# Text Clustern mit Hintergrundwissen

Andreas Hotho

## 1 Einführung und Motivation

Die Clusteranalyse teilt Objekte in aussagefähige, bedeutungsvolle und nützliche Gruppen (Cluster) ein. Heute hat sie sich ihren Platz in vielen Anwendungsbereichen gesichert. Eingesetzt wird die Clusteranalyse z.B. in der Biologie, um Gene und Proteine mit ähnlicher Funktionalität zu finden, oder den Zugriff auf ähnliche Objekte einer Datenbank zu beschleunigen. Das Gruppieren von Kunden im Marketing oder die Unterstützung des Browsens im World Wide Web sind weitere bekannte Anwendungsfelder. Für das Browsen im WWW oder in sehr großen Dokument-Sammlungen in internen Firmennetzen stellt das automatische und effiziente Berechnen von Clustern ein immer wichtigeres Mittel zur *erstmaligen und automatischen Strukturierung* dieser sehr großen Dokumentsammlungen dar.

Bei der Durchführung einer Clusteranalyse arbeiten Spezialisten aus dem Bereich der Statistik oder des Data Minings typischerweise mit Experten aus dem Anwendungsgebiet zusammen. So wird sichergestellt, dass die Ergebnisse auch zu der jeweiligen Aufgabe aus der Praxis passen. Während der Lösung der Aufgabe fließen in diesen Prozess auch viele anwendungsspezifische Informationen ein, die den Erfolg garantieren sollen. Sehr häufig steuert das Wissen der Experten z.B. die Auswahl oder Kombination der zur Unterscheidung der Objekte eingesetzten Merkmale. Für das Clustern oder die Segmentierung ist die Auswahl und Aufbereitung der verwendeten Merkmale sowie ein entsprechendes Domänenwissen essentiell [DHS01]. So schreiben die Autoren in [DHS01, S. 12]: "As with segmentation, the task of feature extraction is much more problem- and domain-dependent [...] Although the pattern classification techniques presented in this book cannot substitute for domain knowledge, [...]" und machen damit in diesem Zusammenhang klar, dass Wissen über die Domäne bei der Segmentierung helfen kann. Sie geben allerdings nicht an, wie dieses Wissen in den Prozess einfließen soll. Neben der trivialen Alternative, auf den Domänenexperten mit seinem Wissen zurückzugreifen und ihn bei jedem Schritt der Analyse zu befragen, wird in der Dissertation [Hot04] das Hintergrundwissen mittels *formaler Repräsentation* in Form von Ontologien automatisch in den Prozess integriert. Damit können erstmals Teile des Benutzerwissens in einem Clusterprozess automatisch verwendet werden. Dieser Schritt kann den Benutzer nicht ersetzen, erlaubt aber eine erste Integration seines Wissens in den Clusterprozess. In dieser Arbeit wird gezeigt, dass Hintergrundwissen ein wichtiger Faktor ist, um erfolgreich Clusterverfahren einsetzen zu können.

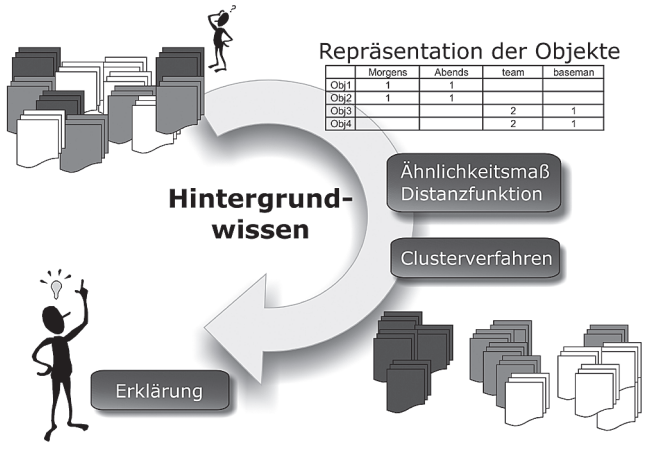


Abbildung 1: Der Clusterprozess

## 2 Problemstellung

Immer wieder kommt es vor, dass bei der Bildung von Gruppen nicht alle wichtigen Merkmale beachtet werden. Auch können zwischen einzelnen Merkmalen so komplexe Beziehungen existieren, dass deren Einfluss auf die Bildung von Gruppen nicht immer von den zu Grunde liegenden mathematischen Modellen korrekt erfasst werden kann. Andere Ursachen für schlechte Clusterergebnisse können die Repräsentation der Objekte oder die Funktionen zur Berechnung der Ähnlichkeiten oder Distanzen sein, die die Beziehung zwischen den Objekten nicht immer korrekt wiedergeben.

Abbildung 1 gibt das typische Vorgehen beim Clustern wieder und zeigt so prinzipielle Ansatzpunkte zur Verbesserung des Clusterprozesses. Die in der linken oberen Ecke symbolisierten Dokumente stellen die Menge der Objekte dar, die in Gruppen einzuteilen sind. Für die Durchführung dieser Aufgabe benötigt man neben einer geeigneten Repräsentation auch ein Maß für die Ähnlichkeit bzw. die Distanz zweier solcher Objekte. Die Tabelle rechts oben in Abbildung 1 repräsentiert die Objekte durch eine Menge von Merkmalen (Spalten), wie z.B. "Morgens" oder "team". Die Merkmale bilden die Grundlage für ein Ähnlichkeitsmaß oder eine Distanzfunktion. Diese Funktionen setzen die Objekte in Beziehung zueinander und geben dafür einen numerischen Wert an. Auf dieser Basis können nun ganz unterschiedliche Verfahren zur Berechnung von Clustern angewendet werden. Ein solches Clusterverfahren liefert die gesuchte Gruppierung entsprechend der gegeben Repräsentation und des Ähnlichkeitsmaßes bzw. der Distanzfunktion der Objekte. Der Prozess endet mit der anschaulichen Präsentation der berechneten Cluster, die gleichzeitig dem Benutzer eine Erklärung der Clusterinhalte durch eine passende Visualisierung liefert.

Der Clusterprozess umfasst mehrere Schritte, die Ansatzpunkte zur Verbesserung der Ergebnisse bieten. Viele Arbeiten präsentieren verbesserte Ergebnisse durch die Modifikation vorhandener oder die Entwicklung neuer Clusterverfahren. Weitere Ansatzpunkte sind die Ähnlichkeitsmaße und die Distanzfunktionen. Die hier vorgestellte Arbeit setzt am dritten möglichen Punkt an, nämlich der Repräsentation der Objekte. *Hintergrundwissen* wird an dieser Stelle in den Prozess eingebracht. Die veränderte Repräsentation führt sowohl zur


Auszug aus: Künstliche Intelligenz, Heft 1/2005, ISSN 0933-1875, Böttcher IT Verlag, Bremen, www.kuenstliche-intelligenz.de

*Steigerung der Verständlichkeit* als auch zur *Verbesserung der Güte* der Ergebnisse. Hintergrundwissen stellt damit in dieser Arbeit einen zentralen Bestandteil dar und beeinflusst durch die Integration in die Repräsentation der Objekte die Bewertung durch Ähnlichkeitsmaße und Distanzfunktionen sowie die berechnete Gruppierung der Clusterverfahren. Dabei können unterschiedlichste Clusterverfahren und Maße mit dieser neuen Repräsentation verwendet werden.

## 3 Lösungsansätze der Arbeit

In der Arbeit werden drei neu entwickelte Methoden zur Verwendung von Hintergrundwissen beim Clustern vorgestellt. Dies spiegelt sich in der Struktur der Arbeit wider, die sich in drei Teile gliedert:

- Die neu entwickelte Methode *Subjektives Clustern* berechnet benutzerbezogene Cluster bei gleichzeitiger Dimensionsreduktion. Damit wird u.a. die Verständlichkeit der Ergebnisse gesteigert (vgl. Abschnitt 3.1).
- Hintergrundwissen kann während der Vorverarbeitung der Dokumente erfolgreich in den Clusterprozess integriert werden (vgl. Abschnitt 3.2).
- Erstmals werden auch Verfahren der Formalen Begriffsanalyse zur Präsentation von Clustern verwendet, die für Menschen leicht verständliche Beschreibungen der berechneten Cluster liefern (vgl. Abschnitt 3.3).

Die entwickelten Methoden wurden in verschiedenen Anwendungsgebieten eingesetzt und evaluiert (vgl. [Hot04]).

### 3.1 Subjektives Clustern

"Subjektives Clustern" verfolgt zwei Ziele. Auf der einen Seite soll dem Benutzer die Möglichkeit eingeräumt werden, mehr Einfluss auf den Clusterprozess zu nehmen. Auf der anderen Seite wird die Dimensionalität des Merkmalsraumes durch die Auswahl von geeigneten Merkmalen und der Aggregation gemäß einer Ontologie reduziert. Ziel ist nicht wie bisher die Berechnung einer Clusterung, sondern mehrerer Clusterungen auf der Basis subjektiver benutzerbezogener Sichten. Die Sichten werden mit Hilfe der Ontologie und der Daten abgeleitet. Sie spiegeln die verschiedenen Präferenzen einzelner Benutzer wider. Der Benutzer hat die Möglichkeit, aus mehreren niedrigdimensionalen Clusterungen mit unterschiedlichen Merkmalen auszuwählen, wobei die Merkmale die Konzepte einer Ontologie sind. Die geringe Dimensionsanzahl erleichtert dem Benutzer auch die spätere Interpretation der Clusterergebnisse. Es konnte gezeigt werden, dass die Clusterungen basierend auf Sichten zu besseren und leichter verständlichen Ergebnissen führen.

Die entwickelten Algorithmen wurden sowohl auf realen Textkorpora als auch auf Kommunikationsdaten von Kunden der Deutschen Telekom AG angewendet und evaluiert. Dazu wurden umfangreiche domainspezifische Ontologien akquiriert, die dann beim Berechnen von Text- und Kundenclustern verwendet wurden (vgl. [HMS02]).

### 3.2 Clustern mit Hintergrundwissen

In dieser Arbeit wird während der Vorverarbeitung der Daten formal repräsentiertes Hintergundwissen in die Repräsentation der Daten integriert und während der Clusterung der Objekte genutzt. Für die Clusterung der Objekte werden bekannte Maße und Verfahren aus der Statistik und dem Maschinellen Lernen eingesetzt. Neben der empirischen Evaluierung wurde mittels Varianzanalyse die Integration des Hintergrundwissens in die vorhandenen und in Klassen eingeteilten Dokumente anhand von drei realen Textkorpora untersucht (vgl. [HSS03b, HSS03c]).

An realen Textdokumentensammlungen wird gezeigt, dass diese neue ontologiebasierte Repräsentation für Textdokumente („Bag of Concepts") gegenüber der herkömmlichen "Bag of Words"-Repräsentation zu einer signifikanten Steigerung der Clustergüte führt. Dazu werden neben verschiedenen Strategien zur Abbildung von Worten eines Textes auf die Konzepte einer Ontologie auch die Nutzung taxonomischer Beziehungen zur Steigerung der Clustergüte anhand dreier Datensätze aus der Praxis untersucht. Einer der Textkorpora besteht aus Nachrichtentexten der Agentur Reuters, einer aus Lernmaterialen der Programmiersprache Java und einer aus Texten landwirtschaftlicher Fachzeitschriften. Die Anwendung der Clusterung mit Hintergrundwissen kann anhand der vorliegenden empirischen Ergebnisse auf alle Fälle empfohlen werden, da die Ergebnisse immer mindestens gleich gut und häufig besser als die Referenzclusterung basierend auf der "Bag of Words"-Repräsentation sind.

### 3.3 Beschreibung der gefundenen Cluster

Die um Hintergrundwissen erweiterte Repräsentation der Dokumente führt nicht nur zu besseren Ergebnissen beim partitionierenden Clusterverfahren, sondern bildet auch die Basis für eine intuitiv verständliche Erklärung der gebildeten Cluster. In der Arbeit wird die Wirkung des Hintergrundwissens durch die Repräsentationsveränderung auf die Erklärungskomponente der Clusterergebnisse untersucht und die Anwendbarkeit anhand von verschiedenen Beispielen demonstriert (vgl. [HSS03a]).

Mit Hilfe der Formalen Begriffsanalyse werden die Ergebnisse in Verbänden visualisiert und liefern so eine für Menschen leicht verständliche Beschreibung der berechneten Textcluster. Grund dafür sind Beziehungen zwischen den Textclustern, die Gemeinsamkeiten und Unterschiede zwischen den Clustern hervorheben. Die in die Textrepräsentation integrierte Ontologie führt zu einer weiteren Verbesserung der Verständlichkeit. Sie strukturiert den Verband durch die bereitgestellten Oberkonzeptbeziehungen und ermöglicht so die einfache Exploration des Verbandes ausgehend von allgemeinen hin zu speziellen Begriffen. Dies wird auf dem Reuters-Korpus gezeigt und untersucht. Experimente auf anderen praxisnahen Textkorpora bestätigten diese Ergebnisse.

## 4 Zusammenfassung und Ausblick

In dieser Arbeit wurden die drei Methoden *Subjektives Clustern*, *Clustern mit Hintergrundwissen* und *Beschreibung von Textclustern mit Hintergrundwissen auf Basis der Formalen Begriffsanalyse* eingeführt. Dabei konnte gezeigt werden, dass die Integration von formal repräsentiertem Hintergrund in Form einer Ontologie die Güte der Clusterergebnisse steigert. Außerdem konnten leicht verständliche Visualisierungen der Textcluster erzeugt werden.

Die Arbeit stellt einen wichtigen Schritt zur Nutzung von formal repräsentiertem Hintergrundwissen in Form von Ontologien im Data, Text und Web Mining dar. Dies führt zur Vision des Semantic Web Mining (vgl. [BHS02]), bei dem es auf der

einen Seite um die Nutzung von Data-Mining-Verfahren zur Unterstützung des Aufbaus des Semantic Web, genannt Ontology Learning geht. Auf der anderen Seite steht die Analyse von strukturierten Daten und Informationen durch die Verfahren und Methoden des Data, Text und Web Minings im Vordergrund. Die hier diskutierte Arbeit bietet einen Beitrag insbesondere zur Erreichung des zweiten Zieles.

## Literatur


[BHS02] B. Berendt, A. Hotho und G. Stumme. Towards Semantic Web Mining. In I. Horrocks und J. A. Hendler (Hg.), *Proceedings of the First International Semantic Web Conference: The Semantic Web (ISWC 2002)*, Bd. 2342 von *Lecture Notes in Computer Science (LNCS)*, S. 264–278. Springer, Sardinia, Italy, 2002.

[DHS01] R. O. Duda, P. E. Hart und D. G. Stork. *Pattern Classification*. John Wiley & Sons, Inc., 2001.

[Hot04] A. Hotho. *Clustern mit Hintergrundwissen*, Bd. 286 von *Diski*. Akademische Verlagsgesellschaft Aka GmbH, Berlin, 2004.

[HMS02] A. Hotho, A. Maedche, S. Staab. Text Clustering Based on Good Aggregations. Künstliche Intelligenz (KI) 16(4), p. 48-54.

[HSS03a] A. Hotho, S. Staab und G. Stumme. Explaining Text Clustering Results using Semantic Structures. In *Proc. of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases, PKDD*, S. 217–228. 2003.

[HSS03b] A. Hotho, S. Staab und G. Stumme. Ontologies Improve Text Document Clustering. In *Proc. of the ICDM 03, The 2003 IEEE International Conference on Data Mining*, S. 541–544. 2003.

[HSS03c] A. Hotho, S. Staab und G. Stumme. WordNet improves text document clustering. In *Proc. of the SIGIR 2003 Semantic Web Workshop*. Toronto, Canada, 2003.



**Kontakt**
Dr. Andreas Hotho
Universität Kassel
Fachbereich Mathematik/Informatik
Fachgebiet Wissensverarbeitung
Wilhelmshöher Allee 73
34121 Kassel
email: hotho@cs.uni-kassel.de


Andreas Hotho studierte bis 1998 Wirtschaftsinformatik an der Technischen Universität Braunschweig. Von 1999 bis 2004 war er wissenschaftlicher Mitarbeiter am Institut für Angewandte Informatik und Formale Beschreibungsverfahren an der Universität Karlsruhe. Er promovierte dort im Bereich Text Mining, Data Mining und Semantic Web und wendete diese Methoden auch zur Kundensegmentierung bei der Deutschen Telekom AG an. Seit April 2004 ist er wissenschaftlicher Assistent an der Universität Kassel und beschäftigt sich dort u.a. mit dem Thema Semantic Web Mining.