



Der SCA - Klassifizierung

kippdata informationstechnologie GmbH, Juni 2008

1 Einleitung

Dieses kurze Dokument beschreibt die Besonderheiten und Alleinstellungsmerkmale von CRIA als Klassifizierungs-System für Texte. Es ist als kurze Orientierung von eher technisch ausgerichteten Personen gedacht.

2 Klassifizierung

Eine zentrale Eigenschaft von CRIA ist die inhaltliche Analyse von E-Mail und ihren Anhängen. Diese Operation wird auch als Kategorisierung oder Klassifizierung bezeichnet und ist die Grundlage für die automatische Gruppierung von Textdokumenten zu thematischen oder organisatorischen Komplexen.

Die inhaltliche Analyse von Texten geht weit über die übliche Suche nach Schlagworten, die auch als Volltextsuche bekannt ist, hinaus. Volltextsuche ist auch nicht ausreichend für die Zuordnung von Dokumenten zu Themenkomplexen: Einerseits werden gleiche thematische Konzepte von verschiedenen Autoren auf sehr unterschiedliche Weise unter Verwendung unterschiedlicher Begriffe ausgedrückt. Andererseits kann ein Wort je nach Kontext unterschiedliche Konzepte ausdrücken.¹

Die Klassifizierung von Texten wie E-Mail erfordert darüberhinaus Verfahren, die über die Anwendung der klassischen Vektorraum-Algorithmen wie kNN oder SVM hinausgehen. Diese Verfahren bilden die semantische Ähnlichkeit von Texten als Abstand von Vektoren ab und sind damit für gewisse Probleme auch gut geeignet. Speziell E-Mails aber haben einige Besonderheiten, die es zu berücksichtigen gilt:

- Eine einzelne E-Mail ist in der Regel sehr klein.
- Der Text der E-Mail eignet sich nicht gut für die Verwendung klassischer Analyseverfahren².
- Die Anhänge müssen berücksichtigt werden.
- Aber: Anhänge können auch irreführend sein.
- ...

Die (unvollständige) Aufzählung der Schwierigkeiten, die zu überwinden sind, legt einen völlig neuen Ansatz nahe, der von CRIA umgesetzt wird.

Zur Klassifizierung einer E-Mail werden unterschiedliche Verfahren herangezogen, die in sogenannten Analysis Engines (AEs) bereitgestellt werden. Jedes dieser Verfahren ist optimiert für die Auswertung spezieller Informationen, die in einer E-Mail vorhanden sind. Neben der üblichen Berücksichtigung der Adressinformationen aus dem Mail-Header und der semantischen Analyse des Inhaltes können Anrede, Grußformel und die im Geschäftsverkehr erforderlichen formalen Informationen über die Organisation des Absenders ausgewertet werden. Darüberhinaus kann auch das kommunikative Umfeld der zu klassifizierenden Mail Berücksichtigung finden.

Die angewandten Einzelverfahren nutzen je nach Teilaufgabe ganz unterschiedliche Ansätze, die von der komplexen Mustererkennung samt Abgleich mit externen Informationsquellen wie Kundendatenbank oder Corporate Directory bis zu den verschiedenen statistischen Verfahren reichen. Alle Einzelanalysen führen jeweils zu Hinweisen für die Klassifizierung.

Die Hinweise werden gesammelt und zu einer globalen Klassifizierung zusammengefaßt. Auch diese Zusammenfassung der Einzelergebnisse zu einer Gesamtanalyse ist eine Besonderheit von CRIA: Die angewandten Verfahren nutzen die wachsende Kenntnis des Systems über gute Klassifizierungen, um die unterschiedlichen Teilergebnisse zu bewerten und zu gruppieren. Auf diese Weise können typische Probleme anderer Systeme vermieden werden: Beispielsweise kann der Wert der Betreff-Information einer Mail a priori nicht festgestellt werden. Es ist vielmehr erforderlich, daß diese Information kontextabhängig berücksichtigt wird. Dies wird von CRIA geleistet.

¹ Ein Beispiel ist das Wort Bank, das sowohl ein Möbelstück als auch ein Geldinstitut bezeichnen kann.

² Diese sind für die Analyse großer Texte wie beispielsweise Zeitungsartikel optimiert.

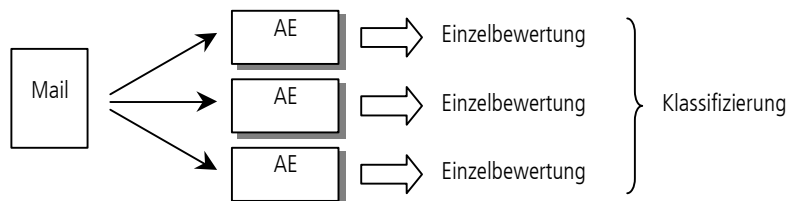


Abbildung 1: Schematische Darstellung der Klassifizierung von E-Mails im SCA

Die vorstehende Abbildung zeigt die Klassifizierung schematisch. Hinter den einzelnen Analysis Engines (AEs) verbergen sich die unterschiedlichen Verfahren; die folgende (auszugsweise) Konfiguration ist typisch für eine Realisierung im Kundenumfeld, wenn eine Klassifizierung nach Kunden und Lieferanten vorgenommen werden soll:

- Identifizierung aller Personennamen und Zuordnung zu Kunden- oder Lieferantendaten gemäß Datenbank
- Identifizierung von bekannten Kunden- oder Lieferantennamen ebenfalls aus der Datenbank
- Identifizierung von Mustern für Angebote, Aufträge sowie Rechnungen und Abgleich mit der Buchhaltung
- Ähnlichkeit einer Mail zu vorhandenen bereits klassifizierten Mails
- Häufigkeit von Begriffen
- ...

Diese Liste läßt sich je nach erforderlicher Präzision und vorhandenen Informationsquellen erweitern. Jede Bewertung liefert einen Beitrag zur Gesamtklassifizierung.

Alle Komponenten können einzeln konfiguriert werden und ermöglichen so die optimale Anpassung an die jeweils vorliegende Aufgabe. Trotzdem ist es bereits mit wenigen Standardeinstellungen möglich, daß gute Ergebnisse erzielt werden.