Reihe: Telekommunikation @ Mediendienste · Band 14

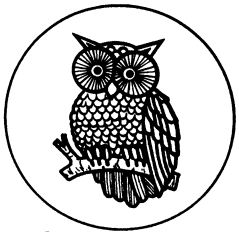
Herausgegeben von Prof. Dr. Dr. h. c. Norbert Szyperski, Köln, Prof. Dr. Udo Winand, Kassel, Prof. Dr. Dietrich Seibt, Köln, Prof. Dr. Rainer Kuhlen, Konstanz, Dr. Rudolf Pospischil, Brüssel, Prof. Dr. Claudia Löbbecke, Köln, und Prof. Dr. Christoph Zacharias, Köln

PD Dr.-Ing. habil. Martin Engelien
Dipl.-Inf. Jens Homann (Hrsg.)

Virtuelle Organisation und Neue Medien 2002

Workshop GeNeMe2002
Gemeinschaften in Neuen Medien

TU Dresden, 26. und 27. September 2002



JOSEF EUL VERLAG
Lohmar · Köln

Die Deutsche Bibliothek – CIP-Einheitsaufnahme

Virtuelle Organisation und Neue Medien 2002 / Workshop GeNeMe 2002 – Gemeinschaften in Neuen Medien – TU Dresden, 26. und 27. September 2002. Hrsg.: Martin Engelien ; Jens Homann. – Lohmar ; Köln : Eul, 2002
(Reihe: Telekommunikation und Medienwirtschaft ; Bd. 14)
ISBN 3-89936-007-9

Josef Eul Verlag GmbH
Brandsberg 6
53797 Lohmar
Tel.: 0 22 05 / 90 10 6-6
Fax: 0 22 05 / 90 10 6-88
http://www.eul-verlag.de
info@eul-verlag.de

Printed in Germany
Druck: RSP Köln

Bei der Herstellung unserer Bücher möchten wir die Umwelt schonen. Dieses Buch ist daher auf säurefreiem, 100% chlorfrei gebleichtem, alterungsbeständigem Papier nach DIN 6738 gedruckt.

Technische Universität Dresden
Fakultät Informatik • Institut für Angewandte Informatik
Privat-Dozentur Angewandte Informatik

PD Dr.-Ing. habil. Martin Engelien

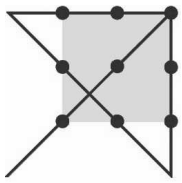
Dipl.-Inf. Jens Homann

(Hrsg.)

an der
Fakultät Informatik der Technischen Universität Dresden

in Zusammenarbeit mit der
Gesellschaft für Informatik e.V.,
GI-Regionalgruppe Dresden

gefördert von der Klaus Tschira Stiftung
gemeinnützige Gesellschaft mit beschränkter Haftung



am 26. und 27. September 2002
in Dresden
http://pdai.inf.tu-dresden.de/geneme
Kontakt: Thomas Müller (geneme@pdai.inf.tu-dresden.de)

F.3. Einsatzmöglichkeiten von Text-Mining zur Unterstützung von internetbasierten Ideenfindungsprozessen

Dirk Krause

Institut für Wirtschaftsinformatik

Universität Leipzig

1. Einleitung

Aktivitäten in Unternehmen und Verwaltungen sind geprägt von Büro- und Projekttätigkeiten, in denen fortlaufend Entscheidungen mit unterschiedlicher Komplexität getroffen werden müssen. Zur Lösung dieser Problemstellungen stehen Methoden, Werkzeuge und Arbeitsanweisungen zur Verfügung, deren Anwendung aber genaue Informationen über Aufgabenstellung und Lösungsweg erfordern. Stehen diese Informationen nicht zur Anwendung bereit, müssen geeignete Wege gefunden werden, diese unstrukturierten Entscheidungsprobleme zu lösen.

Eine Lösungsmethode ist die Konsensbildung durch die Anwendung von Ideenfindungsprozessen in Gruppensitzungen. Mit Hilfe von Netzwerken und geeigneten Technologien können Mitarbeiter und externe Know-How-Träger Entscheidungsprobleme gemeinsam lösen und ihr Informationsdefizit verringern. Zu diesem Zweck können bspw. sitzungsunterstützende Systeme eingesetzt werden, von denen das Werkzeug webSCW ein Vertreter ist. [1]

Mit Ideenfindungsprozessen werden durch vernetzte und verteilte Gruppenarbeit sehr viele Informationen zur Problemlösung gefunden. In nachgelagerten Aktivitäten müssen diese strukturiert und bewertet werden. Diese Vorgänge finden manuell durch die Sitzungsteilnehmer statt. Einige Lösungsvorschläge enthalten jedoch gleiche oder ähnliche Informationen. Mit Hilfe der Methoden des Text-Mining können Ideen automatisch klassifiziert oder zusammengefasst werden. Weitere Nutzeffekte sind automatische Expertenfindung und Summarizing.

In dem Beitrag sollen ausgehend von Problemen traditioneller und computerunterstützter Sitzungen einzelne Methoden des Text-Mining vorgestellt und auf deren Einsatzmöglichkeiten zur Unterstützung von internetbasierten Ideenfindungsprozessen untersucht werden. Aus diesen Überlegungen wird ein Vorgehensmodell abgeleitet, das eine automatische bzw. teilautomatische Weiterbearbeitung von Informationen und Ideen in Textform ermöglicht. Das

Vorgehensmodell wurde erfolgreich im sitzungsunterstützenden System webSCW implementiert und ergänzt die vorhanden Module durch Funktionen der automatischen Informationsgewinnung und -strukturierung. Die Beschreibung der Wirkungsweise und Anwendung dieser Funktionen bildet den Abschluss des Beitrages.

2. Probleme traditioneller und computerunterstützter Sitzungen

In der Vergangenheit entwickelten sich neuen Organisationsformen für Unternehmen, die vor allem gekennzeichnet sind durch Kooperationen zwischen Unternehmen, Organisationen und externen Know-How-Trägern, teilautonome Arbeitsgruppen innerhalb interner Organisationsstrukturen, organisationelles Gedächtnis sowie lernende und flexible Strukturen. Neue Managementkonzepte, flexible Arbeitsorganisation sowie Gruppen- oder Teamarbeit spielen bei der Umsetzung dieser neuen Ansätze eine wichtige Rolle. Ein kritischer Erfolgsfaktor bildet die optimale Nutzung der organisationsinternen und -externen Ressourcen, zu denen auch das Wissen zählt. Die Aktivitäten zur Verwaltung des Wissens werden als Wissensmangement verstanden und durch das Ideenfindungs- sowie Know-How-Management, als Teilbereiche davon, ergänzt. [2]

Tätigkeiten in Unternehmen und Organisationen werden in Projekten strukturiert und sind häufig von folgenden Merkmalen geprägt:

- komplexe Aufgabenstellungen,
- geringe Bearbeitungsdauer einzelner Aktivitäten,
- fehlendes Know-How,
- variierende Teambesetzung,
- externe Kooperationspartner,
- viele (ad hoc) Entscheidungen in kurzer Zeit usw.

Dabei liegen die zu bearbeitenden Aufgabenstellungen in unterschiedlichen Strukturierungsgraden vor und müssen unter Verwendung interner und externer Informationen, Arbeitsanweisungen und Ablaufplänen, IT-Infrastruktur, Software und externen Know-How-Trägern gelöst werden. Für strukturierte Entscheidungsprobleme bieten mathematische, statistische sowie Methoden des Operations Research und der Künstlichen Intelligenz Lösungsvorschriften, während diese für unstrukturierte

Problemstellungen fehlen. Abhilfe bieten Methoden der Gruppenarbeit, um durch Konsens eine Gruppenmeinung abzuleiten, die als Lösungs- bzw. Entscheidungsvorschlag dient. Die Gruppenmeinung kann in Konferenzen und Sitzungen gebildet und durch EMS sowie Kreativ- und Managementtechniken unterstützt werden. Mit dem Einsatz des Internets und dessen Technologien spielen rechnergestützte Problemlösungsprozesse immer mehr eine wichtige Rolle, da sich der Aufwand für Organisation und Durchführung von Gruppensitzungen verringert. Diese können durch internetbasierte EMS zu jeder Zeit von jedem Ort realisiert und auch für Ad-hoc-Entscheidungen eingesetzt werden.

Die einzelnen Aktivitäten in Gruppensitzungen sind Finden, Organisation, Bewerten und Verwaltung von Ideen und Lösungsbausteinen. Der abstrakte, iterative Problemlösungsprozess von unstrukturierten Entscheidungsproblemen durch Ideenfindung ist in Abb. 11 dargestellt.

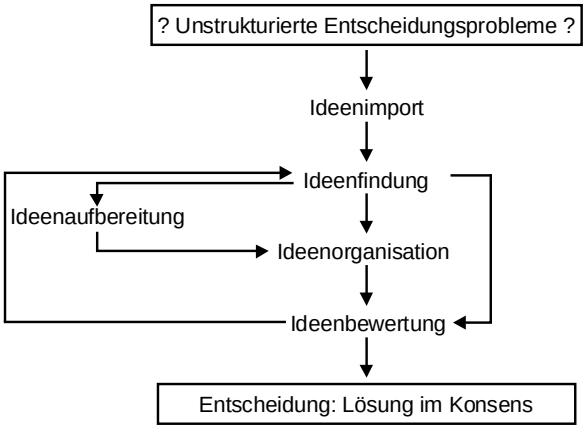


Abb. 1: Problemlösungsprozess durch Ideenfindung

Unterstützend können in den einzelnen Phasen verschiedene Methoden und Hilfsmittel eingesetzt werden, um das Ergebnis zu verbessern. Mit dem Ideenimport, bei der Ideenfindung, -aufbereitung, -organisation und -bewertung werden vorkonfigurierte Lösungsvorschläge angeboten, die als Grundlage für das Ergebnis der jeweiligen Problemlösungsphase dienen. Diese Anregungen können aufbereitet und zugriffsgesteuert aus anderen Ideenfindungsprozessen, Wortschätzen, Begriffsnetzwerken, Dokumentenpools sowie internen und externen Dokumenten stammen. Die Dokumente und Informationen liegen aber in unterschiedlichen Formaten und Strukturierungsgraden vor. Für deren Aufbereitung bieten sich verschiedene Möglichkeiten an, denen bestimmte Methoden zugrunde liegen (vgl. Abschnitt 4).

Mit der Erweiterung der Dimensionen Raum und Zeit in computerunterstützten Sitzungen unter Verwendung von Internettechnologien, steigt auch die Zahl an Problemlösern und Entscheidungsträgern, damit verbunden auch die Zahl der Lösungsvorschläge. Mit der Vielfalt an Lösungsvorschlägen variiert auch die Qualität dieser, wie verschiedene Untersuchungen von Sitzungen am Institut für Wirtschaftsinformatik (IWi) ergeben haben (Tab. 1).

	Personen	Ideen/ Lösungsbausteine	sinnvolle Ideen/ Lösungsbausteine	Detaillierungsgrad der Ideen/ Lösungsbausteine
traditionelle Sitzungen	5-10	30-40	10-20	groß
computerunterstützte Sitzungen	5-20	80-120	40-60	mittel
internetbasierte computerunterstützte Sitzungen	5-... (55)	200-400	100-200	mittel

Tab. 1: Vergleich von Sitzungen am IWi

In traditionellen Sitzungen zur Lösung unstrukturierter Entscheidungsprobleme stand die umfassende Diskussion von wenigen Lösungsvorschlägen im Vordergrund, die von wenigen „Persönlichkeiten“ dominiert wurde. Bedingt durch die freie Diskussion konnte häufig der Zeitplan nicht eingehalten werden.

Mit dem Einsatz von Werkzeugen zur Sitzungsunterstützung konnten Effizienzsteigerungen festgestellt werden. Gründe hierfür lagen in den parallelen Aktivitäten, Optimierung des Organisationsaufwandes, Gleichberechtigung der Beteiligten usw. Aufgrund eines festgelegten Terminplanes mit strikter Überwachung durch das Werkzeug, wurden die Sitzungszeiten fast immer eingehalten. Im Ergebnis entstand ein Problemlösungskatalog mit vielen Vorschlägen, von denen etwa die Hälfte sinnvoll war.

Zeitlich und räumlich verteilte Sitzungen wurden erst durch den Einsatz von Internettechnologien möglich. Mit der gestiegenen Zahl der teilnehmenden Personen stiegen auch die potenziellen Lösungsvorschläge. Die Qualität dieser musste unterschiedlich bewertet werden, da einerseits qualitativ hochwertige externe Ressourcen zur Verfügung standen und andererseits Beiträge sich in schwer kontrollierbare Diskussionen ausweiteten.

In den einzelnen Sitzungsphasen waren unterschiedliche Zeitaufwendungen zu verzeichnen. Mit steigender Teilnehmerzahl stieg der Organisations- und

Bearbeitungsaufwand für nachgelagerte Aktivitäten stark an (Tab. 22), was in der Auswertung bemängelt wurde.

	gesamt	Ideenfindung	Ideenorganisation/ bewertung	Administration während der Sitzung
traditionelle Sitzungen	< 120 min.	< 60 min.	30 - 60 min.	< 5 min.
computerunterstützte Sitzungen	< 90 min.	10 - 20 min.	< 60 min.	< 15 min.
internetbasierte computerunterstützte Sitzungen	< 90 min. ... 150 min. (effektiv)	10 - 20 min.	< 60 ... 120 min.	< 15 min.

Tab. 2: Zeitaufwand für Sitzungen

Die zu verarbeitenden Informationen bieten einen Ansatzpunkt für eine automatische bzw. teilautomatische Weiterverarbeitung mit quantitativen Methoden der Künstlichen Intelligenz, die Gegenstand des nachfolgenden Abschnittes sind. Mit dem Einsatz von Unterstützungsfunktionen kann der Administrationsaufwand von computerunterstützten Sitzungen verringert und die Zufriedenheit der Teilnehmer durch Entlastung von monotonen Aufgaben gesteigert werden

3. Text-Mining

Computerunterstützte Sitzungen bieten viele Vorteile im Gegensatz zu traditionellen Sitzungen. Mit der Möglichkeit, externe Know-How-Träger und Informationsressourcen in die Ideenfindungsprozesse einzubeziehen sowie das Erfahrungswissen abgeschlossener Sitzungen zu nutzen, können viele neue Problemlösungsvorschläge genutzt werden. Mit der Anzahl der Ideen und Lösungsbausteine steigt aber auch der Aufwand zur Aufbereitung und Weiterverarbeitung dieser. Um aus der Vielzahl von semi- und unstrukturierten Textinformationen die für die Lösung relevanten Aussagen und Fakten zu extrahieren, können bspw. Methoden des Text-Mining genutzt werden.

Text-Mining, als Teilgebiet des Data Mining, definiert den Wissensgewinnungsprozess, welcher aus halb- oder unstrukturierten Textdatenbeständen die für den Nutzer interessanten Informationen identifiziert und analysiert. Dieser Vorgang ist sehr komplex und erfordert aufwendige Sprachanalysemethoden, um aus Dokumenten zusammenhängende kontextabhängige Kerninformationen zu extrahieren bzw. weiterzuverarbeiten. [3]

Für das Text-Mining werden verschiedene Methoden und Hilfsmittel, wie statistische und linguistische Verfahren, genutzt. Statistische Verfahren spielen bei der Strukturierung von Texten eine grundlegende Rolle und werden bspw. für die Dokumentenindizierung eingesetzt. Diese Indexverfahren versuchen nicht die tieferliegende Bedeutung eines Wortes zu ermitteln, sondern dienen als semantische Indikatoren. Es werden statistische Maßzahlen der Frequenz des Auftretens eines Terms innerhalb eines Dokumentes bzw. in einer Dokumentensammlung als Anhaltspunkt für eine geringere oder höhere Bedeutung hinsichtlich des Inhalts angesehen. Daraus werden anschließend weiterführende und komplexe Modelle und Methoden zur statistischen Auswertung von Dokumentensammlungen abgeleitet. [4]

Eine Grundidee der Indexierungsverfahren ist im Zipfschen Gesetz beschrieben. Darin wird eine statistische Gesetzmäßigkeit der Sprache über eine konstante Relation C zwischen dem Rang r eines Wortes in einer Häufigkeitsliste und der Frequenz f, mit der es in einem Text vorkommt, nachgewiesen. [5]

Linguistische Verfahren basieren auf der algorithmischen Beschreibung einer Sprache. Dabei wird ein grundlegendes Problem beim Umgang mit einer natürlichen Sprache ersichtlich. Die Aussagen müssen in einem breiten Kontext betrachtet werden, um die richtigen Informationen zu extrahieren. [6] Die Analyse eines Textes erfolgt aus linguistischer Sicht in verschiedenen Ebenen der Textrepräsentation:

- morphologische Ebene,
- lexikalische Ebene,
- syntaktische Ebene,
- semantische Ebene und
- pragmatische Ebene.

Für praktische Anwendungen hat sich eine Kombination der Anwendung verschiedener Ebenen bewährt, um bspw. Mehrdeutigkeiten von Wörtern, wie „Bank“, aufzulösen.

Morphologische Verfahren bestimmen die grammatikalische Funktion eines Satzes, d.h. die Struktur von Wörtern. Dieser Ansatz versucht, Terme nicht als Zeichenketten zu definieren, sondern als bestimmte Formen eines Wortes aufzufassen. Es werden z.B. verschiedene Flexionsformen eines Wortes als zusammengehörig oder sogar als identisch betrachtet. Dabei wird zwischen Grundform- und Stammformreduktion

unterschieden. Die Grundformreduktion führt Wörter auf ihre grammatikalische Grundform zurück. Die Stammformreduktion extrahiert aus den Wortformen den zugehörigen Stamm, der im Allgemeinen keine in der Sprache als Wort vorkommende Form ist und z.B. für ein Verb und ein Substantiv gleich sein kann. Diese Reduktionen werden auch Lemmatisierungen genannt. Sie führen zum einen dazu, dass sehr verschiedene Zeichenketten als gleich angesehen werden. Anderseits werden identische Zeichenketten von verschieden Wortstämmen als verschieden angesehen. Probleme bei der Textanalyse mit morphologischen Verfahren treten bei verschiedenen Sprachen durch die Veränderung von Wortstämmen auf. Diese Sonderformen unterliegen keinem allgemeinen Regelwerk und müssen demzufolge mit Lexika analysiert werden. Vorgehensweisen, die auf Lexika basieren, aber nur für die englische Sprache brauchbare Ergebnisse liefern, sind die lexikografische Grundformenreduktion nach Kuhlen und der Porter-Stemming Algorithmus. [7] [8]

In der lexikalischen Ebene werden einzelne Wörter untersucht. Wörter bilden in vielen Fällen Bedeutungseinheiten, deren Semantik auch ohne Kontextinformationen relativ eingeschränkt ist. Ein Wort ist relativ eindeutig zu identifizieren. Einfache Verfahren sind in der Lage, Texte effizient in seine Worte zu zerlegen. Indexierungsterme als Wortfolgen repräsentieren Texte als Zeichenketten in Sequenzen von Worten. Zusätzlich wird oft die vereinfachende Annahme getroffen, dass die Reihenfolge der Worte vernachlässigt werden kann. Dokumente werden also nicht mehr als Sequenzen von Worten dargestellt, sondern als Multimengen „bags“ von Worten. Diese Repräsentationsform wird deshalb häufig als Bag-of-Words-Ansatz bezeichnet. Die Bag-of-Words-Repräsentation ist konsistent mit der im maschinellen Lernen benutzten Attribut-Wert-Darstellung von Beispielen. Jedes unterschiedliche Wort ist ein Attribut. Der Wert eines Attributs für ein Dokument ist die Anzahl der Vorkommen des entsprechenden Wortes. Die Anzahl wird als Term Frequenz TF(w,d) des Wortes w im Dokument d bezeichnet. Die Repräsentation von Dokumenten als Multimengen ist eine gebräuchliche Technik im Bereich des Information Retrieval. Es ist zu beachten, dass durch die beschriebene Transformation von Text in diese Repräsentation Informationen über das Dokument verloren gehen.

Die syntaktische Ebene untersucht die Struktur von Sätzen. Die Idee besteht darin, Indexierungsterme nicht nur aus einzelnen, sondern aus mehreren Wörtern bestehen zu lassen, die eine bestimmte syntaktische Funktion im Satz haben. Nominalphrasen wurden hierbei am intensivsten untersucht, der Ansatz wird auch als Syntactic Phrase Indexing [9] bezeichnet. Sätze werden durch einen Parser in ihre Bestandteile zerlegt und die Beziehungen zwischen den Wörtern analysiert. Mit einem Mehr-Wort-Index ist

es möglich, relevante Beziehungen zu einem Wort zu finden, in die Suche von Begriffen einzubeziehen bzw. komplexe Suchanfragen in ihre Bestandteile zu zerlegen. Als Verfahren sind bspw. Syntaxzerlegung in Baumstrukturen, syntagmatische Beziehung oder Nachbar-Analyse zu nennen.

Ein Text-Mining-Algorithmus würde optimal arbeiten, wenn dieser die Semantik von Dokumenten erfassen könnte. Mit dem aktuellen Stand der Forschung können aber nur eingeschränkte semantische Aussagen über die Bedeutung der Dokumente mit Hilfe von statistischen Analysen getroffen werden. Ausgehend von einer Bag-of-Words-Repräsentation werden in der semantischen Ebene automatisch semantische Kategorien gebildet. Die Methoden des Term Clustering identifizieren durch statistische Untersuchungen semantisch ähnliche Terme und fassen diese zusammen. [10]

Die pragmatische Ebene erweitert die semantische, indem die Bedeutung des Textes in Bezug auf den sprachlichen und außersprachlichen Kontext erweitert wird.

Begriffsorientierte Verfahren versuchen, die Bedeutung von Wörtern zu erkennen. Zu diesen Verfahren zählen auch die Wortwahlanalyse und die Thesauriverfahren. In den Thesauri sind Wörter, Therme und Ausdrücke aus bestimmten Fachgebieten aufgelistet und durch Beschreibungen bzw. Verknüpfungen zueinander in Beziehung gesetzt.

Weitere Text-Mining-Verfahren, wie das Pattern-Matching, spielen in der Praxis nur eine untergeordnete Rolle und beschränken sich bei der Anwendung auf wenige Spezialgebiete.

4. Vorgehensmodell zur Unterstützung von Ideenfindungsprozessen durch Text-Mining

In der Literatur sind weder einheitliche Prozessmodelle noch eine identische Abgrenzung zwischen den Methoden zum Text Mining und den verwandten Disziplinen Information Retrieval, Informationsextraktion und Textvisualisierung beschrieben. [3] Da in den einzelnen Phasen des Problemlösungsprozesses unterschiedliche Ziele mit dem Einsatz von Text-Mining verfolgt werden, wurde ein differenziertes Prozessmodell (vgl. Abb. 22) entwickelt, das sich an der allgemeinen Funktionalität und Vorgehensweise von Text Mining-Tools anlehnt.

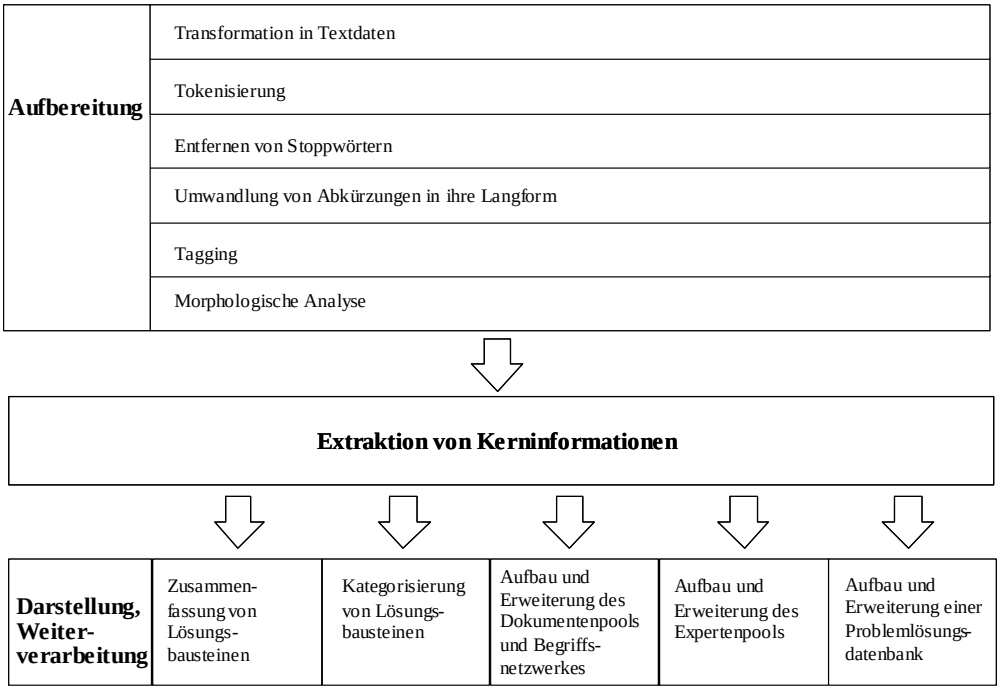


Abb. 2: Prozessmodell zur Anwendung von Text Mining in Ideenfindungsprozessen

Das Prozessmodell gliedert sich in die drei Hauptphasen Aufbereitung, Extraktion von Kerninformationen und Darstellung sowie Weiterverarbeitung der gewonnenen Daten.

In der Hauptphase der Aufbereitung werden interne bzw. externe Dokumente für vordefinierte Problemlösungsvorschläge analysiert, um das Format und die Sprache zu bestimmen. Dabei fungieren OCR- und Texterkennungsverfahren als Transformatoren von Bild- und Audioinformationen in Textdaten.

Die Phase der Tokenisierung basiert auf mehreren Teilphasen, die aufeinander aufbauen und einzelne sinnbehaftete Wortgruppen erzeugen. Im ersten Arbeitsschritt werden die Whitespaces, wie z.B. Tabulatoren, Leerzeichen und Zeilenumbrüche eliminiert. Zeichenketten, die am Zeilenende mit dem Zeichen '-' enden, werden markiert, um zu einem späteren Zeitpunkt mögliche Trennungen aufzuheben. Jedes erkannte Token ist durch einen Zeilenumbruch getrennt. [11]

Im nachfolgenden Schritt werden zusammengeschriebene Varianten von Wörtern getrennt. In der deutschen Sprache werden zwei Wörter mit Schrägstrich getrennt betrachtet. Ausnahmen entstehen, wenn ein Wort oder beide nur aus einem Buchstaben

bestehen. Weiterhin werden Zahlen mit einem Schrägstrich nicht getrennt. Zum Abschluss erfolgt eine durch Zeilenumbrüche getrennte Ausgabe.

Satzzeichen erzeugen im nächsten Abschnitt einen separaten Token. Punkte am Wortende werden nicht abgetrennt, da ihre Bedeutung erst zu einem späteren Zeitpunkt in der Punktdisambiguierung bestimmt wird und dazu eine vollständige Tokenisierung notwendig ist.

Trennungsmarkierungen werden im nächsten Schritt analysiert. Bei einem Trennungszeichen am Zeilenende können folgende Phänomene vorliegen:

- Silbentrennung,
- Trennung eines mit Bindestrich geschriebenen Wortes,
- Bindestrich ist ein Ergänzungszeichen eines zusammengesetzten Wortes.

Bei den aktuellen Textverarbeitungsprogrammen wird nur noch selten eine Trennung der Silben vorgenommen. Wird dennoch ein Bindestrich am Zeilenende vorgefunden, erfolgt eine morphologische Analyse der beiden Wortsilben. Ein Problem stellt die Unterscheidung zwischen einem zusammengesetzten Wort und einem Ergänzungswortteil dar. Zuerst wird getestet, ob das nächste Zeichen nach dem Bindestrich eine Zahl oder ein Buchstabe ist. Ist das nicht der Fall, dann ist das Wort ein Einzelbegriff und wird mit einem Token eingeschlossen. Andernfalls wird getestet, ob das eingeschlossene Wort zwischen Bindestrich und ergänztem Wort in einer vorher definierten Liste mit häufigen Verbindungswörtern vorkommt. Trifft das nicht zu, werden beide Worte zusammengeschrieben, als Mehrwortbegriff gekennzeichnet, und als ein Token ausgegeben. Ist das Wort in der Liste vorhanden, werden beide Wörter als einzelne Token markiert. [11]

In einem nächsten Schritt der Tokenisierung müssen Zahlen zusammengeführt werden. Zahlen, die aus mehr als drei Ziffern bestehen und im Originaltext Leerzeichen enthalten, werden zusammengeführt. Wichtig ist dabei eine genaue Definition der Zahlenmuster, um Verschmelzungen zu vermeiden. Eine fehlerhafte Tokenisierung kann unterschiedliche Zahlenangaben ergeben.

Der komplexeste Schritt der Tokenisierung besteht im Disambiguieren von Punkten am Wortende. Bei einem Punkt kann es sich um einen Satzpunkt, einen Abkürzungspunkt, den Punkt einer Ordinalzahl oder um Satzpunkt und einen der beiden anderen Fälle gleichzeitig handeln. Um eine Entscheidung zu treffen, ist daher das Erkennen von

Abkürzungen notwendig. Die Entscheidung über die Abtrennung von Punkten am Wortende arbeitet dabei mit folgender vereinfacht dargestellten Heuristik:

- Das folgende Wort ist kleingeschrieben.
- Das Wort besteht aus nur einem Buchstaben.
- Das Wort besteht aus Initialen.
- Das Wort besteht nur aus Konsonanten.
- Das folgende Wort ist ein Satzzeichen, welches nur innerhalb eines Satzes auftreten kann.
- Das Wort wurde im Abkürzungslexikon gefunden.
- Das Ende des Wortes wurde im Suffixlexikon gefunden.
- Das folgende Wort ist ebenfalls eine Abkürzung (in diesem Fall wird im Abkürzungslexikon nach einer mehrteiligen Abkürzung gesucht. Mehrteilige Abkürzungen werden als ein Token ausgegeben). [11]

Sind die Kriterien nicht erfüllt, dann handelt es sich um das Ende eines Satzes.

Der nächste Schritt eliminiert aus den Textdaten sogenannte Stoppwörter. Diese Wörter, wie Pronomen, Präpositionen und Konjunktionen, leisten keinen Beitrag zur Aussage oder kommen zu häufig vor. Es müssen aber auch problembezogene Wörter entfernt werden, falls diese zu allgemein und wenig informativ sind. Dazu wird eine Liste erstellt, die sachverhaltbezogene Stoppwörter enthält.

Ein weiterer Schritt behandelt Akronyme und Abkürzungen in verschiedenen Varianten, die schon bei der Tokenisierung erkannt wurden. Nach deren Identifizierung müssen die korrekten Langformen zugeordnet werden, um Mehrdeutigkeiten auszuschließen. Dies erfolgt ebenfalls über ein Verzeichnis mit allgemeinen Abkürzungskonventionen. Die Abkürzung wird anschließend durch die inhaltlich passende Langform ersetzt.

Durch Tagging werden gleich geschriebene Wörter mit unterschiedlicher Bedeutung identifiziert. In zwei Phasen wird die Analyse durchgeführt. Zu Beginn wird jeder Wortform eine Anzahl von möglichen Tags zugeordnet. Im zweiten Schritt erfolgt die Disambugierung zugeordneter Tags, so dass ein eindeutig „getaggter“ Text entsteht.

Für die automatisierte Verarbeitung von Texten ist die Vereinheitlichung von Begriffen wichtig. Alle Wortvarianten sind auf eine Stammform, den sogenannten „kanonischen Namen“, zurückzuführen. Ein kanonischer Name ist der eindeutigste Name, der aus verschiedenen im Dokument gefundenen Varianten eines Begriffs generiert wird. Dazu werden lexikonbasierte, regelbasierte oder statistische Verfahren eingesetzt, die Gegenstand der Ausführungen des dritten Abschnittes waren.

Ein Ansatz zur Extraktion von Kerninformationen besteht darin, Merkmale und Relationen von Texten herauszufiltern. Merkmale sind beispielsweise Eigennamen, Fachbegriffe, zusammengesetzte Ausdrücke, Datumsangaben, Währungen und Zahlen. Zusammengesetzte Fachausdrücke können mit heuristischen Funktionen identifiziert werden. Dabei wird das Dokument nach der charakteristischen Substantivstruktur von Fachbegriffen durchsucht und die gefundenen zusammengesetzten Begriffe durch Platzhalterzeichen ersetzt. Durch eine relationale Extraktionsfunktion können mit Hilfe einer Heuristik Muster erkannt werden, die anzeigen, dass eine Person, eine Firma oder ein Objekt eine bestimmte Relation zu einem anderen Objekt hat. Weitere Verfahren zur Informationsextraktion analysieren die Semantik von Sätzen eines Dokuments. Diese semantischen Verfahren vergleichen sie dann mit gespeicherten Regeln und versuchen, daraus Schlüsse auf die enthaltenen Aussagen zu erhalten. Lexikalische Analyseverfahren ermitteln die Häufigkeiten der Stammformen. Mit Hilfe der Termfrequenz werden die gewonnenen Informationen in Indizes gespeichert. Je nach Anwendungsfall erfolgt die Generierung von spezifischen Zusatzinformationen. Das können Kollokationen, Beziehungen zu linken und rechten Nachbarn, und Klassifikationen, Ähnlichkeiten zu anderen Textdokumenten, sein.

Die Weiterverarbeitung der gewonnenen Informationen erfolgt in Bezug auf den jeweiligen Anwendungsfall. Für Ideenfindungsprozesse ist dieser Gegenstand des nachfolgenden Abschnittes.

5. Einsatzmöglichkeiten von Text-Mining in computerunterstützten Sitzungen

Neben den Vorteilen des Computereinsatzes in Sitzungen sind auch einige Probleme zu erkennen (vgl. zweiter Abschnitt). Diese bilden in den verschiedenen Phasen des Ideenfindungsprozesses (vgl. Abb. 1) viele Ansatzpunkte, den Einsatz von bestimmten Methoden und Technologien zu optimieren. Da die zu verarbeitenden Informationen in Textform vorliegen und in verschiedenen Aktivitäten aufbereitet werden müssen, können die Methoden des Text-Mining eingesetzt werden.

Mit steigenden Teilnehmerzahlen und bei zeitlich verteilten Sitzungen wächst auch die Zahl der Ideen und Lösungsvorschläge, die sich „lähmend“ auf den weiteren Sitzungsverlauf auswirken kann. Lösungsvorschläge können ähnliche Informationen enthalten und demzufolge zusammengefasst werden. Um den dargestellten Problemlösungsprozess durch Ideenfindung zu optimieren, können automatisch Problemlösungsbausteine durch Zusammenfassung und Informationsextraktion aufbereitet werden. Im Ergebnis entstehen Vorschläge, die manuell durch Moderatoren und Sitzungsteilnehmer ergänzt werden können, um das optimale Ergebnis zu erreichen.

Ein weiterer Ansatzpunkt für die Anwendung von Text-Mining bei Ideenfindungsprozessen liegt im Filtern von „leeren“ Ideen. Durch das anonyme Handeln der Sitzungsteilnehmer kann sich die Ideenfindung zu einer Diskussion ohne Themenbezug und Inhalte wandeln. Die gewonnen Lösungsvorschläge besitzen für das Problem keine Relevanz. Der Moderator muss diese Beiträge herausfiltern, um den Verlauf zu optimieren. Mit Hilfe des Text-Mining kann dieser Vorgang automatisch realisiert und durch den Diskussionsleiter ergänzt werden. Für einen schnellen Überblick gewonnener Problemlösungsvorschläge können Summarizing-Funktionen genutzt werden. Das Extrakt wichtiger Informationen dient als Überblick für nachgelagerte Aktivitäten und kann die Qualität dieser positiv beeinflussen.

Eine Phase im Ideenfindungsprozess dient zur Strukturierung von Sitzungsbeiträgen. Dazu werden Klassifikationen festgelegt, die eine Zuordnung nach bestimmten Aspekten ermöglichen. Diese Tätigkeit wird von Moderator mit und ohne Unterstützung der Teilnehmer durchgeführt. Bei einer großen Zahl von Ideen entstehen lange Wartezeiten, um die entsprechenden Rubriken zu finden. Neben der automatischen Bestimmung von Klassifikationsmerkmalen kann eine automatische Zuordnung von Ideen und Lösungsbausteinen erfolgen. Dadurch wird der Sitzungsverlauf um jeweils eine Aktivität verkürzt.

Für Sitzungen mit einer ähnlichen Problemstellung können je nach Zugriffsberechtigung Ergebnisse aus anderen Ideenfindungsprozessen importiert werden. Dabei besteht die Möglichkeit, alte Problemstellungen neu aufzugreifen oder Ergebnisse auf neue Aufgabenstellungen zu übertragen. Die gleiche Problematik gilt für interne und externe Dokumente, die Lösungsvorschläge enthalten. Diese müssen aber zuvor aufbereitet werden, um die Kerninformationen in Form von Textdaten zu gewinnen. Falls keine Dokumente zur Verfügung stehen, können diese bspw. durch intelligente Agenten in internen und externen Informationsbasen recherchiert werden. Die Suche erfolgt autonom und die Ergebnisse werden durch die Extraktion von

Kerninformationen und Begriffsverknüpfungen in Dokumentenpools strukturiert abgelegt.

Internetbasierte sitzungsunterstützende Systeme bieten die Möglichkeit, neben externen Ressourcen auch externe Know-How-Träger und Experten in die Aktivitäten von Problemlösungsprozessen einzubinden. Diese können über das Netzwerk ihr Wissen einbringen. Falls keine Experten bekannt sind, können externe Expertenpools abgefragt oder interne Sitzungen nach entsprechenden Personen durchsucht werden. Das EMS verwaltet dazu parallel eine Datenbank mit Schlagwörtern und zugeordneten Personen, die an entsprechenden Problemlösungsprozessen mitgewirkt haben. Dieser Vorgang erfolgt anonym und automatisch. Bei Anfrage wird vom sitzungsunterstützenden System die entsprechende Person nach Mitarbeit anonym und unverbindlich angefragt. Ist diese Person an einer Mitarbeit interessiert, kann sie Kontakt mit dem Moderator aufnehmen, um an einzelnen Aktivitäten der Sitzung teilzunehmen.

Die Ergebnisse von abgeschlossenen Ideenfindungsprozessen können je nach Zugriffsrechten für zukünftige Sitzungen aufbereitet und genutzt werden. Ähnlich wie bei der Verwaltung eines internen Expertenpools wird eine interne Problemlösungsdatenbank aufgebaut, die als Know-How-Datenbank für zukünftige Sitzungen dient. Neben den spezifischen Problemstellungen und gewonnenen Ergebnissen erfolgt die Verwaltung von Sitzungsdetails, wie Anzahl der Aktivitäten und Sitzungsdauer, um den Moderator bei der Planung von zukünftigen Ideenfindungsprozessen zu unterstützen. Diese Problemlösungsdatenbank kann zusätzlich interne und externe Dokumente beinhalten, die wichtige Aussagen und Informationen für mögliche Ergebnisse beinhalten.

6. Ausblick

In dem Beitrag wurde ein Lösungsansatz vorgestellt, mit dem Informationen aus Ideenfindungsprozessen durch Text-Mining automatisch aufbereitet und verarbeitet werden können. Ausgehend von verschiedenen Problemen traditioneller und computerunterstützter Sitzungen wurde ein Prozessmodell vorgestellt, mit Hilfe dessen Ideen und Lösungsvorschläge für Problemstellungen automatisch bzw. teilautomatisch aufbereitet werden können, um bspw. Kerninformationen zu extrahieren und zusammenzufassen sowie Lösungsbausteine zu präsentieren und weiterzuverarbeiten. Diese Lösung wurde erfolgreich in einzelnen Modulen des sitzungsunterstützenden Systems webSCW umgesetzt und in verschiedenen Sitzungen angewandt. Durch die automatische Informationsgewinnung und -strukturierung können internetbasierte

Ideenfindungsprozesse optimaler gestaltet werden. Für eine weitere Unterstützung besteht die Möglichkeit, externe Wissens- und Know-How-Datenbanken einzubinden, um den problembezogenen Austausch von Informationen zwischen verschiedenen Systemen zu gewährleisten. Bei der Umsetzung der Extraktion von Kerninformationen wurden hauptsächlich statistische Verfahren des Text-Mining eingesetzt, da zusätzliche Werkzeuge, wie geeignete Thesauri, in der frühen Realisierungsphase nicht zur Verfügung standen oder zu hohe Aufwendungen verursacht hätten. In der Literatur und in eigenen Anwendungsbeispielen wurde jedoch die Leistungsfähigkeit der umgesetzten Algorithmen nachgewiesen.

7. Literatur

[1] Krause, D. (2001): Internetgestützte Ideenfindungsprozesse mit webSCW, in: Engelien, M.; Homann, J. (Hrsg.): Virtuelle Organisation und Neue Medien 2001 / Workshop GeNeMe2001 – Gemeinschaften in Neuen Medien, Josef Eul Verlag Lohmar Köln, S. 373 - 390.

[2] Ehrenberg, D.; Krause, D. (2002): Potenzial von internetgestütztem Know-How-Management zur Problemlösung in flexiblen Unternehmensstrukturen, in: Indutrie Management 18(2002)3, GITO-Verlag, Berlin, 2002, S. 36 - 39.

[3] Meier, M., Beckh, M. (2000): Text Mining. In: Wirtschaftsinformatik 42(2000)2, S. 165 - 167.

[4] Knorz, G. (1994): Automatische Indexierung, http://www.iud.fh-darmstadt.de/iud/wwwmeth/publ/skript/autind94/paper1.htm.

[5] Zipf, G. K. (1935): The psycho-biology of language. An introduction to dynamic philology; Cambridge/Mass., M.I.T. Press, 1935.

[6] Blair, D. (1992): Information Retrieval and the Philosophy of Language, The Computer Journal, 35(3), 1992.

[7] Kuhlen, R. (1977): Experimentelle Morphologie in der Informationswissenschaft, München Verlag Dokumentation, 1977.

[8] Porter, M. (1980): An algorithm for suffix stripping, Automated Library and Information Systems, vol. 14, no. 3, 1980, S. 130 - 137.

[9] Lewis, D. (1992): Representation and Learning in Information Retrieval, PhD thesis, Department of Computer and Information Science, University of Massachusetts, 1992.

[10] Heyer, G. (2001): Text Mining, http://wortschatz.uni-leipzig.de/asv/vortraege/materialien/ 001115_GH_GKTextMining.ppt, 2001.

[11] Zierl, M. (1997): Entwicklung und Implementierung eines Datenbanksystems zur Speicherung und Verarbeitung von Textkorpora, http://www.linguistik.uni-erlangen.de/tree/pdf/corsica/zierl97.pdf, 1997.