Klassifikationsaufgaben mit der SENTRAX. Konkreter Fall: Automatische Detektion von SPAM

Dirk T. Frobese

Universität Hildesheim
Institut für Mathematik und Angewandte Informatik
Marienburger Platz 22
31141 Hildesheim
dfrobese@frobese.de

Zusammenfassung

Die Suchfunktionen des SENTRAX-Verfahrens werden für die Klassifizierung von Mails und im Besonderen für die Detektion von SPAM eingesetzt. Die Eigenschaften einer kontextähnlichen Suche und die Fehlertoleranz sollen genutzt werden, um SPAM Nachrichten treffsicher aufzuspüren.

Abstract:

This article introduces the SENTRAX-engine for classification of E-Mails and detection of SPAM.

1 Einleitung

Das Internet und seine Dienste sind mittlerweile fester Bestandteil der Gesellschaft geworden. Ein technisches System, entstanden als Netz zur Unterstützung wissenschaftlicher Arbeiten auf weltweiter Basis, ist auch aus dem privaten und beruflichen Leben nicht mehr wegzudenken. Die Kernfunktionalität ist immer noch die Bereitstellung von Informationen und der Informationsaustausch weltweit, auf Basis von Standards, inhaltlich lediglich einer Selbstkontrolle unterworfen. Als eine der wichtigsten Dienste ist die elektronische Nachricht, die E-Mail oder Mail, anzusehen. Sie hat den traditionellen Briefverkehr dezimiert, und gäbe es nicht andere Angebote aus dem Internet wie zum Beispiel Versandhäuser oder Auktionshäuser (ebay), dann würde die Post kaum noch für den Privatkunden in Erscheinung treten. Laut einer Studie von ARD Online nutzen 78 Prozent der Anwender das Internet hauptsächlich zum Senden und Empfangen von E-Mails. Dies führt aber auch zu anderen, unangenehmen Begleiterscheinungen. Da es so einfach ist E-Mails zu schreiben, bekommt man auch sehr viele davon. Das führt dazu, daß man als aktiver, am Berufsleben aktiv beteiligter Mensch am Tag durchaus an die hundert E-Mails bekommen kann und mehr. Es gibt mittlerweile Ergebnisse aus Untersuchungen, daß der elektronische Mailverkehr nicht mehr ohne Hilfsmittel wie Suchmaschinen und Filter zu bewältigen ist. Hervorzuheben sind die Werbe- oder Müllbotschaften, so genannte SPAM oder Junk-Mails, die häufig auch Computer-Viren mit sich führen. Beispielsweise die Norddeutsche Landes-

bank in Hannover (NORD/LB) erhält bis zu 800 SPAM- Mails pro Stunde [Artikel:NORD/LB 1]. Aber es sind nicht nur SPAMs, sondern auch die Vielzahl der ernsthaften Nach- richten benötigt Zeit zur Bewältigung. Besonders Unternehmen und Dienstleistungsfirmen werden von ihren Kunden verstärktüber den Informationskanal Mail kontaktiert. Eine Veranstaltung wie die Tour de France führt dazu, daß der Berichterstatter und Fernsehsender ARD am Tag ca. 2.000 Mails erhält. Ein weiters Beispiel ist die Bahn [Artikel:xtramind]. Sie erhält pro Jahr ca. 2 Millionen Mail-Anfragen von ihren Kunden. Anbieter von Mail- und Callcenter-Lösungen gehen davon aus, dass eine E-Mail in der Vollkostenrechnung zwischen 3,- bis 5,- Euro je nach Beantwortung und Un- ternehmensstruktur verursachen [Website:www.vera.ag]. Ausgehend von diesen Zahlen ergibt sich ein Kostenpotential von 6 bis 10 Millionen Euro pro Jahr. Daher werden die Verfahren zum Kategorisieren bzw. Klassifizieren der E-Mails immer wichtiger. Nun bietet das von Prof. Bentz an der Universität Hildesheim entwickelte Such- verfahren einen alternativen Ansatz. Es soll nun vorgestellt werden, in wie weit sich das SENTRAX-Verfahren für ein E-Mail Management eignet als die derzeit bekannten und ob sich ein verbesserter Nutzen für die geschilderte Problematik ergibt.

2 Aufgabenstellung

Aus den Ergebnissen des Fachbereich III Informations- und Kommunikationswissenschaften der Universität Hildesheim unter der Leitung von Prof. Bentz ist eine Implementierung assoziativer Suchverfahren (SENTRAX) enstanden. Dieses System wird zur Klassifizierung von Mails eingesetzt. Das SENTRAX-Verfahren kam zur Textklassifikation bisher noch nicht zum Einsatz. Die genaue Funktionsweise wird in Bentz [2006] dargestellt. Die Aufgaben des E-Mail Management fokussieren sich auf folgende Aufgabengebiete:

1. Die Kategorisierung von Nachrichten und damit die kontextabhängige Zuleitung von Mails in vorgegebene Kategorien mit SENTRAX
2. Klassifizierung und Erkennung von Werbenachrichten (SPAM)

Weitergehend soll eine maschinelle Bestimmung möglicher Kategorien betrachtet werden. Die bisherigen Aufgabenstellungen beschränkten sich auf Klassen, die durch repräsentative Mails oder Texte durch den Benutzer in ausreichender Anzahl festgelegt wurden. Nun soll das Verfahren selbst aus einem unstrukturierten Bestand die möglichen Kategorien bestimmen.

3 Vorgehen

Die Durchführung der Untersuchung zur Kategorisierung von exemplarischen Mail-Beständen erfolgt in folgenden Schritten:

1. *Export:* Die E-Mail Bestände werden aus den Client-Anwendungen bereitgestellt.
2. *Normalisierung:* Header und Body der E-Mails werden für den Lernvorgang auf notwendigen Umfang und Darstellung umgeformt, insbesondere die Anhänge.
3. *Lernphase:* Der Großteil des Bestandes wird für den Lernvorgang verwendet.
4. *Detektion:* Auf Basis der vorangegangenen Lernphase werden die übrigen Mails ver-

wendet, um eine Klassifizierung durchzuführen. Inhaltlich sind die E-Mails manuell kategorisiert worden. Natürlich ist es nicht auszuschliessen, daß es fehlerhafte Zuordnungen durch den Anwender gibt. Dies ist aber zu vernachlässigen.

Betrachtet man Aufbau und Inhalt der Junk-Mails, die man so täglich erhält genauer, dann erkennt man immer wiederkehrende Worte bzw. Zusammenhänge zu den Absendern. Es wird daher davon ausgegangen, daß das SENTRAX-Verfahren bei dieser Art von Nachrichten besonders gute Ergebnisse liefert.

Zur Untersuchung werden die folgenden Algorithmen des SENTRAX-Verfahrens eingesetzt:

- Mindmap(Kontext): Darstellung aller exakter Treffer sowie im Kontext verwendete Alternativen
- Lexikomap: Darstellung von Begriffen mit exakter und ähnlicher Schreibweise.
- Trefferliste: Eine Auflistung der Dateien, in der die Suchbegriffe gefunden worden sind, sortiert aufsteigend nach der Häufigkeit der Treffer in Prozent.
- Fehlertolerante Trefferliste: Mit dieser Funktion wird das Verfahren der Lexikomap und der Trefferliste kombiniert. In einer Liste aufsteigend nach Trefferhäufigkeit werden die gefundenen Begriffe mit exakter oder ähnlicher Schreibweise dargestellt.
- Ähnliche: Es wird eine Trefferliste mit Dokumenten ähnlichen Inhalts geliefert.

Alle Verfahren, mit Ausnahme der Lexicomap, liefern eine Liste der Trefferhäufigkeit in den gefundenen Dokumenten. Damit ist die Lexikomap für die Klassifizierung nicht geeignet. An Stelle eines Suchbegriffes wird eine Mail in ihrer vollen Länge als Kombination komplexer Suchbegriffe verwendet. Dabei werden, wie bei der Indizierung auch, nicht relevante Wörter durch eine Stopwortliste entfernt. Anders als bei der Indizierung werden bei der SENTRAX-Suche nicht einzelne Mails indiziert, sondern alle Nachrichten innerhalb einer Kategorie zusammengefasst und dann indiziert. Die oben aufgeführten Verfahren liefern als Ergebnis damit immer eine Liste der Kategorien in der Reihenfolge der größten Trefferhäufigkeit.

4 Ergebnisse

Mit einem exemplarischen Mailbestand wurden die oben genannten Verfahren von SENTRAX angewendet. Durch eine Reihe von Messungen ergibt sich folgende Ergebnistabelle 1.

Die linke Spalte stellt die Anzahl der Kategorien dar, die für eine Messreihe verwendet wurden. Es wurden 12, 9 und 6 Kategorien aus einem Testbestand von Mails verwendet, die sich inhaltlich nur geringüberschnitten. Die folgende Spalte zeigt die Anzahl der gelernten Mails pro Kategorie. Aus den Mails wurden jeweils 50, 100 und 150 Mails als Basis für den Index erlernt. Die Klassifizierung wurde dann mit 10, 20 ,30 und 50 Samples pro Kategorie durchgeführt.

Kat.	Mails		Trefferliste		Fehlertolerant		Kontext		Ähnliche	
	gel.	prob.	alle	Spam	alle	Spam	alle	Spam	alle	Spam
12	50	10	82%	100%	79%	100%	80%	100%	82%	90%
12	50	20	80%	95%	79%	95%	77%	95%	82%	90%
12	50	30	78%	96%	77%	96%	76%	96%	83%	96%
12	50	50	73%	84%	72%	92%	71%	94%	80%	92%
9	50	10	80%	90%	80%	90%	78%	90%	80%	90%
9	50	20	79%	85%	79%	90%	76%	85%	78%	95%
9	50	30	77%	87%	77%	87%	75%	87%	79%	93%
9	50	50	73%	88%	74%	90%	71%	88%	71%	94%
9	100	10	83%	80%	82%	70%	83%	80%	85%	90%
9	100	20	86%	80%	85%	75%	86%	80%	85%	95%
9	100	30	85%	80%	84%	73%	84%	80%	85%	96%
9	100	50	82%	82%	81%	78%	81%	82%	81%	94%
6	50	10	86%	100%	85%	100%	83%	100%	83%	80%
6	50	20	85%	95%	85%	95%	82%	95%	85%	85%
6	50	30	83%	93%	84%	97%	81%	93%	86%	83%
6	50	50	84%	94%	83%	92%	82%	94%	85%	80%
6	100	10	91%	90%	93%	90%	90%	90%	90%	90%
6	100	20	90%	90%	91%	85%	89%	80%	90%	80%
6	100	30	90%	93%	91%	90%	89%	93%	91%	85%
6	100	50	89%	94%	90%	92%	88%	94%	91%	88%
6	150	10	80%	100%	85%	100%	80%	100%	86%	90%
6	150	20	85%	100%	89%	100%	85%	100%	88%	95%
6	150	30	87%	100%	88%	100%	87%	100%	88%	98%
6	150	50	88%	100%	89%	100%	88%	100%	88%	94%

Tabelle 1: Ergebnisse Kategorisierung Mails privat/geschäftlich

Die Ergebnisse der Klassifikation werden in Prozent dargestellt. Das Ergebnis über alle Kategorien wird in der Spalte "alle" dargestellt. Die Quote bzgl. der Detektion von SPAM in der Zeile daneben. Die aufbereiteten Ergebnisse lassen folgende Rückschlüsse zu:

- Optimale Ergebnisse liefern Kategorien, die eine geringe inhaltliche Überschneidung haben
- Eine Verringerung der Kategorien wirkt sich günstig auf das Ergebnis aus
- Schon ab 70 indizierten Mails wird eine Trefferquote von 80% erreicht
- Das SENTRAX-Verfahren liefert Trefferquoten besser als 70% in der Regel 85%
- SPAM-Mails werden besonders gut erkannt. Die Quote ist meist besser als 90%
- die SENTRAX-Funktion ”ähnliche“ liefert die besten Ergebnisse

Damit liefert SENTRAX sehr gute Ergebnisse für die Klassifikation von Mails und ist besonders für die Erkennung von SPAM geeignet.

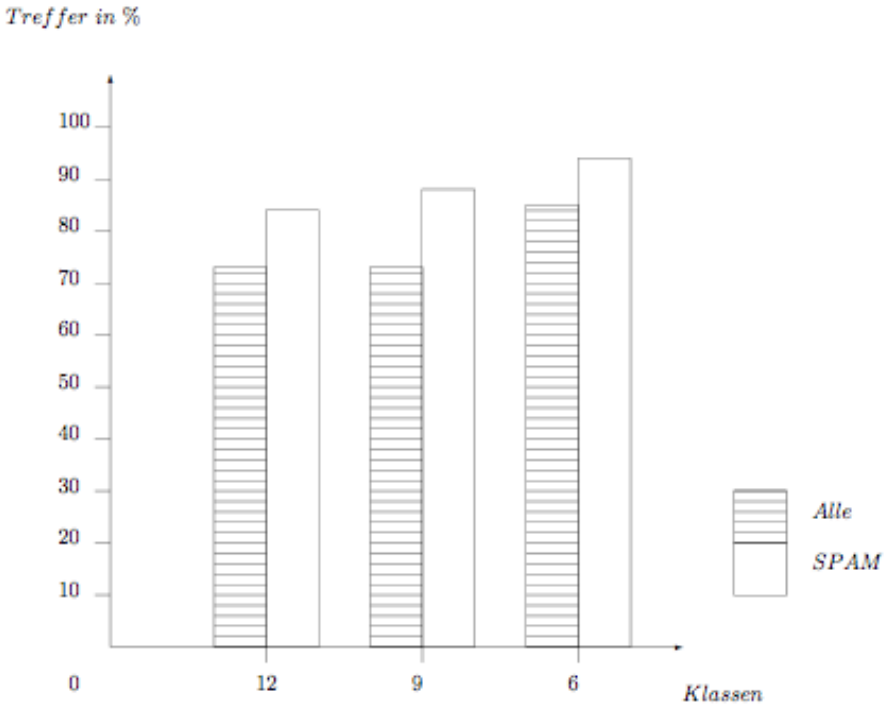

Abbildung 1: Trefferquote SPAM abhängig von der Anzahl der Kategorien

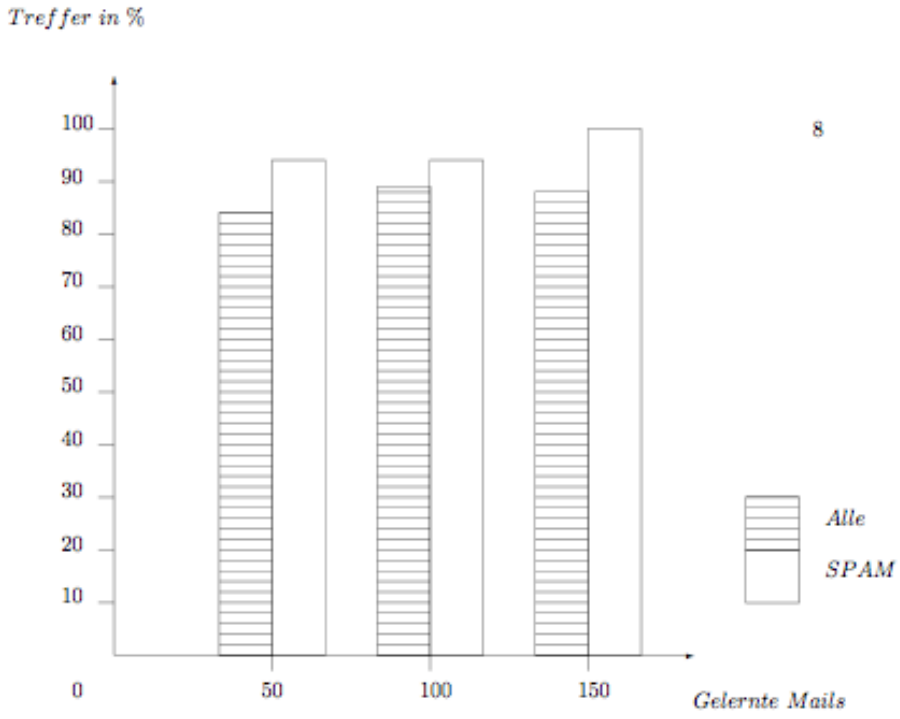

Abbildung 2: Vergleich der Trefferquoten mit SPAM

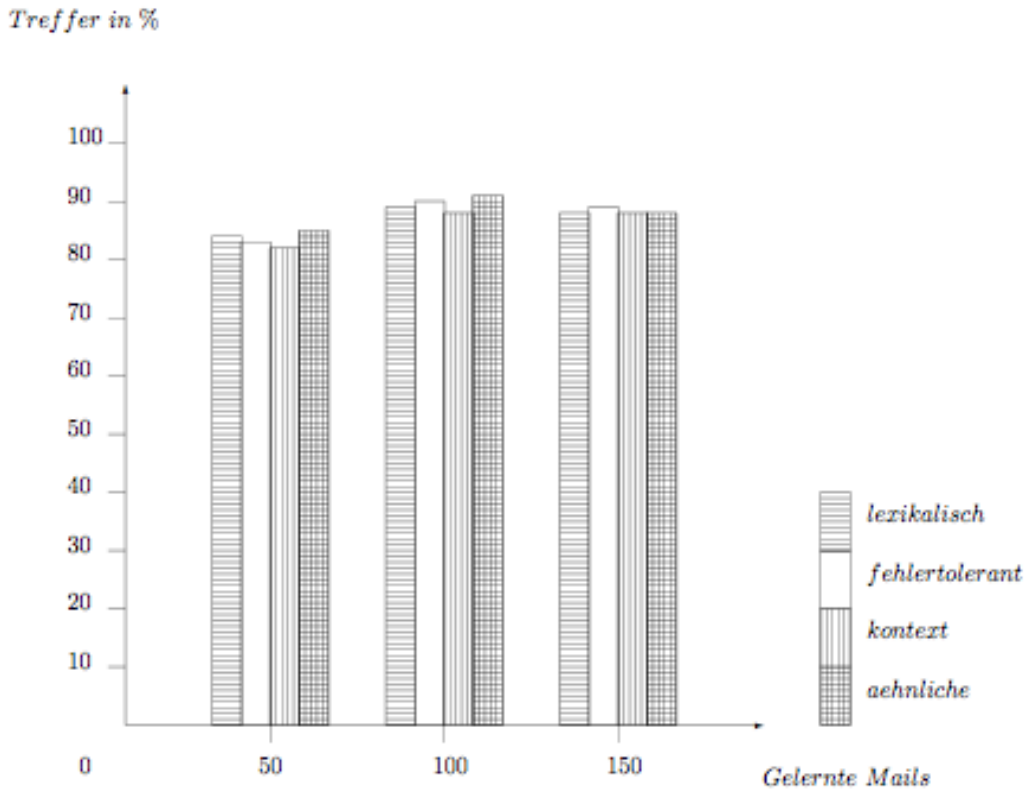

Abbildung 3: Vergleich der Trefferquoten

Literaturverzeichnis

Bentz, Hans-Joachim [2006] Die Suchmaschine SENTRAX. Grundlagen und Anwendungen dieser Neuentwicklung. In diesem Band.

Müller, Karen: Automatische Klassifikation von Textdokumenten, Universität Hildesheim, Dezember 2002

Suriya Na Nhongkai, Hans-Joachim Bentz: Bilinguale Suche mittels Konzeptnetzen, Universität Hildesheim, August 2005

[NORD/LB 1] Artikel Durchblick im Postfach in New Spirit, 1/2004 LITERATUR 11

[Website:www.vera.ag] Hompage der VERA Callcenter Lösung

[Website:www.xtramind.de] Hompage der xtramind E-Mail Management Lösung

[xtramind] xtramind: Success Story DB Dialog Telefonservice GmbH, 8/2004