features within each group separately (Section 5.3.3). Following, in Section 5.3.4, we conducted an analysis of the effectiveness of our method in the four collections on a per-class basis. In Section 5.3.5, we compare our method to several baselines published in the literature. Finally, in Section 5.3.6, we discuss some implications of our findings for other data and information quality problems over the Internet.

5.3.1 Analysis of the Groups of Features. To analyze the impact of each group of features, we divided our feature set into 6 groups: Structure, Length, Style, Revision History, Network, and Readability. We then conducted two series of experiments. First, we applied each feature group alone to the dataset in order to determine its individual impact. Second, we applied all groups, leaving one out at a time.

Table V. SVR Effectiveness by Using Feature Groups Taken in Isolation (Isolated) or Excluded from the Complete Set (Excluded) for Each Dataset. Features Are Ranked According to Their MSE

Sample	Group	Isolated		Excluded	
		error (MSE)	loss	error (MSE)	loss
WIKIPEDIA	All	0.851	–	–	–
	Structure	0.893*	4.92%	0.892*	4.75%
	Length	0.968*	13.72%	0.858*	0.82%
	Style	1.200*	41.16%	0.846	-0.56%
	Revision History	1.224*	43.74%	0.982*	15.30%
	Network	1.256*	47.57%	0.854	0.35%
	Readability	2.351*	176.21%	0.848	-0.36%
MUPPET	All	1.644	-	–	–
	Revision History	1.746*	6.15%	1.780*	8.22%
	Structure	1.747*	6.23%	1.656*	0.72%
	Length	1.795*	9.18%	1.672*	1.69%
	Style	1.820*	10.65%	1.682	2.28%
	Network	1.848*	12.37%	1.635*	-0.58%
	Readability	1.877*	14.11%	1.693*	2.93%
STWR_3CLASS	All	0.071	-	-	-
	Structure	0.079*	10.69%	0.083*	16.88%
	Style	0.100*	36.88%	0.071	0.62%
	Length	0.109*	52.27%	0.071*	0.35%
	Network	0.142*	99.82%	0.073	2.20%
	Revision History	0.148*	108.37%	0.074	4.18%
	Readability	0.280*	294.29%	0.072	1.56%
STWR_5CLASS	All	1.670	-	-	-
	Revision History	1.740*	4.18%	1.72*	3.04%
	Structure	1.750*	4.96%	1.69*	1.11%
	Network	1.770*	5.91%	1.69*	1.69%
	Style	1.790*	7.42%	1.67	0.4%
	Length	1.820*	9.44%	1.67	0.01%
	Readability	1.880*	13.08%	1.67	-0.01%

Table V presents the results obtained with each test dataset for each of the feature groups. These groups were evaluated when taken in isolation (column "Isolated") and when excluded from the full feature set (column "Excluded"). For each group, we present the MSE and the percentage of loss over the MSE result obtained by the baseline, that is, the method that uses all the features. A "*" indicates a statistically significant difference from the baseline.

We start by analyzing results obtained for WIKIPEDIA. As we can observe in Table V, the Structure group, when taken in isolation, achieved the best effectiveness, whereas the Readability group presented the worst. This suggests that indicators associated with article organization such as sections, images, and citations are more useful for distinguishing article quality. On the other hand, Readability features taken in isolation are not good indicators of the quality of the articles. This is not surprising since Readability features do not take into consideration how well the article covers its topic. Thus, even a draft can be considered a good article if it is readable according to the metrics of Section 4.3.4.

Length and Style features presented reasonable effectiveness when taken in isolation. However, when excluded from the whole set of features, they presented low or no impact at all on effectiveness. This can be partially explained by the fact that some Style features are redundant with length features, since they provide similar information. For instance, the number of phrases that start with a preposition (a Style feature) is generally proportional to the document length. We can also see that effectiveness is much worse when Revision History features are not taken into consideration, which indicates the importance of features associated with the maturity and with the frequency which an article is edited.

Some of the results obtained for the MUPPET sample were similar to those obtained with WIKIPEDIA. The best indicators in both collections were Revision History and Structure features. In fact, Revision History features presented good effectiveness both when taken in isolation and when excluded. Their effectiveness was the best for datasets with star-based taxonomies. This may be due to the fact that older articles, with more revisions and which are more stable, tend to have more reviews by more people. This consequently implies a more precise evaluation of the article, since, in the case of star-based taxonomies, we use the average of the ratings. Structure features, when taken in isolation, were the best or second to best group in all the datasets. Readability features were always the worst.

When comparing all the datasets, we note that the STWR_3CLASS sample presented the smallest MSE values for all the feature groups. Such effectiveness can be partially attributed to the characteristics of the classes in this dataset, since the STWR_3CLASS taxonomy includes the *SB* class, which can be easily identified due to its usually very small article length. Thus, errors will occur mostly between classes *FA* and *GA*, yielding small error values, since the distance between these classes from which the error is computed is 1.

We also note that errors are larger for star-based taxonomies. This is probably due to the lack of precise criteria to characterize them. Different from the Wikipedia-based taxonomy, the number of stars is not associated with any standard mandatory criteria. As a consequence, criteria used in star-based taxonomies are personal and much more subjective. For instance, citations in Wikipedia have to be present for an article to be classified as *FA*, but no similar criterion would be required to classify a Wookieepedia article as five stars. Thus, a user can give a high rating to a Wookieepedia article about a certain character only because he/she likes that character. These factors motivated us to make a detailed per-class effectiveness analysis presented in Section 5.3.4.

By analyzing the results obtained with the STWR_5CLASS sample, we observe that the removal of the Revision History feature has the highest impact on the final effectiveness. Nevertheless, this is the dataset which was less affected by removing feature groups. This may be explained by the fact that no particular feature group is much more effective than the others in this sample, and the attributes that remain may have somewhat similar discriminative power than those removed.

5.3.2 Ranking of Features. For each sample, we also produced a general ranking containing all features ordered by infogain. Table VI shows the distribution of features for each group at the top 10 to 60 ranking positions. Results confirm the importance of Text and Revision features for assessing the quality of the articles. Textual features occupy 8 of the 10 first positions in the rank on the samples of WIKIPEDIA and STWR_3CLASS. Moreover, in all samples, features related to text size occupy the three top positions of the ranking. Thus, as mentioned before, length seems to be one important aspect to discriminate between levels of quality. Interestingly, as seen in Table V, text features that explore other aspects of the text, such as Structure, have produced better effectiveness in terms of MSE. This may be due to the fact that the infogain

Table VI. Number of Features at Top Positions, Ranked Using Infogain

Sample	Group	# of features at top...					
		10	20	30	40	50	60
WIKIPEDIA	**Length**	3	3	3	3	3	3
	Structure	3	6	9	13	15	16
	Style	2	4	6	9	10	13
	Revision History	2	6	8	10	12	13
	Network	-	1	4	5	9	10
	Readability	-	-	-	-	1	5
STWR_3CLASS	**Length**	3	3	3	3	3	3
	Structure	2	5	9	13	13	14
	Style	3	6	7	9	13	14
	Revision History	1	5	7	10	11	13
	Network	1	1	4	4	6	9
	Readability	-	-	-	1	4	7
STWR_5CLASS	**Length**	3	3	3	3	3	3
	Structure	1	3	5	10	14	16
	Style	-	4	7	10	11	14
	Revision History	5	8	10	11	12	14
	Network	1	2	5	6	10	11
	Readability	-	-	-	-	-	2
MUPPET	**Length**	3	3	3	3	3	3
	Structure	2	4	5	8	13	15
	Style	1	3	6	7	12	14
	Revision History	3	6	10	13	13	14
	Network	1	4	6	9	9	10
	Readability	-	-	-	-	-	4

metric does not take into consideration dependencies among the features. In other words, the Length features are good when analyzed independently, but other types of features, like Structure, are better when combined with each other. Revision History features also obtained good ranking values. They occupy 5 out of the 10 first positions on STWR_5CLASS and 3 out of 10 on MUPPET and STWR_3CLASS.

Finally, we note that although the majority of the Network features appear only in the top 30 in WIKIPEDIA, in the remaining samples they presented better effectiveness. In others words, features related to the importance and popularity of the article are better in samples in which the quality criteria are less clear, implying that there is a higher chance that popular articles will receive higher rates. In all of the samples, Readability features did not obtain good ranking values. These results, together with those from Table V, allow us to conclude that they may be unconsidered when computing article quality.

5.3.3 Feature Analysis Within Groups. Although the first experiment conducted was useful to evaluate the relative importance of each group of features, we were not able to determine how redundant the features within each group are. To accomplish this goal, we computed the infogain value for each feature. More specifically, we evaluated our quality predictor using the top N% most discriminating features according to infogain and observed the impact of the error as we varied the value of N. The results of this experiment are presented in Figure 2.

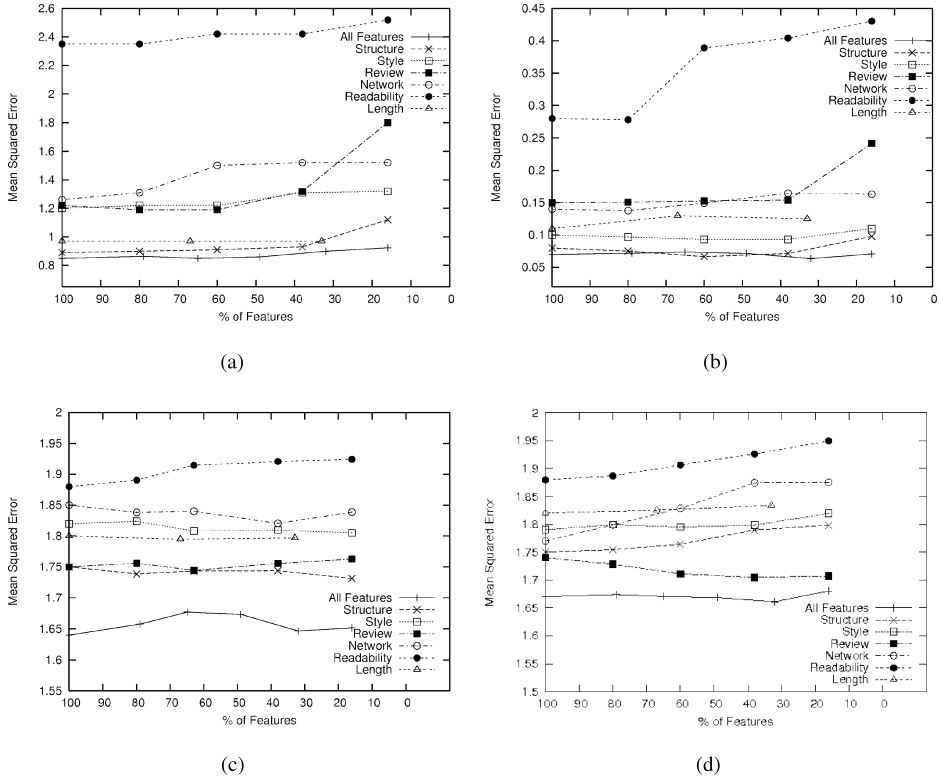

Fig. 2. Mean Squared Error obtained when using only the $N\%$ features with the highest infogain in WIKIPEDIA (a), STWR_3CLASS (b), MUPPET (c), and STWR_5CLASS (d).

As we can see in Figure 2, in all samples the curves for Style and Length do not seem to be much affected by the removal of features, which is an indication that the features in these groups are very redundant, capturing similar quality aspects. As a consequence, we also note that, for all datasets, a large number of the features in each group could be removed without significant impact on the overall effectiveness.

Regarding our best two groups, Structure and Review History, we note that large losses are observed only for WIKIPEDIA and STWR_3CLASS when just the top 20% of features were used. Regarding the Structure group, this corresponds to the removal of features Section Count (*ts-sc*), Citation Count per Section (*ts-ctcs*), and Mean Section Size (*ts-mss*) in WIKIPEDIA. In STWR_3CLASS, it corresponds to the removal of features Mean Section Size (*ts-mss*), Abstract Size, (*ts-as*), and Links per Text Length (*ts-ltl*). These results highlight the significance of features that indicate the importance of organizational aspects of the article, such as sections, abstract, citations and phrases.

In the case of the Review group, this corresponds to the removal of features Reviews per User Standard Deviation (*r-rsdpu*), Review Count (*r-rc*), and ProbReview (*r-pr*), in WIKIPEDIA. It corresponds to the removal of the same features together with the feature Review Per Day (*r-pd*) in STWR_3CLASS. All these features are useful to verify the frequency of updates in the articles and the quality of the users that make such updates.

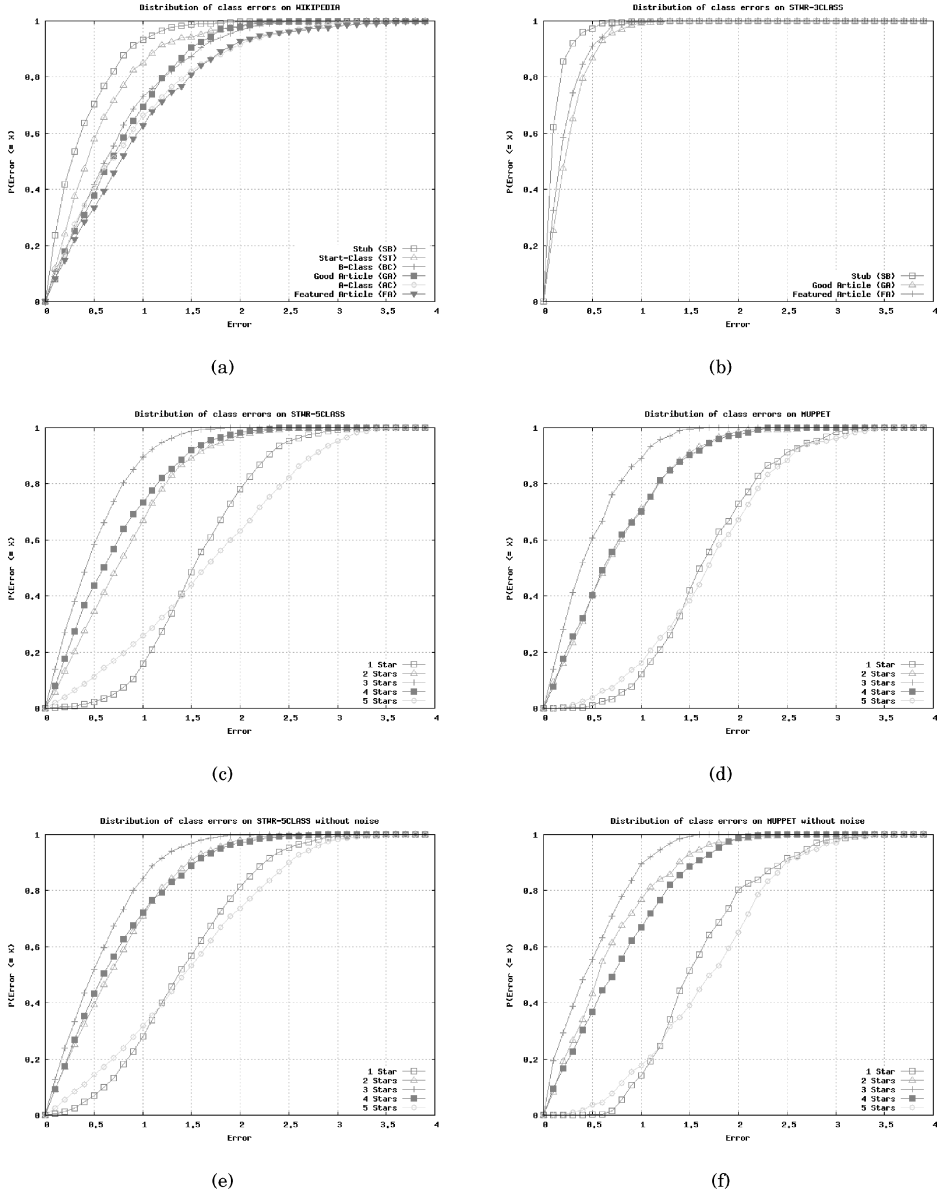

Fig. 3. Cumulative class distribution function of articles with errors less than or equal to a certain threshold for WIKIPEDIA (a), STWR_5CLASS (b), STWR_5CLASS (c), and MUPPET(d). Figures (e) and (f) correspond to datasets STWR_5CLASS and MUPPET after removing noisy articles.

5.3.4 Evaluation of Effectiveness per Class. To verify the distribution of errors in each class, in Figure 3 we show the percentage of articles with error values inferior to a given threshold. As we can see, for the WIKIPEDIA sample, articles from *SB*, and *ST* classes were the easiest to classify. More than 85% of them presented an error inferior to 1, they were incorrectly classified to their closest class. We also note that the higher the quality level, the greater the error. On the other hand, articles from

low-quality levels can be more clearly distinguished. For instance, articles from the *ST* class can be easily separated from articles of the *SB* class since they are usually less structured and contain fewer references. Such a distinction is increasingly harder for higher quality articles.

The same fact is observed in the STWR_3CLASS sample where the easiest class to estimate was *SB*. It is interesting to note that, in this collection, good quality articles were easier to distinguish than in Wikipedia. This is probably due to the smaller taxonomy since the largest error distance for this collection is 2, while it is 5 for WIKIPEDIA.

For the datasets using star-based taxonomies, effectiveness was weaker for 1 and 5-star classes, that is, for the worst and the best articles. After a careful analysis of these samples, we noted some anomalies, such as the fact that 16% of the 5-star articles have only 5 sentences in STWR_5CLASS. This anomaly is probably related to incomplete articles incorrectly evaluated by the users since in sample STWR_3CLASS, which was extracted from the same collection, the *FA* and *GA* articles have an average of 110 sentences, and there are no articles with only 5 sentences. Furthermore, 41% of the Stub articles have no more than 5 sentences. Clearly, such characteristics contribute to the poor effectiveness rating of the predictor.

In order to analyze the impact of these cases, we performed another experiment. We removed articles that were rated as 5 stars but had no more than 5 sentences. The results are shown on the Figure 3(e) and 3(f) for the datasets STWR_5CLASS and MUPPET, respectively. As expected, this change consistently reduced the MSE by 6.5% in MUPPET and by 15.4% in STWR_5CLASS. Also, we note that the largest improvements were in classes with 1 and 5 stars, which explains the higher error rates observed. This is likely due to the higher subjectiveness of the evaluations.

5.3.5 Comparison to Previous Work. In this section, we compare our method to previous work published in the literature. We start with the PageRank algorithm whose goal is to capture the popularity of an article [Brin and Page 1998]. The idea is to verify how popularity relates to quality. The second baseline is ProbReview [Hu et al. 2007], described in Section 4.1. Notice that these two methods are also used as features by us. In addition, we compare our method to Dondio et al. [2006], P. Dondio and Weber [2006], and Rassbach et al. [2007] which are state-of-the-art methods that, like ours, combine different types of evidence for article quality assessment. In the remaining of this paper, we will refer to these methods as PAGE_RANK, PROB_REVIEW, DONDIO and RASSBACH, respectively. Since the original implementations of the algorithms of PROB_REVIEW, DONDIO and RASSBACH are not publicly available, we implemented them ourselves.

For the algorithms DONDIO and RASSBACH, we implemented two versions. The first ones, which we call DONDIO-LINEAR and RASSBACH-MAXENT, correspond to the methods as originally described in P. Dondio and Weber [2006] and Rassbach et al. [2007]. In particular, in the case of DONDIO-LINEAR, we combine the pieces of evidence necessary to compose the rankings using a nonweighted linear combination, or in other words, the average of their results. The RASSBACH-MAXENT algorithm uses a Maximum Entropy classification method to combine the proposed set of features. For evaluation and comparative analysis, the classification results obtained with this method were transformed into a ranking using the output probability of an article belonging to a given class. For the second versions, which we call DONDIO-SVR and RASSBACH-SVR, we simply use the features originally proposed by the authors as input to our SVR predictor. In this way, we can verify the effectiveness of our proposed set of features when directly compared with the features proposed by P. Dondio and Weber [2006] and Rassbach et al. [2007].

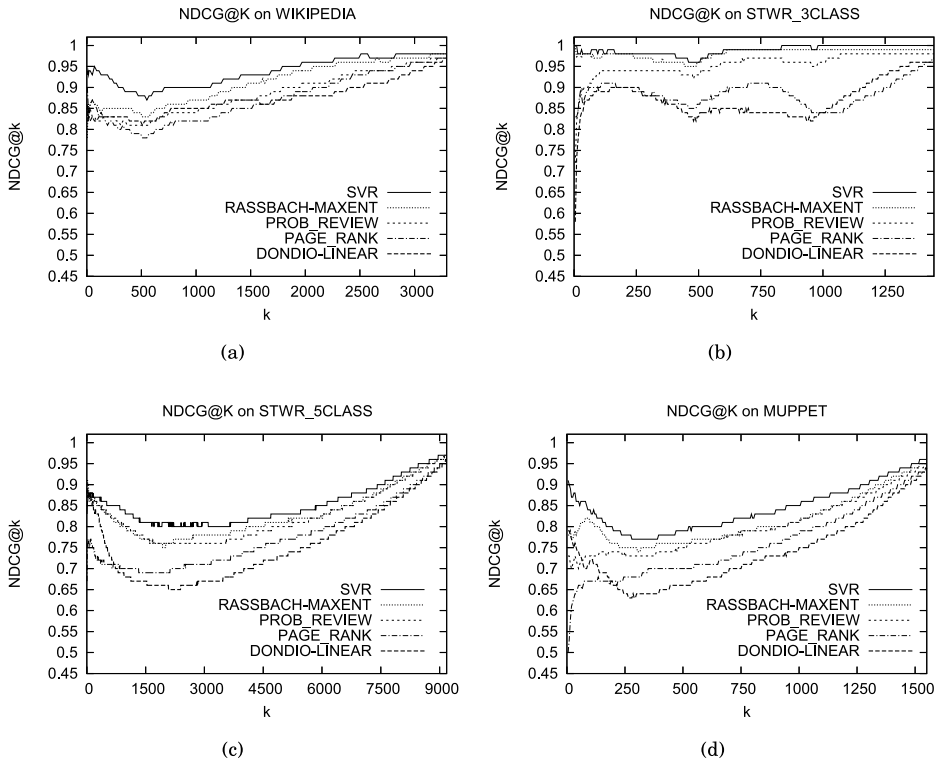

Fig. 4. NDCG@k obtained by the DANTE-LINEAR, RASSBACH-MAXENT, and SVR methods in WIKIPEDIA (a), STWR_3CLASS (b), STWR_5CLASS (c), and MUPPET (d).

Figure 4 shows the NDCG@k figures for DONDIO-LINEAR, RASSBACH-MAXENT, PAGE_RANK, PROB_REVIEW and our SVR-based method, using the best article representations proposed for each method in the samples WIKIPEDIA, STWR_3CLASS, STWR_5CLASS, and MUPPET, respectively. In Figures 4(a) and 4(b), we observe that the methods based on machine learning approaches (RASSBACH-MAXENT and SVR) outperform the remaining methods for all values of k in the respective samples. SVR was the best in all samples, with a clear advantage in WIKIPEDIA. In the case of STWR_3CLASS, the effectiveness of SVR and RASSBACH-MAXENT was basically the same, with no significant difference. On the others samples, SVR consistently outperformed all methods, while RASSBACH-MAXENT and PROB_REVIEW yielded similar NDCG results on STWR_5CLASS. All the differences pointed out are statistically significant according to the Wilcoxon test.

Figure 5 shows the NDCG@k figures for SVR, DONDIO-SVR, and RASSBACH-SVR methods using the best article representations proposed for each method. By observing both Figures 4 and 5, we can note that the adoption of the SVR approach improved DONDIO effectiveness in all the samples and RASSBACH in WIKIPEDIA. More importantly, in all samples, SVR presents superior or equivalent effectiveness when compared to the other two methods, which can be attributed to a better set of features. All of the results are statistically significant.

5.3.6 Implications For Other Data and Information Quality Problems over the Internet. Although we have focused on the quality assessment of articles in online collaborative encyclopedias, some of our results may have implications for other data and quality problems on

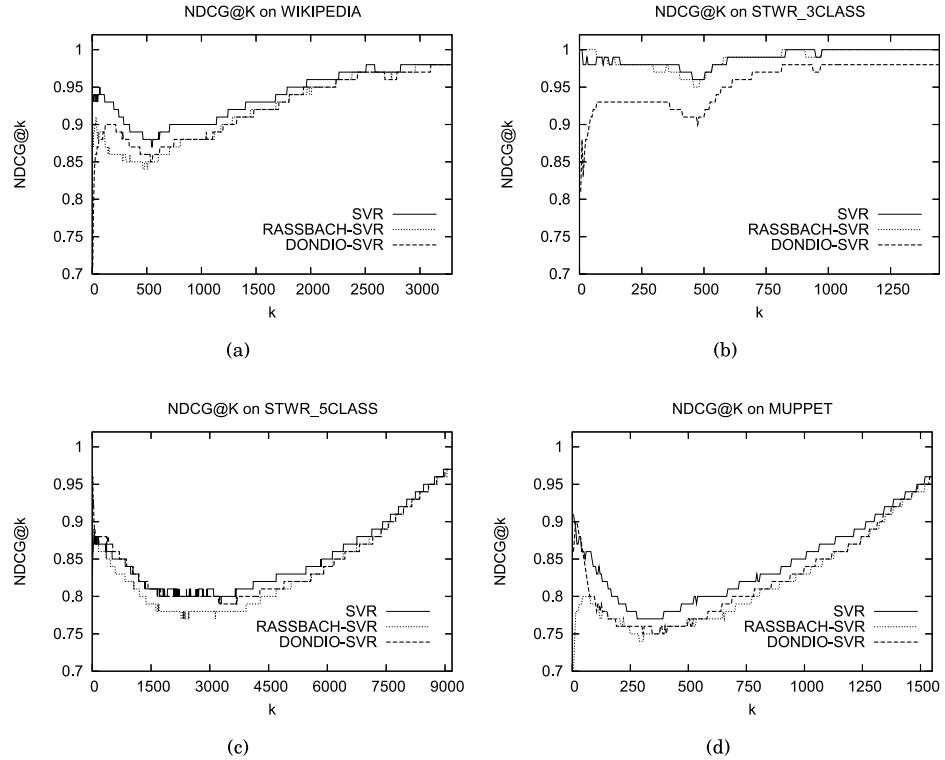

Fig. 5. NDCG@k obtained by the DANTE-SVR, RASSBACH-SVR, and SVR methods on WIKIPEDIA (a), STWR_3CLASS (b), STWR_5CLASS (c), and MUPPET (d).

the Internet. For example, our consistent results for Wikipedia and the Wikia samples indicate that similar results may be obtained using our proposed techniques and set of features for other collaborative Wiki-like projects or even for some other collaborative projects such as community-oriented blogs or Question and Answer (Q&A) sites such as Yahoo!Answers [Agichtein et al. 2008].

The fact that much simpler text-based features produced comparable or even better results than much more complex and costly features, such as those related to Revision History and Citation Network, may inspire simplifications in methods based on, for example, mutual reinforcement (e.g., hubs and authorities) or random surfers models over large networks (e.g., PageRank).

On the other hand, the usually poor effectiveness of Readability features indicates that these may have to be adapted to each specific collection, mainly in regard to the constant values used in the respective formulas. Moreover, for cases like the assessment of quality in very short and informal messages like those issued in microblogs such as Twitter[12], these features may have to be completely rethought. For example, instead of a static equation, we could use machine learning to combine differents aspects of the text to infer the readability of a post.

Using inverse reasoning, lack of quality may also indicate other types of problems such as spam or the presence of attacks [Chin et al. 2010]. Thus, methods for detecting these problems could benefit from our proposal.

[12]For example, users trying to influence the opinions of others may have a tendency to create better written messages [Bigonha et al. 2010].

Further, our results can benefit designers of applications and tools that try to estimate or communicate content quality issues to users, such as the work of Chevalier et al. [2010] and Pirolli et al. [2009].

Finally, content quality is a very subjective concept and depends on aspects that can be different among particular communities. As such, it depends on different evaluation criteria. In Wikipedia, for instance, different project groups can evaluate the quality of the same article considering aspects as diverse as its adherence to specific structural style and content coverage [13]. Thus, for any system which aims at automatically assessing content quality, it would be very useful to learn about its target community and that community specific quality criteria. This is even more important when we consider much more diversified content, such as those found in blogs, twitter, and News pages, to cite a few.

6. CONCLUSIONS

In this work, we studied the impact of different features on the estimation of content quality on three collaborative digital libraries. We studied features derived from the textual content (Length, Structure, Readability, and Style), from the Revision History, and from the article Citation Graph. From this analysis, we observed that the most useful feature group was the one associated with the text Structure. Interestingly, these features are also those easiest to extract from freely available collaborative encyclopedias. We also noted that the best results are achieved when Structure features are combined with Network and Revision features.

Through experiments, we were also able to conclude that the articles in Wikipedia whose quality level is harder to distinguish are those of higher quality. For Wikia collections using star-based taxonomies, we found that the more ambiguous articles were the extreme cases, that is, those with very good and those with very bad quality. In general, we also noted that collections using a star-based taxonomy were harder to classify than those using the Wikipedia-based taxonomy, which is probably due to the lack of criteria associated with their quality rating.

There are many possibilities for future work. For example, we aim to extend our study to consider articles with multiple quality classifications. We also intend to study ways to remove noisy examples (outliers) from the training data in order to improve the quality assessment. Other features could be designed and explored, for example, based on assigning ratings to both users and articles. Such features can be used in mutually reinforcing strategies such as "good users tend to submit good articles" or exploring the (social) connections among users. The parameters of some features, such as the values used by Readability features, can also be learned on a per collection basis. The analysis of the discriminative power of the features could be improved by considering methods that exploit correlations and dependencies among features, such as those described in Guyon and Elisseeff [2003]. While some preliminary studies using different machine learning methods have not yielded results better than the ones shown in this article, we intend to continue testing different learning approaches, including metalearning strategies. Tools could be designed to assist users with the process of quality classification and assessment of articles. Furthermore, user studies where human users manually evaluate the quality of articles could be performed and the results contrasted with those produced by the automatic methods. Finally, one could also study the impact of quality on several information services such as searching and recommendation.

[13]In Wikipedia, articles about people, places, and insects are expected to fit specific structural organization. The coverage of an article (e.g., the biography of the statistician and geneticist Ronald Fisher) can be considered very adequate for one group (Biology), while imprecise for another (Statistics).

REFERENCES

ADLER, T. B. AND DE ALFARO, L. 2007. A content-driven reputation system for the Wikipedia. In *Proceedings of the 16th International Conference on the World Wide Web (WWW'07)*. 261–270.

AGICHTEIN, E., CASTILLO, C., DONATO, D., GIONIS, A., AND MISHNE, G. 2008. Finding high-quality content in social media. In *Proceedings of the International Conference on Web Search and Web Data Mining (WSDM'08)*. ACM, New York, NY, 183–194.

ALEXANDER, J. E. AND TATE, M. A. 1999. *Web Wisdom; How to Evaluate and Create Information Quality on the Web*. L. Erlbaum Associates Inc., Hillsdale, NJ.

ARGAMON, S., KOPPEL, M., FINE, J., AND SHIMONI, A. R. 2003. Gender, genre, and writing style in formal written texts. *TEXT 23*, 321–346.

BENEVENUTO, F., RODRIGUES, T., ALMEIDA, V., ALMEIDA, J., AND GONÇALVES, M. 2009. Detecting spammers and content promoters in online video social networks. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'09)*. 620–627.

BETHARD, S., WETZER, P., BUTCHER, K., MARTIN, J. H., AND SUMNER, T. 2009. Automatically characterizing resource quality for educational digital libraries. In *Proceedings of the Joint International Conference on Digital Libraries (JCDL'09)*. ACM, 221–230.

BIGONHA, C., CARDOSO, T. N., MORO, M. M., ALMEIDA, V., AND GONÇALVES, M. A. 2010. Detecting evangelists and detractors on twitter. In *Simpósio Brasileiro de Sistemas Multimídia e Web - Webmedia 2010*, Belo Horizonte, Minas Gerais, Brazil, 107–114.

BJÖRNSSON, C. 1968. *Lesbarkeit durch Lix*. Stockholm: Pedagogiskt Centrum.

BOLDI, P. AND VIGNA, S. 2004. The webgraph framework I: Compression techniques. In *Proceedings of the 13th International Conference on the World Wide Web (WWW'04)*. ACM, New York, NY, 595–601.

BORTHWICK, A., STERLING, J., AGICHTEIN, E., AND GRISHMAN, R. 1998. Exploiting diverse knowledge sources via maximum entropy in named entity recognition. In *Proceedings of the 6th Workshop on Very Large Corpora*.

BRIN, S. AND PAGE, L. 1998. The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst. 30,* 1-7, 107–117.

BROWN, R. 2009. Does fundamentalist religion cause the rejection of evolution? or is it the other way around? `http://karmatics.com/docs/evolution-and-wisdom-of-crowds.html`.

CHANG, C. C. AND LIN, C. J. 2001. LIBSVM: A library for support vector machines. `http://www.csie.ntu.edu.tw/ cjlin/libsvm`.

CHEVALIER, F., HUOT, S., AND FEKETE, J.-D. 2010. Wikipediaviz: Conveying article quality for casual wikipedia readers. In *Proceedings of the Pacific Visualization Symposium (PacificVis)*. IEEE, 49–56.

CHIN, S.-C., STREET, W. N., SRINIVASAN, P., AND EICHMANN, D. 2010. Detecting wikipedia vandalism with active learning and statistical language models. In *Proceedings of the 4th Workshop on Information Credibility (WICOW'10)*. ACM, New York, NY, 3–10.

CHU, W., KEERTHI, S. S., AND ONG, C. J. 2001. A unified loss function in bayesian framework for support vector regression. In *Proceedings of the 18th International Conference on Machine Learning (ICML'01)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, 51–58.

CHU, Y. 1997. Trust management for the World Wide Web. M.S. thesis, MIT, Cambridge, MA.

COLEMAN, M. AND LIAU, T. L. 1975. A computer readability formula designed for machine scoring. *J. Appl. Psych. 60,* 2, 283–284.

CUSINATO, A., DELLA MEA, V., DI SALVATORE, F., AND MIZZARO, S. 2009. QuWi: Quality control in Wikipedia. In *Proceedings of the 3rd Workshop on Information Credibility on the Web (WICOW'09)*. ACM, 27–34.

DALIP, D. H., GONÇALVES, M. A., CRISTO, M., AND CALADO, P. 2009. Automatic quality assessment of content created collaboratively by web communities: A case study of Wikipedia. In *Proceedings of the Joint International Conference on Digital libraries (JCDL'09)*. 295–304.

DE LA CALZADA, G. AND DEKHTYAR, A. 2010. On measuring the quality of Wikipedia articles. In *Proceedings of the 4th Workshop on Information Credibility (WICOW'10)*. ACM, New York, NY, 11–18.

DONDIO, P., BARRETT, S., WEBER, S., AND SEIGNEUR, J. 2006. Extracting trust from domain analysis: A case study on the Wikipedia project. In *Autonomic and Trusted Computing*, Springer, Berlin, 362–373.

DOROGOVTSEV, S. N. AND MENDES, J. F. F. 2003. *Evolution of Networks: From Biological Nets to the Internet and WWW (Physics)*. Oxford University Press.

DRUCKER, H., BURGES, C. J. C., KAUFMAN, L., SMOLA, A. J., AND VAPNIK, V. 1996. Support vector regression machines. In *Advances in Neural Information Processing Systems (NIPS)*. M. Mozer, M. I. Jordan, and T. Petsche Eds., MIT Press, 155–161.

FLESCH, R. 1948. A new readability yardstick. *J. Appl. Psych.*, 221–235.

FOGG, B. J., MARSHALL, J., LARAKI, O., OSIPOVICH, A., VARMA, C., FANG, N., PAUL, J., RANGNEKAR, A., SHON, J., SWANI, P., AND TREINEN, M. 2001. What makes web sites credible?: A report on a large quantitative study. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'01)*. 61–68.

FOGG, B. J., SOOHOO, C., DANIELSON, D. R., MARABLE, L., STANFORD, J., AND TAUBER, E. R. 2003. How do users evaluate the credibility of web sites?: A study with over 2,500 participants. In *Proceedings of the Conference on Designing for User Experiences (DUX'03)*. 1–15.

GILES, J. 2005. Internet encyclopaedias go head to head. *Nature 438,* 7070, 901–902.

GUNNING, R. 1952. *The Technique of Clear Writing*. McGraw-Hill International Book Co.

GUYON, I. AND ELISSEEFF, A. 2003. An introduction to variable and feature selection. *J. Mach. Learn. Res. 3*, 1157–1182.

HSU, C.-W., CHANG, C.-C., AND LIN, C.-J. 2000. A practical guide to support vector classification. *Bioinformatics 1*, 1.

HU, M., LIM, E.-P., SUN, A., LAUW, H. W., AND VUONG, B.-Q. 2007. Measuring article quality in Wikipedia: Models and evaluation. In *Proceedings of the 16th ACM Conference on Conference on Information and Knowledge Management (CIKM'07)*. 243–252.

JÄRVELIN, K. AND KEKÄLÄINEN, J. 2000. IR evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'00)*. 41–48.

KITTUR, A. AND KRAUT, R. E. 2008. Harnessing the wisdom of crowds in Wikipedia: Quality through coordination. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW'08)*. ACM, New York, NY, 37–46.

KORFIATIS, N., POULOS, M., AND BOKOS, G. 2006. Evaluating authoritative sources using social networks: An insight from wikipedia. *Online Inf. Rev. 30,* 3, 252–262.

KROWNE, A. 2003. Building a digital library the commons-based peer production way. *D-Lib Mag. 9,* 1082.

MAGED, K. B., MARAMBA, I., AND WHEELER, S. 2006. Wikis, blogs and podcasts: A new generation of web-based tools for virtual collaborative clinical practice and education. *BMC Medical Educ. 6*, 41+.

MCLAUGHLIN, G. H. 1969. Smog grading: A new readability formula. *J. Read.*, 639–646.

MINGUS, B. 2008. personal communication.

MITCHELL, T. M. 1997. *Machine Learning*. McGraw-Hill Higher Education.

MUPPET. 2010. Statistics - Muppet wiki. `http://muppet.wikia.com/wiki/Special:Statistics`.

P. DONDIO, S. B. AND WEBER, S. 2006. Calculating the trustworthiness of a Wikipedia article using dante methodology. In *Proceedings of the IADIS International Conference on e-Society*.

PIROLLI, P., WOLLNY, E., AND SUH, B. 2009. So you know you're getting the best possible information: A tool that increases Wikipedia credibility. In *Proceedings of the 27th International Conference on Human Factors in Computing Systems (CHI'09)*. ACM, New York, NY, 1505–1508.

RASSBACH, L., PINCOCK, T., AND MINGUS, B. 2007. Exploring the feasibility of automatically rating online article quality.
`http://upload.wikimedia.org/wikipedia/wikimania2007/d/d3/RassbachPincockMingus07.pdf`.

RESSLER, S. 1993. *Perspectives on Electronic Publishing: Standards, Solutions, and More*. Prentice-Hall, Inc., Upper Saddle River, NJ.

RUBIO, R., MARTÍN, S., AND MORÁN, S. 2010. Collaborative web learning tools: Wikis and blogs. *Comput. Appl. Engin. Educ. 18*, 502–511.

SMITH, E. A. AND SENTER, R. J. 1967. Automated readability index. *Aerospace Medical Division*.

VAPNIK, V. N. 1995. *The Nature of Statistical Learning Theory*. Springer-Verlag, Berlin, Germany.

VELTMAN, K. H. 2005. Access, claims and quality on the internet – future challenges. *Progress Inform. PI 2*, 17–40.

WEISS, G. M. AND PROVOST, F. 2003. Learning when training data are costly: The effect of class distribution on tree induction. *J. Artif. Intell. Res. 19*, 315–354.

WIKIPEDIA. 2010a.
`http://en.wikipedia.org/wiki/Wikipedia`.

WIKIPEDIA. 2010b. Featured article candidates.
`http://en.wikipedia.org/wiki/Wikipedia:Featured_article_candidates`.

WIKIPEDIA. 2010c. Good article nominations.
`http://en.wikipedia.org/wiki/Wikipedia:Good_article_nominations`.

WIKIPEDIA. 2010d. Version 1.0 editorial team/assessment.
http://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team/Assessment.

WIKIPEDIA. 2010e. Version 1.0 editorial team/release version criteria.
http://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team/Release_Version_Criteria.

WIKIPEDIA. 2010f. Wikipedia:database download - wikipedia, the free encyclopedia.
http://en.wikipedia.org/wiki/Wikipedia_database.

WIKIPEDIA. 2010g. Wikipedia:wikiproject.
http://en.wikipedia.org/wiki/Wikipedia:WikiProject.

WILCOXON, F. 1945. Individual comparisons by ranking methods. *Biometrics*, 80–83.

WILKINSON, D. M. AND HUBERMAN, B. A. 2007. Cooperation and quality in wikipedia. In *Proceedings of the International Symposium on Wikis (WikiSym'07)*. ACM, New York, NY, 157–164.

WÖHNER, T. AND PETERS, R. 2009. Assessing the quality of Wikipedia articles with lifecycle based metrics. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration (WikiSym'09)*. ACM, New York, NY, 16:1–16:10.

WOOKIEEPEDIA. 2010a. Statistics - wookieepedia, the star wars wiki.
http://starwars.wikia.com/wiki/Special:Statistics.

WOOKIEEPEDIA. 2010b. Wookieepedia: Featured article nominations.
http://starwars.wikia.com/wiki/Wookieepedia:Featured_article_nominations.

WOOKIEEPEDIA. 2010c. Wookieepedia: Good article nominations.
http://starwars.wikia.com/wiki/Wookieepedia:Good_article_nominations.

ZENG, H., ALHOSSAINI, M., DING, L., FIKES, R., AND MCGUINNESS, D. L. 2006a. Computing trust from revision history. In *Proceedings of the International Conference on Privacy, Security and Trust.*

ZHENG, R., LI, J., CHEN, H., AND HUANG, Z. 2006b. A framework for authorship identification of online messages: Writing-style features and classification techniques. *J. Amer. Soc. Inf. Sci. Technol. 57*, 378–393.

Received December 2010; revised April 2011; accepted July 2011