Measuring the Quality of Web Content using Factual Information

Elisabeth Lex
Know-Center GmbH
elex@know-center.at

Michael Voelske
Bauhaus-Universität Weimar
michael.voelske@uni-weimar.de

Marcelo Errecalde
Edgardo Ferretti, Leticia Cagnina
Universidad Nacional de San Luis
{merreca|ferretti|lcagnina}@unsl.edu.ar

Christopher Horn
Graz University of Technology
christopher.horn@tugraz.at

Benno Stein
Bauhaus-Universität Weimar
benno.stein@uni-weimar.de

Michael Granitzer
University of Passau
Michael.Granitzer@uni-passau.de

ABSTRACT

Nowadays, many decisions are based on information found in the Web. For the most part, the disseminating sources are not certified, and hence an assessment of the quality and credibility of Web content became more important than ever. With *factual density* we present a simple statistical quality measure that is based on facts extracted from Web content using Open Information Extraction. In a first case study, we use this measure to identify featured/good articles in Wikipedia. We compare the factual density measure with word count, a measure that has successfully been applied to this task in the past. Our evaluation corroborates the good performance of word count in Wikipedia since featured/good articles are often longer than non-featured. However, for articles of similar lengths the word count measure fails while factual density can separate between them with an F-measure of 90.4%. We also investigate the use of *relational features* for categorizing Wikipedia articles into featured/good versus non-featured ones. If articles have similar lengths, we achieve an F-measure of 86.7% and 84% otherwise.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.3 Information Search and Retrieval—*Information filtering*

1. INTRODUCTION

People use the Web as a basis for their decisions and beliefs. Due to lacking quality control, Web-based information sources often contain inaccurate and false information. Thus, in addition to the content itself, measures are needed to capture credibility and quality aspects. In this work, we propose a statistical quality measure called *factual density*, which assesses the quality of content with respect to facts. We define the factual density of a document as the number of facts found in this document in relation to the document length. Consequently, factual density indicates a document's informativeness. We also propose to use binary relations, i.e. triples of the form (argument1, relation, argument2) [3], as features to distinguish between high-quality factual content and non-factual content. Our hypothesis is that a document's content is of higher quality if it is both factual and informative.

WebQuality '12, April 16, 2012, Lyon, France

1.1 Related Work

The quality of Web content has mainly been assessed with metrics capturing content quality aspects like objectivity [6], content maturity and readability [10]. A key aspect here is to determine an appropriate set of features. In [6], it is proposed to use stylometric features to assess content quality. Lipka and Stein [7] exploit character trigrams distributions to identify high quality featured/good articles in Wikipedia. Blumenstock [2] suggests to simply use word count as indicator for the quality of Wikipedia articles.

To assess the factual accuracy of Web content, more complex, semantic features are needed. A common approach is to employ Open Information Extraction [4] or methods that use background knowledge on semantic relations available in ontological resources such as Wordnet [5] and Yago [9]. These approaches extract relational information about entities named in a particular text (e.g., facts like f = *(Mozart, was_born_in, Salzburg)*). Besides, they exploit defined semantic relationships such as meronymy and hypernymy, and others to infer relational information between entities,which is not given explicitly in the text. In this work, we refer to such features as *relational features*.

2. MEASURING THE QUALITY USING FACTUAL INFORMATION

In order to measure information quality based on factual information, we propose three approaches: (i) using simple statistics about the facts obtained from a text, (ii) exploiting relational information contained in facts, and (iii) exploiting semantic relationships like meronymy and hypernymy.

In this work, we focus on the first two approaches. In the first approach, we resort to simple statistical features about facts to determine the informativeness of a document. We denote this kind of features as *fact frequency-based features*.

2.1 Fact Frequency-Based Features

Fact frequency-based features are of a simple structure; they require only direct 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 paragraph, a document, a corpus), and F_t is the collection of facts extracted from t by an information extraction method IE, a possible measure could be the number of facts extracted by IE from t. This quantity will be referred to as the *fact count* of t and is defined below.

Definition 1. Let t be an arbitrary textual resource and F_t be the collection of facts extracted from t by an arbitrary information extraction method IE. The *fact count* of t, denoted $fc(t)$, is defined as the total number of facts obtained from t by IE, $fc(t) = |F_t|$.

Obviously the fact count depends on the size of the textual resource t: long texts have more facts than short texts. We hence relate the fact count to the size of t and refer to this quantity as the *factual density* of t:

Definition 2. Let t be an arbitrary textual resource and $fc(t)$ be the fact count of t. Let $size(t)$ be a measure intended to quantify the *size* of t[1]. The *factual density* of t, denoted $fd(t)$, is defined as $fd(t) = \frac{fc(t)}{size(t)}$.

2.2 Relational Features

In this work, we explore the possibility of representing a document as a bag-of-relations, where a textual resource t is represented as the multi-set of relations r_{ij} occurring in all facts extracted from t. This is somewhat analogous to the bag-of-words model commonly applied in Information Retrieval.

The relations we use represent binary relations between entities. At this point, we consider facts of the form $f = (e_i, r_{ij}, e_j)$, where e_i and e_j are strings that denote *entities*, while r_{ij} is a string denoting a *relationship* between them. In this context, a sentence such as "Mozart was born in Salzburg." is written as fact *(Mozart, was_born_in, Salzburg)*. In this respect, this setting is similar to the ones used in recent works on Open Information Extraction and contradiction detection [4, 8].

3. DATASET AND PREPROCESSING

Our dataset consists of 2000 Wikipedia articles, 1000 featured/good and 1000 non-featured articles randomly selected from the snapshot of the English Wikipedia from October 2011. Due to the lack of a standard corpus related to our work, we used Wikipedia, because its editors annotate articles with respect to information quality. We focus on the "Featured Article" and "Good Article" annotations. Per definition, featured/good articles are of high information quality [1]. Also, they provide a comprehensive coverage of the major facts in the context of the article's subject[2] which makes them perfectly suited as positive class for our task. Featured/Good articles were identified by searching for files in the dump that contained the featured article or good article template in the Wikitext. As negative class, we used non-featured articles that were randomly selected from the remaining articles in the dump. We use the corresponding binary classification problem to evaluate our factual information quality measures.

In [2], word count is proposed as a simple but effective heuristic to distinguish featured/good and non-featured articles. We use this measure as a baseline. Typically, featured/good articles are longer than non-featured articles, which introduces a bias since longer documents probably contain more facts. To evaluate this, we created a balanced corpus from our dataset. "Balanced" means that featured/good and non-featured articles were selected with almost similar document lengths, which left us with 740 articles in each category.

Figures 1(a) and 1(b) show the distributions of the documents of both corpora according to document length. Due to computational constraints, we counted the article length on the Wikitext. Given that some articles contain only templates, their content is empty after removing the Wikitext.

Most of the non-featured articles from the unbalanced dataset have fewer than 2000 words, thus word count is highly discriminative. On the balanced corpus, the document distributions of non-featured and featured/good articles overlap between 500 and 2000 words, which weakens the discriminative power of word count.

The plain text has been extracted from the articles using the Sweble Wikitext parser.[3] From the plain text, we extracted facts and relational features using the ReVerb Open Information Extraction framework.[4] For normalization, the texts were split into sentences with the OpenNLP framework.[5] Finally, we used the facts to compute the factual density measure. The relational features have been used to classify Wikipedia articles into featured/good versus non-featured.

4. EXPERIMENTS AND RESULTS

First, we evaluated the factual density measure on the unbalanced corpus. For this, we first computed the word count baseline on this corpus. A tokenization based on whitespaces resulted in an average word count of 200 words for the lower quality articles and 1400 for the high quality articles. Naturally, the number of words per article directly influences the number of extracted facts per article. In this dataset, the featured/good articles contain on average 159 facts, while non-featured articles contain only 27 facts on average. From this observation, we conclude that the number of facts per document is a good feature to distinguish between featured/good and non-featured articles. To verify this, we performed two experiments on the unbalanced and balanced corpora, as reported below.

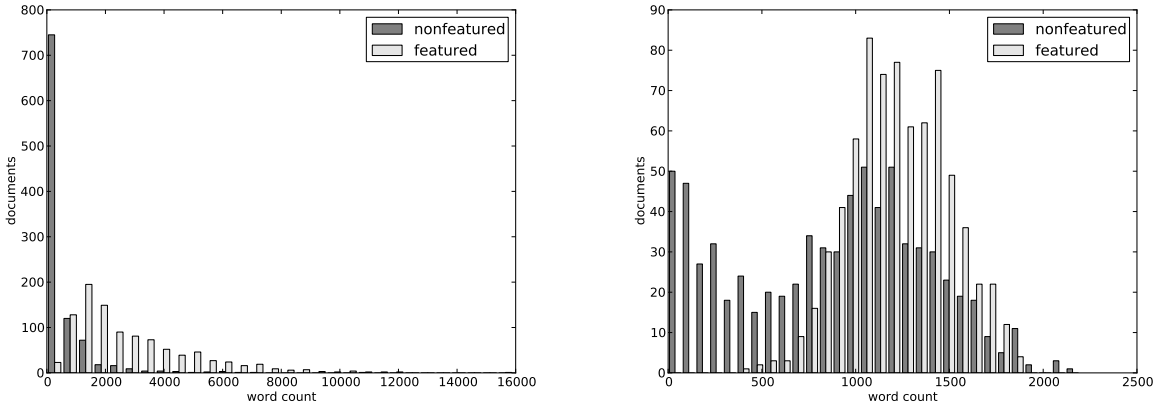
[1] For instance, it could be the number of words or sentences in t or the length in number of characters of t.

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

[3] `http://sweble.org/wiki/Sweble_Wikitext_Parser`

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

[5] `http://opennlp.sourceforge.net/projects.html`



(a) Unbalanced (b) Balanced

Figure 1: Histograms of Wikipedia corpora for unbalanced dataset and balanced dataset.

Figure 2(a) shows three precision-recall curves for the featured/good versus non-featured categorization task on the unbalanced corpus. The plain line plots the results achieved with the word count measure. The dotted line with circles represents the results obtained with factual-density/sentence-count, while the line with squares illustrates the results obtained with factual-density/word-count. Here, "Factual-density/word-count" refers to the factual density measure derived from the formula $fd(t)$ in Definition 2 where $size(t)$ is the word count of t, and t is a Wikipedia article. The same holds for "Factual-density/sentence-count".

The word count measure outperforms the factual density measure normalized to sentence count as well as the word count on the unbalanced corpus. Apparently, word count is a strong feature on the unbalanced corpus. We then evaluated the factual density measure on the balanced corpus where both featured/good and non-featured articles are more similar in respect to document length. The results for this experiment are shown in Figure 2(b) as precision-recall curves.

On the balanced corpus, factual density normalized to sentence count as well as word count performs much better than on the unbalanced corpus, while word count, as expected, performs worse. There is not much difference between the normalization to word or sentence count since here, the number of words per document has a smaller influence on the result.

We also analyzed the distributions of featured/good and non-featured articles if factual density is used as measure, as depicted in Figure 3. We found that the distribution of the featured/good articles is clearly separated from the distribution of the non-featured articles, with peaks at two different factual density values (0.06 and 0.03 respectively).

This finding is in contrast to the fact that the distributions of featured/good articles and non-featured articles have a high degree of overlap if word count is used, as shown in Figure 1(b). Consequently, on the balanced corpus, factual density clearly outperforms our baseline word count.

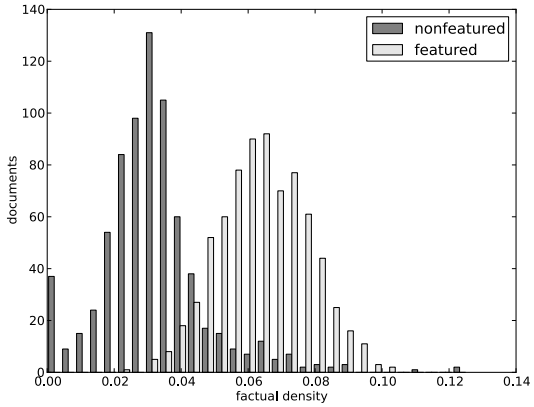


Figure 3: Distribution of articles by factual density.

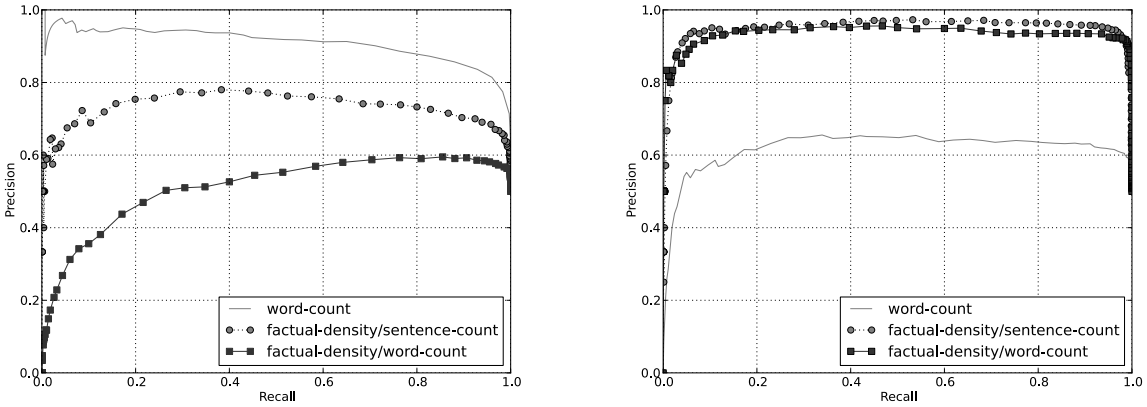
In a related experiment, we investigated the relational information contained in the binary relationships ReVerb extracts from sentences. We used the relations, i.e. only the predicates from the extracted triples as a vocabulary to represent the documents. We then tested the discriminative power of these features by training a classifier to solve the binary classification problem of distinguishing featured/good from non-featured articles. The results reported in Table 1 were obtained using the WEKA[6] implementation of a Naive Bayes Classifier in combination with feature selection based on Information Gain (IG). From 40 000 relations, we selected the 10% best features in terms of IG.

We achieved similar results for both corpora. Apparently, relational features are more robust when the document length varies. However, we need to investigate this in more detail.

5. CONCLUSIONS

In this work, we propose to exploit facts for assessing Web content quality. We suggest a simple fact-related quality

[6] `http://www.cs.waikato.ac.nz/~ml/weka/`



(a) Unbalanced (b) Balanced

Figure 2: Factual Density on Wikipedia corpora.

Table 1: Classification results using relational features on both corpora.

	Unbalanced	Balanced
Measure	Value [%]	Value [%]
Accuracy	84.01	87.14
F-Measure	84	86.7
Precision	84	89.2
Recall	84	87.1

measure, *factual density*. Factual density measures the relative number of document facts and thus indicates a document's informativeness. Our experiments on a subset of the English Wikipedia reveal that, based on factual density, featured/good articles can be separated from non-featured articles with a high confidence even if the articles are similar in length. If the articles differ in terms of length, our experiments corroborate previous work indicating that word count is a good estimator of article quality in Wikipedia, since featured/good articles are often longer than non-featured. In the future, we aim to evaluate a combination of word-count and factual-density to outperform wordcount also on unbalanced datasets. Besides, we plan to incorporate the edit history: we believe articles with more editors are denser in terms of facts. We also describe preliminary experiments employing relational features to solve the featured/good versus non-featured classification problem. While the initial results are very promising, more in-depth investigations of these features are needed. In the long run, we aim to exploit defined semantic relationships such as meronymy and hypernymy to infer relational information between entities. We expect these to unlock several new information quality dimensions, such as generality/specificity and consistency.

6. ACKNOWLEDGMENTS

This work has been funded by the European Commission as part of the WIQ-EI project (project no. 269180) within the FP7 People Programme. The authors M. Errecalde, E. Ferretti and L. Cagnina thank the Universidad Nacional de San Luis from which receive continuous support. The Know-Center is funded within the Austrian COMET Program under the auspices of the Austrian Ministry of Transport, Innovation and Technology, the Austrian Ministry of Economics and Labor and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG.

7. REFERENCES

[1] M. Anderka, B. Stein, and N. Lipka. Towards automatic quality assurance in wikipedia. In *Proc. of the 20th int. conf. on World wide web*, pages 5–6, 2011.

[2] J. E. Blumenstock. Size matters: word count as a measure of quality on wikipedia. In *Proc. of the 17th int. conf. on World Wide Web*, pages 1095–1096, 2008.

[3] M. J. Cafarella, J. Madhavan, and A. Halevy. Web-scale extraction of structured data. *SIGMOD Rec.*, 37:55–61, 2009.

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

[5] C. Fellbaum, editor. *WordNet: an electronic lexical database*. MIT Press, 1998.

[6] E. Lex, A. Juffinger, and M. Granitzer. Objectivity classification in online media. In *Proc. of the 21st ACM conf. on Hypertext and hypermedia*, pages 293–294, 2010.

[7] N. Lipka and B. Stein. Identifying featured articles in wikipedia: writing style matters. In *Proc. of the 19th int. conf. on World wide web*, 2010.

[8] A. Ritter, S. Soderland, D. Downey, and O. Etzioni. It's a contradiction - no, it's not: A case study using functional relations. In *EMNLP*, pages 11–20. ACL, 2008.

[9] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: a core of semantic knowledge. In *Proc. of the 16th int. conf. on World Wide Web*, pages 697–706, 2007.

[10] N. Weber, K. Schoefegger, J. Bimrose, T. Ley, S. Lindstaedt, A. Brown, and S.-A. Barnes. Knowledge maturing in the semantic mediawiki: A design study in career guidance. In *Learning in the Synergy of Multiple Disciplines*, pages 700–705. Springer Berlin/Heidelberg, 2009.