# A Multi-view Approach for the Quality Assessment of Wiki Articles

Daniel Hasan Dalip[1], Thiago Cardoso[1], Marcos André Gonçalves[1], Marco Cristo[2], Pável Calado[3]

[1] Universidade Federal de Minas Gerais
[2] Universidade Federal do Amazonas
[3] Instituto Superior Técnico/INESC-ID
{hasan,thiagon,mgoncalv}@dcc.ufmg.br, marco.cristo@icomp.ufam.edu.br,
pavel.calado@tagus.ist.utl.pt

**Abstract.** Wikipedia is a great example of a very large 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. To deal with this problem, some studies attempt to assess the quality of articles in Wikipedia automatically. In these studies, a large number of quality indicators is usually collected and then combined in order to obtain a single value representing the quality of the article. In this work, we propose to group these indicators in semantically meaningful *views* of quality and investigate a new approach to combine these views based on a meta-learning method, known as stacking. Particularly, we grouped the indicators into three views (textual, review history and citation graph), and demonstrated that it is possible to use this approach in collaborative encyclopedias such as Wikipedia and Wikia. In our experimental evaluation, we obtained gains of up to 18% compared the state-of-the-art quality assessment method that considers all indicators at once.

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

Keywords: Quality Assessment, Wikipedia, Machine Learning, SVM, Multi-View

## 1. INTRODUCTION

The Web 2.0 phenomenon and its highly collaborative nature are currently giving rise to a new type of repository for 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 [Krowne 2003; Dondio et al. 2006].

However, such freedom without any type of control raises an important question: *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, published using physical media, such as books or journals, is still naturally seen as being of higher quality and more trustworthy [Dondio et al. 2006].

To deal with this problem, Digital Libraries (DLs) such as Wikipedia, rely on human judgment. For instance, members of the Wikipedia community constantly review articles labeling them according to (some of) its qualitative aspects. However, given the huge size and growth rate of such collections, a manual revision process will eventually cease to be feasible [Voß 2005]. Moreover, manual reviews are subject to human bias, which can be influenced by varying backgrounds, expertises, and even a tendency for abuse [Hu et al. 2007].

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.

A possible solution for this problem is to automatically estimate the quality of the documents in the digital library. To accomplish this, some approaches have been proposed in the literature. Among these, the one which presented the best results was suggested by the authors in [Dalip et al. 2009; 2011]. In that approach, the authors have exploited several quality indicators (e.g., article length, revision, and citation count) and combined them using a machine learning technique. Experimental results showed significant gains over other baselines.

In this article, we propose to group the indicators used in [Dalip et al. 2009; 2011] in semantically meaningful "views" of quality. Particularly, we grouped the indicators into three views: textual, review history and citations graph. This idea was motivated by work such as [Muslea et al. 2002; Kakade and Foster 2007], which demonstrated that the combination of views may improve the performance of machine learning methods. Since views represent different perceptions of a same concept (in our case, the relative quality of an article), the combination of models created specifically for each view may improve results in a way similar to the combination of the opinions of different experts. Thus, in this work, we propose to assess the quality by (1) organizing this indicators in different views, and then (2) combining these views by means of a meta-learning method known as stacking [Wolpert 1992].

Using our proposed approach, we were able to assess the quality of the articles in three collaboratives encyclopedias (Wikipedia[1], Star Wars[2] and Muppets[3]). Experimental results show gains of up to 18% when compared to the state-of-the-art baseline [Dalip et al. 2009; 2011]. In summary, our main contributions are: (i) a proposal of a view-based approach to automatically estimate the quality of Wiki articles; and (ii) its successful application in three collaboratives encyclopedias.

This article is organized as follows. Section 2 covers related work. Section 3 described in details the proposed approach. Section 4 presents and discusses our experimental evaluation. Finally, Section 5 concludes the paper.

## 2. RELATED WORK

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 a content, as well as the degree of certainty and importance of its claims. An example of such mechanism is proposed by [Chu 1997], through which 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 requirement may be very hard to comply. 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, i.e., 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 literature, some in different contexts, include for example, hyperlinks [Alexander and Tate 1999],writing style [Zheng et al. 2006], network features [Korfiatis et al. 2006], history of reviews [Adler and de Alfaro 2007; Hu et al. 2007; Cusinato et al. 2009; Wöhner and Peters 2009], etc. 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.

Based on these previous studies, several authors have proposed to combine different sources of evidence into a unique value to represent quality. For instance, [Dondio and Weber 2006; Dondio et al. 2006] combine several pieces of evidence to build an article ranking that tries to jointly capture

[1]http://en.wikipedia.org/
[2]http://starwars.wikia.com
[3]http://muppet.wikia.com

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 approaches proposed by [Dondio and Weber 2006; Dondio et al. 2006], which used simple linear combination methods, a few other efforts were proposed to combine the available evidence using machine learning techniques. This is the case of [Rassbach et al. 2007], which suggested the use of a Maximum Entropy Model [Borthwick et al. 1998] to estimate the quality of the articles. In addition, the authors also proposed some new text-based sources of evidence for the problem, like the number of phrases, auxiliary verbs, and the Kincaid readability index [Ressler 1993] giving two views of the same instance. Other work that uses textual features to predict the quality of article is [Xu and Luo 2011], in which lexical clue words and a decision tree are used. Further, [De la Calzada and Dekhtyar 2010] proposed a machine learning approach to estimate the quality of articles regarding two categories: stabilized articles and controversial articles.

In [Dalip et al. 2009; 2011], we have proposed to treat quality estimation as a regression problem. In other words, we estimated the quality of articles in Wikipedia as a grade in a continuous quality scale. To accomplish that, we used a Support Vector Regression method [Vapnik 1995]. Our main contribution in that work was 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 overall better results than the best approaches previously proposed in literature.

All these cited methods have created a unique model to combine the proposed indicators. However, by observing these indicator, we note that they represent three distinct views of quality, namely, by the perspective (a) of the written text, (b) from the history of reviews, and (c) the linkage among the articles. As mentioned in [Muslea et al. 2002; Blum and Mitchell 1998] multiple views can be utilized to obtain many independent opinions about a classification process. For example, a video can be classified using its contents from two different views: audio and video.

Differently from previous methods, we intend to learn a model for each view and, then combine the models. To accomplish this, we will use a meta-learning technique based on stacking [Wolpert 1992]. In this technique, a meta-classifier learns the relation between the output of distinct learning algorithms and the target class. In our case, instead of using models generated by distinct algorithms, we will use models generated from distinct views. In this sense, our proposed technique is slightly different from the stacking method as originally proposed.

## 3. ASSESSING ARTICLE QUALITY

Suppose three experts assess the quality of a wiki article, each one according to a different perspective (or "view") of quality. In our case, they assess according to the textual content, the review history, and the citation graph. The final quality assessment should be a combination of these multiple assessments. Particularly, if each opinion is given with a degree of certainty, its possible to learn its global quality from the certainty related to each view. Likewise, this certainty can be learned from the various indicators that constitute the view.

Thus, the problem of estimating quality can occur in two learning phases. In the first phase, or *learning level 0*, each article is represented by a set of indicators related to each view. The articles will thus have three representations (i.e., sets of features). One quality model is then learned for each view. Using this model, we can obtain an assessment of the quality of each view of the article. In the second phase, or *learning level 1*, each article is represented by the quality prediction from each view. A global model of quality is then learned and, as a result, we can obtain a final (combined) assessment of quality. This process is depicted in Figure 1. The next sections detail each of those representations.

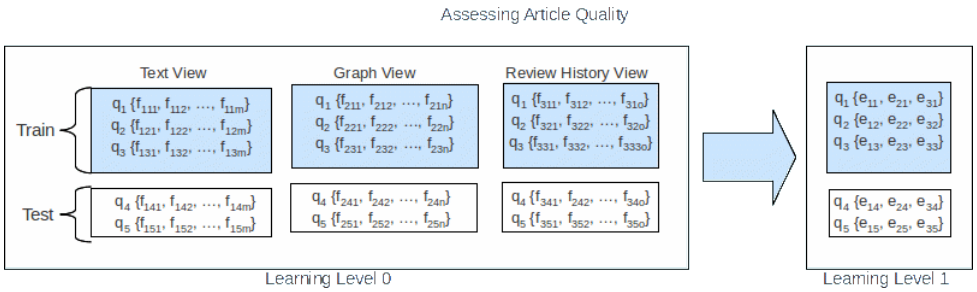


Fig. 1. The quality assessment process. $q_i$ is the target value and represents the quality value given for each article. $f_{vij}$ is the feature $j$ for the article $i$ in the view $v$. Finally, $e_{vi}$ represents the quality which has been estimated in each view $v$ for the article $i$ in the learning level 0 and used as a feature in the level 1.

### 3.1 Learning level 0

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" (see a more complete explanation about these classes in Section 4). We note, however, that in general quality can be seen as a value in 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, class AC articles are defined as those that: (a) have recently been promoted and await expert evaluation; (b) have been evaluated by experts and await corrections; or (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.

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

3.1.1 *Quality assessment with SVR.* To apply SVR to the quality estimation task, we represent the articles to be classified as follows. Given a view $v$, let $A_v = \{a_{v1}, a_{v2}, ..., a_{vn}\}$ be a set of articles. Each article $a_{vi}$ is represented by a set of $m$ features $F_v = \{F_{v1}, F_{v2}, ..., F_{vm}\}$, such that $a_{vi} = (f_{vi1}, f_{vi2}, ..., f_{vim})$ is a vector representing article $a_{vi}$, where each $f_{vij}$ is the value of feature $F_{vj}$ in $a_{vi}$. In this work, the term *feature* describes a statistic value that represents a measurement of some quality indicator associated with an a view $v$ and an article. For instance, $f_{vij}$ could represent the length of article $a_{vi}$ in the textual view.

In our proposal, we assume that we have access to some *training data* of the form $A_v \times \mathbb{R} = \{(a_{v1}, q_1), (a_{v2}, q_2), ..., (a_{vn}, q_n)\}$, where each pair $(a_{vi}, q_i)$ represents an article $a_{vi}$ and its corresponding quality assessment value $q_i$, such that if $q_1 > q_2$, then the quality of article $a_{v1}$, as perceived by the user, is higher than the quality of article $a_{v2}$. The solution we propose to this problem consists in: (1) determining the set of views $v$; (2) determining the set of features$\{F_{v1}, F_{v2}, ..., F_{vm}\}$ used to represent the articles in $A_v$; and (3) applying a regression method to find the best combination of the features, for each view $v$, to predict the quality value $q_i$ for any given article $a_{vi}$.

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, $A_v$, 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 $(f(a_v) - q), q \in \mathbb{R}, a_v \in A_v$, 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 $\epsilon$-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 $\epsilon$ 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 Fig. 2(a). In that figure,

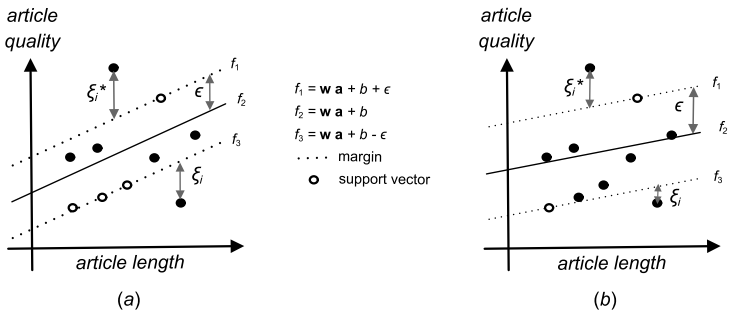


Fig. 2. Regression problem with one numeric target (article quality) and ten articles (points), represented by their lengths. Note that two articles, in both graphics, are considered examples of error because they lie outside the area delimited by the margins. Their distances to the margins are given by $\xi_i^*$ and $\xi_i$, respectively. Figures (a) and (b) represent regressions performed using two different $\epsilon$-values.

$f_1$ and $f_3$ represent the margins around hypothesis function $f_2$. Thus, our goal is to find a function $f : A_v \rightarrow \mathbb{R}$ that has at most $\epsilon$ deviation from the actual targets $q \in \mathbb{R}$ for all the training data.

In SVR, the input article $a_v$ is first mapped onto an $m$-dimensional feature space using some nonlinear mapping $\Phi$. Then, a linear model is constructed in this feature space. More formally, the linear model $f(\mathbf{a_v}, \mathbf{w})$ is given by $f(\mathbf{a_v}, \mathbf{w}) = \langle \mathbf{w}, \Phi(\mathbf{a_v}) \rangle + b$, where $\mathbf{w}$ is a weight vector of $m$ feature values, $b$ is the bias term, and $\langle \mathbf{w}, \Phi(\mathbf{a_v}) \rangle$ denotes the inner product between $\mathbf{w}$ and $\Phi(\mathbf{a_v})$. The quality of estimation is measured by the $\epsilon$-intensive loss function $L^\epsilon(q, f(\mathbf{a_v}, \mathbf{w}))$ defined in Eq. 1:

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

SVR performs a linear regression in the high-dimension feature space using the $\epsilon$-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 $(q_i - f(\mathbf{a_{vi}}, \mathbf{w}))$ and $(f(\mathbf{a_{vi}}, \mathbf{w}) - q_i)$, measured, respectively, by non–negative slack variables $\xi_i^*$ and $\xi_i$. If we consider $f_1$ the margin above $f$ and $f_3$ the margin below $f$, $\xi_i^*$ measures deviations above $f_1$ whereas $\xi_i$ measures deviations below $f_3$, as shown in Fig. 2. Thus, SVR can be formulated as the convex optimization problem of minimizing:

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

subject to:

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

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:

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

where $n_{\mathrm{SV}}$ is the number of support vectors (vectors lying on the margins, depicted as white circles in Fig. 2) and $\kappa$ is an inner product function (*kernel function*) in a given vector space, given by $\kappa(\mathbf{a_{vi}}, \mathbf{a_v}) = \sum_{j=1}^{m}(\Phi_j(\mathbf{a_{vi}})\Phi_j(\mathbf{a_v}))$.

Note that the SVR estimation accuracy depends on a good setting for $C$, $\epsilon$ and the kernel param-

eters. $C$ determines the trade-off between the model complexity (flatness) and the degree to which deviations larger than $\epsilon$ are tolerated. If $C$ is too large, the objective becomes simply to minimize $\frac{1}{n}\sum_{i=1}^{n} L^{\epsilon}\left(q_i, f\left(\mathbf{a_{vi}}, \mathbf{w}\right)\right)$. Parameter $\epsilon$ controls the width of the $\epsilon$-insensitive zone, used to fit the training data. Bigger $\epsilon$-values use fewer support vectors, at the expense of providing more "flat" estimates, as we can see in Fig. 2(a) and Fig. 2(b).

We have chosen SVR due to its advantages over other methods, such as the presence of a global minimum solution resulting from the minimization of a convex programming problem, relatively fast training speed, and the capability of dealing with sparseness [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 $\kappa$. 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 will present the utilized combination and features used to represent the articles.

3.1.2 *Article Representation.* Determining which features should be used to represent an article is a key decision in a regression-based quality assessment. Such features were extracted based on the criteria used by Wikipedia Guidelines [Wikipedia 2008] to manually assess the quality of an article. The features we were divided into 3 "natural" views: textual, review, and network features obtained from the citation graph between articles. See Figure 1.

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 can further divided them into four subgroups, although this subdivision is not exploited as views in our work: length, style, structure, and readability. Examples of these features are the number of characters of an article (length), the ratio between short and large paragraphs (style), the distribution of sections (structure) and the Flesch reading ease [Flesch 1948], which indicates the reading complexity level (readability) of an English text.

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 [Dondio and Weber 2006]. It can be expected that good quality articles have reached a maturity level in which no extensive corrections are necessary. Some examples of these features are the age, number of reviews per day, and ProbReview[Hu et al. 2007], which estimates the quality of an article based on the quality of its authors.

Network features are those extracted from the interconnections among the articles. In this case, we see the collection as a graph, where nodes are articles and edges are the citations between them. Examples of such features are the clustering coefficient of an article [Dorogovtsev and Mendes 2003], its PageRank [Brin and Page 1998], and the number of in- and out-links.

In this work, we used all the 68 features of [Dalip et al. 2009; 2011]. For a detailed description of the features used in this work, we refer the reader to [Dalip et al. 2009; 2011].

### 3.2 Learning level 1

Once the quality of articles has been predicted for each view, we can describe them with a new representation. Thus, each item $a_i$ is represented by a set of three attributes $\{e_{i1}, e_{i2}, e_{i3}\}$. At this level, each feature represents the article quality estimates given by each view. Thus, $e_{i1}$ represents the prediction of quality for the article $i$ according to the textual view, $e_{i2}$ the prediction according to the review, and $e_{i3}$ to the network. Given the training set $\{(a_1, q_1), (a_2, q_2), ..., (a_n, q_n)\}$, where each pair $(a_i, q_i)$ represents the article $a_i$ and its quality $q_i$, the quality of the article can be learned by applying SVR as described in the previous section. Note that, by doing this, we are in fact learning to combine the estimation obtained from each different view of quality.

## 4. EXPERIMENTS

Using the features described in Section 3.1.2, we performed a set of experiments using three different test collections. We now describe our experimental design, the collections, and the results.

### 4.1 Datasets

In our experiments, we used a sample of Wikipedia in English and samples of two other collaborative encyclopedias provided by the Wikia service. 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 2011]. Furthermore, the full content of Wikipedia is freely available for download allowing the extraction of its features [Wikipedia 2010]. From now on, we refer to this encyclopedia as WIKIPEDIA. Any user can evaluate a Wikipedia article, according to the following quality taxonomy[4] [Wikipedia 2011]:

—Featured Article (*FA*): These are, according to the 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.

—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.

From the Wikia service, we selected the encyclopedias *Wookieepedia*[5], about the Star Wars universe, and *Muppet*[6], 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[7]. Their repositories are freely available for download [Wookieepedia 2010; Muppet 2010].

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, i.e., 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, which we refer to as MUPPET, provides only a star-based taxonomy.

The size of each sample is presented in Table I. 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

[4]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 by the time we performed our crawling.
[5]`http://starwars.wikia.com/`
[6]`http://muppet.wikia.com/`
[7]To obtain the article evaluations we used the APIs provided at `http://starwars.wikia.com/api.php` and `http://muppet.wikia.com/api.php`.

number of articles from the remaining classes.

For all datasets, we also collected the links between the articles, in order to extract network attributes. These links were extracted through an import file available for download[8]. Table I presents information about the total number of articles and revisions of each sample and about the network graphs derived from the datasets. In the table, edges correspond to links between pages and the nodes correspond to the article pages and redirections to articles of the complete collection. We used the *Web Graph* library [Boldi and Vigna 2004] to create the graph and extract all the Network attributes.

| Dataset | # Articles | # Reviews | # 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 |

Table I. Sample size for each dataset used in our experiments.

### 4.2 Evaluation Methodology

Our experiments aim at performing a comparative analysis between different methods for combining views, as well as a comparison with our baseline. Since we proposed a regression based method, its effectiveness was evaluated using the *mean squared error* measure (MSE). MSE is defined as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n} e^2 \tag{4}$$

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 perform the comparative experiments, we used a 10-fold cross validation method [Mitchell 1997]. Each dataset was randomly split into ten parts, such that, in each run, a different part was used as a test set while the remaining were 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 ten runs. Note that, different partitions were used in each learning level, as in [Wolpert 1992]. Thus, in order to obtain the predictions to the level 1, we performed a cross-validation in the training set.

For all comparisons reported in this work, we used the signed-rank test of Wilcoxon [Wilcoxon 1945] to determine if the differences in effectiveness were statistically significant. This is a nonparametric paired test that does not assume any particular distribution on the tested values. In all cases, we only draw conclusions from results that were considered statistically significant with a 90% confidence level.

### 4.3 Results

Table II presents the results of the experiments for each collection. Besides the general result with all articles, we also present separately results considering only articles whose predictions of all views were closer to the same integer (View Agreement) and when the predictions for the articles had any disagreement among the views (View Disagreement). From now on we will call SVR the *baseline* method. ML_VIEW is the meta-learning method described in the Section 3. And in ML_VIEW_ARTICLE, in the level 1, in addition to the results of the views, we used all the 68 features that represent the article (the union of features of the 3 views). An "*" on MSE value indicates a statistically significant difference when compared to the baseline.

[8]In Wikipedia: `http://en.wikipedia.org/wiki/Wikipedia_database`. For the others collections, the graph was created using the link structure of the article.

| General Result | | | | View Agreement | | View Disagreement | |
|---|---|---|---|---|---|---|---|
| Sample | Method | MSE | % Improvement | MSE | % articles | MSE | % articles |
| WIKIPEDIA | SVR | 0.856 | - | 0.739 | 24.35 % | 0.900 | 75.65% |
| | ML_VIEW | 0.84* | 1.02 % | 0.733* | | 0.879* | |
| | ML_VIEW_ARTICLE | 0.809* | 5.8 % | 0.709* | | 0.849 | |
| MUPPET | SVR | 1.685 | - | 1.650 | 46.19 % | 1.716 | 53.81% |
| | ML_VIEW | 1.676 | 0.5 % | 1.688* | | 1.657* | |
| | ML_VIEW_ARTICLE | 1.682 | 0.2 % | 1.689* | | 1.677* | |
| STWR_5CLASS | SVR | 1.681 | - | 1.669 | 44.04 % | 1.690 | 55.96% |
| | ML_VIEW | 1.665* | 0.9 % | 1.639* | | 1.686 | |
| | ML_VIEW_ARTICLE | 1.661* | 1.2 % | 1.638* | | 1.676 | |
| STWR_3CLASS | SVR | 0.075 | - | 0.045 | 72.68 % | 0.151 | 27.32% |
| | ML_VIEW | 0.061* | 18.6 % | 0.034* | | 0.129 | |
| | ML_VIEW_ARTICLE | 0.067* | 10.6 % | 0.039* | | 0.139 | |

Table II. Mean Squared Error by method for each sample and result by agreement of views

As we can observe, the meta-learning was able to improve the result in all samples, except in MUPPET. Furthermore, when we used the article information in the level 1, we obtained an even better result in the WIKIPEDIA sample. Thus, it was possible to observe that the meta-learning can be useful for the article quality assessment in several wiki collections as well as the importance of using the whole article representation in combination with the views in this task.

Considering the agreement among views in Table II, we verified that it was higher for collections with fewer classes. In addition, stacking improved the performance more when the views agree with each other. However, this does not occur in the MUPPET collection. This may explain why results did not improve by using stacking in that collection. We hypothesize that the manual assessment process used in MUPPET is not as reliable as in the other collections making it difficult to assign quality ratings. As future work, we intend to perform a user study to verify this hypothesis.

## 5. CONCLUSIONS

In this work we proposed a multiview approach for quality assessment of Wiki articles. A large number of features was organized into three views of quality, related to the text of the article (e.g. its organization, length, readibility), its revision history and network properties. These views were combined using a meta-learning strategy. Experiments with several collaborative encyclopedias showed that the proposed method was able to reduce the error when compared to a state-of-the-art method, of the assessment of quality in all the collaboratives encyclopedias, except one.

As future work, we intend to explore and compare other view combination methods and analyse performance issues by reducing the training set used for the representation of multi-view problem. We also want to study the impact of quality in other information services such as searching and recommendation, besides analyzing how reliable is the quality labeling provided by the reviewers in each different dataset.

## REFERENCES


Adler, T. B. and de Alfaro, L. A content-driven reputation system for the wikipedia. In *Proceedings of the International Conference on World Wide Web*. Banff, Canada, pp. 261–270, 2007.

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

Blum, A. and Mitchell, T. Combining labeled and unlabeled data with co-training. In *Proceedings of the Annual Conference on Computational Learning Theory*. New York, NY, USA, pp. 92–100, 1998.

Boldi, P. and Vigna, S. The webgraph framework I: Compression techniques. In *Proceedings of the International Conference on World Wide Web*. New York, NY, USA, pp. 595–601, 2004.

Borthwick, A., Sterling, J., Agichtein, E., and Grishman, R. Exploiting diverse knowledge sources via maximum entropy in named entity recognition. In *Proceedings of Workshop on Very Large Corpora*, 1998.

Brin, S. and Page, L. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems* 30 (1-7): 107–117, April, 1998.

Chang, C. C. and Lin, C. J. LIBSVM: a library for support vector machines. `http://www.csie.ntu.edu.tw/~cjlin/libsvm`, 2001.

Chu, W., Keerthi, S. S., and Ong, C. J. A unified loss function in bayesian framework for support vector regression. In *Proceedings of the International Conference on Machine Learning*. San Francisco, USA, pp. 51–58, 2001.

Chu, Y. *Trust Management for the World Wide Web*. M.S. thesis, MIT, USA, 1997.

Cusinato, A., Della Mea, V., Di Salvatore, F., and Mizzaro, S. QuWi: quality control in Wikipedia. In *Proceedings of the Workshop on Information Credibility on the Web*. Madrid, Spain, pp. 27–34, 2009.

Dalip, D. H., Gonçalves, M. A., Cristo, M., and Calado, P. 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*. Austin, TX, USA, pp. 295–304, 2009.

Dalip, D. H., Gonçalves, M. A., Cristo, M., and Calado, P. Automatic assessment of document quality in web collaborative digital libraries. *ACM Journal of Data and Information Quality* 2 (13), 2011.

De la Calzada, G. and Dekhtyar, A. On measuring the quality of wikipedia articles. In *Proceedings of the Qorkshop on Information Credibility*. New York, NY, USA, pp. 11–18, 2010.

Dondio, P., Barrett, S., Weber, S., and Seigneur, J. Extracting trust from domain analysis: A case study on the wikipedia project. In *Autonomic and Trusted Computing*. Springer Berlin / Heidelberg, pp. 362–373, 2006.

Dondio, Pierpaolo, S. B. and Weber, S. Calculating the trustworthiness of a wikipedia article using dante methodology. In *IADIS International Conference on e-Society*. Dublin, Ireland, 2006.

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

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A. J., and Vapnik, V. Support vector regression machines. In *NIPS*, M. Mozer, M. I. Jordan, and T. Petsche (Eds.). MIT Press, pp. 155–161, 1996.

Flesch, R. A new readability yardstick. *Journal of Applied Psychology*, 1948.

Hsu, C.-W., Chang, C.-C., and Lin, C.-J. A practical guide to support vector classification, 2000.

Hu, M., Lim, E.-P., Sun, A., Lauw, H. W., and Vuong, B.-Q. Measuring article quality in wikipedia: models and evaluation. In *Proceedings of the ACM Conference on Conference on Information and Knowledge Management*. Lisbon, Portuga, pp. 243–252, 2007.

Kakade, S. M. and Foster, D. P. Multi-view regression via canonical correlation analysis. In *Proceedings of Conference on Learning Theory*, 2007.

Korfiatis, N., Poulos, M., and Bokos, G. Evaluating authoritative sources using social networks: An insight from wikipedia. *Online Information Review* 30 (3): 252–262, 2006.

Krowne, A. Building a digital library the commons-based peer production way. *D-Lib magazine* 9 (1082), 2003.

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

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

Muslea, I., Minton, S., and Knoblock, C. A. Active semi-supervised learning = robust multi-view learning. In *Proceedings of International Conference on Machine Learning*. pp. 435–442, 2002.

Rassbach, L., Pincock, T., and Mingus, B. Exploring the feasibility of automatically rating online article quality. `http://upload.wikimedia.org/wikipedia/wikimania2007/d/d3/RassbachPincockMingus07.pdf`, 2007.

Ressler, S. *Perspectives on electronic publishing: standards, solutions, and more*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993.

Vapnik, V. N. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., 1995.

Veltman, K. H. Access, claims and quality on the internet – future challenges. *Progress in informatics : PI* vol. 2, pp. 17–40, 2005.

Voss, J. Measuring wikipedia. In *Proceedings of International Conference of the International Society for Scientometrics and Informetrics*. Number PREPRINT 2005-04-12, 2005.

Wikipedia. Version 1.0 editorial team/release version criteria. `http://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team/Release_Version_Criteria`, 2008.

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

Wikipedia. Version 1.0 editorial team/assessment. `http://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team/Assessment`, 2011.

Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics*, 1945.

Wöhner, T. and Peters, R. Assessing the quality of wikipedia articles with lifecycle based metrics. In *Proceedings of the International Symposium on Wikis and Open Collaboration*. New York, NY, USA, pp. 16:1–16:10, 2009.

Wolpert, D. H. Stacked generalization. *Neural Networks* vol. 5, pp. 241–259, 1992.

Wookieepedia. Statistics - wookieepedia. `http://starwars.wikia.com/wiki/Special:Statistics`, 2010.

Xu, Y. and Luo, T. Measuring article quality in wikipedia: Lexical clue model. In *Proceedings of Symposium on Web Society*. pp. 141 –146, 2011.

Zheng, R., Li, J., Chen, H., and Huang, Z. A framework for authorship identification of online messages: Writing-style features and classification techniques. *J. Am. Soc. Inf. Sci. Technol.* vol. 57, pp. 378–393, February, 2006.