Featured Article Identification in Wikipedia

- Thesis Defense -

Christian Fricke

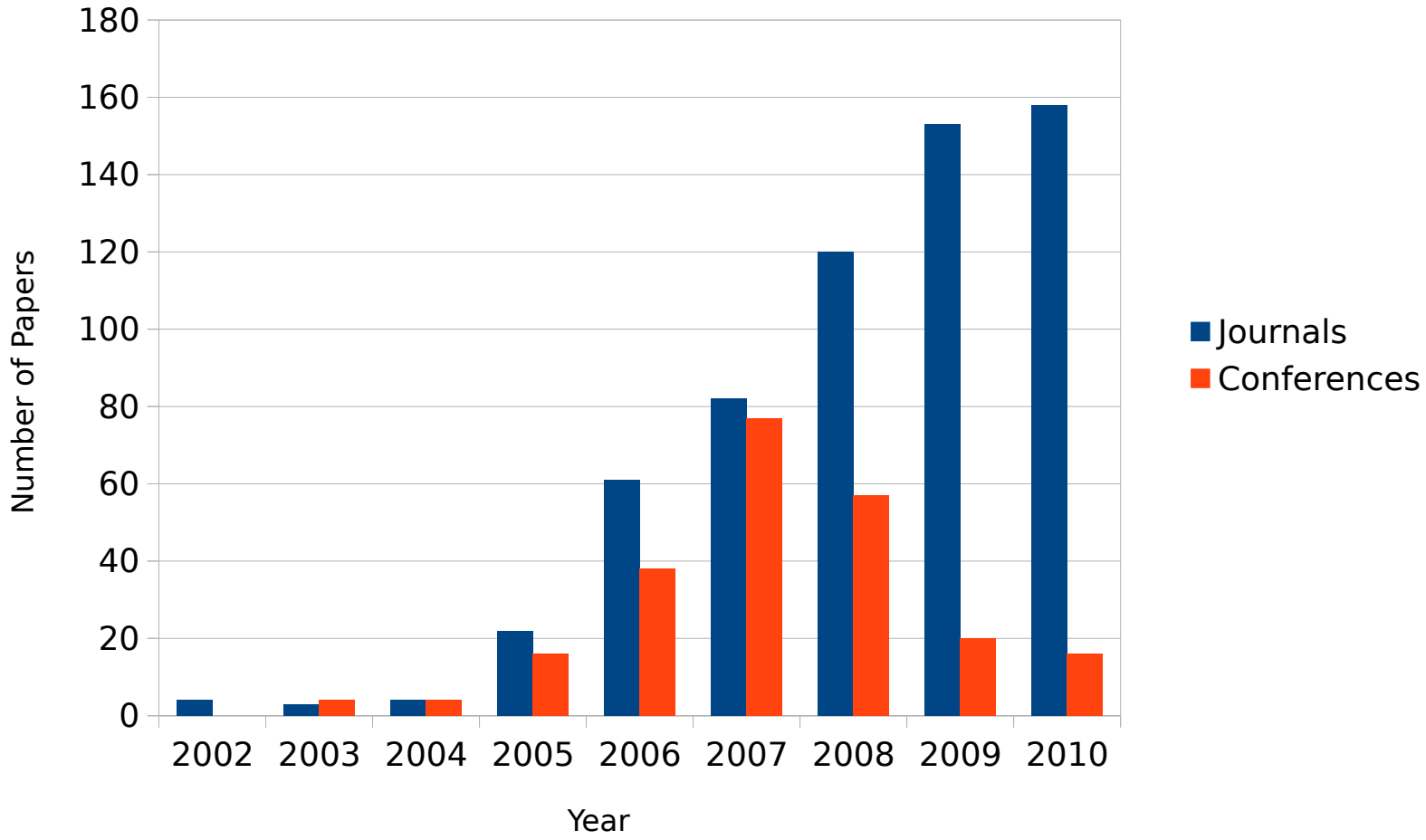
christian.fricke@uni-weimar.de

Faculty of Media / Media Systems
Bauhaus-Universität Weimar, Germany

October 18, 2012

Why is Wikipedia relevant?

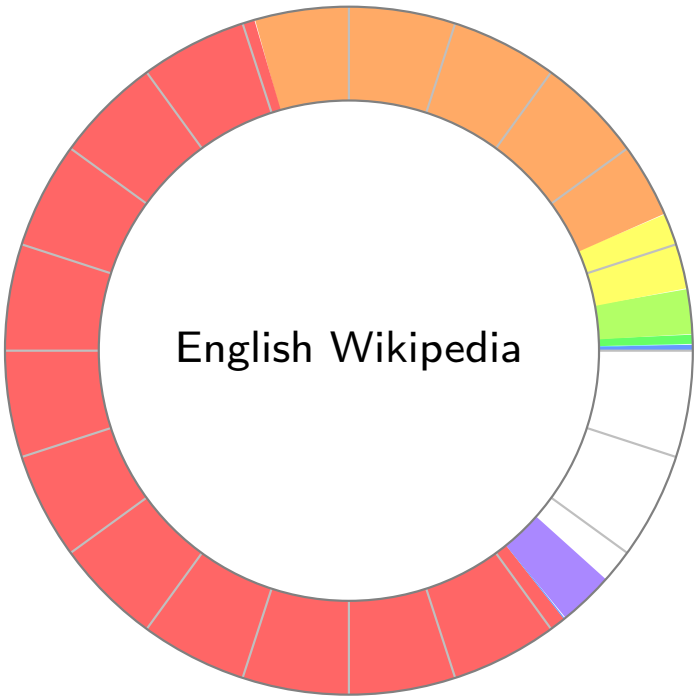
Millions of people use Wikipedia, including authors, readers, researchers, and data analysts.



Source: `http://en.wikipedia.org/w/index.php?title=File:Growth_of_Academic_Interest_in_Wikipedia.svg`

Wikipedia Statistics

The quality assessment of articles is manually unmanageable for the ever-growing encyclopedia.



FA-Class:	4 213	(0.01%)
A-Class:	991	(0.03%)
GA-Class:	16 508	(0.43%)
B-Class:	82 787	(2.15%)
C-Class:	129 483	(3.36%)
Start-Class:	881 813	(22.9%)
Stub-Class:	2 169 051	(56.4%)
FL-Class:	1 781	(0.01%)
List-Class:	100 812	(2.62%)
Unassessed:	461 818	(12.0%)
Total:	3 849 257	(100%)

Automated Solution

- Quality judgement of articles as indicator for improvement
- Most common method: binary classification of *featured* and *non-featured* articles represented as vectors of feature values

Featured:	FA-Class
Non-featured:	all other articles

Outline

Binary Classification Approaches

(1) Blumenstock [WWW 2008]
(2) Dalip et al. [JDIQ 2011]
(3) Lipka and Stein [WWW 2010]
(4) Stvilia et al. [IQ 2005]

Problem: Extenuation of results through customized data sets

(1) Blumenstock [WWW 2008]

Features	A single metric, the length (word count) of an article as its sole representation
Dataset	Unbalanced, random
	Featured: 1 554
	Non-featured: 9 513
Classifier	Multi-Layer Perceptron

(2) Dalip et al. [JDIQ 2011]

Features	54 features ranging from simple counts to complex graph-based metrics
Dataset	Unbalanced, random
	Featured: 549
	Non-featured: 2 745
Classifier	Support Vector Machine

(3) Lipka and Stein [WWW 2010]

Features	Character trigram vector—mapping from substrings of three tokens to their respective frequencies
Dataset	Balanced, domain-specific
	Featured: 380
	Non-featured: 380
Classifier	Support Vector Machine

(4) Stvilia et al. [IQ 2005]

Features Seven distinct metrics based on variable groupings that contain 19 features

Dataset Unbalanced, random

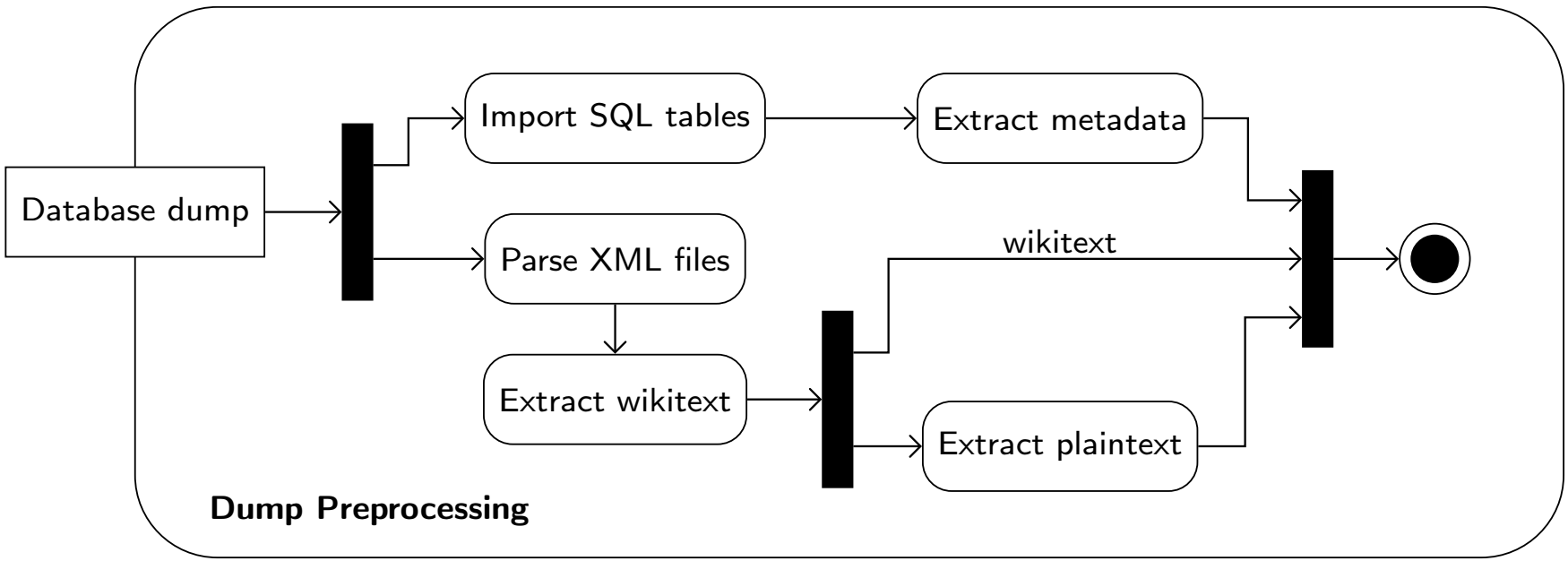
Featured: 236
Non-featured: 834

Classifier C4.5 Decision Tree

Outline

Data Preparation

The January 2012 snapshot of the English Wikipedia constitutes 8TB of text data and is processed in less than two hours using the optimized Webis Hadoop cluster.



Feature Categories

Features are organized in four categories:

Content	Length and part of speech rates, readability indices, trigrams ...
Structure	Lead rate, section distribution, counts for categories, files, images, lists, tables, and templates ...
Network	Link counts and PageRank ...
History	Age, currency, counts for edits, editors, and reverts ...

Feature Computation

The runtime for the computation of each feature for all articles depends on its source and complexity.

Category	**Features**	**Runtime**	**Source**
Content	35	< 1h	plaintext
Structure	23	< 1h	wikitext
Network	8	< 12h	metadata
History	9	< 12h	all
Total:	75	~ 1d	

Experiment Reconstruction

- Implemented most features to accurately replicate results in an easy to use framework incorporating data extraction, feature computation, dataset construction, and model definitions
- Employed WEKA to train and evaluate the classifiers
- Biased dataset selections made exact reproduction difficult

Outline

Evaluation Measures

Precision (Proportion of correctly identified negatives)

Recall (Proportion of correctly identified positives)

F-Measure (Harmonic mean of *Precision* and *Recall*)

Reconstruction Results

Model	Featured	Non-featured	Average
	Precision/ *Recall*/ *F*-Measure	*Precision*/ *Recall*/ *F*-Measure	*F*-Measure
(1)	0.871 / 0.936 / 0.902	0.989 / 0.977 / 0.983	0.970
	0.781 / 0.877 / 0.826	0.980 / 0.960 / 0.970	0.949
(2)	$\perp$	$\perp$	$\perp$
	0.903 / 0.900 / 0.901	0.980 / 0.981 / 0.980	0.967
(3)	0.966 / 0.961 / 0.964	$\perp$	$\perp$
	0.949 / 0.939 / 0.944	0.940 / 0.950 / 0.945	0.944
(4)	0.900 / 0.920 / 0.910	0.980 / 0.970 / 0.975	0.957
	0.859 / 0.907 / 0.882	0.973 / 0.958 / 0.965	0.947

(1) Blumenstock (2) Dalip et al. (3) Lipka and Stein (4) Stvilia et al.

Reconstruction Results

Model	Featured	Non-featured	Average
	Precision/ *Recall*/ *F*-Measure	*Precision*/ *Recall*/ *F*-Measure	*F*-Measure
(1)	0.871 / 0.936 / 0.902 0.781 / 0.877 / 0.826	0.989 / 0.977 / 0.983 0.980 / 0.960 / 0.970	0.970 0.949
(2)	$\perp$ 0.903 / 0.900 / 0.901	$\perp$ 0.980 / 0.981 / 0.980	$\perp$ 0.967
(3)	0.966 / 0.961 / 0.964 0.949 / 0.939 / 0.944	$\perp$ 0.940 / 0.950 / 0.945	$\perp$ 0.944
(4)	0.900 / 0.920 / 0.910 0.859 / 0.907 / 0.882	0.980 / 0.970 / 0.975 0.973 / 0.958 / 0.965	0.957 0.947

(1) Blumenstock (2) Dalip et al. (3) Lipka and Stein (4) Stvilia et al.

Reconstruction Results

Model	Featured	Non-featured	Average
	Precision/ *Recall*/ *F*-Measure	*Precision*/ *Recall*/ *F*-Measure	*F*-Measure
(1)	0.871 / 0.936 / 0.902 0.781 / 0.877 / 0.826	0.989 / 0.977 / 0.983 0.980 / 0.960 / 0.970	0.970 0.949
(2)	$\perp$ 0.903 / 0.900 / 0.901	$\perp$ 0.980 / 0.981 / 0.980	$\perp$ 0.967
(3)	0.966 / 0.961 / 0.964 0.949 / 0.939 / 0.944	$\perp$ 0.940 / 0.950 / 0.945	$\perp$ 0.944
(4)	0.900 / 0.920 / 0.910 0.859 / 0.907 / 0.882	0.980 / 0.970 / 0.975 0.973 / 0.958 / 0.965	0.957 0.947

(1) Blumenstock (2) Dalip et al. (3) Lipka and Stein (4) Stvilia et al.

Reconstruction Results

Model	Featured	Non-featured	Average
	Precision/ *Recall*/ *F*-Measure	*Precision*/ *Recall*/ *F*-Measure	*F*-Measure
(1)	0.871 / 0.936 / 0.902 0.781 / 0.877 / 0.826	0.989 / 0.977 / 0.983 0.980 / 0.960 / 0.970	0.970 0.949
(2)	⊥ 0.903 / 0.900 / 0.901	⊥ 0.980 / 0.981 / 0.980	⊥ 0.967
(3)	0.966 / 0.961 / 0.964 0.949 / 0.939 / 0.944	⊥ 0.940 / 0.950 / 0.945	⊥ 0.944
(4)	0.900 / 0.920 / 0.910 0.859 / 0.907 / 0.882	0.980 / 0.970 / 0.975 0.973 / 0.958 / 0.965	0.957 0.947

(1) Blumenstock (2) Dalip et al. (3) Lipka and Stein (4) Stvilia et al.

Uniform Dataset

We define four datasets to fairly compare the performance of each proposed model and propose an additional model that combines every implemented feature.

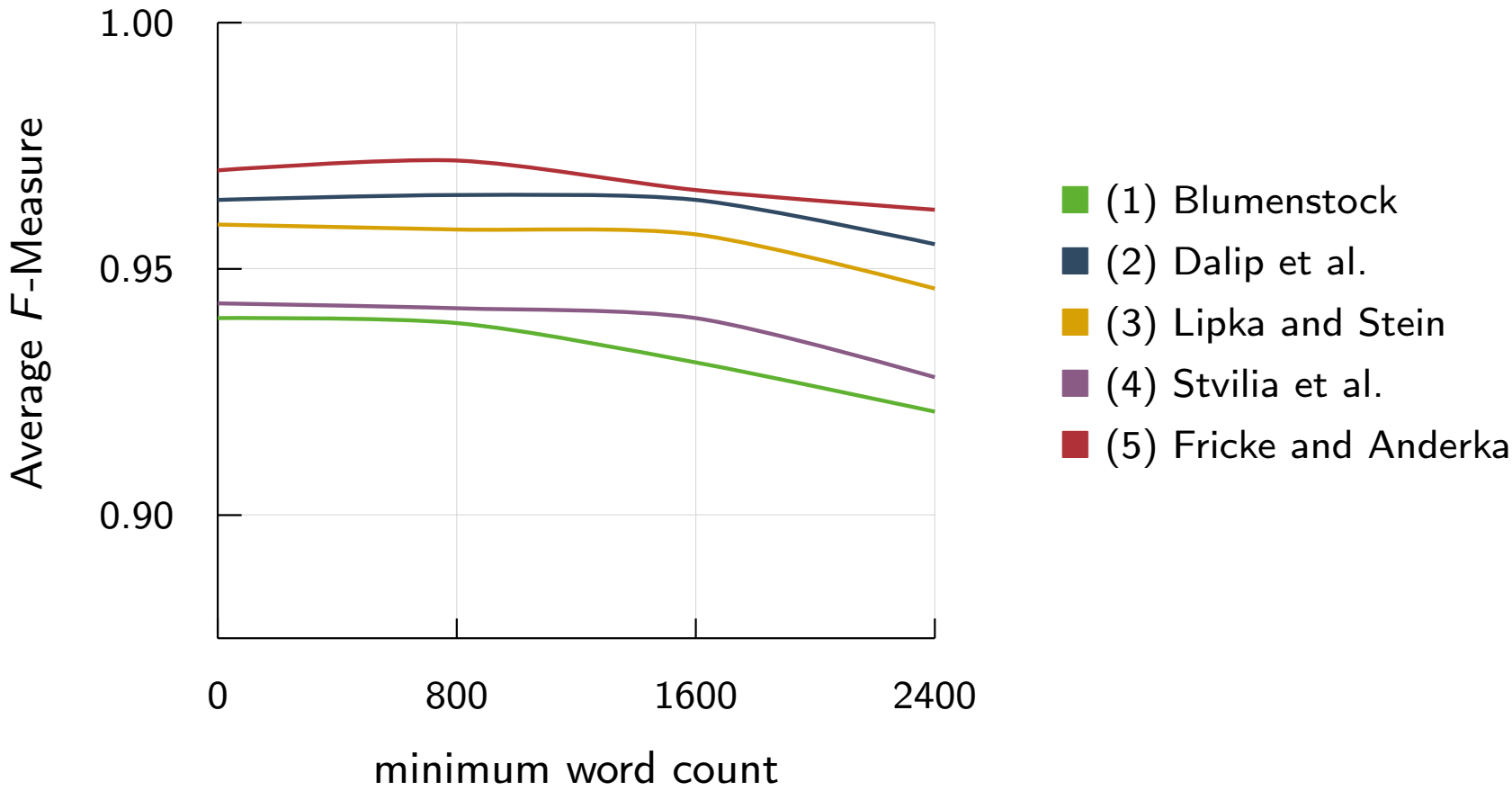
Dataset Balanced, random, corresponding to minimum word counts of 0, 800, 1600, and 2400

Featured:	3 000
Non-featured:	3 000

(5) Fricke and Anderka:

Features All 75 features from every category
Classifier Support Vector Machine

Uniform Evaluation



Conclusion and Outlook

- A framework for convenient and consistent evaluation
- A new model utilizing every implemented quality indicator
- The most comprehensive collection of article features to date

Conclusion and Outlook

- A framework for convenient and consistent evaluation
- A new model utilizing every implemented quality indicator
- The most comprehensive collection of article features to date

- Exploration of novel quality indicators
- Combination with flaw detection algorithms
- Application to other classes (e.g. Start)