Themenentdeckung und -verfolgung und ihr Einsatz bei Informationsdiensten für Nachrichten

Wolfgang G. Stock, Düsseldorf

Themenentdeckung und -verfolgung (topic detection and tracking; TDT) fasst unterschiedliche Dokumente zu einem Thema zusammen und bietet dem Nutzer zunächst das Thema und erst in einem zweiten Schritt die einzelnen Dokumente zur Anzeige an. TDT ist nützlich bei Nachrichten sowie bei Blog-Einträgen. Ein bekanntes Beispiel ist Google News. Der Artikel bespricht den Forschungsstand zu TDT und diskutiert den Einsatz von TDT bei News-Informationsdiensten wie z.B. Factiva oder LexisNexis. Dort ist TDT zweifach wichtig: Erstens ist es (analog zu Google News) bei Profildiensten einsetzbar, zweitens ist es möglich, Dokumente zum gleichen Thema in einen Ordner zu klassieren, um bei einer retrospektiven Recherche die Treffermenge geordnet nach unterschiedlichen Themen anzubieten (und dies ohne Nutzung einer Dokumentationssprache).

Topic Detection & Tracking and its Application in News Information Services.

Topic detection and tracking (TDT) integrates different documents on the same topic. First it offers the topic to the users and in a second step the specific documents. TDT is useful for presenting news articles and blog postings. A well known example is Google News. The article summarizes the research on TDT and discusses the application of TDT in news information providers (e.g., Factiva or LexisNexis). For these services TDT is important for two reasons: (1) analogous to Google News, TDT works within alerting services; (2) TDT allows for classifying sets of records according to topics (without using a documentation language).

Thematisches Klassieren von Nachrichtendokumenten

Im World Wide Web, im Deep Web (etwa bei Informationsdiensten von Nachrichtenagenturen) sowie in anderen Medien (z.B. Hörfunk oder Fernsehen) liegen Dokumente vor, die einen aktuellen Bezug haben und deren Inhalt sich häufig (ggf. leicht abgewandelt) in unterschiedlichen Quellen findet. Dies trifft vor allem für Nachrichten, aber auch für gewisse Einträge in Weblogs zu (Peters & Stock 2006). Im Gegensatz zum „normalen" Information Retrieval startet hier die Recherche nicht mit einem erkannten Informationsbedarf eines Nutzers, sondern mit einem neuen Ereignis, das es zu erkennen und darzustellen gilt. Dem Nutzer als Pushdienst angeboten werden die Informationen zu den Ereignissen über spezialisierte Nachrichtensysteme wie beispielsweise Google News. Es ist so, als ob ein Nutzer ein SDI in Auftrag gegeben hätte, ihn laufend über alle neuen Ereignisse zu informieren. Dabei ist es durchaus möglich, die Neuigkeiten thematisch auszuwählen. Auch im Nachhinein stehen die erkannten Themen für Recherchen oder informetrische Analysen bereit. Yang, Shi und Wie (2006) können die Bedeutung analytischer Methoden am Beispiel der Ereignisse und Themen terroristischer Anschläge herausstellen.

James Allan, der das Forschungsgebiet maßgeblich geprägt hat (Allan 2002a; Allan 2002b; Allan 2003; Allan, Carbonell, Doddington, Yamron, & Yang 1998; Allan, Feng, & Bolivar 2003; Allan, Harding, Fisher, Bolivar, Guzman-Lara, & Amstutz 2005; Allan, Lavrenko, & Connell 2003; Allan, Lavrenko, & Jin 2000; Allan, Lavrenko, & Swan 2002; Allan, Papka, & Lavrenko 1998; Allan, Wade, & Bolivar 2003; Feng & Allan 2005; Frey, Gupta, Khandelwal, Lavrenko, Leuski, & Allan 2001; Kumaran & Allan 2005; Lavrenko, Allan, DeGuzman, LaFlamme, Pollard, & Thomas 2002; Leuski & Allan 2002; Manmatha, Feng, & Allan 2002; Papka & Allan 2000), nennt diesen Teilbereich des Information Retrieval „topic detection and tracking" (TDT), was wir mit „Themenentdeckung und -verfolgung" übersetzen wollen (Stock 2007, 425-436). Allan definiert den Forschungsbereich:

> *„Topic Detection and Tracking (TDT) is a body of research and an evaluation paradigm that addresses event-based organization of broadcast news. The TDT evaluation tasks of tracking, cluster detection, and first story detection are each information filtering technology in the sense that they require that 'yes or no' decisions be made on a stream of news stories before additional stories have arrived" (Allan 2002b, 139).*

Google News beschränkt sein Angebot auf im WWW vorliegende Dokumente, die Nachrichtenagenturen oder Online-Redaktionen von Zeitungen unentgeltlich bereitstellen. Nicht berücksichtigt werden kommerzielle Angebote der News Wires, die meisten Artikel der Druckausgaben der Zeitungen und Zeitschriften (und dies ist die überwältigende Mehrzahl aller Nachrichten) sowie alle nicht-digital (über Rundfunk) verteilten Informationen. Google News richtet seinen Fokus auf neue Artikel (der Beobachtungszeitraum beträgt wenige Tage) und solche, die von allgemeinem Interesse sind. Krishna Bharat berichtet:

> *„Specifically,* freshness – *measurable from the age of articles, and* global editorial interest – *measurable from the number of original articles published worldwide on the subject, are used to infer the importance of the story at a given time. If a story is fresh and has caused considerable original reporting to be generated it is considered important. The final layout is determined based on additional factors such as (i) the fit between the story and the section being populated, (ii) the novelty of the story relative to other stories in the news, and (iii) the interest within the country, when a country specific edition is being generated" (Bharat 2003, 9).*

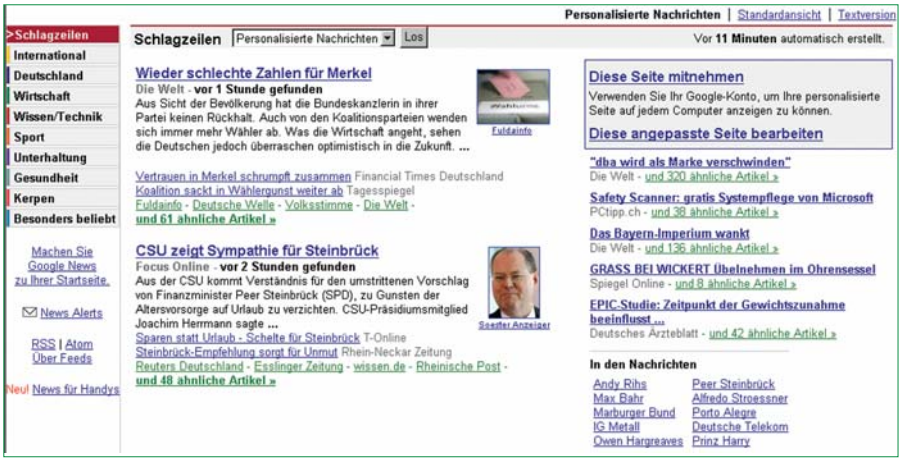


Abbildung 1: Themenentdeckung und -verfolgung am Beispiel von Google News: Anzeige der aktuellen Themen. Quelle: news.google.de

Abbildung 2: Themenentdeckung und -verfolgung am Beispiel von Google News: Anzeige der Dokumente zu einem Thema. Quelle: news.google.de

Abbildung 3: Themenentdeckung und -verfolgung am Beispiel von Google News: Personalisierung. Quelle: news.google.de

Abbildung 4: Trefferliste bei LexisNexis Wirtschaft mit der Option, die Dokumente zu gruppieren. Quelle: LexisNexis

Auf der Einstiegsseite von Google News werden ausschließlich die Themen aufgelistet (Abbildung 1), z.B. der Topic „dba wird als Marke verschwinden“ (rechts oben). Wir erfahren hier zusätzlich, dass etwa 320 Dokumente unter dieses Thema fallen. Erst der Klick auf „und 320 ähnliche Artikel“ führt zur kompletten Liste aller Dokumente zum Thema (Abbildung 2). Die Liste aus Abbildung 2 zeigt die „klassischen“ dokumentarischen Bezugseinheiten (die Dokumente), die Liste aus Abbildung 1 führt thematisch verwandte dokumentarische Bezugseinheiten zu einer Meta-Bezugseinheit (das Thema) zusammen. Es ist möglich, die Anzeigeseite zu „personalisieren“, indem vom Nutzer selbst gewählte Themenkomplexe (z.B. „Informationswissenschaft“ oder „Heinrich-Heine Universität“) mit jeweils neuesten Themen bestückt werden (Abbildung 3).

Themenentdeckung und -verfolgung dürfte nicht nur bei kostenlosen Nachrichtensuchmaschinen von Vorteil sein, sondern auch bei solchen kommerziellen Informationsanbietern, die Nachrichten vertreiben, also beispielsweise ASV Infopool (M.Stock 2002b), GBI-GENIOS (Stock & Stock 2003a; 2003d), Dialog NewsEdge (Stock & Stock 2003c), Factiva (M.Stock 2002a; Stock & Stock 2003b) oder LexisNexis (Stock & Stock 2005). Wir wollen dies (unter Fortführung unseres dba-Beispiels) bei LexisNexis Wirtschaft demonstrieren! Die erste Anwendung von Themenentdeckung und -verfolgung, quasi die Normalvariante, arbeitet bei Profildiensten und damit analog zu Google News in der personalisierten Anwendung. Der Unterschied zu Google liegt einzig in den elaborierten Optionen der Kreation eines Informationsprofils bei den kommerziellen Informationsanbietern. Angezeigt wird das Thema (die Meta-Dokumentationseinheit), und erst im zweiten Schritt die Liste der Dokumente zum Thema.

LexisNexis verfügt über eine Option, Treffermengen zu gruppieren, in Abbildung 4 haben wir etwa nach Quellen gruppiert. Hiermit hat der Nutzer die Möglichkeit, große Trefferlisten kontextspezifisch einzuschränken. Derzeit bietet LexisNexis keine Gruppierung

nach Themen an. Hier liegt die zweite Möglichkeit für Themenentdeckung und -verfolgung, eine Klassierungsoption nach Topics. Solch eine Gruppierung fasst alle Dokumente der Trefferliste zu gleichen Themen zusammen. Unsere 19 Dokumente aus Abbildung 4 würden so zu einem einzigen Topic verschmelzen. Insbesondere bei großen Treffermengen, die bei News-Recherchen nicht unüblich sind, ist dies eine brauchbare Option, Dokumente thematisch zu klassieren (und dies ohne Klassifikationssystem oder Thesaurus – einzig mittels TDT).

Aufgaben der Themenentdeckung und -verfolgung

Vier grundlegende Begriffe sind im Kontext von TDT wichtig:

- Eine „Story" ist eine abgrenzbare Textstelle (oder ein ganzes Dokument), in der (oder in dem) ein Ereignis besprochen wird.
- Ein „Thema" (topic) ist die Beschreibung eines Ereignisses in den jeweiligen Stories, gemäß Allan „a set of news stories that are strongly related by some seminal real-world event" (Allan 2002a, 2). Für James Allan ist ein „Topic" demnach nur die Beschreibung eines spezifischen Ereignisses (wie etwa die Übernahme von dba durch Air Berlin im August 2006), nicht aber ein Thema ohne Orts- und Zeitbezug (wie beispielsweise das Wachsen von Blumen an schattigen Standorten). M.E. lässt sich TDT jedoch auf alle Arten von Themen anwenden.
- Das Gruppieren aktueller Stories zu einem *neuen* Thema ist die „Themenentdeckung" (topic detection).
- Das Hinzufügen von Stories zu einem *bekannten* Thema ist die „Themenverfolgung" (topic tracking).

Die Themenentdeckung und -verfolgung besteht aus mehreren Einzelaufgaben (die ersten fünf folgen Allan 2002a):

- Textstellenzerlegung (story segmentation): Isolation derjenigen Textstellen (Stories), die das jeweilige Thema beinhalten (bei Dokumenten, die mehrere Ereignisse besprechen),
- Themenentdeckung bzw. Erkennung eines neuen Ereignisses (new event detection): Identifikation der ersten Story, die ein neues Ereignis thematisiert,
- Cluster-Erkennung (cluster detection): Zusammenfassung aller Stories, die dasselbe Thema beinhalten,
- Themenverfolgung (topic tracking): Analyse des laufenden Nachrichtenstroms auf bereits bekannte Themen,
- Link-Erkennung (link detection): Analysewerkzeug zur Bestimmung der thematischen Ähnlichkeit zweier Stories,
- Zuordnung eines Titels zu einem Cluster: entweder Titel der ersten Story oder Zuordnung der ersten n nach Gewichtung geordneten Terme aus allen Stories, die dem Cluster angehören,
- Abstract: Verfassen einer kurzen Zusammenfassung des Themas (als Form automatischen Abstracting) oder – weitaus einfacher – Übernahme des ersten Abschnittes der ersten Story,
- Rangordnung der Stories: sofern ein Cluster mehrere Stories umfasst, Sortierung der Texte nach Wichtigkeit.

Einen Überblick zu den Arbeitsschritten vermittelt Abbildung 5.

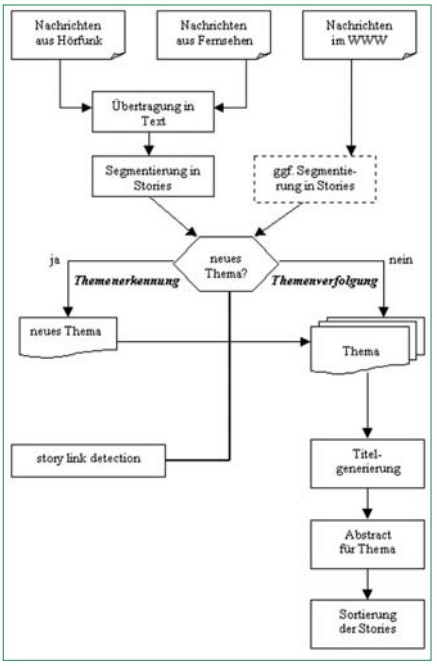


Abbildung 5: Arbeitsschritte der Themenentdeckung und -verfolgung.
Quelle: Stock 2007, 427.

Themenentdeckung

Bei Nachrichten aus Hörfunk und Fernsehen muss zunächst eine Übertragung der Audio-Signale in (geschriebenen) Text vorgenommen werden. Dies geschieht entweder durch intellektuelle Transkription oder unter Nutzung von Spracherkennungssystemen (Allan 2003; zu Problemen vgl. McCarley & Franz 2000). In einer Nachrichtensendung (z.B. einer „Tagesschau" mit einer Länge von 15 Minuten) werden mehrere singuläre Stories gebracht, die durch eine Segmentierung des Gesamttextes als jeweils einzelne Einheiten gezählt werden (Allan, Carbonell, Doddington, Yamron, & Yang 1998, 196 ff.).

Analog wird mit dem Nachrichtenstrom einer Agentur umgegangen. Zur Vereinfachung kann man in diesem Fall annehmen, dass jedes Nachrichtendokument genau eine Story beinhaltet.

Die entscheidende Frage bei der Analyse des Nachrichtenstromes ist: Behandelt die gerade angekommene Story ein neues Thema oder thematisiert sie etwas, was bereits als Thema bekannt ist (Allan, Lavrenko, & Jin 2000)? Ein (erkanntes) Thema wird durch den Durchschnittsvektor (Zentroid) seiner Stories ausgedrückt. Bei der *Themenentdeckung* errechnen wir die Ähnlichkeit bzw. Unähnlichkeit zwischen der aktuellen Story und allen anderen aus der Datenbank. Lässt sich keine Ähnlichkeit feststellen (liegt also ein neues Thema vor), so gilt diese erste Story als Repräsentant des neuen Themas. Lassen sich demgegenüber Ähnlichkeiten zwischen aktueller Story und alten Stories ausmachen, so ist der Fall der Themenverfolgung gegeben. Es folgt ein Vergleich zwischen der aktuellen Story und allen bereits erkannten Themen. Eine zentrale Rolle bei der Themenentdeckung bzw. -verfolgung spielt die „story link detection", deren Algorithmus feststellt, ob es sich bei der aktuellen Story um ein neues oder um ein bekanntes Thema handelt.

Allan, Harding, Fisher, Bolivar, Guzman-Lara, & Amstutz (2005) setzen an dieser Stelle das Vektorraummodell und eine Variante von TF*IDF (Termhäufigkeit * inverse Dokumenthäufigkeit) zur Bestimmung der Termgewichtung ein. Für jede Story aus dem Nachrichtenstrom wird für jeden Term ein Gewichtungswert errechnet. $tf_{t,s}$ ist die (absolute) Auftretenshäufigkeit eines Terms t in der Story s, df_t zählt alle Stories in der Datenbank, die den Term t enthalten, und N ist die Anzahl der Stories in der Datenbank. Das Termgewicht w von t in s errechnet sich wie folgt:

$$w_{t,s} = [tf_{t,s} * \log((0{,}5 + N) / df_t)] / [\log(N + 1)].$$

Allan et al. schlagen vor, die ersten 1000 nach Gewichtung sortierten Terme in einem Story-Vektor zu berücksichtigen. Mit der Ausnahme weniger langer Nachrichtentexte sollten dabei alle Worte einer Story Berücksichtigung finden. Anhand der Berechnung des Cosinus wird die Ähnlichkeit des Story-Vektors mit allen anderen Story-Vektoren der Datenbank analysiert. Die Autoren haben empirisch einen Wert von $Sim(s1,s2) = 0{,}21$ bestimmt, der neue von alten Stories trennt. Liegt die höchste Ähnlichkeit zwischen der neuen und einer beliebigen alten Story unter 0,21, so wird die aktuelle Story als neues Thema aufgefasst; liegt sie darüber, wird eruiert, zu welchem bekannten Thema die neue Story gehört.

Es zeigt sich in der Praxis, dass dieser allgemeine Ansatz nicht ausreicht, um genügend zuverlässig neue von alten Stories zu trennen. Mittels zusätzlicher Faktoren wird die Leistung von TDT-Systemen besser.

Eigennamen und weitere Themen

Bei der Identifizierung eines Themas spielen Eigennamen („named entities") auf der einen Seite und die weiteren Worte („topic terms") auf der anderen Seite eine wichtige Rolle. Zwei Stories behandeln dann das gleiche Thema, wenn in ihnen sowohl dieselben „named entities" als auch dieselben restlichen Worte übereinstimmen. Giridhar Kumaran und James Allan begründen diesen Ansatz so:

> *„The intuition behind using this features is that we believe every event is characterized by a set of people, places, organizations, etc. (named entities), and a set of terms that describe the event. While the former can be described as the* who, where *and* when *aspects of an event, the latter relates to the* what *aspect. If two stories were on the same topic, they would share both named entities as well as topic terms. If they were on different, but similar, topics, then either named entities or topic terms will match but not both" (Kumaran & Allan 2005, 123).*

Ein anschauliches Beispiel eines „Fehlläufers" zeigt Abbildung 6. Das System hat hierbei keinen Gebrauch von der Unterscheidung „named entities" – „topic terms" gemacht. Durch den hohen WDF-Wert von „Turkey" bzw. „Turkish" sowie den sehr hohen IDF-Wert von „Ismet Sezgin" behauptet die Vektorraummaschine, dass die obere Story ähnlich zur unteren und damit keineswegs neu sei. Tatsächlich ist der obere Text jedoch neu. Kein einziger „topic term" kommt in beiden Meldungen gemeinsam vor, so dass Kumaran und Allan schlussfolgern:

> *„Determining that the topic terms didn't match would have helped the system to avoid this mistake" (Kumaran & Allan 2005, 124).*

Es scheint demnach sinnvoll, die Ähnlichkeit (Cosinus) zwischen zwei Stories getrennt für „named entities" und „topic terms" zu berechnen. Nur wenn beide Ähnlichkeitswerte einen Schwellenwert überschreiten, wird eine Story als einem „alten" Thema zugehörig eingestuft.

Topic not seen before

Turkey has sent 10,000 troops to its southern border with Syria amid growing tensions between the two neighbors, newspapers reported Thursday. Defense Minister **Ismet Sezgin** denied any troop movement along the border, but said **Turkey's** patience was running out. **Turkey** accuses Syria of harboring Turkish Kurdish rebels fighting for autonomy in **Turkey's** southeast; it says rebel leader Abdullah Ocalan lives in Damascus.

Closest Story due to Named Entities

A senior **Turkish** government official called Monday for closer military cooperation with neighboring Bulgaria. After talks with President Petar Stoyanov at the end of his four-day visit, Turkish Deputy Premier and National Defense Minister **Ismet Sezgin** expressed satisfaction with the progress of bilateral relations and the hope that Bulgarian-**Turkish** military cooperation will be promoted.

Abbildung 6: Die Rolle von „named entities" und „topic terms" bei der Identifikation eines neuen Themas. Quelle: Kumaran & Allan 2005, 124.

Zeit- und Ortsbezug von Nachrichten

Ein sehr wichtiger Aspekt von Nachrichten ist deren Zeit- und Ortsbezug. Makkonen, Ahonen-Myka und Salmenkivi (2004, 354 ff.; vgl. auch Makkonen, Ahonen-Myka, & Salmenkivi 2003) arbeiten zusätzlich zur „allgemeinen Ähnlichkeit" mit „zeitlicher" und „räumlicher Ähnlichkeit". Zur Bestimmung des Zeitbezugs ist zunächst erforderlich, aus den Angaben im Text exakte Daten abzuleiten (Makkonen & Ahonen-Myka 2003; Kim & Myaeng 2004; Li, Li, & Lu 2006). Nehmen wir an, eine Nachricht trägt das Datum des 27. Mai 2003 (Abbildung 7). Im Text vorkommende Formulierungen wie „letzte Woche", „letzter Mittwoch", „am nächsten Donnerstag" usw. verlangen nach einer Auflösung in Datumsangaben, also in „2003-05-19: 2003-05-26", „2003-05-21" bzw. „2003-05-29". Stories, die ansonsten ähnlich sind, sich aber im Datum nicht überschneiden, sind wahrscheinlich unterschiedlichen Themen zuzuordnen. Berichte über Karnevalszüge in Köln aus den Jahren 2005 und 2006 unterscheiden sich kaum in ihren Termen („Millionen Zuschauer", „Prinzenwagen", „Kamelle"), aber durch das Datum.

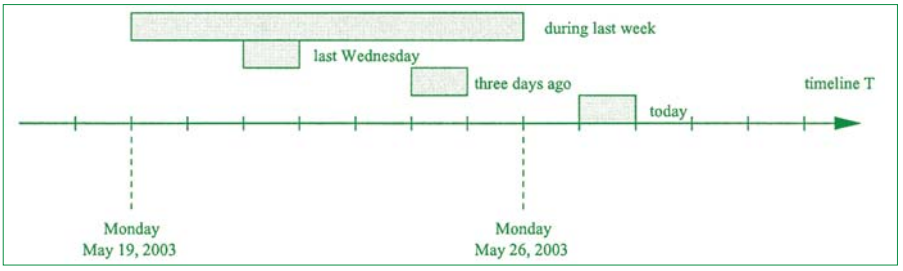


Abbildung 7: Auflösung von zeitlichen Ausdrücken. Quelle: Makkonen, Ahonen-Myka, & Salmenkivi 2004, 355.

Themenverfolgung

Bei der Themenverfolgung gehen wir von einer Menge bekannter Themen aus. Die Ähnlichkeitsberechnung erfolgt nunmehr durch den Abgleich zwischen den Themenvektoren, also dem jeweiligen Namen- und Topic-Zentroiden des Themas, und den jeweils in die Datenbank aufgenommenen neuen Stories.

Dem Ortsbezug (Jin, Myaeng, Lee, Oh, & Jang 2005) wird mit Hilfe eines geographischen Begriffssystems nachgegangen (für ein einfaches Beispiel siehe Abbildung 8). Die Ähnlichkeit zwischen unterschiedlichen Ortsangaben in zwei Stories kann durch die Pfadlänge ausgedrückt werden (Makkonen, Ahonen-Myka, & Salmenkivi 2004, 357 f.). Redet eine Quelle etwa von Puchheim und eine andere, thematisch verwandte, vom Landkreis Fürstenfeldbruck (FFB), so werden beide Stories anhand der Pfadlänge von 1 als ähnlich eingestuft. Ein anderes Paar von Nachrichten behandelt ebenfalls ähnliche Themen, wobei die eine Nachricht wiederum über Puchheim spricht und die andere über Venlo. Da die Pfadlänge nun aber 8 beträgt, spricht nichts dafür, dass es wirklich um dasselbe Ereignis geht.

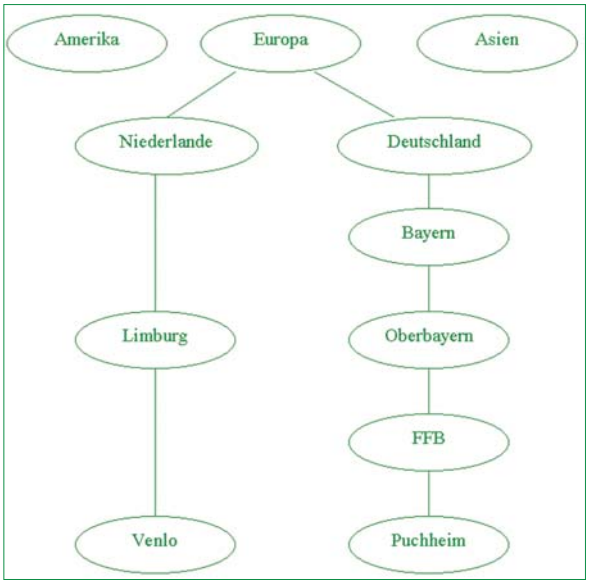


Abbildung 8: Geographische Begriffsordnung als Hilfsmittel der Themenentdeckung und -verfolgung. Quelle: Stock 2007, 432.

Hinzutreten müssen Vergleiche von Raum- und Zeitbezug, sofern gegeben. Bei der ersten Story ist der Zentroid identisch mit dem Vektor dieser Story. Ab einer zweiten Story können wir erst sinnvoll von einem „Durchschnittsvektor" reden. Der Zentroid ändert sich solange, wie noch weitere Stories zum Thema identifiziert werden. Benutzt man den Zentroiden zur Bestimmung des Titels, indem man beispielsweise die ersten zehn nach Gewichtung sortierten Terme des Zentroiden als „Titel" auszeichnet, so kann sich der Titel durchaus ändern, insofern neue Stories dem Thema zugeordnet werden. Ebenfalls über den Zentroiden lässt sich das Abstract zum Topic herstellen (Radev, Jing, Stys, & Tam 2004). Die Berichterstattung über Ereignisse erfolgt – je nach internationalem Interesse – in mehreren Sprachen. Möchte man über Sprachgrenzen hinweg erkannte Themen verfolgen, so ist dies eine Aufgabe für die multilinguale Themenverfolgung (Levow & Oard 2002; Schultz & Liberman 2002; Ma, Yang, & Rogati 2005). Larkey, Feng, Connell und Lavrenko (2003) arbeiten mit automatischer Übersetzung, die jedoch nicht zu zufriedenstellenden Resultaten führt. Für den einsprachigen Fall liegt ein breites Wissens des TDT-Einsatzes in der englischen Sprache vor. Über andere Sprachen wird mit Ausnahmen zu experimentellen Ansätzen in Chinesisch (Wayne 2000; Chen & Ku 2002; Cieri, Strassel, Graff, Martey, Rennert, & Liberman 2002), Hindi (Allan, Lavrenko, & Connell 2003) und Spanisch (Yang, Carbonell, Brown, Lafferty, Pierce, & Ault 2002) kaum etwas berichtet.

Verfügt ein Thema über mehrere Stories, so müssen diese in eine Rangfolge gebracht werden. Curtiss, Bharat und Schmitt verfolgen in einer Patentanmeldung von Google den Weg, Qualitätskriterien für die jeweiligen Quellen zu entwickeln:

> *„(T)he group of metrics may include the number of articles produced by the news source during a given time period, an average length of an article from the news source, the importance of coverage from the news source, a breaking news score, usage patterns, human opinions, circulation statistics, the size of the staff associated with the news source, the number of news bureaus associated with the news source, the number of original named entities the source news produces within a cluster of articles, the breath of coverage, international diversity, writing style, and the like" (Cutiss, Bharat, & Schmitt 2003, 3).*

Umfasst ein TDT-System alle einschlägigen Quellen, so liegt es nahe, der ersten Story die „Ehre" zu erweisen, prominent an der Top-Stelle genannt zu werden. Die übrigen Stories können dann durchaus den Google News-Kriterien gemäß geordnet werden.

Verfügt eine Story über mehrere Themen, kann sie also unterschiedlichen Topics zugeordnet werden, so steht TDT vor einer Herausforderung (Chali 2005). Eine sinnvolle Lösung ist, die Story allen erkannten Topics einzuordnen, wobei das „Erkennen" eine Frage der Einstellung der Grenzwerte bei der Ähnlichkeit sowie dem Zeit- und Raumbezug ist. Hängen Themen hierarchisch zusammen, insofern ein Topic einen oder mehrere Sub-Topic(s) hat, so liegt es nahe, die Themen auch hierarchisch anzuzeigen (Pons-Porrata, Berlanga-Llavori, & Ruiz-Shulcloper 2003).

Fazit

- Bei der Themenentdeckung und -verfolgung analysiert man den Strom an Nachrichten aus WWW, Deep Web (vor allem Datenbanken der Nachrichtenagenturen und Zeitungen) sowie Rundfunk (Hörfunk wie Fernsehen). Analog kann mit Beiträgen in Weblogs vorgegangen werden. Ziele sind, (1) neue Ereignisse zu identifizieren und (2) Stories zu bereits erkannten Themen zuzuordnen.
- Themenentdeckung und -verfolgung wird bereits (bezogen auf kostenlose Webseiten) als Profildienst für Neues von Nachrichtensuchmaschinen (z.B. Google News) angeboten.
- Bei der Entdeckung eines neuen Themas errechnet man die Ähnlichkeit zwischen einer aktuellen Story und allen bereits gespeicherten Stories in der Datenbank. Ergibt sich keine Übereinstimmung, so wird die Story als erster Repräsentant eines neuen Themas eingeführt. Ergeben sich Ähnlichkeiten mit gespeicherten Stories, so wird die Ähnlichkeit der aktuellen Story mit bekannten Themen errechnet und die Story einem Thema zugeordnet.
- Zur konkreten Berechnung von Ähnlichkeiten zwischen Stories sowie zwischen Story und Thema bieten sich TF*IDF sowie das Vektorraummodell an. Ein Thema wird durch den Zentroiden aller Stories, die das jeweilige Ereignis besprechen, dargestellt.
- In Nachrichten spielen „named entities" eine große Rolle. Es erweist sich als sinnvoll, pro Dokument zwei Vektoren zu bestimmen, einen für Eigennamen und einen für die „topic terms". Nur wenn bei beiden Vektoren Ähnlichkeiten mit anderen Stories festzustellen sind, dürfte die aktuelle Story zu einem bekannten Thema gehören. Zusätzlich ist es nötig, konkrete Zeit- und Raumbezüge in Stories als diskriminierende Merkmale heranzuziehen.
- Titel wie Abstract des Topic lassen sich aus dem Term- bzw. Satzmaterial des Zentroiden, dem Durchschnittsvektor aller Stories des Topic, automatisch erstellen. Alternativ kann man mit Titel und erstem Abschnitt der ersten Story arbeiten.
- Liegen zu einem Thema mehrere Stories vor, so werden diese in eine Rangfolge gebracht. An erster Stelle sollte diejenige Story gelistet werden, die als erstes über das Ereignis berichtet hat. Danach kann das Ranking anhand von Qualitätskriterien der Quellen gebildet werden.
- Kommerzielle Informationsanbieter für Nachrichten (wie ASV Infopool, GBI-GENIOS, Dialog NewsEdge, Factiva oder LexisNexis) können von Methoden der Themenentdeckung und -verfolgung profitieren. Hier werden *erstens* personalisierte Pushdienste zu neuen Themen, die das Informationsprofil eines Nutzers befriedigen, sowie *zweitens* Klassierungsoptionen für Dokumente in einer Trefferliste nach Themen (unabhängig von Dokumentationssprachen) möglich.

Literatur

Allan, J. (2002a): Introduction to topic detection and tracking. – In: Allan, J., Hrsg.: Topic Detection and Tracking: Event-based Information Organization. – Boston: Kluwer, S. 1-16.

Allan, J. (2002b): Detection as multi-topic tracking. – In: Information Retrieval 5, S. 139-157.

Allan, J. (2003): Robust techniques for organizing and retrieving spoken documents. – In: EURASIP Journal of Applied Signal Processing Nr. 2, S. 103-114.

Allan, J.; *Carbonell*, J.; *Doddington*, G.; *Yamron*, J.; *Yang*, Y. (1998): Topic detection and tracking pilot study. Final report. – In: Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, S. 194-218.

Allan, J.; *Feng*, A.; *Bolivar*, A. (2003): Flexible intrinsic evaluation of hierarchical clustering for TDT. – In: Proceedings of the 12th International Conference on Information and Knowledge Management. – New York: ACM, S. 263-270.

Allan, J.; *Harding*, S.; *Fisher*, D.; *Bolivar*, A.; *Guzman-Lara*, S.; *Amstutz*, P. (2005): Taking topic detection from evaluation to practice. – In: Proceedings of the 38th Annual Hawaii International Conference on System Sciences.

Allan, J.; *Lavrenko*, V.; *Connell*, M.E. (2003): A month of topic detection and tracking in Hindi. – In: ACM Transactions on Asian Language Information Processing 2(2), S. 85-100.

Allan, J.; *Lavrenko*, V.; *Jin*, H. (2000): First story detection in TDT is hard. – In: Proceedings of the 9th International Conference on Information and Knowledge Management. – New York: ACM, S. 374-381.

Allan, J.; *Larvenko*, V.; *Swan* R. (2002): Explorations within topic tracking and detection. – In: Allan, J., Hrsg.: Topic Detection and Tracking: Event-based Information Organization. – Boston: Kluwer, S. 197-224.

Allan, J.; *Papka*, R.; *Lavrenko*, V. (1998): On-line new event detection and tracking. – In: Proceedings of the 21th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. – New York: ACM, S. 37-45.

Allan, J.; *Wade*, C.; *Bolivar*, A. (2003): Retrieval and novelty detection at the sentence level. – In: Proceedings of the 26th International ACM SIGIR Conference on Research and Development in Information Retrieval. – New York: ACM, S. 314-321.

Bharat, K. (2003): Patterns on the Web. – In: Lecture Notes in Computer Science 2857, S. 1-15.

Brants, T.; *Chen*, F. (2003): A system for new event detection. – In: Proceedings of the 26th International ACM SIGIR Conference on Research and Development in Information Retrieval. – New York: ACM, S. 330-337.

Chali, Y. (2005): Topic detection of unrestricted texts: Approaches and evaluations. – In: Applied Artificial Intelligence 19(2), S. 119-136.

Chen, H.H.; *Ku*, L.W. (2002): An NLP and IR approach to topic detection. – In: Allan, J., Hrsg.: Topic Detection and Tracking: Event-based Information Organization. – Boston: Kluwer, S. 243-264.

Cieri, C.; *Strassel*, S.; *Graff*, D.; *Martey*, N.; *Rennert*, K.; *Liberman*, M. (2002): Corpora for topic detection and tracking. – In: Allan, J., Hrsg.: Topic Detection and Tracking: Event-based Information Organization. – Boston: Kluwer, S. 33-66.

Curtiss, M.; *Bharat*, K.; *Schmitt*, M. (2003): Systems and methods for improving the ranking of news articles. Patentanmeldung Nr. US 2005/0060312 A1. – (Eingereicht am: 16.9.2003).

Del Vorso, G.M.; *Gulli*, A.; *Romani*, F. (2005): Ranking a stream of news. – In: Proceedings of the 14th International World Wide Web Conference. – New York: ACM, S. 97-106.

Feng, A.; *Allan*, J. (2005): Hierarchical topic detection in TDT-2004. Technical Report. – Center for Intelligent Information Retrieval. University of Massachusetts, Amherst.

Fiscus, J.G.; *Doddington*, G.R. (2002): Topic detection and tracking evaluation overview. – In: Allan, J., Hrsg.: Topic Detection and Tracking: Event-based Information Organization. – Boston: Kluwer, S. 17-31.

Flynn, C.; *Dunnion*, J. (2004): Event clustering in the news domain. – In: Lecture Notes in Computer Science 3206, S. 65-72.

Franz, M.; *Ward*, T.; *McCarley*, J.S.; *Zhu*, W.J. (2001): Unsupervised and supervised clustering for topic tracking. – In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. – New York: ACM, S. 310-317.

Frey, D.; *Gupta*, R.; *Khandelwal*, V.; *Lavrenko*, V.; *Leuski*, A.; *Allan*, J. (2001): Monitoring the news: A TDT demonstration system. – In: Proceedings of the 1st International Conference on Human Language Technology Research. – Morristown, NJ: Association for Computational Linguistics, S. 1-5.

Fukumoto, F.; *Suzuki*, Y. (2000): Extracting key paragraph based on topic and event detection: towards summarization. – In: NAACL-ANLP 2000 Workshop on Automatic Summarization – Vol. 4. – Morristown, NJ: Association for Computational Linguistics, S. 31-39.

Jin, Y.; *Myaeng*, S.H.; *Lee*, M.H.; *Oh*, H.J.; Jang, M.G. (2005): Effective use of place information for event tracking. – In: Lecture Notes in Computer Science 3689, S. 410-422.

Jones, G.J.F.; *Gabb*, S.M. (2002): A visualization tool for topic tracking analysis and development. – In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. – New York: ACM, S. 389-390.

Kim, P.; *Myaeng*, S.H. (2004): Usefulness of temporal information automatically extracted from news articles for topic tracking. – In: ACM Transactions on Asian Language Information Processing 3(4), S. 227-242.

Kumaran, G.; *Allan*, J. (2005): Using names and topics for new event detection. – In: Proceedings of Human Language Technology Conference/ Conference on Empirical Methods in Natural Language Processing, Vancouver, S. 121-128.

Kurtz, A.J.; *Mostafa*, J. (2003): Topic detection and interest tracking in a dynamic online news source. – In: Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries. – Washington, DC: IEEE Computer Society, S. 122-124.

Larkey, L.S.; *Feng*, F.; *Connell*, M.; *Lavrenko*, V. (2004): Language-specific models in multilingual topic tracking. – In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. – New York: ACM, S. 402-409.

Lavrenko, V.; *Allan*, J.; *DeGuzman*, E.; *LaFlamme*, D.; *Pollard*, V.; *Thomas*, S. (2002): Relevance models for topic detection and tracking. – In: HTL San Diego, S. 104-110.

Leek, T.; *Schwartz*, R.; *Sista*, S. (2002): Probabilistic approaches to topic detection and tracking. – In: Allan, J., Hrsg.: Topic Detection and Tracking: Event-based Information Organization. – Boston: Kluwer, S. 67-83.

Leuski, A.; *Allan*, J. (2002): Improving realism of topic tracking evaluation. – In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development on Information Retrieval. – New York: ACM, S. 89-96.

Levow, G.A.; *Oard*, D.W. (2002): Signal boosting for translingual topic tracking. – In: Allan, J., Hrsg.: Topic Detection and Tracking: Event-based Information Organization. – Boston: Kluwer, S. 175-194.

Li, B.; *Li*, W; *Lu*, Q. (2006): Enhancing topic tracking with temporal information. – In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. – New York: ACM, S. 667-668.

Ma, N.; *Yang*, Y.; *Rogati*, M. (2005): Applying CLIR techniques to event tracking. – In: Lecture Notes in Computer Science 3411, S. 24-35.

Makkonen, J. (2003): Investigations on event evolution in TDT. – In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Proceedings of the HLT-NAACL 2003 Student Research Workshop – Vol. 3. – Morristown, NJ: Association for Computational Linguistics, S. 43-48.

Makkonen, J.; *Ahonen-Myka*, H. (2003): Utilizing temporal information in topic detection and tracking. – In: Lecture Notes in Computer Science 2769, S. 393-404.

Makkonen, J.; *Ahonen-Myka*, H.; *Salmenkivi*, M. (2004): Simple semantics in topic detection and tracking. – In: Information Retrieval 7, S. 347-368.

Makkonen, J.; *Ahonen-Myka*, H.; *Salmenkivi*, M. (2003): Topic detection and tracking with spatio-temporal evidence. – In: Lecture Notes in Computer Science 2633, S. 251-265.

Manmatha, R.; *Feng*, A.; *Allan*, J. (2002): A critical examination of TDT's cost function. – In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development on Information Retrieval. – New York: ACM, S. 403-404.

Mayer, Y. (2004): System and method for improved searching on the Internet or similar networks and especially improved metanews and/or improved automatically generated newspapers. Patentanmeldung US 2005/ 0114324 A1. – (Eingereicht am: 14.9.2004).

McCarley, J.S.; Franz, M. (2000): Influence of speech recognition errors on topic detection. – In: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. – New York: ACM, S. 342-344.

Nallapati, R. (2003): Semantic language models for topic detection and tracking. – In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Proceedings of the HLT-NAACL 2003 Student Research Workshop – Vol. 3. – Morristown, NJ: Association for Computational Linguistics, S. 1-6.

Otterbacher, J.; *Radev*, D. (2006): Fact-focused novelty detection: A feasibility study. – In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. – New York: ACM, S. 687-688.

Papka, R.; *Allan*, J. (2000): Topic detection and tracking: Event clustering as a basis for first story detection. – In: Croft, W.B. (Hrsg.): Advances in Information Retrieval. Recent Research from the Center for Intelligent Information Retrieval. – Boston: Kluwer, S. 97-126.

Peters, I.; *Stock*, W.G. (2006): Corporate Blogs im Wissensmanagement. – In: Wissensmanagement Nr. 6, S. 40-41

Pons-Porrata, A.; *Berlanga-Llavori*, R.; *Ruiz-Shulcloper*, J. (2003): Building a hierarchy of events and topics for newspaper digital libraries. – In: Lecture Notes in Computer Science 2633, S. 588-596.

Radev, D.; *Jing*, H.; *Stys*, M; *Tam*, D. (2004): Centroid-based summarization of multiple documents. – In: Information Processing & Management 40, S. 919-938.

Schultz, J.M.; *Liberman*, M.Y. (2002): Towards a „Universal Dictionary" for multi-language information retrieval applications. – In: Allan, J., Hrsg.: Topic Detection and Tracking: Event-based Information Organization. – Boston: Kluwer, S. 225-241.

Stock, M. (2002a): Factiva.com: Neuigkeiten auf der Spur. Searches, Tracks und News Pages bei Factiva. – In: Password Nr. 5, S. 31-40.

Stock, M. (2002b): ASV Infopool. Boulevard online. – In: Password Nr. 10, 22-27.

Stock, M; *Stock*, W.G. (2003a): GBI – the content-machine: Wirtschaftsinformationen für Hochschulen, Unternehmen und Internetsurfer. – In: Password Nr. 2, 8-17.

Stock, M.; *Stock*, W.G. (2003b): Von Factiva.com zu Factiva Fusion: Globalität und Einheitlichkeit mit Integrationslösungen – auf dem Weg zum Wissensmanagement. – In: Password Nr. 3, S. 19-28.

Stock, M; *Stock*, W.G. (2003c): Dialog Profound / NewsEdge: Dialogs Spezialmärkte für Marktforschung und News. – In: Password Nr. 5, 42-49.

Stock, M; *Stock*, W.G. (2003d): GENIOS Wirtschaftsdatenbanken: Bündelung deutscher und internationaler Informationen als Wettbewerbsvorteil. – In: Password Nr. 6, 14-22.

Stock, M; *Stock*, W.G. (2005): Digitale Rechts- und Wirtschaftsinformationen bei LexisNexis. – In: JurPC. Zeitschrift für Rechtsinformatik, Web-Dok. 82/2005, Abs. 1-105.

Stock, W.G. (2007): Information Retrieval. – München; Wien: Oldenbourg.

Wayne, C.L. (2000): Topic detection and tracking in English and Chinese. – In: Proceedings of the 5th International Workshop on Information Retrieval with Asian Languages. – New York: ACM, S. 165-172.

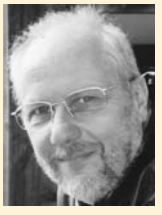
Yang, C.C.; *Shi*, X.; *Wei*, C.P. (2006): Tracing the event evolution on terror attacks from on-line news. – In: Lecture Notes in Computer Science 3975, S. 343-354.

Yang, Y.; *Carbonell*, J.; *Brown*, R.; *Lafferty*, J.; *Pierce*, T.; *Ault*, T. (2002): Multi-strategy learning for topic detection and tracking. – In: Allan, J., Hrsg.: Topic Detection and Tracking: Event-based Information Organization. – Boston: Kluwer, S.85-114.

Themenentdeckung und -verfolgung, TDT, Google News, Nachrichten, Themenentdeckung, Themenverfolgung, elektronischer Informationsdienst, LexisNexis, Profildienst

Der Autor

Prof. Dr. Wolfgang G. Stock



ist Leiter der Abteilung für Informationswissenschaft der Heinrich-Heine-Universität Düsseldorf. Der Artikel ist eine erweiterte Fassung von Kapitel 25 seiner aktuellen Publikation „Information Retrieval" (Oldenbourg Verlag).

stock@phil-fak.uni-duesseldorf.de
www.phil-fak.uni-duesseldorf.de/infowiss