Identifying Featured Articles in Wikipedia

Writing Style Matters

Nedim Lipka and Benno Stein
Bauhaus-Universität Weimar
99421 Weimar, Germany
<first name>.<last name>@uni-weimar.de

ABSTRACT

Wikipedia provides an information quality assessment model with criteria for human peer reviewers to identify featured articles. For this classification task "Is an article featured or not?" we present a machine learning approach that exploits an article's character trigram distribution. Our approach differs from existing research in that it aims to writing style rather than evaluating meta features like the edit history. The approach is robust, straightforward to implement, and outperforms existing solutions. We underpin these claims by an experiment design where, among others, the domain transferability is analyzed. The achieved performances in terms of the F-measure for featured articles are 0.964 within a single Wikipedia domain and 0.880 in a domain transfer situation.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval; H.5.3 [Information Interfaces]: Group and Organization Interfaces

General Terms: Algorithms, Experimentation.

Keywords: Wikipedia, Information Quality, Domain Transfer.

1. INTRODUCTION

The automatic assessment of information quality (IQ) will become a key factor in information retrieval. Whether this is possible in its generality is an open question since the quality of a text is subjectively perceived: it depends on a user's context, her expectations, and on prior knowledge. Wikipedia provides a controlled situation, where high-quality articles are labeled as *featured*, after being run through an extensive human peer review process. The Wikipedia community characterizes featured articles among others as well-written, comprehensive, well-researched, neutral, and stable.[1] The paper in hand focuses on the automatic identification of featured articles in Wikipedia.

Related Work. Several researchers develop metrics that are suitable to capture quality indicators, but that are demanding in computational respects: Zeng et al. [10, 5] compute an article's trustworthiness using revision history features and citation features. Stvilia et al. [9] develop metrics that are based on edits, editors, links, article length, age, and readability indices. Brandes et al. [3] indicate structural parameters of the edit network. Stein and Hess [8] as well as Adler and Alfaro [1] develop authorship-based quality ratings, which concern the amount of the authors contributions in an article and a reputation estimate. Hu et al. [4] take the reviewership into account, which relies on the assumption that unedited content is reviewed by an author who edits the respective article.

Blumenstock [2] proposes the word count of an article, which is a simple metric but which works significantly better than several of the aforementioned metrics. His approach, the classification of articles with more than 2000 words as featured, is doing well for an unbalanced corpus with a large amount of small articles. However, our experiments show a performance decline for balanced corpora as well as when only articles with 1500–2500 words are used.

Contributions. We employ various trigram vector representations along with a classifier in order to identify featured articles. In particular, we examine their robustness and generalizability in domain transfer experiments. Especially character trigram vectors, which are not yet considered in IQ research, are a promising representation: they are comparable to word counts in simplicity but gain a higher discriminability. The following sections explain the rationale of those trigrams and report on the experiments and results.

2. AUTOMATIC IQ ASSESSMENT

Starting point is the classification task "Is an article featured or not?". For this purpose we apply two established learning algorithms, namely linear support vector machines (SVM) and Naïve Bayes (NB) [6]. Our study deals with writing-style-related representations of articles and their binarizations: (1) character trigram vectors and (2) part of speech (POS) trigram vectors.

An n-gram vector of a text t is an $\ell 1$-normed numeric vector, where each dimension specifies the frequency of its associated n-gram in t. An n-gram in turn is a substring of n tokens of t, where a token can be a character, a word, or a POS tag. The vector is called binarized if the occurrence or non-occurrence of an n-gram is counted as 1 and 0, respectively.

POS n-gram vectors and character n-gram vectors are writing-style-related since they capture intrinsics of an author's text synthesis traits. POS n-grams unveil sentence construction preferences; character n-grams unveil preferences for sentence transitions as well as the utilization of stopwords, adverbs, and punctuation—all of which are important authorship indicators. To illustrate how writing style matters with respect to our classification task, Table 1 compiles the most discriminative character trigrams, ranked by information gain on our evaluation corpora. Note that authorship indicators are more important than topic indicators, such as word stems.

Table 1: The top 27 most discriminative character trigrams.

'ing'	'ng '	', a'	'at '	'e, '	'er '	' an'	'ed '	'd a'
' be'	'ter'	's a'	' re'	'as '	'ted'	'g a'	'tha'	'n t'
' a '	'ly '	'to '	' th'	'nd '	'. a'	'on '	'sed'	't t'

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

Copyright is held by the author/owner(s).
WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

Table 2: Identification performance for featured articles, within and across domains (P/R/F ∼ Precision / Recall / F-measure). Maximum F-measure values are shown in bold.

Representation	Classifier	Identification of featured articles (P/R/F)	
Cross Validation.		*within Biology*	*within History*
bin char trigram	SVM	0.966 / 0.961 / **0.964**	0.888 / 0.955 / **0.920**
bin POS trigram	SVM	0.949 / 0.933 / 0.941	0.889 / 0.925 / 0.907
word count	SVM	0.755 / 0.600 / 0.669	0.874 / 0.870 / 0.872
bag of words	NB	0.832 / 0.989 / 0.904	0.860 / 0.950 / 0.903
Domain Transfer.		*History → Biology*	*Biology → History*
bin char trigram	SVM	0.800 / 0.978 / **0.880**	0.886 / 0.855 / **0.870**
bin POS trigram	SVM	0.799 / 0.883 / 0.839	0.898 / 0.790 / 0.840
word count	SVM	0.772 / 0.733 / 0.752	0.878 / 0.830 / 0.853
bin bag of words	SVM	0.800 / 0.889 / 0.842	0.930 / 0.665 / 0.776

3. EXPERIMENTS AND RESULTS

The English Wikipedia domains Biology and History are used as sources for the compilation of two corpora: given the extracted plain texts with more than 800 words per article from a domain, all available featured texts and the same number of non-featured texts are added to the respective corpus. Altogether 180+180 articles belong to Biology, and 200+200 articles belong to History. We run three kinds of experiments:

Cross Validation. Evaluate a classifier c by tenfold cross validation within a single domain. Rationale of the experiment is to minimize the influence of topical discrimination, which can occur when articles of more than one domain are shuffled.

Domain Transfer. Construct a classifier c with articles from a source domain (training), and apply c to a different target domain (test). The experiments, denoted as "*source domain → target domain*", show both the potential of transfering relations about IQ across domains and the generalization ability of c.

Length Sensitivity. Apply a classifier c constructed within the domain transfer experiment to the three sets that contain those articles with less than 1500 words, those with 1500–2500 words, and those with more than 2500 words. The interesting questions are:

1. Is the article length sufficient for robust feature computation?
2. Is it sensible to combine a word-count-based classifier with an n-gram-based classifier?

Table 2, Cross Validation and Domain Transfer, and Table 3, Length Sensitivity, summarize the results of the trigram vector representations and, as baselines, the bag of words and the word count representations. Only the best performing representations, binarized or non-binarized, and classifiers, SVM or NB, are mentioned in the tables. Here, binarized trigram vectors outperform the non-binarized: about +0.2 averaged F-measure in the cross validation experiments and +0.3 in the domain transfer experiments.

The binarized character trigram vectors are most effective. As well as that, the length sensitivity analysis shows that the combination of a word-count-based classifier with an n-gram-based classifier achieves no improvement.

The Word Count Discrimination Rule. In [2] a discrimination rule is used: articles with more (less) than 2000 words are classified as featured (non-featured), yielding an accuracy of 0.96 for an unbalanced corpus (ratio 1:6, featured : non-featured). Figure 1 shows the probability densities over word count for our balanced corpora, and also here the 2000 word threshold is close to the optimum discrimination rule. However, we achieve via length discrimination an accuracy of 0.79 within Biology and 0.89 within History.

Table 3: Identification performance for featured articles across domains, broken down with respect to article lengths (F ∼ F-measure). Classification technology are SVMs. ⊥ indicates Precision=Recall=0.

Representation	Identification of featured articles (F)		
	< 1500 words	1500–2500 words	> 2500 words
Length Sensitivity.	*History → Biology*		
	1% featured articles	22% featured articles	77% featured articles
bin char trigram	1.000	0.860	0.885
word count	⊥	0.677	0.852
Length Sensitivity.	*Biology → History*		
	3% featured articles	8% featured articles	89% featured articles
bin char trigram	⊥	0.316	0.888
word count	⊥	⊥	0.905

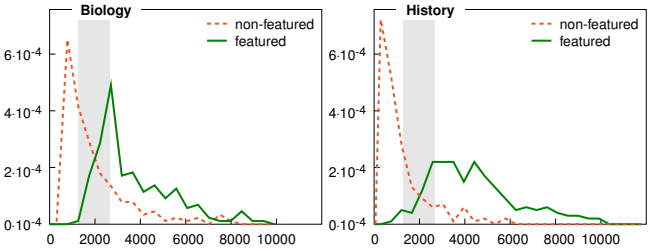

Figure 1: Probability density over absolute word count.

But the binarized character trigram vector representation combined with an SVM yields an accuracy of 0.96 within Biology and 0.92 within History.

4. CONCLUSION

This paper deals with IQ assessment of Wikipedia articles. We present the character trigram feature, originally applied for writing style analysis [7], which has not been considered for IQ assessment. We study existing research and new solutions that combine different text representations and learning algorithms. Altogether, the combination of a linear SVM with a binarized character trigram vector representation has convincing properties: it yields a high identification performance of featured articles—even across domains, it works with plaintext, and it is computationally efficient.

5. REFERENCES

[1] B. T. Adler and L. de Alfaro. A content-driven reputation system for the Wikipedia. In *Proc. of WWW'07*, 2007.
[2] J. E. Blumenstock. Size matters: word count as a measure of quality on Wikipedia. In *Proc. of WWW'08*, 2008.
[3] U. Brandes, P. Kenis, J. Lerner, and D. van Raaij. Network analysis of collaboration structure in Wikipedia. In *Proc. of WWW'09*, 2009.
[4] M. Hu, E.-P. Lim, A. Sun, H. W. Lauw, and B.-Q. Vuong. Measuring article quality in Wikipedia: models and evaluation. In *Proc. of CIKM'07*, 2007.
[5] D. L. McGuinness, H. Zeng, P. P. da Silva, L. Ding, D. Narayanan, and M. Bhaowal. Investigations into trust for collaborative information repositories: A Wikipedia case study. In *Proc. of MTW'06*, 2006.
[6] F. Sebastiani. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1), 2002.
[7] E. Stamatatos. A survey of modern authorship attribution methods. *JASIST*, 60, 2009.
[8] K. Stein and C. Hess. Does it matter who contributes: a study on featured articles in the german Wikipedia. In *Proc. of HT'07*, 2007.
[9] B. Stvilia, M. B. Twidale, L. C. Smith, and L. Gasser. Assessing information quality of a community-based encyclopedia. In *Proc. of ICIQ'05*, 2005.
[10] H. Zeng, M. A. Alhossaini, L. Ding, R. Fikes, and D. L. McGuinness. Computing trust from revision history. In *Proc. of PST'06*, 2006.