On the Feasibility of External Factual Support as Wikipedia's Quality Metric

Sobre la Factibilidad del Soporte Factual Externo como Métrica de Calidad para Wikipedia

Carlos G. Velázquez[1], **Leticia C. Cagnina**[1,2], **Marcelo L. Errecalde**[1]
[1] LIDIC - Universidad Nacional de San Luis. San Luis, Argentina.
[2]Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)
e-mails: carvear20@yahoo.com.ar , {lcagnina,merreca}@unsl.edu.ar

Abstract: Developing metrics to estimate the information quality of Wikipedia articles is an interesting and important research area. In this article, we propose, and analyze the feasibility, of a new quality metric based on the "external factual support" of an article. The rationale behind this metric is identified, a formal definition of the metric is presented and some implementation aspects are introduced. Preliminary results show the feasibility of our proposal and its potential to discriminate high quality versus low quality Wikipedia's articles.
Keywords: Quality Metrics, Wikipedia, Featured Articles, External Support

Resumen: El desarrollo de métricas para estimar la calidad de información de los artículos de Wikipedia es un área de investigación intersante e importante. En este artículo, se propone una nueva métrica de calidad basada en el "soporte factual externo" de un artículo y se analiza su viabilidad. Los motivos que dan sustento a esta métrica son identificados, se presenta una definición formal de la misma y también se dan detalles de su implementación. Los resultados preliminares obtenidos, muestran la viabilidad de nuestra propuesta y su potencial para discriminar entre artículos de alta y baja calidad en Wikipedia.
Palabras clave: Métricas de Calidad, Wikipedia, Artículos Destacados, Soporte Externo

1 Introduction

Automatic assessment of Information Quality (IQ) is a topic of growing interest, mainly due to the increasing popularity of user-generated Web content and the unavoidable divergence of the delivered content's quality (Baeza-Yates, 2009). In this context, Wikipedia, the largest and most popular user generated knowledge source on the Web, presents different challenges related to quality assurance. In particular, its size and its dynamic nature render a manual quality assurance completely infeasible. This has resulted in an increasing number of articles related to automatic IQ assessment in Wikipedia that can be categorized into three research lines: (a) Featured articles identification (Blumenstock, 2008; Lipka and Stein, 2010); (b) Development of quality metrics (Lih, 2004; Stvilia et al., 2005); and (c) Quality flaws detection (Anderka, Stein, and Lipka, 2012; Ferretti et al., 2014).

In this paper we will focus on the second task, development of quality metrics for Wikipedia, an area where several methods have been recently proposed (Lex et al., 2012; Ingawale et al., 2013). A distinctive characteristic of most of those works is that they exclusively rely on "local" information directly obtained from the Wikitext content of the article or its edition history. However, in many cases, this information alone would seem to be insufficient to capture some IQ aspects which are intuitively related to "external information". Our hypothesis is that the *external support* of the information contained in Wikipedia articles can be useful to identify quality aspects of those articles. In order to start working on this hypothesis, we propose a quality metric named "external factual support". To this end, we first introduce in Section 2 some general concepts on quality metrics for Wikipedia. Then, in Section 3, motivations for the proposed metric and its formal definition are presented. Section 4 gives implementation details of the metric, a description of the data sets and experimental results validating our proposal. Finally, in Section 5 some conclusions are drawn and possible future work is discussed.

2 Quality metrics for Wikipedia

In a nutshell, a quality metric is a quantitative *estimation* of *to what extent* a textual resource (a Wikipedia article in this case) satisfies a specific property, such as *informativeness*, *reputation*, *generality*, *completeness*, etc. As we can see, quality metrics are *subjective*, in the sense that different people could define them in different ways. That contrasts with other "objective" properties such as article's *length* or *number of pictures* in the article, which are usually termed *quality measures*. Quality measures are directly *measured* with a suitable computer program while quality metrics are *estimated* by using some arbitrary formula. As an example, assume d is an arbitrary Wikipedia article, $len(d)$ the measure representing the length of d and $nuin(d)$ another measure that gives the number of images in d. One could represent the (abstract) property "*informativeness*" by means of a metric $info$ defined as: $info(d) = len(d) + 4 \times nuin(d)$. Obviously, another person might use another criteria to define the same quality metric. Stvilia in (Stvilia et al., 2005), for instance, proposes 7 arbitrary quality metrics which are based on 19 quality measures. The proposed IQ metrics showed to be successful in discriminating high quality Wikipedia articles.

Quality metrics can be used for ranking (and visualizing) documents according to the property represented by the metric. For instance, Wikipedia articles could be shown in decreasing order according to their estimated informativeness. On the other hand, they can also be integrated as part of other more general processing systems, such as text categorization or text clustering systems. In those cases, quality metrics can be used alone as features for representing the documents or combined with other arbitrary features.

As far as we know, the first works that specifically addressed the definition of quality metrics in Wikipedia date back to 2004 (Lih, 2004; Viégas, Wattenberg, and Dave, 2004), where concepts likes "*reputation*" of an article are defined by using the article edition history. In contrast, in (Emigh and Herring, 2005) different features are proposed to identify "formal language", which are directly derived from the article *content* (*POS* tags, for instance).

An aspect recurrently used in definitions of quality metrics for Wikipedia is the social/colaborative structure generated between article *editors* and the *articles* been edited. Results obtained by Wilkinson and Huberman in (Wilkinson and Huberman, 2007) agree with those presented in (Anthony, Smith, and Williamson, 2009; Lih, 2004) about the influence of qualified and occasional collaborators in the quality of the articles. Hu et. al. (Hu et al., 2007) also analyse collaborative models for measuring quality aspects based on relations between "good collaborators" and "good articles". Finally, in (Ingawale et al., 2013) the interaction among editors and articles is visualized as a *network* (or *graph*) and graph theory is used to infer *structural properties* associated to quality of articles.

In (Lex et al., 2012) is recognized that to assess factual accuracy of Web content, more complex, semantic features are needed. A common approach is employing Open Information Extraction (Etzioni et al., 2008) or methods that use background knowledge on semantic relations available in ontological resources. These methods extract relational information about entities, i.e. facts like $f = (Mozart, was\_born\_in, Salzburg)$. Besides, they exploit semantic relationships such as meronymy and hypernymy to infer relational information between entities not explicitly given in the text. In order to measure information quality based on factual information, different approaches are identified. Afterwards, they propose very simple metrics, named *fact frequency-based features*, which attempt to determine the informativeness level of a document. These features are the closest antecedent and the basis for the proposal presented in the paper in hand. Therefore, they will be described in this section with more details in order to make easier the understanding of the "external factual support" concept presented in Section 3.

Fact frequency-based features only require information about the number of facts obtained by an information extraction process from a textual resource. For instance, if t is an arbitrary textual resource (e.g. a para-graph, a document, a corpus), and F_t is the collection of facts extracted from t by an arbitrary information extraction method IE, it is direct computing the *fact count* of t, denoted $fc(t)$. It is simply defined as the total number of facts obtained from t by IE, $fc(t) = |F_t|$. Obviously the fact count directly depends on the size of the textual resource t, so it is usually normalized according to the size of t. This quantity is referred in (Lex et al., 2012) as the *factual density of* t, and denoted $fd(t)$. In that case, if $size(t)$ is a measure intended to quantify the size of t,[1] the factual density of t, is defined as $fd(t) = \frac{fc(t)}{size(t)}$. As it will be seen in Section 3, facts from the F_t collection will be used to compute the external factual support of t, where t corresponds to a Wikipedia article.

3 External Factual Support

Most of the above-mentioned approaches assume that all the relevant information to determine the Wikipedia articles' quality is present in the content of an article or in its edition history. However, that is not always the case. For instance, let consider the *original research* (OR) aspect, one of three core content policies that, along with "Neutral point of view" and "Verifiability", determines the type and quality of material acceptable in Wikipedia articles.[2] OR refers to a problem (flaw) exhibited by material such as facts, allegations, and ideas for which no reliable, published sources exist. To demonstrate that you are not adding OR, you must be able to cite reliable, published sources that are directly related to the topic of the article, and directly support the material being presented. However, checking for the absence of inline citations of sources does not guarantee that OR will be detected because all the statements might involve well known information. For example: the statement "Paris is the capital of France" needs no source, because no one is likely to object to it and we know that sources exist for it. The statement is *attributable*, even if *not attributed*. As it can be seen, a Wikipedia article that violates the "No Original Research" principle will directly affect its chances of being a "featured article". However, the necessary information

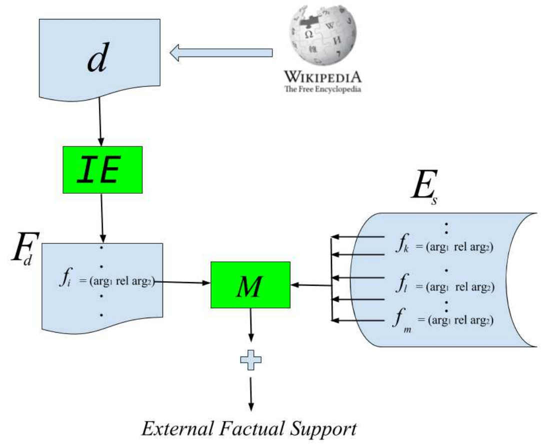

Figure 1: Computation of the EFS metric.

to determine this aspect cannot be realistically obtained if only the article's content is considered and some kind of extra "external information" is required.

Our main aim in this paper is defining a measure that estimates the *external support* of a document d, i.e., how much information in an external source E_s contributes to show that the content in d is either true, important, well known or all of them together. To do this, we will take as basis (the same as in (Lex et al., 2012)) the set of facts F_d, that is, the collection of facts extracted from d by an arbitrary information extraction method IE (for instance, the ReVerb Open Information Extraction framework[3]). Our idea in the present work is taking a closer look to the information available about each fact $f_i \in F_d$ and estimating the external support $s_e(f_i)$ that this fact has in the external source E_s. The computation of the external support of f_i will be based on a *matching* mechanism $\mathcal{M}$ in charge of deciding if a fact in E_s "matches"[4] f_i (or not). Then, the external support $\mathcal{S}_e(d)$ of the whole document d will be a weighted sum of the support $s_e(f_i)$ of each fact $f_i \in F_d$. That intuitive idea of the *external factual support* (EFS) of a document is illustrated in Figure 1 and more formally defined below.

Definition 1. (External Factual Support) Let d be a document and $F_d = \{f_1, \ldots, f_n\}$ the collection of facts extracted from d by an arbitrary information extraction

[1] For instance, it could be the number of words or sentences in t or the number of characters of t.

[2] `http://en.wikipedia.org/wiki/Wikipedia:No_original_research`

[3] `http://reverb.cs.washington.edu/`

[4] Initially, that will mean that both facts are the same. Then, we will see that other (more relaxed) types of pairings between facts will be allowed.

method IE. The **external factual support of** d, denoted $\mathcal{S}_e(d)$, is defined as

$$\mathcal{S}_e(d) = \sum_{i=1}^{n} w_i s_e(f_i) \tag{1}$$

where w_i is the *weight* that fact f_i is given in document d and $s_e(f_i)$ is the *external factual support* of f_i.

The idea of using weights w_i's to give different "importance" to the facts f_i's (and their associated external factual supports $s_e(f_i)$) is intuitively simple. It is motivated by the idea that in specific situations some information is available about which facts could be more relevant than others in a document d. In (Magdy and Wanas, 2010), for instance, facts obtained from sentences appearing earlier in the document are given a higher weight.

We will use a different approach that consists in using information directly provided by the information extraction method IE. For instance, fact-extraction systems like Reverb associate with each extracted fact f_i a *trust* or *confidence* value c_i. Typically, c_i indicates how confident is the extractor about the accuracy of the extracted fact f_i. In that way, a direct method to determine the weight w_i is simply taking the confidence value of f_i, $w_i = c_i$. However, other alternatives to set w_i are also valid like, for instance, considering some type of "threshold" t, such that $w_i = c_i$ only in those cases where c_i is greater than t. Thus, for example, if a threshold $t = 0.8$ were considered, the w_i formula in that case would be:

$$w_i = \begin{cases} c_i & \text{if } c_i \geq 0.8 \\ 0 & \text{si } c_i < 0.8 \end{cases} \tag{2}$$

It is also clear here, that a trivial setting for w_i is giving the same uniform value to all the extracted facts (for instance $w_i = 1$).

From Equation 1 we can see that another key component to compute $\mathcal{S}_e(d)$ is the EFS of f_i, $s_e(f_i)$. Intuitively, this quantity should give some information about how many times the fact f_i was found in the external source E_s. Thus, if f_i appears N_i times in E_s, a direct option is using $s_e(f_i) = N_i$ as external factual support of f_i. However, we also could be interested in the *boolean case*, that is, only evaluating if f_i was found in E_s or not. In that case, $s_e(f_i)$ might be defined as:

$$s_e(f_i) = \begin{cases} 1 & \text{if } f_i \in E_s \\ 0 & \text{in other case.} \end{cases} \tag{3}$$

Another aspect that must be taken into account in the support computation is the *size* of a document d. Intuitively, we can speculate that a greater size of d will result in a higher value of $\mathcal{S}_e(d)$. Thus, some kind of "normalization" in our metric definition could be desirable. Therefore, instead of directly considering the $\mathcal{S}_e(d)$ formula shown in Equation 1, we will use a more general equation that allows to specify that no normalization is required, or different normalization units when the results need to be normalized. Thus, our EFS formula for a document now is defined as:

$$\widehat{\mathcal{S}_e}(d) = \frac{\mathcal{S}_e(d)}{nor} \tag{4}$$

with the *normalization factor nor* taking one of the following values: a) $nor = 1$ (no normalization), b) $nor = NL_d$ (number of lines in d, c) $nor = NW_d$ (number of words in d), $nor = |F_d|$ (number of facts extracted from d).

In summary, if the different options for w_i are identified as: C when $w_i = c_i$, T when Equation 2 is used and U when $w_i = 1$; we identify the alternatives for $s_e(f_i)$ as: N when $s_e(f_i) = N_i$ and B for the "boolean case" (Equation 3), and the normalization alternatives are denoted as: N (no normalization), L (lines-based normalization), W (words-based normalization) and F (facts-based normalization), we can see that different methods for computing the external factual support are obtained by simply considering different combinations of the weight w_i, the external support of the facts ($s_e(f_i)$) and the used normalization (if any). Following the above specified naming convention, each of those components will be assigned a character in a "code" that will identify the used support. Thus, for instance, an EFS identified as "CNW" will correspond to the case in which w_i is the confidence level assigned by the fact-extraction system (Reverb in our case) to f_i, the external support of f_i is the number of occurrences of f_i in E_s and the results are normalized taking into account the number of words in each document d. Table 1 summarizes different support codifications that result from using different alternatives for w_i, $s_e(f_i)$ and nor.

Codification	w_i	$s_e(f_i)$	nor
CNN	c_i	N_i	1
CNL	c_i	N_i	NL_d
CNW	c_i	N_i	NW_d
CNF	c_i	N_i	$\|F_d\|$
CBN	c_i	Equation 3	1
CBL	c_i	Equation 3	NL_d
CBW	c_i	Equation 3	NW_d
CBF	c_i	Equation 3	$\|F_d\|$
TNN	Equation 2	N_i	1
TNL	Equation 2	N_i	NL_d
TNW	Equation 2	N_i	NW_d
TNF	Equation 2	N_i	$\|F_d\|$
TBN	Equation 2	Equation 3	1
TBL	Equation 2	Equation 3	NL_d
TBW	Equation 2	Equation 3	NW_d
TBF	Equation 2	Equation 3	$\|F_d\|$
UNN	1	N_i	1
UNL	1	N_i	NL_d
UNW	1	N_i	NW_d
UNF	1	N_i	$\|F_d\|$
UBN	1	Equation 3	1
UBL	1	Equation 3	NL_d
UBW	1	Equation 3	NW_d
UBF	1	Equation 3	$\|F_d\|$

Table 1: EFS codifications.

There is an aspect that has not been analyzed yet but, as it will be seen in the next section, deserves a lot of attention: the process used to "match" a fact f_i with the facts in the external source E_s when the $s_e(f_i)$ value needs to be computed. Up to now, we have assumed that a fact f_i is "found" in E_s when there is a "perfect" matching with the external fact, that is to say, they are the same fact. However, we will see later that this "strict matching" approach produces low recall values and the matching process needs to be relaxed.

4 Implementation aspects and experimental results

To test the feasibility of the proposed quality metric it is necessary generating adequate data sets with high and low quality Wikipedia's articles. Intuitively, the EFS metric should help discriminating in these data sets between both types of articles. Wikipedia has a definite concept of information quality standard represented by the concepts of "Featured articles" and "Good articles". Its editors annotate articles with respect to these information quality criteria which makes them perfectly suited as positive examples of the highest quality articles that one would expect to find in Wikipedia. Featured/Good articles were identified by searching for files in a Wikipedia dump that contained the featured article or good article template in the Wikitext. As low quality examples, we used non-featured articles that were randomly selected from the remaining articles in the dump or taken from a set of articles that had a specific "flaw", as it will be explained below.

Our dataset consists of 2445 Wikipedia articles, 1000 featured/good and 1445 non-featured articles. They will be referred from now on as the "featured article" (FA) set and the "non-featured article"(NF) set respectively. In fact, we can differentiate in the NF set two subsets: one, the subset that we will name NF_R, formed by 939 "regular" non-featured articles randomly selected from the snapshot of the English Wikipedia from October 2011; the other one, that will be called NF_{OR}, consists of 506 articles that have been tagged as having the "original research" flaw in the corpus used in the PAN'12 competition on "Quality Flaw Prediction in Wikipedia" (Anderka and Stein, 2012). The rationale of having those subsets separated is simple. The external support proposed in the present work is intended to detect some of the characteristics that are distinctive of original research. In that way, if both NF subsets are differentiated in the experimental work, we will be able to detect to what extend the original research affects our proposed measure and the other ones used in the experiments.

The whole dataset was processed in order to obtain 24 EFS measures that correspond to the 24 codifications described in Table 1. We used as external source E_s the ReVerb ClueWeb Extractions data set (Fader, Soderland, and Etzioni, 2011). This data set contains approximately 15 million binary assertions from the Web. It is a subset of ReVerb's output run on the English portion of the ClueWeb09 corpus.[5]

As it was pointed out above, the "strict matching" approach used to determine the EFS of each fact produced very low recall values. In fact, for many arbitrary Wikipedia articles d_j, all the extracted facts $F_{d_j} = \{f_{j_1}, \ldots, f_{j_n}\}$ will have a EFS $s_e(f_{j_i}) = 0$, for $i = 1 \ldots n$ and, in consequence, the external factual support of d_j, $\widehat{\mathcal{S}}_e(d_j)$ will be 0 for all the codifications shown in Table 1. Thus, for instance, if only the articles d that

[5]More information on the ReVerb homepage at: http://reverb.cs.washington.edu/

have $\widehat{\mathcal{S}}_e(d) \neq 0$ are considered, a reduction in the number of articles is observed in all the (sub)-sets of our dataset: from $|FA| = 1000$ to 346, from $|NF_R| = 939$ to 78 and from $|NF_{OR}| = 506$ to 75. We will denote $FA^\star$, $NF_R^\star$ and $NF_{OR}^\star$ those "reduced" (non-zero external factual support) sets of articles (see Table 2).

It is interesting to notice that, despite the low recall problem that introduces the "strict matching" approach, we can already see some "discriminative" capabilities of the external factual support. The percentage of FA documents with external support $\neq 0$: $346/1000 = 34.6$ %, is considerably higher than the percentage of non-featured articles with external support $\neq 0$ in the NF set: $\frac{|NF_R^\star| \cup |NF_{OR}^\star|}{|NF_R| \cup |NF_{OR}|} = 153/1445 = 10.59\%$.

This is an encouraging reason for keep working on the external support measures and also poses a challenging scenario to be addressed in the experimental work. That is to say, $FA^\star$, $NF_R^\star$ and $NF_{OR}^\star$ constitute by themselves a difficult dataset to test our EFS measures. It represents a sub-collection of the original dataset where the negative class $(NF_R^\star \cup NF_{OR}^\star)$ includes those examples that are the nearest to the positive examples because they have at least some minimum EFS (with respect to "strict matching" approach).

Obviously, to obtain a metric that gives more information on all the considered documents, it is necessary to define alternative (more relaxed) matching criteria than the exact matching of facts. We have a lot of possibilities to do this and, in fact, they will be considered in future works. However, in the present work we decided to start with two very simple matching approaches that we called the *local* and *global* matching approaches.

The local matching approach simply measures the component-by-component degree overlapping of each part of a fact and the (external) fact we are comparing to. More formally, let $f_i = (s_{i_1}, s_{i_2}, s_{i_3})$ be the fact we are computing the external factual support, and let $f_e = (s_{e_1}, s_{e_2}, s_{e_3})$ be a fact in the external source, $f_e \in E_s$. The *local matching* of f_i with respect to f_e will be:

$$\mathcal{M}_l(f_i, f_e) = J(s_{i_1}, s_{e_1}) \times J(s_{i_2}, s_{e_2}) \times J(s_{i_3}, s_{e_3})$$

where $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$ is the *Jaccard similarity coefficient* of sets A and B. That is to say, $\mathcal{M}_l(f_i, f_e)$ computes the product of the component-by-component overlapping degree of both facts, considering that each part of a fact is a set of *terms*. Those terms are not the *words* directly found in the original documents (Wikipedia's articles) but a "processed" version of those words. In a few words, that processing is required because we use the "lemmatized" version of the external source and, in that way, a similar format is required for the components of the facts found in Wikipedia's articles.[6]

It is important to see that any "empty" overlapping between two parts of f_i and f_e will cause a $\mathcal{M}_l(f_i, f_e) = 0$ value.

The global matching approach only differs from the local one in that it considers all the parts in a fact as a single set, that is

$$\mathcal{M}_g(f_i, f_e) = J(s_{i_1} \cup s_{i_2} \cup s_{i_3}, s_{e_1} \cup s_{e_2} \cup s_{e_3})$$

The only aspect to decide in each case is which would be an appropriate threshold value t, such that when $\mathcal{M}(f_i, f_e) \geq t$ it will be considered that f_i and f_e "match". In the experimental work we empirically determine two different thresholds $t_l = 0.3$ and $t_g = 0.4$ for $\mathcal{M}_l$ and $\mathcal{M}_g$ respectively, that produced fairly reasonable matches between facts.

In Table 3 a summary of the number of documents of each sub-group of the data set is shown and also of the reduced version ($DS^\star$) that results of considering non-zero EFS measures when different matching approaches are used. There, we can see that an almost perfect recall is obtained for featured articles for the "relaxed" matches, with

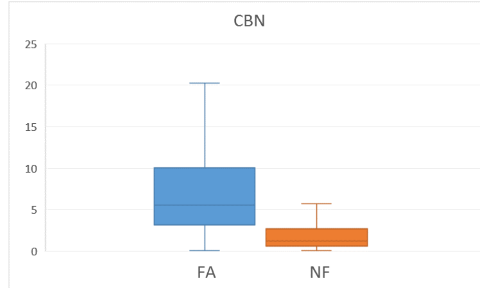

Figure 2: FA's vs NF's CBN values.

[6]This lemmatization task was carried out with the Wordnet lemmatizer provided by the Java Wordnet API named CICWN, `fviveros.gelbukh.com/wordnet.html`.

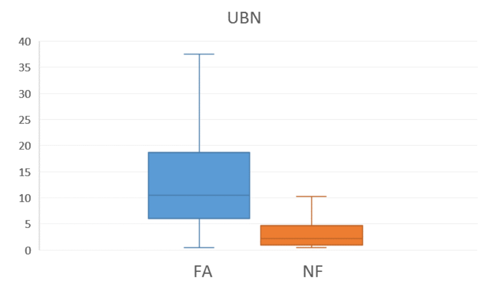

Figure 3: FA's vs NF's UBN values.

$\frac{|FA^{\star}|}{|FA|} = 960/1000 = 96\%$ of external support $\neq 0$ for the local matching, and $\frac{|FA^{\star}|}{|FA|} = 999/1000 = 99.9\%$ of external support $\neq 0$ for the global matching approach. Following a similar analysis to the one carried out with strict matching, we can see that these recall values are higher for FA articles than for the ones obtained with NF articles in both, local and global matching: $96\% > 890/1445 = 61.6\%$ and $99.9\% > 1234/1445 = 85.4\%$ respectively.

Results showed in Table 3 are indicative of the "coarse grained" capabilities of EFS to discriminate Wikipedia's articles according to quality criteria. Other more detailed analysis, as the one presented in (Ingawale et al., 2013), consists in comparing how a metric evaluates in terms of min, max and average values when applied to featured versus non-featured articles. Here, we will constrain this analysis to the two codifications of the metric (UBN and CBN) that obtained the highest information gain in relation to both categories (featured and non-featured) when the global matching is used. For both codifications, we present this information by showing their box plots in Figures 2 and 3. In those graphs, it can be seen that values for both codifications of the metric are consistently higher in FA than in NF articles.

For a comprehensive analysis of the metric it also would be interesting analysing its performance as feature for Wikipedia's article representation in standard text categorization tasks. Space constraints prevent us from doing an exhaustive study of this type but, in a similar way to the analysis performed in (Stvilia et al., 2005), we present some preliminary results of its performance in a simple binary (FA versus NF) supervised task. With documents represented with the 24 EFS codifications (24 features), standard 10-fold cross validation tests with decision trees (DTs) [7] and (backpropagation) multi-layer neural networks (NNs) produced the following results: DT obtained an accuracy of 85.9382 with 1919 correctly classified instances out of 2233 Wikipedia articles; NN, on the other hand, correctly classified 1970 instances with an accuracy of 88.2221. We also tested the capabilities of our metric in non-supervised (clustering) categorization tasks, with a simple k-means algorithm with $k = 2$ and the five codifications with the highest information gain (UBN, CBN, TBL, TBF and TBW). In this case, these codifications were able of generating good groupings of FA and NF articles with only 26.562% of instances incorrectly classified in the wrong group.

5 *Conclusions and Future Works*

Using "external" information to assess the IQ of a document seems to be an interesting idea already posed by Juffinger et al. (Juffinger, Granitzer, and Lex, 2009) in the context of a blog credibility ranking task. Magdy et al. in (Magdy and Wanas, 2010) also measure the support of textual documents by using very basic facts derived from Noun-to-Noun phrases of a document. These facts are compared to those obtained from the information retrieved by a well known search engine (Bing). The procedure used to obtain facts, how the match between facts is determined and the used external resource differ from the ones used in this article. However, it could be considered as the previous work closest to our idea of "external factual support".

In the present article, the motivations behind our EFS metric, its formal definition and the main implementation aspects were introduced. Different data sets for research in quality metrics for Wikipedia were generated, described and made available for other researchers. They include plain texts of high and low quality Wikipedia articles and data sets with the values of the proposed metric in its 24 variants (see Table 1).[8] In this context, preliminary statistics obtained with the EFS metric show its capability to (coarse-grained) filtering and more fine numeric analysis of featured versus non-featured articles. This

[7]Weka's J4.8 implementation of the decision tree learning algorithm C4.5.

[8]Those interested readers can contact the first author of the article to have access to these collections.

	Featured Articles	Regular non-feat. articles	Original Research
Data Set (DS)	$\|FA\| = 1000$	$\|NF_R\| = 939$	$\|NF_{OR}\| = 506$
Reduced Data Set ($DS^\star$)	$\|FA^\star\| = 346$	$\|NF_R^\star\| = 78$	$\|NF_{OR}^\star\| = 75$

Table 2: Data sets description - Strict matching.

	DS	$DS^\star$ - Strict Matching	$DS^\star$ - Local Matching	$DS^\star$ - Global Matching
FA	$\|FA\| = 1000$	$\|FA^\star\| = 346$	$\|FA^\star\| = 960$	$\|FA^\star\| = 999$
NF_R	$\|NF_R\| = 939$	$\|NF_R^\star\| = 78$	$\|NF_R^\star\| = 514$	$\|NF_R^\star\| = 757$
NF_{OR}	$\|NF_{OR}\| = 506$	$\|NF_{OR}^\star\| = 75$	$\|NF_{OR}^\star\| = 376$	$\|NF_{OR}^\star\| = 477$

Table 3: Data sets description - Strict, Local and Global matching.

good performance, was also observed in some basics experiments using EFS codifications as representation features for categorization tasks of Wikipedia's articles.

As future work, we want to compare the performance of EFS codifications with the obtained with other state of the art proposals in the area (Blumenstock, 2008; Lipka and Stein, 2010) and other approaches based on factual information (Lex et al., 2012). In the last case, the focus will be on determining to what extent EFS information extend/complement "internal" factual information present in the analysed Wikipedia article. Besides, other more elaborated matching mechanisms and external sources will be considered. In this context, the feasibility of using the proposed EFS metric in other domains beyond the Wikipedia encyclopedia will also be considered.

References

Anderka, M. and B. Stein. 2012. Overview of the 1st int, competition on quality flaw prediction in wikipedia (CLEF 2012).

Anderka, M., B. Stein, and N. Lipka. 2012. Predicting Quality Flaws in User-generated Content: The Case of Wikipedia. In *35rd Annual Int. ACM SIGIR Conf. on Research and Development in Inf. Retrieval*. ACM.

Anthony, D., S. Smith, and T. Williamson. 2009. Reputation and reliability in collective goods: The case of the online encyclopedia wikipedia. *Rationality & Society*, 21(3):283–306.

Baeza-Yates, R. 2009. User generated content: how good is it? In *3rd Workshop on Information Credibility on the Web (WICOW'09)*, pages 1–2. ACM.

Blumenstock, J. E. 2008. Size matters: word count as a measure of quality on Wikipedia. In *17th Int'l Conference on World Wide Web*. ACM.

Emigh, W. and S. Herring. 2005. Collaborative authoring on the Web: a genre analysis of online encyclopedias. In *Proc. of the 38th annual Hawaii int. conference on system sciences*. IEEE CS.

Etzioni, O., M. Banko, S. Soderland, and D. Weld. 2008. Open information extraction from the web. *Communications of the ACM*, 51(12):68–74.

Fader, A., S. Soderland, and O. Etzioni. 2011. Identifying relations for open information extraction. In *Proc. of the Conf. of Empirical Methods in Natural Language Processing*, Scotland.

Ferretti, E., M. L. Errecalde, M. Anderka, and B. Stein. 2014. On the use of reliable-negatives selection strategies in the PU learning approach for quality flaws prediction in wikipedia. In *25th International Workshop on Database and Expert Systems Applications, DEXA 2014, Munich, Germany, September 1-5, 2014*, pages 211–215.

Hu, M., E. Lim, A. Sun, H. Lauw, and B. Vuong. 2007. Measuring article quality in Wikipedia: models and evaluation. In *16th ACM International CIKM'07*, pages 243–252. ACM.

Ingawale, M., A. Dutta, R. Roy, and P. Seetharaman. 2013. *Network analysis of user generated content quality in Wikipedia. Online Information Review*, 37(4):602–619.

Juffinger, A., M. Granitzer, and E. Lex. 2009. Blog credibility ranking by exploiting verified content. In *Proc. of WICOW 2009*, pages 51–58. ACM.

Lex, E., M. Völske, M. Errecalde, E. Ferretti, L. Cagnina, C. Horn, B. Stein, and M. Granitzer. 2012. Measuring the quality of web content using factual information. In *2nd joint WICOW/AIRWeb Workshop on Web quality*. ACM.

Lih, A. 2004. Wikipedia as participatory journalism: reliable sources? Metrics for evaluating collaborative media as a news resource. In *5th Int. Symposium on Online Journalism*, pages 16–17.

Lipka, N. and B. Stein. 2010. Identifying featured articles in wikipedia: writing style matters. In *Proc. of the 19th int. conference on World wide web*, WWW '10, pages 1147–1148, NY, USA. ACM.

Magdy, A. and N. Wanas. 2010. Web-based statistical fact checking of textual documents. In *Proc. of SMUC '10*, pages 103–110, NY, USA. ACM.

Stvilia, B., M. Twidale, L. Smith, and L. Gasser. 2005. Assessing information quality of a community-based encyclopedia. In *10th Int. Conference on Information Quality*, pages 442–454. MIT.

Viégas, F., M. Wattenberg, and K. Dave. 2004. *Studying Cooperation and Conflict Between Authors with History Flow Visualizations*. In *Proc. of the SIGCHI Conf.*, pages 575–582. ACM.

Wilkinson, D. and B. Huberman. 2007. *Cooperation and Quality in Wikipedia*. In *Proc. of the 2007 Int. Symposium on Wikis*, pages 157–164. ACM.