

Automatische Klassifizierung und Visualisierung im Archiv der Süddeutschen Zeitung

MARKUS SCHEK



■ **Markus Schek**
Süddeutscher Verlag Mediengruppe
DIZ München GmbH – Informationstechnik
Sendlinger Str. 8
D-80331 München
markus.schek@diz-muenchen.de

1 Einleitung

Die Süddeutsche Zeitung (SZ) verfügt seit ihrer Gründung 1945 über ein Pressearchive, das die Texte der eigenen Redakteure und zahlreicher nationaler und internationaler Publikationen dokumentiert und auf Anfrage für Recherchezwecke bereitstellt. Die Einführung der EDV begann Anfang der 90er Jahre mit der digitalen Speicherung zunächst der SZ-Daten. Die technische Weiterentwicklung ab Mitte der 90er Jahre diente zwei Zielen: (1) dem vollständigen Wechsel von der Papierablage zur digitalen Speicherung und (2) dem Wandel von einer verlagsinternen Dokumentations- und Auskunftsstelle zu einem auch auf dem Markt vertretenen Informationsdienstleister. Um die dabei entstehenden Aufwände zu verteilen und gleichzeitig Synergieeffekte zwischen inhaltlich verwandten Archiven zu erschließen, gründeten der Süddeutsche Verlag und der Bayerische Rundfunk im Jahr 1998 die Dokumentations- und Informationszentrum (DIZ) München GmbH, in der die Pressearchive der beiden Gesellschafter und das Bildarchiv des Süddeutschen Verlags zusammengeführt wurden. Die gemeinsam entwickelte Pressedatenbank ermöglichte das standortübergreifende Lektorat, die browserbasierte Recherche für Redakteure und externe Kunden im Intra- und Internet und die kundenspezifischen Content Feeds für Verlage, Rundfunkanstalten und Portale. Die DIZ-Pressedatenbank enthält zur Zeit 6,9 Millionen Artikel, die jeweils als HTML oder PDF abrufbar sind. Täglich kommen ca. 3.500 Artikel hinzu, von denen ca. 1.000 lektoriert werden.

Das Lektorat erfolgt im DIZ nicht durch die Vergabe von Schlagwörtern am Dokument, sondern durch die Verlinkung der Artikel mit "virtuellen Mappen", den Dossiers. Diese stellen die elektronische Repräsentation einer Papiermappe dar und sind das zentrale Erschließungsobjekt. Im Gegensatz zu statischen Klassifikationssystemen ist die Dossierstruktur dynamisch und auf-

kommensabhängig, d.h. neue Dossiers werden hauptsächlich anhand der aktuellen Berichterstattung erstellt. Insgesamt enthält die DIZ-Pressedatenbank ca. 90.000 Dossiers, davon sind 68.000 Sachthemen (Topics), Personen und Institutionen. Die Dossiers sind untereinander zum "DIZ-Wissensnetz" verlinkt.

DIZ definiert das Wissensnetz als Alleinstellungsmerkmal und wendet beträchtliche personelle Ressourcen für die Aktualisierung und Qualitätssicherung der Dossiers auf. Nach der Umstellung auf den komplett digitalisierten Workflow im April 2001 identifizierte DIZ vier Ansatzpunkte, wie die Aufwände auf der Inputseite (Lektorat) zu optimieren sind und gleichzeitig auf der Outputseite (Recherche) das Wissensnetz besser zu vermarkten ist:

1. (Teil-)Automatische Klassifizierung von Presstexten (Vorschlagwesen)
2. Visualisierung des Wissensnetzes (Topic Mapping)
3. (Voll-)Automatische Klassifizierung und Optimierung des Wissensnetzes
4. Neue Retrievalmöglichkeiten (Clustering, Konzeptsuche)

Die Projekte 1 und 2 "Automatische Klassifizierung und Visualisierung" starteten zuerst und wurden beschleunigt durch zwei Entwicklungen:

- Der Bayerische Rundfunk (BR), ursprünglich Mitbegründer und 50%-Gesellschafter der DIZ München GmbH, entschloss sich aus strategischen Gründen, zum Ende 2003 aus der Kooperation auszusteigen.
- Die Medienkrise, hervorgerufen durch den massiven Rückgang der Anzeigenerlöse, erforderte auch im Süddeutschen Verlag massive Einsparungen und die Suche nach neuen Erlösquellen.

Beides führte dazu, dass die Kapazitäten im Bereich Pressedokumentation von ursprünglich rund 20 (nur SZ, ohne BR-Anteil) auf rund 13 zum 1. Januar 2004 sanken und gleichzeitig die Aufwände für die Pflege des Wissensnetzes unter verstärktem Rechtfertigungsdruck gerieten. Für die Projekte 1 und 2 ergaben sich daraus drei quantitative und qualitative Ziele:

- Produktivitätssteigerung im Lektorat
- Konsistenzverbesserung im Lektorat

- Bessere Vermarktung und intensivere Nutzung der Dossiers in der Recherche

Alle drei genannten Ziele konnten erreicht werden, wobei insbesondere die Produktivität im Lektorat gestiegen ist. Die Projekte 1 und 2 "Automatische Klassifizierung und Visualisierung" sind seit Anfang 2004 erfolgreich abgeschlossen. Die Folgeprojekte 3 und 4 laufen seit Mitte 2004 und sollen bis Mitte 2005 abgeschlossen sein.

Im folgenden wird in Abschnitt 2 die Produktauswahl und Arbeitsweise der Automatischen Klassifizierung beschrieben. Abschnitt 3 schildert den Einsatz der Wissensnetz-Visualisierung in Lektorat und Recherche. Abschnitt 4 fasst die Ergebnisse der Projekte 1 und 2 zusammen und gibt einen Ausblick auf die Ziele der Projekte 3 und 4.

2 Automatische Klassifizierung

Nach einem mehrstufigen Auswahlverfahren beauftragte DIZ Anfang 2003 vier Firmen, einen Prototypen zur überwachten (halb-)automatischen Klassifizierung von Presstexten zu erstellen: Xtramind (Saarbrücken), Amenotec (Bocholt), Temis (Heidelberg) und die Partnerfirmen brainbot (Mainz) / intelligent views (Darmstadt).

Ausgangspunkt für die automatisch zu generierenden Lektoratsvorschläge war das DIZ-Wissensnetz. Den Anbietern wurden deshalb mehrere Jahrgänge lektorierte SZ-Artikel sowie alle Dossiers für ein initiales Training der Prototypen zur Verfügung gestellt. Für die laufende Aktualisierung der Trainingsmenge (Retraining/Update) wurden tagesaktuell die lektorierten SZ-Artikel und alle geänderten Dossiers bereitgestellt.

Jeder Prototyp wurde mindestens 2 Wochen mit Daten aus der laufenden Produktion getestet. Im täglichen Lektorat wurden die Vorschläge durch Dokumentare geprüft und daraufhin übernommen, verworfen oder durch weitere Klassifizierungen ergänzt. Der Abgleich zwischen den automatisch erzeugten Vorschlägen und den intellektuell vorgenommenen Klassifizierungen ergab Auswertungen über die Qualität der einzelnen Prototypen.

Der Produktivtest musste für die einzelnen Prototypen zeitlich nacheinander ablaufen. Die Ergebnisse wurden durch die täglich wechselnde Nachrichtenlage und unerwartet auftretende Großereignisse (wie z.B. den Irak-Krieg 2003) beeinflusst. Ergänzend zum Produktivtest wurde deshalb ein Objektivtest gefahren, d.h. ein einheitlicher Testkorpus von unlektorierten Artikeln aus einem 8-Wochen-Zeitraum vor Beginn der Produktivtests wurde automatisch klassifiziert und die Vorschläge direkt



Schlüsselbegriffe

Automatische Klassifizierung – Topic Mapping – Wissensnetz – Clustering – Ähnlichkeitsanalyse – Visualisierung

mit den – unabhängig vom Test – manuell vergebenen Klassifizierungen verglichen.

Die Funktionalitäts-Bewertung der Prototypen erfolgte nach den Kriterien Performance, Robustheit und Fehlerbehebung, für die Bewertung der Leistungsfähigkeit wurden Effizienz (Kosten, Zeitaufwand, Speicherkapazität etc.) und Effektivität (die Fähigkeit möglichst viele relevante Informationseinheiten zu liefern und unbrauchbare auszusortieren) ermittelt. Die Effektivität wurde bei der Anbieterauswahl am stärksten gewichtet.

Die wichtigsten Maße für die Effektivitätsbewertung sind Recall und Precision. Recall bezeichnet das Verhältnis der Anzahl der relevanten und gefundenen Informationseinheiten gemessen an der Gesamtzahl der relevanten Informationseinheiten, d.h. vereinfacht ausgedrückt: Wie viele der korrekten (d.h. manuell vergebenen) Klassifizierungen hat die Software vorgeschlagen? Precision gibt das Verhältnis der relevanten und gefundenen Einheiten gemessen an der Gesamtzahl der gefundenen Einheiten an, d.h.: Wie viele der Vorschläge sind korrekt?

Die Gewichtung der Maße hängt von der geplanten Einsatzweise ab:

1. Teilautomatische Klassifizierung als Vorschlagwesen mit Bestätigung der korrekten Vorschläge und Hinzufügen weiterer Klassifizierungen durch Fachpersonal: Recall
2. Teilautomatische Klassifizierung als Vorschlagwesen mit Entfernen fehlerhafter Vorschläge durch Fachpersonal: Precision
3. Vollautomatische Klassifizierung mit Vergabe der (reduzierten) Klassifizierung ohne fachliche Nachprüfung bzw. lediglich Stichprobenkontrolle: Precision

DIZ plante zunächst nur Einsatzweise 1. Die Prototypen waren dementsprechend Recall-optimiert, die Recall-Auswertungen wurden am stärksten gewichtet. Mit Blick auf eine weitere Automatisierung wurde parallel die Precision ausgewertet, allerdings geringer gewichtet. Auf weitere Auswertungen (z.B. nach Accuracy, Error-Maß, Fallout etc.) musste angesichts der strengen Zeit- und Budgetvorgaben des Projekts gemäß Kosten-Nutzen-Analyse verzichtet werden. Die besten Recall- und Precision-Werte wurden vom Produkt der Firmen intelligent views und brainbot erreicht. Die Software wurde im Dezember 2003 bei DIZ installiert und abgenommen und ging im Januar 2004 in Produktion.

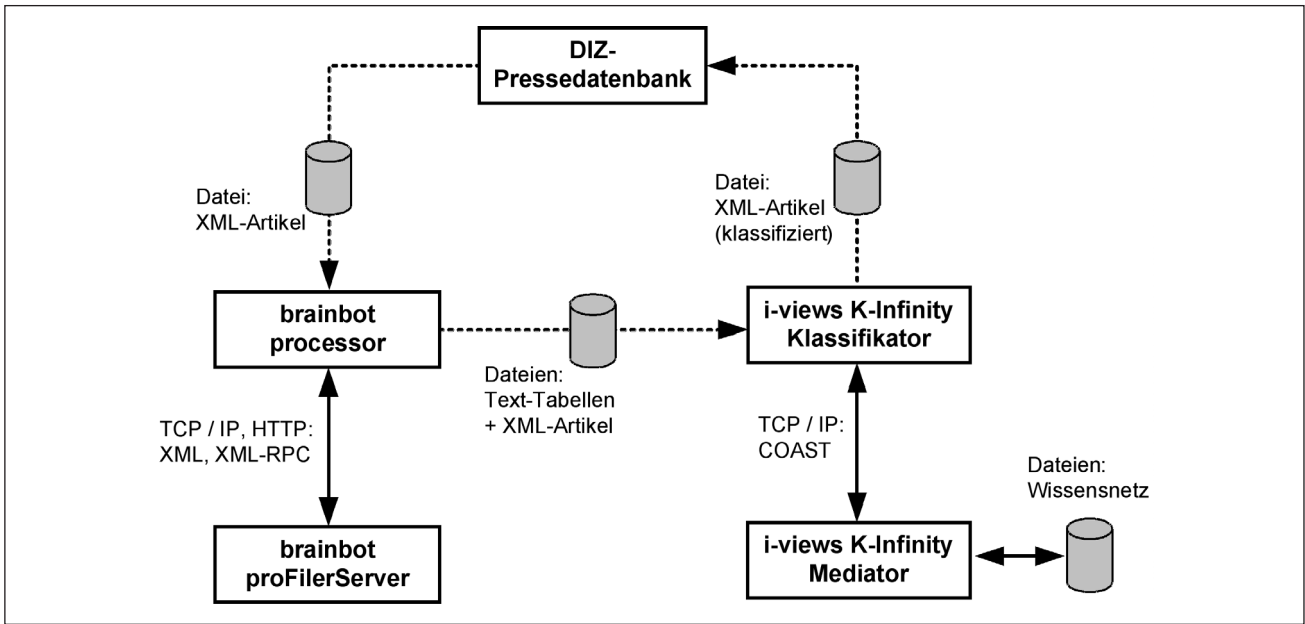


Abbildung 1: Ablauf Automatische Klassifizierung

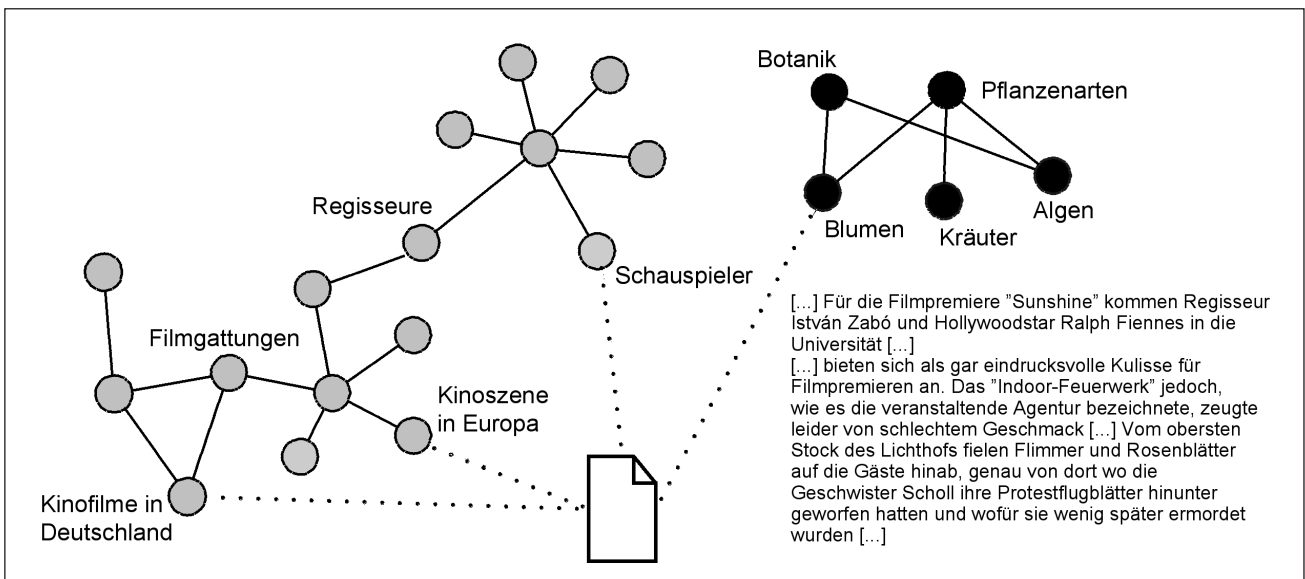


Abbildung 2: Cluster-Analyse

Die Automatische Klassifizierung bei DIZ verknüpft Komponenten der Firmen brainbot und intelligent views (Abbildung 1). Die gesamte Verarbeitung ist mittels einer Reihe von Schwellenwerten (z.B. für die Mindestgröße des Texts, die maximale Anzahl der Vorschläge, das Ranking des jeweiligen Vorschlags) frei parametrierbar.

In der ersten Verarbeitungsstufe werden durch den brainbot processor mittels statistischer Verfahren die in einer Datei übergebenen Textdaten analysiert und Klassifizierungsvorschläge erzeugt.¹ In der zweiten Verarbeitungsstufe werden durch den intelligent

views-K-Infinity-Klassifikator die Vorschläge mit dem Wissensnetz abgeglichen, d.h. die Vorschläge werden innerhalb ihres systematischen Umfelds mit allen horizontal und vertikal nahe verbundenen Dossiers analysiert. Im Falle von völlig unverbundenen Themenclustern ergibt sich dadurch die Möglichkeit, das relativ am weitesten entfernte bzw. isolierteste Thema aus der Vorschlagsliste zu entfernen. Im unten gezeigten Beispiel (Abbildung 2) würde die fehlende Verbindung zwischen dem Themencluster "Botanik" und den Themenclustern "Film" eine Überarbeitung der Vorschlagsliste in Bezug auf "Botanik" auslösen.

Die Ergebnisse der zweiten Verarbeitungsstufe werden als Dossier-Vorschläge – zusammen mit einem Ranking-Wert für die vermutete Relevanz – in die Artikel-

¹ Die interne Hypothese, dass eine Kombination aus statistischen und linguistischen Verfahren die besten Ergebnisse liefern würde, wurde durch die Prototypen-Evaluation falsifiziert. Das rein statistisch arbeitende System von brainbot lag auch ohne die intelligent views-Komponente deutlich vor dem Nächstplatzierten, einer Anwendung mit Statistik-/Linguistik-Kombination.

The screenshot displays the DIZ-Pressdatenbank interface. At the top, there are navigation tabs: 'Navigation', 'Wissensnetz', 'Recherche', 'Erfassung', 'Ablage', 'Extras', and 'Hilfe'. Below these are search options: 'Artikel-Schnellsuche', 'Erweiterte Suche', 'Experten-Suche', 'Dossier-Suche', and 'Recherche-Profile'. The main content area shows a search for 'EU-Erweiterung' with a 'Suchen' button. Below the search bar, there are filters for 'Topic', 'Person', 'Institution', 'Media', 'Critics', and 'Detail'. The central part of the screen features a knowledge network visualization for 'Erweiterung der EU'. A dialog box is open over the central node, offering options: 'Artikel anzeigen', 'in die Ablage legen', and 'ins Lektorat übernehmen'. The network includes nodes for various EU expansion topics, such as 'Erweiterung der EU 2004', 'Erweiterung der EU 2003', 'Erweiterung der EU 2002', 'Erweiterung der EU 2001', 'Erweiterung der EU 1992 - 2000', 'Finanzierung der EU-Erweiterung', 'Auswirkungen der EU-Erweiterung auf die Agrarwirtschaft / Wirtschaft', 'Auswirkungen der EU-Erweiterung auf Deutschland', 'Beitritt der neuen EU-Mitgliedsstaaten zur EWU', 'Freizügigkeit für osteuropäische Arbeitnehmer in der EU', 'Außenpolitik der EU', 'Stabilitätspakt für den Balkan', 'Gemeinsame Außen- und Sicherheitspolitik der EU', 'Europäische Union (EU)', 'taz-Serie Armes Europa - Reiches Europa', 'Stern-Serie Das neue Europa', 'Zeit-Serie EU-Osterweiterung', and 'SZ-Serie Europas Mitte'. The bottom of the interface shows a copyright notice: 'Copyright (c) 2004 by Intelligent Views' and a status bar with 'Appllet nn started' and 'Lokales Intranet'.

Abbildung 3: Wissensnetz der DIZ-Pressdatenbank

Datei geschrieben und im Lektoratsdialog der DIZ-Pressdatenbank angezeigt. Die Dokumentare bestätigen, verwerfen oder ergänzen die Vorschläge des Systems.

3 Visualisierung

Das Projekt 2 "Visualisierung des DIZ-Wissensnetzes" wurde parallel zum Projekt 1 "Automatische Klassifizierung" realisiert. Das überzeugendste Angebot wurde von der Firma intelligent views vorgelegt. Die Visualisierungs-Software ist innerhalb der DIZ-Pressdatenbank seit Februar 2004 für die Dokumentation frei geschaltet und wurde seitdem sukzessive weiteren Nutzergruppen (Redakteure, externe Kunden) zugänglich gemacht.

Im Lektorat bildet die Automatische Klassifizierung zusammen mit der Visualisierung einen integrierten Workflow. Jeder Dossier-Vorschlag kann bei Bedarf (d.h. wenn der Vorschlag nicht direkt übernommen wird bzw. überprüft werden soll) als Einsprungstelle in die Visualisierung dienen. Das vorgeschlagene Dossier wird durch die Visualisierung innerhalb seines systematischen Umfelds angezeigt und zum Lektorat angeboten. Der

Dokumentar erkennt alternative Zuordnungsmöglichkeiten und kann diese per Mausklick übernehmen (Abbildung 3). Falls weitere Dossiers zugeordnet werden sollen, die nicht in der Vorschlagsliste enthalten sind bzw. mit Weiternavigation nicht erreicht werden, kann über die Suchfunktion der Visualisierung im Dossierbestand recherchiert werden und direkt aus der Dossier-Trefferliste heraus lektoriert werden.

DIZ definiert das Wissensnetz als Alleinstellungsmerkmal. Durch Kundenbefragungen hat sich herausgestellt, dass zwar die Nutzer von den Recherchemöglichkeiten mit Dossiers begeistert sind, aber nur eine Minderheit der Kunden die Dossierrecherche überhaupt nutzt. Ursächlich dafür ist die Komplexität des Wissensnetzes – der Aufwand für Schulung und Einarbeitung bildet eine Einstiegshürde für die Recherchekunden. Durch die Visualisierung sollen die Vorteile und Strukturen des Wissensnetzes als machtvolle Ergänzung der klassischen Volltextrecherche unmittelbar verdeutlicht und direkt nutzbar gemacht werden. Leitmotiv aller Bemühungen ist dabei, eine intuitive Navigation und assoziative Recherche zu ermöglichen.

4 Ergebnisse

Die Ziele des Projekts "Automatische Klassifizierung und Visualisierung des Wissensnetzes" wurden erreicht.

Die Wissensnetz-Visualisierung erleichtert laut Rückmeldungen aus der SZ-Redaktion und von externen Kunden die Nutzung der Dossiers. Die Nutzungshäufigkeit steigt – ausgehend von einem niedrigen Niveau – langsam aber kontinuierlich an.

Das Ziel der Qualitätsverbesserung im Lektorat ist statistisch kaum erfassbar, wurde laut Rückmeldung aus der Dokumentation aber erreicht, indem häufig Dossiers vorgeschlagen werden, die aufgrund der verschiedenen Wissens- und Erfahrungshorizonte der Dokumentare sonst nicht oder weniger spezifisch gefunden worden wären, d.h. die Konsistenz der Klassifizierungen steigt.

Beim wichtigsten Ziel, der Produktivitätssteigerung im Lektorat, wurden die größten Fortschritte erreicht. Die wesentlichste Kennzahl dafür ist die Steigerung der Produktivität je Lektoratskapazität. Beginnend bei 20% hat sie sich im Laufe des ersten halben Jahres bis auf einen Wert von 45% gesteigert, der seitdem kontinuierlich gehalten wird. Für die positive Entwicklung sind drei Faktoren maßgeblich:

1. Nochmals gesteigerte Qualität der Klassifizierungsvorschläge (Recall) von 69% im Prototypentest auf 75% in der Produktion
2. Zunehmende Routiniertheit der Dokumentare bei der Nutzung des Vorschlagwesens
3. Optimierung des Workflows durch die Wissensnetz-Visualisierung

Die gesteigerte Produktivität im Lektorat ermöglicht DIZ mit stark reduzierten Kapazitäten das Quantitäts- und Qualitätsniveau zu erreichen, welches für die Weiterentwicklung des Wissensnetzes nötig ist.

Aufgrund der positiven Erfahrungen bei der Implementierung der Automatischen Klassifizierung testet die DIZ München GmbH im Rahmen von Projekt 3 die vollautomatische Klassifizierung, indem mit einer Precision-optimierten Klassifizierungs-Software diejenigen Teilbereiche klassifiziert werden, welche Recall-Werte von über 80% erreichen. Ergänzend dazu werden die Strukturen des Wissensnetzes mittels der Visualisierung und automatisierter Verfahren analysiert und erkannte Inkonsistenzen beseitigt. Das Ziel dabei ist, die Qualität der automatischen Klassifizierung weiter zu steigern und dadurch den manuellen Lektoratsaufwand so weit zu reduzieren, dass für bestimmte Teilbereiche nur noch die fehlerhaften Vorschläge entfernt werden müssen oder – im Idealfall – bis auf Stichproben keinerlei fachliche Nachbearbeitung mehr nötig ist.

Projekt 4 greift ebenfalls auf Techniken der automatischen Klassifizierung zurück. Die Klassifizierungssoftware erstellt ihre Vorschläge größtenteils aufgrund einer statistischen Ähnlichkeitsanalyse zwischen dem zu klassifizierenden Text und einem Modell der bereits vorhandenen Klassifizierungen. Diese Ähnlichkeitsanalyse kann als sogenanntes "Clustering" auch in der Recherche eingesetzt werden. Dabei werden Gruppen (Cluster) von ähnlichen Objekten zusammengefasst. Das Clustering kann auf großen Treffermengen nach einer Volltextsuche erfolgen oder als Clustering von Artikeln innerhalb umfangreicherer Dossiers eingesetzt werden. Zwei Vorgehensweisen sind dabei möglich: Das "ungerichtete Clustering" vollzieht eine Ähnlichkeitsanalyse innerhalb beliebiger (und meist zu großer) Treffermengen und ermöglicht dem Nutzer, über ein Relevance Feedback (d.h. Bewertung zutreffender / nicht zutreffender Objekte) die Treffer zu verfeinern und zu strukturieren. Beim "gerichteten Clustering" erfolgt die Ähnlichkeitsanalyse anhand bestehender Strukturen, z.B. entlang des DIZ-Wissensnetzes. Der Nutzer wird auf bereits vorhandene spezifische Dossiers aufmerksam gemacht, die ein weiteres Navigieren erlauben. Zielgruppen für das Clustering sind vor allem Redakteure und externe Kunden mit unspezifischen Fragen, die über die Möglichkeit eines intuitiveren Recherchierens das Wissensnetz stärker nutzen sollen.

Die DIZ München GmbH hat innerhalb eines straffen Zeit- und Kostenrahmens die (halb-)automatische Klassifizierung und die Visualisierung realisiert und arbeitet an der vollautomatischen Klassifizierung und dem Clustering. Dreh- und Angelpunkt dabei ist die Qualitätserwartung der Kunden unter erhöhtem Kostendruck und die Notwendigkeit, bei verminderten Kapazitäten den Lektoratsdurchsatz zu stabilisieren.

Nur wenn die Dokumentations- und Informationsstellen von sich aus die Potentiale der neuen Techniken nutzen, haben sie eine Chance, die damit immer verbundenen Rationalisierungsbegehren in eine für die inhaltliche Qualität wünschenswerte Richtung zu lenken.



Kernpunkte

Im Archiv der Süddeutschen Zeitung werden mit hohem Zeit- und Personalaufwand Presstexte inhaltlich erschlossen und ein Wissensnetz gepflegt. Der Einsatz von Automatischer Klassifizierung und Wissensnetz-Visualisierung hat Rationalisierungspotentiale erschlossen und die Vermarktungschancen der Archivdienstleistungen gesteigert. Darauf aufbauend sollen die Wissensnetz-Strukturen optimiert und Clustering-Verfahren für die Recherche erprobt werden.