Automatische Erkennung von Vandalismus in Wikipedia mit Hilfe maschineller Lernverfahren

Robert Gerling
robert.gerling(at)medien.uni-weimar.de

Bauhaus Universität Weimar
Web Technology & Information Systems

6. Mai 2008

Outline

Wikipedia

Wiki [Cunningham, 1994]

- strukturierte Ansammlung von Webseiten (Wikiseiten)
- einfaches Anlegen und Bearbeiten

Wikipedia [Wales, 2001]

- freie Online-Enzyklopädie auf Basis der Wiki-Software
- kollaboratives Erstellen und Pflegen von Artikeln
- nahezu alle Artikel sind editierbar (auch anonym)
- ca. 9.719.049 Artikel in 256 Sprachen[1]

[1] Stand: 10.02.2008

Vandalismus in Wikipedia

„Vandalism is any addition, removal, or change of content made in a deliberate attempt to compromise the integrity of Wikipedia.“

[Wikipedia Vandalism Policy, 2008]

Beispiele:

- http://en.wikipedia.org/w/index.php?diff=136908072&oldid=136903030
- http://en.wikipedia.org/w/index.php?diff=136771311&oldid=136768815
- http://en.wikipedia.org/w/index.php?diff=134880408&oldid=134710659
- http://en.wikipedia.org/w/index.php?diff=136685101&oldid=136665646

Statistik:

- ca. 4-5 % Vandalismus [Wikipedia Studie: Study1, 2006] , [Priedhorsky, 2007]
- ca. 280.000 Edits pro Tag in der englischen Wikipedia

Motivation:

- nur wenig automatisierte Unterstützung
- maschinelles Lernen zur automatischen Klassifikation kann helfen (analog Email-SPAM Filterung)

Wikipedia-Edits

Korpus von Wikipedia-Edits:

- 940 Edits: 639 konstruktive Edits + 301 Vandalismus-Edits.
- manuell überprüft, gekennzeichnet und strukturiert in XML gespeichert

Typologie von Vandalismus:

Form der Manipulation	**veränderter Inhalt**			
	Text	Link	Media	Struktur
Einfügen	43,9 % Charakteristik: Anstößig, themenfremd, Nonsens, bewertend, Duplikate, Kauderwelsch	6,9 %	0,7 %	14,6 % Charakteristik: Formatierung, Hervorhebung
Ersetzen	45,8 %	4,7 %	2 %	15,5 %
Löschen	31,6 %	22,9 %	19,4 %	20,3 %

Vandalismuserkennung als Klassifikationsproblem I

Reale Situation:

- Menge von Edits E
- Menge von Klassen $C = \{\textit{Vandalismus}, \textit{konstruktiver Edit}\}$
- Zuordnung $\gamma : E \to C$, tatsächliche Klassenzugehörigkeit der Edits
- Klassifizieren: für ein gegebenen Edit $e \in E$ ermitteln der Klasse $\gamma(e) \in C$

Vandalismuserkennung als Klassifikationsproblem I

Reale Situation:

- Menge von Edits E
- Menge von Klassen $C = \{\textit{Vandalismus}, \textit{konstruktiver Edit}\}$
- Zuordnung $\gamma : E \to C$, tatsächliche Klassenzugehörigkeit der Edits
- Klassifizieren: für ein gegebenen Edit $e \in E$ ermitteln der Klasse $\gamma(e) \in C$

Problem: $\gamma : E \to C$ unbekannt

Vandalismuserkennung als Klassifikationsproblem I

Reale Situation:

- Menge von Edits E
- Menge von Klassen $C = \{\textit{Vandalismus}, \textit{konstruktiver Edit}\}$
- Zuordnung $\gamma : E \rightarrow C$, tatsächliche Klassenzugehörigkeit der Edits
- Klassifizieren: für ein gegebenen Edit $e \in E$ ermitteln der Klasse $\gamma(e) \in C$

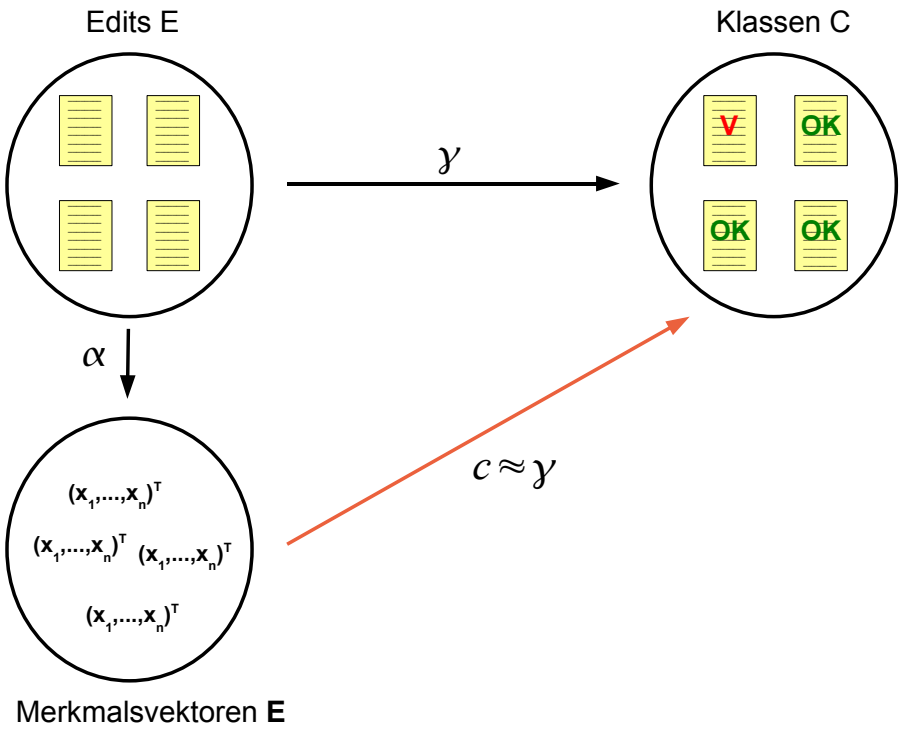
Problem: $\gamma : E \rightarrow C$ unbekannt

Modellbildung und Automatisierung

- Abstraktion der Edits $e \in E$ zu Merkmalsvektoren $\mathbf{e} = \alpha(e)$
- Erstellen einer Menge von Beispielen $(\alpha(e), \gamma(e))$
- Approximation einer Funktion $c(\mathbf{e}) \approx \gamma(e)$ mit $\mathbf{e} = \alpha(e)$

[Stein, 2007]

Vandalismuserkennung als Klassifikationsproblem II



Gegeben:

- γ : reale Klassenzugehörigkeit der Edits
- α : Modellbildungsfunktion

Gesucht:

- c : Approximation für γ

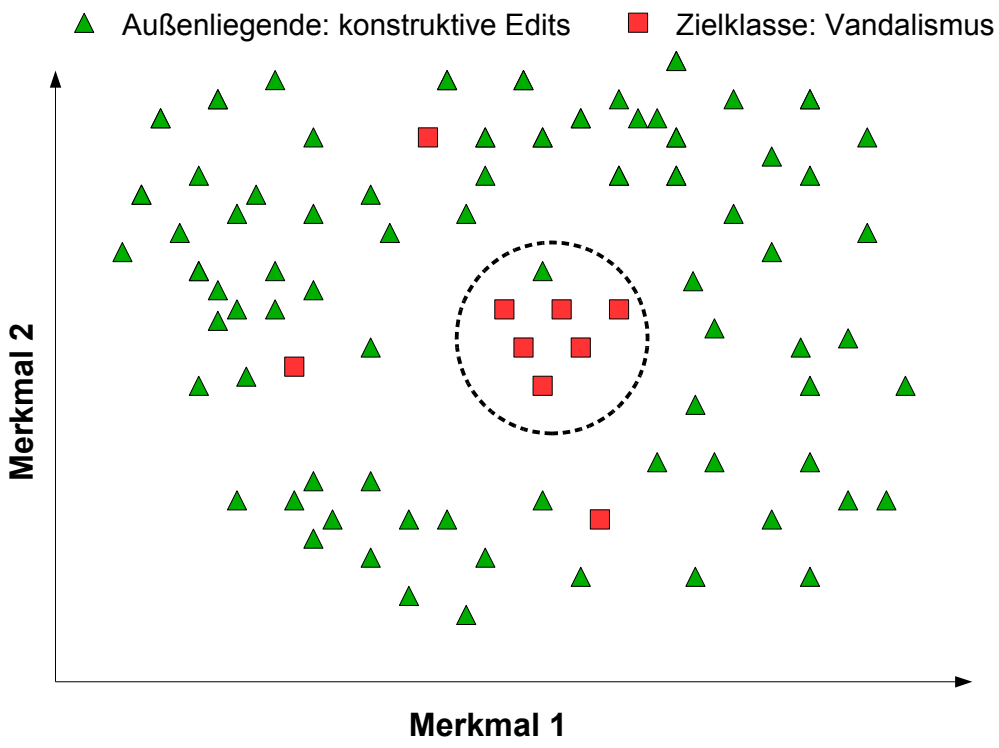
Lösung:

- verschiedene Verfahren der Statistik und des maschinellen Lernens

[Vgl. Stein, 2007]

Vandalismuserkennung als Klassifikationsproblem III

Vandalismuserkennung als One-Class-Klassifikationproblem:



Zielklasse:

- steht im Fokus des Interesses
- möglichst genaue Beschreibung durch quantifizierbare Merkmale

Außenliegenden:

- Objekte die nicht der Zielklasse angehören

[Vgl. Tax, 2001]

Klasse der konstruktiven Edits ist überrepräsentiert und schwer formal zu beschreiben

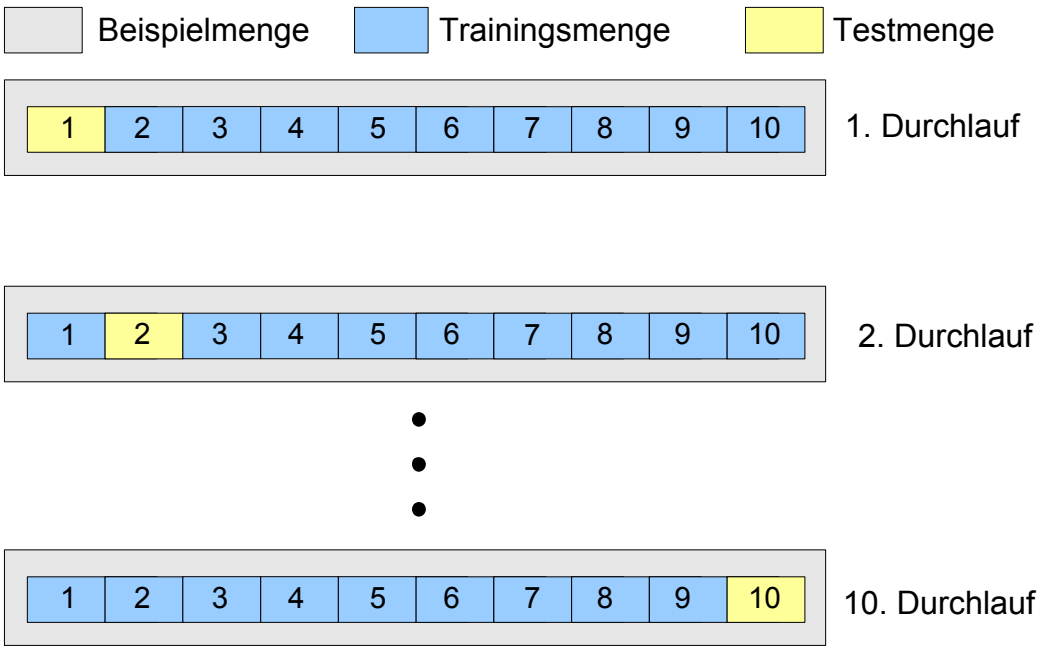
Feature I

Feature f	Kurzbeschreibung
Zeichenebene	
Char-Distribution	Abweichung der relativen Zeichenhäufigkeiten des Edits vom sprachcharakteristischen Wert
Char-Sequence	Längste zusammenhängende Wiederholung eines Zeichens im Edit
Compressibility	Kompressionsrate des Edittextes
Upper-Case-Ratio	Verhältnis von Großbuchstaben zu allen Buchstaben im Edit
Termebene	
Term-Impact	Durchschnitt der relativen Häufigkeiten der Terme des Edits in der neuen Artikelrevision
Longest-Word	Länge des längsten Wortes im Edit
Pronoun-Frequency	Verhältnis von Pronomen, der ersten und zweiten Person, zu allen Termen des Edits
Pronoun-Impact	Anteil der Edit Pronomen an der Anzahl der Pronomen der neuen Artikelrevision
Vulgarism-Frequency	Verhältnis von vulgären Worten zu allen Worten des Edits
Vulgarism-Impact	Anteil der vulgären Worte des Edits an der Anzahl der vulgären Worte der neuen Artikelrevision
Artikelebene	
Size-Ratio	Verhältnis von alter zu neuer Revision, bezogen auf den Umfang
Replacement-Similarity	Ähnlichkeit von gelöschtem und dafür eingefügtem Text
Context-Relation	Ähnlichkeit von der alten Artikelrevision und Artikeln Wikipedias zu extrahierten Schlüsselwörtern des Edittextes
Metaebene	
Anonymity	Gibt an, ob ein Edit anonym gemacht wurde
Comment-Length	Umfang des Kommentars zu einem Edit
Edits-Per-User	Anzahl früherer Edits durch den selben Benutzer oder IP

Methodik der Experimente

Problem: Trainings- und Testmenge erforderlich, aber nur eine Beispielmenge vorhanden

Lösung: Ten-Fold-Cross-Validation



- Beispielmenge wird in 10 Teilmengen aufgeteilt
- 10 Durchläufe von Training und Test, wobei jede Teilmenge genau einmal als Testmenge dient
- Teilergebnisse der einzelnen Durchläufe werden zu aussagekräftigem Gesamtergebnis zusammengefasst

[Kohavi, 1995]

Vergleichsmaße

Wahrheitsmatrix

klassifiziert als →	konstruktiv	Vandalismus
konstruktiv	$n_{k \to k}$	$n_{k \to v}$
Vandalismus	$n_{v \to k}$	$n_{v \to v}$

Precision:

- Verhältnis von korrekt klassifizierten Vandalismus-Edits zu allen als Vandalismus eingestuften Edits.

$$\frac{n_{v \to v}}{n_{k \to v} + n_{v \to v}} \tag{1}$$

Recall:

- Verhältnis von korrekt klassifizierten Vandalismus-Edits zu allen Vandalismus-Edits innerhalb der Testmenge.

$$\frac{n_{v \to v}}{n_{v \to k} + n_{v \to v}} \tag{2}$$

[Van Rijsbergen, 1979]

Ergebnisse I

Vergleich mit regelbasierten Bots *AntiVandalBot (AVB)* [Wiki-User: Twaker, 2006] und *ClueBot* [Wiki-User: Cobi, 2007]

Manipulationsform	**AVB**		**ClueBot**		**Lernalgorithmus**	
	Precision	Recall	Precision	Recall	Precision	Recall
Einfügen	0,67	0,35	1,0	0,03	0,82	0,87
Ersetzen	0,69	0,53	1,0	0,29	0,86	0,76
Löschen	0,85	0,61	1,0	0,49	0,90	0,89
Zusammen	0,71	0,43	1,0	0,16	0,83	0,77

- ClueBot erreicht Precision von 1,0, jedoch sehr geringen Recall
- Lernalgorithmus beiden Bots überlegen, besonders beim wichtigeren Recall

Ergebnisse II

- Sprachabhängigkeit einiger Feature erschweren die Portierung des Ansatzes auf nicht-englischsprachige Wikipedia-Ableger
- Wie gut ist die Klassifikationsgüte ohne die sprachabhängigen Feature?

Manipulationsform	Set $\mathcal{F}_{alle}$		Set $\mathcal{F}_{su}$	
	Precision	Recall	Precision	Recall
Zusammen	0,83	0,77	0,76	0,75

- mit geringem Verlust kann auf sprachabhängige Feature verzichtet werden

Ausblick & Kritik

Ausblick:

- Wikipedia-Vandalismus:
  - speziell abgestimmte Feature-Sets für die Manipulationsformen
  - evaluieren von Maßnahmen gegen nicht-textuellen Vandalismus wie z.B. Link-SPAM
  - sicherstellen der Skalierbarkeit für den praktischen Einsatz
- Missbrauch Sozialer Software:
  - Anwendung des Ansatzes auf ähnlich gelagerten Probleme wie z.B. Flaming oder Trolling in Foren

Kritik:

- Ansatz basiert auf Vergleich zweier Revisionen $\longrightarrow$ Annahme, dass alte Revision frei von Vandalismus war
- Class-Imbalance-Problem

Diskussion und Fragen

Vielen Dank für Ihre Aufmerksamkeit!
Fragen?