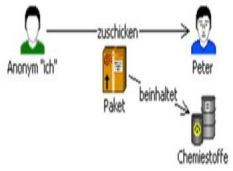
Im Einsatz – im Thema.
POLIZEI PRAXIS

BIG DATA: GROSSES POTENTIAL DURCH SEMANTISCHE DATENANALYSE DES BKA

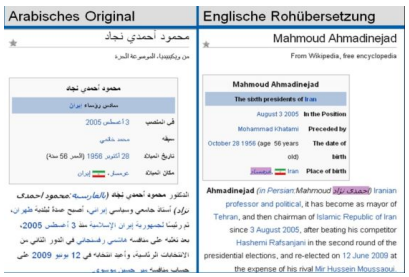
Entität	Typ
Anton Alfsberger	Personenname
Mark Beyer	Personenname
Thomas Rittelkopf	Personenname
zippi43@yahoo.de	Emailadresse
T.Aufberg@gmx.de	Emailadresse
Mike-Marser@yahoo.de	Emailadresse
Südtirol	Geographischer Ort
12.05.11	Zeit
7 Uhr morgens	Zeit
2,08 Billionen Euro	Währung
540 Millionen Euro	Währung
Sechs Milliarden Euro	Währung
Basel	Stadt
Berlin	Stadt
Bendorf	Stadt
2010	Datum
2016	Datum
2. Dezember 2011	Datum
27. Okt	Datum
Finanzminister	Position
Institut für Lebensmittelwissenschaft	Organisation
Vereinte Nationen	Organisation
1,3 Gramm	Maße
100 g	Maße

Beispiel mit fiktiven Entitäten und Testdaten

Sprache	Personenname	Transkription
Russisch	Джона Кеннеди	John Kennedy
Russisch	Дмитрия Медведева	Dmitry Medvedeva
Russisch	Дэвид Бекхэм	David Beckham
Russisch	Лари Джордан	Larry Jordan
Russisch	Нил Армстронг	Neil Armstrong
Russisch	Сэм Перкинс	Sam Perkins
Arabisch	أبراهام لينكولن	Abraham Lincoln
Arabisch	إرنست همنغواي	Ernest Hemingway
Arabisch	باراك أوباما	Barak Obama
Arabisch	تيودور روزفلت	Theodore Roosevelt
Arabisch	ريتشارد نيكسون	Richard Nixon
Arabisch	مارتن لوتر كينغ	Martin Luther King



Ich schicke Peter das Paket mit den Chemikalien morgen.



Der Bereich der Informationstechnologien ist einer der schnelllebigsten unserer Zeit. Nicht nur die Rechenleistung von Prozessoren steigt stetig weiter an, auch die Kapazitäten von Datenspeichern werden bei sinkenden Preisen immer größer. Es werden immer mehr Daten produziert, gespeichert und auch verarbeitet.

Diese Entwicklungen wirken sich massiv auf die Arbeit der Polizei- und Strafverfolgungsbehörden aus. Im Rahmen von Ermittlungsverfahren sind sichergestellte Datenmengen im Terrabytebereich längst keine Seltenheit mehr. In

einigen Deliktsbereichen sind sie sogar zum Normalfall geworden. Dabei müssen beispielsweise tausende elektronische Dokumente zeitnah ausgewertet und auf Verfahrensrelevanz geprüft werden.

Besonders in Zeiten begrenzter personeller und finanzieller Ressourcen stellen diese Aspekte die Sicherheitsarchitektur permanent vor große Herausforderungen.

Die Tatsache, dass in vielen Fällen auch fremdsprachige Daten eine Rolle spielen, macht die Auswertung noch um ein Vielfaches schwieriger. Ohne Übersetzer ist in vielen Fällen keine Aussage zum Inhalt und damit zur Relevanz einzelner Dateien, wie bspw. Audio- oder Textdateien, möglich.

Das Bundeskriminalamt (BKA) hat als Zentralstelle zur Unterstützung der Polizeien des Bundes und der Länder den Auftrag, polizeiliche Methoden und Arbeitsweisen der Kriminalitätsbekämpfung zu erforschen und zu entwickeln.

Diesem Auftrag folgend betreibt das Kriminalistische Institut das *Technische Entwicklungs- und Servicezentrum, Innovative Technologien*. Neben vielen anderen Spezialgebieten, werden hier auch Methoden und Lösungen gesucht, erprobt und entwickelt, die eine effektive und effiziente polizeiliche Auswertung von großen Datenmengen gewährleisten.

„Wie kann ich zeitnah in 30.000 Emails diejenigen finden, in denen mein Beschuldigter eine Rolle spielt?“
„Wie finde ich heraus, in welchen von den 17.435 russischen Dokumenten auf der sichergestellten Festplatte relevante Informationen für mein Ermittlungsverfahren enthalten sind? Handelt es sich überhaupt um russische Dokumente? Oder brauche ich stattdessen doch eher einen Übersetzer für Bulgarisch... oder Rumänisch... oder Kasachisch?“

Solche Fragen stellen sich den Ermittlern und Auswertern tagtäglich.

Konventionelle Methoden wie eine einfache Schlagwortsuche reichen bei weitem nicht mehr aus. Ein händisches Auswerten tausender Einzeldokumente ist ineffizient und nicht zielführend. Natürlich können die unzähligen zu übersetzenden Texte an Übersetzer übergeben werden. Nur an welchen? Steht überhaupt ein passender Übersetzer zur Verfügung? Wie ist seine derzeitige Arbeitsauslastung und wie lange wird eine Übersetzung dauern? Vor allem im Kontext von Haftprüfungsterminen oder Gefährdungssachverhalten sind solche Aspekte als besonders kritisch zu bewerten.

Um diese Fragen und viele andere zu beantworten, ist es notwendig, sich neuer und intelligenter Methoden der inhaltlichen Datenauswertung zu bedienen.

Im BKA wurde daher vor einigen Jahren unter der Überschrift „Semantische Textanalyse“ damit begonnen, sich unter anderem mit modernsten Methoden der Analyse textueller Daten auseinanderzusetzen. Kommerzielle Produkte wurden auf ihre Geeignetheit für die polizeiliche Verwendung untersucht und von Experten eigene Lösungen entwickelt um sich dieser und zukünftiger Herausforderungen Rechnung tragend aufzustellen.

Fremdsprachenerkennung

Bei der Datenauswertung ist der erste notwendige Schritt oftmals die Bestimmung der Sprache in der eine Textdatei verfasst wurde. Im BKA wurde hierzu die Anwendung FREDI (FRemdsprachenErkennungsDIenst) entwickelt. Mit dieser Software ist es möglich, auch mehrere Sprachen innerhalb eines Dokumentes konturenscharf zu erkennen (Bild 1). Einzelne Sprachabschnitte eines Dokumentes können so für eine automatisierte sprachspezifische Weiterverarbeitung separiert werden. Eine Weiterleitung an den richtigen Übersetzer wird dadurch enorm beschleunigt. FREDI erkennt derzeit insgesamt 57 verschiedene Sprachen. Darunter befinden sich die gängigen wie Englisch, Deutsch, Französisch, Arabisch, Russisch und Spanisch. Es werden jedoch auch „exotische“ Sprachen wie bspw. Kinyarwanda, Kurdisch-Sorani, Baskisch, Paschto, Urdu oder Farsi erkannt. FREDI ist bei Bedarf beliebig auf weitere Sprachen trainierbar. Mit der Anwendung ist sowohl die Einzel-, als auch die Stapelverarbeitung einer Vielzahl von elektronischen Dokumenten möglich. Dabei kann die Software im Rahmen der Bürokommunikation betrieben oder auf einem Stand-Alone-PC genutzt werden.

Automatisiertes Filtern von Kerninformationen - Entitätenextraktion

Um im Rahmen von Ermittlungsverfahren tausende (auch fremdsprachige) Texte zeitnah inhaltlich auszuwerten, sind bisher große personelle Ressourcen notwendig. Durch den Einsatz intelligenter bzw. semantischer Datenanalysemethoden sollen die Ermittler einen schnelleren Einblick in die Inhalte der Dokumente erhalten. Durch hochentwickelte Algorithmen und statistische Analysen werden bei dieser Vorgehensweise Kerninformationen (sog. Entitäten), wie z.B. Personennamen, Firmennamen, Telefonnummern, Bankdaten, E-Mail-Adressen und vieles mehr, automatisiert in den Dokumenten erkannt und gefiltert. (Bild 2) Darüber hinaus bieten gegenwärtig verfügbare Softwarelösungen an, gefilterte Personennamen und Ortsangaben innerhalb des Analyseprozesses zu übersetzen oder zu transkribieren (Übertragung ins lateinische Alphabet anhand der Laute eines Begriffs – Bild 3). Des Weiteren können Personen und Ortsnamen auch sprachübergreifend in den Dokumenten gefunden werden, selbst wenn Sie in einer anderen Sprache, beispielsweise mit arabischen Buchstaben, geschrieben wurden. Auch variierende Schreibweisen eines Namens werden dabei erfolgreich gefiltert. Besonders hier zeigt sich der Vorteil gegenüber einer einfachen Schlagwortsuche, bei der ein Suchwort im Vorfeld bekannt sein muss und auch nur dann gefunden wird, wenn es mit identischer Schreibweise im Text enthalten ist.

Mit diesen Methoden werden somit fremdsprachige Texte für eine erste Relevanzeinschätzung handhabbar, ohne dass zwingend ein Übersetzer hinzugezogen werden muss.

Das Aufdecken von Verbindungsgeflechten

Wenn nun Informationen in Form von Entitäten (wie Personennamen oder E-Mail-Adressen) aus den Daten gefiltert wurden, lassen sich durch weitere automatisierte Analyseschritte auch Verbindungen / Relationen zwischen diesen Kerninformationen extrahieren. So können Fragen wie: *„Wer hat wann, mit wem und wie häufig kommuniziert*?“, bspw. in einem großen E-Mail Datenbestand oder Web-Forum, beantwortet werden. Diese können anschließend mit entsprechender Software anschaulich als Kommunikationsnetzwerk visualisiert werden.

Bisherige Erfahrungen zeigen großes Potential

Erfolgreich durchgeführte Unterstützungsmaßnahmen mittels semantischer Datenanalyse haben ein enorm großes Potential gezeigt. Dabei können die Methoden und Anwendungen auch deliktsunabhängig eingesetzt werden. Personennamen, Bankdaten, Ortsangaben, Telefonnummern oder E-Mail-Adressen - solche und weitere Kerninformationen kommen in jeglicher Art von Ermittlungsverfahren vor. Sie bilden das Gerüst eines jeden Ermittlungskonzeptes.

So sollte z.B. in einem unterstützten Ermittlungsverfahren die Auswertung von ca. 20.000 Emails den Nachweis des illegalen Verkaufs von Arzneimitteln erbringen. Eine Schlagwortsuche war nicht zielführend und eine händische Sichtung der Emails hätte mehrere Tage vielleicht sogar Wochen gedauert. Die automatisierte Erkennung von Geldbeträgen in den Texten und Anhängen der Emails machte es binnen weniger Minuten möglich die Emails zu filtern, in denen Rechnungsbelege und somit Verkaufsnachweise enthalten waren.

Ein anderes Beispiel entstammt dem Bereich Betrugsdelikte. Den Ermittlern lagen in einem Ermittlungskomplex ca. 18.000 Dokumente zur Auswertung vor. Diese beinhalteten Textbereiche in insgesamt acht unterschiedlichen Sprachen. Ziel war es, Muster und Zusammenhänge in den Texten zu erkennen und auf diesem Wege Ermittlungsansätze zu generieren. Hier wurden im Rahmen der technischen Unterstützung u.a. Personen- und Organisationsnamen, Telefonnummern und Emailadressen gefiltert. Diese wurden anhand der Häufigkeit ihres Vorkommens sortiert um festzustellen, welche Namen, Firmen und Kommunikationsdaten immer wieder auftraten. In enger Zusammenarbeit zwischen Ermittlern und technischen Experten im Bereich Datenanalyse wurde die Datenmenge immer weiter verringert. Im Ergebnis erreichte man durch diese Vorgehensweise eine Reduktion der händisch auszuwertenden Dokumente um 95 Prozent.

Ein Blick nach vorn

Da auch in Zukunft die Herausforderungen im Bereich der fremdsprachigen Massendatenauswertung fortwährend wachsen werden, muss der Weg der intelligenten Datenanalyse konsequent weiter verfolgt werden. Die hier bisher aufgezeigten Möglichkeiten stellen dabei nur die ersten vielversprechenden Schritte dar.

Großes Potential bietet die Weiterentwicklung der automatisierten Extraktion von Verbindungen/Relationen zwischen Kerninformationen. Führt man sich vor Augen, welche Fülle von Informationen in Texten enthalten ist, lässt sich erkennen, welche Aussagekraft eine komplexere „Informationsvorschau“ zur Relevanzeinschätzung besitzt.

„Person A hat am 25. September 2013 ein Paket mit Chemikalien an Person B gesendet.“
„Person C und Person D weisen eine Verbindung zum Fahrzeug mit dem Kennzeichen KT – XX – 9999 auf.“

Werden beispielsweise nicht „nur“ Personennamen, Bankdaten und Firmennamen, sondern auch solche Verbindungen zwischen ihnen extrahiert, lässt sich auf diesem Wege deutlich mehr Information zur Bewertung von Texten anbieten (Bild 4).

Neben der Extraktion von Kerninformationen und deren Transkription/Translation ist die Maschinelle Rohübersetzung ein weiteres wertvolles Hilfsmittel bei der Erschließung von fremdsprachigen Texten.

Erste erfolgreiche Tests haben gezeigt, dass eine grobe, automatisierte Übersetzung zwar nicht auf dem Qualitätsniveau einer menschlichen professionellen Übersetzung anzusiedeln ist, das Ergebnis jedoch in vielen Fällen ausreicht um den Inhalt eines Dokuments grob zu erfassen (Bild 5). Für viele aus polizeilicher Sicht relevante Sprachen existieren bereits Lösungen, die eine intensive Auseinandersetzung mit dem Thema rechtfertigen.

Fazit

Um der hier behandelten Problematik aus polizeilicher Sicht Herr zu werden und sich zukunftsorientiert aufzustellen führt kein Weg an der Verwendung modernster Methoden und Lösungen vorbei. Sowohl kommerzielle, als auch Open Source Produkte bieten hierzu sehr gute Möglichkeiten an.

Natürlich hängt im Bereich automatisierter Datenverarbeitung die Qualität der Analyseergebnisse immer auch von der Qualität der zu verarbeitenden Daten ab. Eine Kommunikation in einem Webchat weist in der Regel im Vergleich zu einer geschäftlichen Email oder einem formellen Bericht Schwächen bezüglich Grammatik und Interpunktion auf, die eine automatisierte Analyse erschweren können. Auch sind kriminalistische Erfahrung und Gespür der Ermittlerinnen und Ermittler durch nichts zu ersetzen. Jedoch stellen die vorgestellten Methoden höchst wertvolle Hilfsmittel dar, wenn es darum geht der Datenflut an sich, vor allem jedoch auch im fremdsprachigen Kontext, Herr zu werden. Die erzielten Ergebnisse belegen eindeutig, dass Lösungen zur Fremdsprachenerkennung, Entitäten- und Relationenextraktion sowie Maschinelle Rohübersetzung in Zeiten knapper Ressourcen die Effektivität und vor allem die Effizienz polizeilicher Auswertung von großen Datenmengen bereits heute deutlich verbessern können.

K. Kessler, BKA
T. Lenzen, BKA

Alle Artikel dieser Kategorie

Folgen Sie uns!