

# Wie man mit der Wikipedia semantische Verfahren verbessern kann

*Das automatische Zuweisen von Themengebieten zu beliebigen Dokumenten ist eine der anspruchsvollsten Aufgaben in der Computerlinguistik. Um dies technisch überhaupt bewerkstelligen zu können, setzt es ein gewisses »Verständnis« eines Textes voraus. Üblicherweise werden bei solchen Verfahren große – von Hand erstellte – thematisch vorsortierte Datenbanken benutzt. In dieser Arbeit wird zusammen mit statistischen Datenanalysen die »Datenbank« Wikipedia verwendet, um mit ihren semantischen Strukturen automatisch passende Themen von Dokumenten zu identifizieren und anschließend zuzuordnen. Darüber hinaus wird mit einem weiteren Verfahren gezeigt, wie das Auffinden ähnlicher Dokumente verbessert werden kann.*

## Inhaltsübersicht

- 1 Automatische Zuordnung von Themen
- 2 Nutzen der Wikipedia-Strukturen
- 3 WMTrans-Produkte
  - 3.1 WMTrans-Technologie
  - 3.2 Produktbereiche
  - 3.3 Der WMTrans-Lemmatizer
- 4 TF-IDF
- 5 Semantische Kategorisierung und themenbasierte Verschlagwortung von Dokumenten mit der Wikipedia
  - 5.1 Das Auffinden ähnlicher Dokumente
  - 5.2 Automatisches Kategorisieren von Dokumenten
- 6 Schlussbetrachtung und Ausblick
- 7 Literatur

## 1 Automatische Zuordnung von Themen

Diese Arbeit beschreibt einen neuen Ansatz, um beliebige Dokumente semantisch zu

strukturieren und automatisch Themengebieten zuzuordnen (Taggen). Aufgrund dieser Zuordnung kann eine automatisierte Alternative für »normale« Suche in Datenbanken oder Webseiten bereitgestellt werden, bei der nicht nur nach einzelnen Wörtern, sondern auch nach Themengebieten gesucht werden kann. Darüber hinaus bietet der Ansatz die Möglichkeit, qualitativ besser verwandte Dokumente zu einem beliebigen Dokument zu finden.

Umgesetzt wurde diese Arbeit mit drei verschiedenen »Ansätzen«:

- a) Die Wikipedia ist so aufgebaut, dass Artikel semantisch kategorisiert sind. Diese vorhandenen Wikipedia-Kategorien der einzelnen Artikel wurden benutzt, um bei beliebigen Dokumenten herauszufinden, welchen Themen ein Dokument zuzuordnen ist.
- b) Um Dokumente besser analysieren zu können, wurden in einem weiteren Schritt die Sprachanalysetools – WMTrans – der Schweizer Softwarefirma Canoo AG verwendet. Diese helfen beispielsweise, das »Rauschen« in Dokumenten zu verringern. So werden etwa konjugierte Wörter auf deren Grundform zurückgeführt, sodass z.B. die beiden Wörter »gingen« und »gehst« von einem Computer als ein Wort – nämlich »gehen« – begriffen werden können.
- c) Um Ähnlichkeiten zwischen Dokumenten zu finden, wurde ein dafür üblicher Algorithmus »TF-IDF« (term frequency - inverse document frequency) verwendet. Hierbei werden spezielle Terme (Schlüsselwörter) in einem Dokument festgestellt und dann besonders gewichtet und daraufhin nach deren Vorkommen in anderen Dokumenten gesucht.

## 2 Nutzen der Wikipedia-Strukturen

Die Wikipedia ist seit ihrer Entstehung einem rasanten linearen Wachstum unterworfen (siehe Abb. 1)<sup>1</sup>. Alleine in der deutschen Version gibt es mittlerweile 956.531 Artikel (Stand: 17.9.2009), während die englische Version der Wikipedia sogar 3.038.561 Artikel (Stand: 22.9.2009) vorzuweisen hat.<sup>2</sup>

Dies stellt nicht nur eine umfangreiche Dokumentation unseres heutigen Wissens dar, sondern die Wikipedia kann durch diesen Informationsgehalt auch zum Analysieren von Dokumenten eingesetzt werden, und zwar aus mehreren Gründen:

1. Viele Wörter eines beliebigen Textes besitzen innerhalb der Wikipedia einen eigenen Artikel und dessen Metastruktur.
2. Diese Artikel sind logisch mit anderen Artikeln verbunden, d.h., es bestehen kausale Verlinkungen zu anderen Artikeln.
3. Fast alle Artikel in der Wikipedia sind semantisch kategorisiert und anhand dieser Kate-

gorien wieder mit anderen Themen und Artikeln verbunden.

4. Die Wikipedia ist äußerst effizient beim Auflösen ambiguer (mehrdeutiger) Wortformen. Mehrdeutige Wörter kommen häufig in Texten vor und stellen jede maschinelle Verarbeitung vor immense Probleme.

Diese Erkenntnisse sind jedoch nicht neu. So wurde die Wikipedia schon des Öfteren zum Untersuchungsgegenstand computerlinguistischer Forschung. Zu nennen sind hierbei vor allem [Bunescu & Pasca 2006], [Cucerzan 2007] und [Gabrilovich & Markovitch 2006]. Vor der Entstehung der Wikipedia wurden vor allem Korpora (Sammlungen von Texten oder Äußerungen in einer Sprache) zur Analyse von Dokumenten verwendet, die über mehrere Jahre hinweg aufwendig erstellt wurden. Beispiele hierbei sind WordNet<sup>3</sup> oder GermaNet<sup>4</sup>. [Gabrilovich & Markovitch 2007] haben bei einer Untersuchung verschiedener Korpora im Vergleich zu Wikipedia festgestellt, dass mit der durch [Finkelstein et al. 2002] aufgestellten

1. <http://de.wikipedia.org/wiki/Wikipedia:Meilensteine>

2. [http://en.wikipedia.org/wiki/Main\\_Page](http://en.wikipedia.org/wiki/Main_Page)

3. <http://wordnet.princeton.edu/>

4. [www.sfs.uni-tuebingen.de/lzd/](http://www.sfs.uni-tuebingen.de/lzd/)

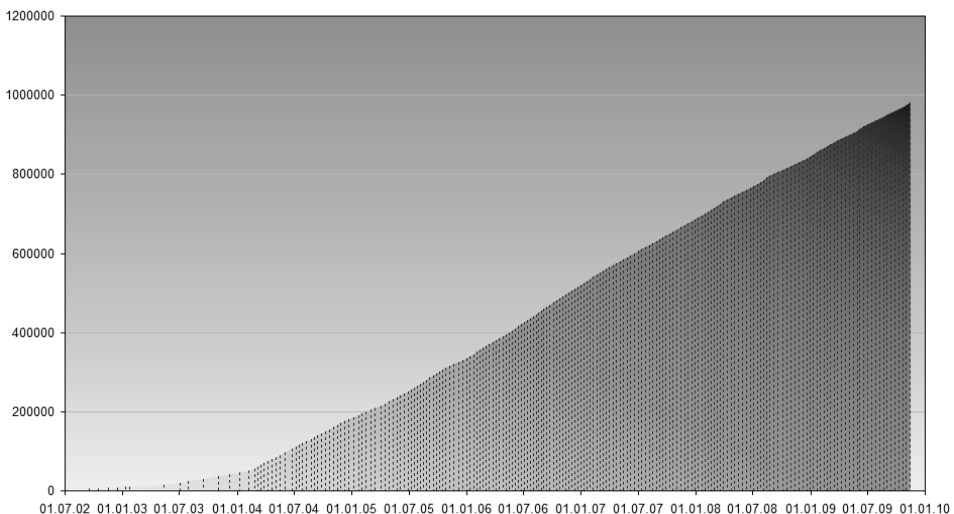


Abb. 1: Artikelwachstum der Wikipedia seit 2002

WordSimilarity-353 collection die Wikipedia der menschlichen Semantik am nächsten kommt.

Das Besondere an der Wikipedia ist zudem, dass für viele Eigennamen, die in den »normalen« Korpora nicht erfasst sind, die komplette Struktur ebenso vorliegt wie zu gebräuchlicheren Wörtern. Dies betrifft sowohl semantisches Tagging (die automatische Zuweisung eines »Themas« zu einem Dokument) als auch kausale Linkstrukturen.

Gleichzeitig stellt die Fülle an Wörtern den Benutzer der Wikipedia wieder vor ein neues Problem: *Je mehr Wörter es in der Wikipedia gibt, desto größer wird die Anzahl der mehrdeutigen Wörter.* Kann ein Wort in einem Dokument nicht eindeutig Einträgen in der Wikipedia zugeordnet werden, wird es nicht weiterverwendet, da die Bedeutung des Wortes für eine weitere Analyse nicht eindeutig bestimmt werden kann.

[Milne & Witten 2008] haben ebenfalls mithilfe der Wikipedia für dieses Problem eine elegante Lösung gefunden. Abbildung 2 zeigt einen Text, bei dem das Wort *tree* die Bedeutung »tree (data structure)« hat und nicht etwa »tree« im allgemeinen Sinne von »Baum«. Wenn man mit Standardgewichtungsverfahren (wie etwa PageRank<sup>5</sup>) alle möglichen Kategorien zu dem Wort *tree* innerhalb der Wikipedia gewichten würde (wie etwa: *tree*, *tree (data structure)*, *tree (graph theory)*, *tree network*, ...),

5. <http://ilpubs.stanford.edu:8090/422/>

dann würde der allgemeinen Bedeutung »tree« die höchste Relevanz zugesprochen werden, weil sie am meisten mit anderen Artikeln verlinkt ist. Die – in diesem Fall – korrekte Kategorie »tree (data structure)« wäre erst auf Platz 3.

Dieses Problem kann aufgelöst werden, indem man verfolgt, wohin alle anderen Wörter im Text innerhalb der Wikipedia verlinken. In diesem Beispiel stellt man fest, dass die Wörter *algorithm*, *tree structure*, *uniformed search* und *LIFO stack* auf »tree (data structure)« verlinken. Diese kontextsensitive Relevanz kann mit folgender Formel bestimmt werden:

$$\text{relatedness}(a, b) = \frac{\log(\max(|A \cup B|)) - \log(|A \cap B|)}{\log(|W|) - \log(\min(|A \cup B|))}$$

a und b sind die ausgewählten Wikipedia-Artikel, A und B sind die Sets aller Artikel, die entweder auf a oder b verlinken, und W ist die Anzahl aller Artikel in der Wikipedia.

Allerdings ist der alleinige Einsatz der Wikipedia zur Analyse von Dokumenten nicht zufriedenstellend, da es bei vielen Konjugationen (z.B. *Bank*, *Bänke*, *Banken*) keine Auflösung auf die richtige Wortform bzw. Bedeutung in der Wikipedia gibt. Auch der Umgang mit zusammengesetzten Wörtern oder mit »unbekannten« Wortformen ist für jedes System – auch für die Wikipedia – schwer. Um mit diesem Problem umzugehen, wurde auf die WM-Trans-Produktpalette zurückgegriffen.

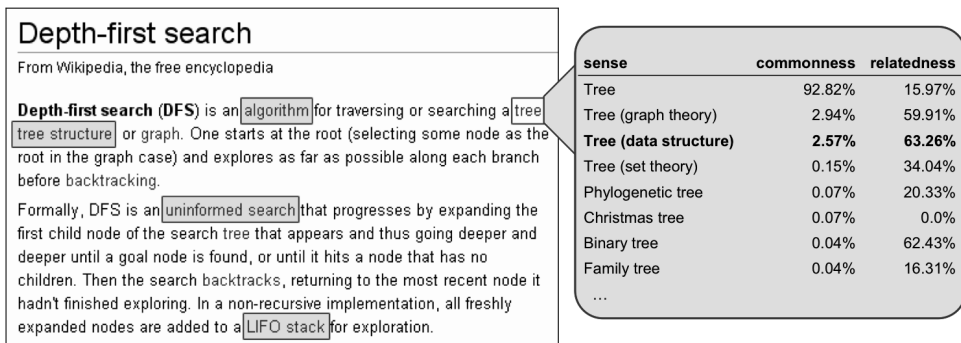


Abb. 2: Textausschnitt, bei dem mithilfe der Wikipedia Mehrdeutigkeiten aufgelöst werden

## Fazit

Die Wikipedia ist eine äußerst große und umfangreiche »Datenbank«, deren semantische Struktur und kausale Verlinkung auf andere Artikel hilfreich ist bei der Analyse von Texten. Darüber hinaus ist zu erwähnen, dass die Wikipedia für die Dokumente vieler Sprachen<sup>6</sup> angewendet werden kann.

## 3 WMTrans-Produkte

WMTrans ist eine Sammlung verschiedener Analysewerkzeuge, die auf den gleichen wortbasierten morphologischen Informationen aufbauen. Diese Informationen basieren ihrerseits auf den generierten und gesammelten Daten von WordManager [Hacken 2009], einem Autorensystem für Wortformen und Flexionsregeln (Regeln zur Änderung der Gestalt eines Wortes).

Mit dem Autorensystem werden Wörter mit den dazugehörigen Wortbildungsregeln erfasst und verwaltet. WordManager unterstützt den Linguisten mit entsprechenden Benutzerschnittstellen in der komplexen Aufgabe, die Flexions- und Wortformationsregeln zu spezifizieren. Die Daten werden in einer zentralen, speziell strukturierten Datenbank verwaltet, aus der mit Hilfsprogrammen die benötigte Information für die verschiedenen Aspekte und Formate der Sprachprodukte generiert werden kann. Zudem bietet es Mittel und Werkzeuge zum Auffinden und Beheben von Fehlern in den Regeln und der Konsistenzprüfung der Daten für den Linguisten an.

Die Basis der Daten für die Sprachprodukte bilden somit die erfassten Morphologie-Wörterbücher. Verschiedene Linguisten haben bis heute für die Sprachen Deutsch, Französisch und Italienisch Wörter mit den entsprechenden Wortbildungsregeln erfasst. Zurzeit können mithilfe der erfassten Daten folgende Anzahl Wortformen in den drei Sprachen generiert werden:

6. Die Wikipedia ist in vielen Sprachen verfügbar (<http://de.wikipedia.org/wiki/Wikipedia:Sprachen>).

- Für Deutsch sind 300.000 Einträge erfasst, aus denen mehr als drei Millionen deutsche Wortformen generiert werden können.
- Für Englisch sind 50.000 Einträge erfasst und es können damit 115.000 englische Wortformen generiert werden.
- Für Italienisch können aus den 50.000 Einträgen 460.000 italienische Wortformen generiert werden.

### 3.1 WMTrans-Technologie

Die WMTrans-Produkte werden für die Analyse von einzelnen Wörtern in einem Text verwendet und sind ausgestattet mit nützlichen Word-Manager-Informationen, wie den Flexionsregeln, den Wortfamilien, den Wortableitungen und den Wortkompositionen. Typische Anwendungsbereiche sind Informationsextraktion, intelligente Suche, automatische Indexierung, Text Mining, Spracherwerb, Hyperlink-Generierung, Rechtschreibprüfer, Grammatikprogramme und maschinelle Übersetzung.

Alle WMTrans-Produkte basieren auf einer weitverbreiteten *Finite-State*-Technologie. Diese Technologie verwendet einfache Finite-State-Automaten oder *Transducer*, die in Bezug auf Speicherverbrauch und Performanz als eine effiziente Implementation für die Analyse von Wortformen und Wortgeneration gelten. Weitere detaillierte Informationen zu dieser Technologie sind in [Koskenniemi 1983] und [Karttunen 1994] beschrieben.

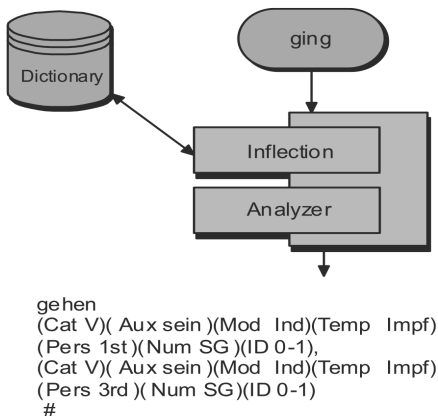
Die Schweizer Softwarefirma Canoo bietet unterschiedliche WMTrans-Produkte an. Der Unterschied in den Produkten besteht im Wesentlichen darin, wie die Information im Produkt codiert wird und wie diese Information für verschiedene Sprachanwendungen aufbereitet werden kann.

### 3.2 Produktbereiche

Für das Erkennen der Flexionsinformation steht ein einfacher »Recognizer« (WMTrans-Recognizer) zur Verfügung. Dieser erkennt mit einem yes/no-Resultat, ob ein Wort in einer

Sprache eine gültige Form hat. Der WMTrans-Lemmatizer kann zudem als »Part-of-speech Tagging« (POS) einerseits den analysierten Wörtern ihre Wortarten (Verb, Nomen usw.) zuordnen, aber auch ihre Zitatform ausgeben. Die ausführlichste Information über ein analysiertes Wort gibt der WMTrans-Analyser aus. Für ein zu analysierendes Wort werden entsprechende Muster (linguistische Paradigmen) und morphosyntaktische Informationen generiert. In Abbildung 3 wird ein vereinfachtes Diagramm mit der generierten Ausgabe des WMTrans-Analyzers für den Begriff »ging« gezeigt. Dabei wird das analysierte Wort der Grundform (»gehen«) mit den entsprechenden linguistischen Paradigmen ausgegeben. Für das angegebene Beispiel sind dies zwei Möglichkeiten: »ich ging« und »er ging«. Die maschinell verarbeitbare Ausgabe bedeutet: »Ging« hat die Grundform »gehen«, gehört zur Kategorie Verb (*Cat V*), das Perfekt wird mit dem Hilfsverb sein (*Aux sein*) gebildet, es handelt sich um die erste oder dritte Person Singular (*Pers 1st*) (*Num SG*) / (*Pers 3rd*) (*Num SG*) in der Vergangenheitsform (Imperfekt) (*Temp Impf*) und dem Modus (Aussageweise) Indikativ (Wirklichkeitsform) (*Mod Ind*). Die Angabe (*ID 0-1*) ist eine systeminterne Bezeichnung.

Im Gegensatz zum WMTrans-Analyser generiert der WMTrans-Generator anhand der ein-



**Abb. 3: Vereinfachtes Diagramm des WMTrans-Analyzers**

gegebenen Zitatform alle möglichen Schreibweisen eines Wortes.

Analog zur Flexionsinformation gibt es Produkte für die Aspekte der Wortformationen. Der Wortformen-Analyser (WMTrans/WF-Analyser) liefert den Ursprung eines zusammengesetzten Wortes oder den Ursprung von dessen Ableitung. Rekursiv angewendet, können ganze Ableitungsbäume eines Wortes generiert werden.

Für die umgekehrte Funktionalität kann der WMTrans/WF-Generator verwendet werden. Damit können z.B. alle möglichen Ableitungen und Wordformationen eines Wortes generiert werden.

### 3-3 Der WMTrans-Lemmatizer

Von den WMTrans-Produkten ist der Lemmatizer ein wichtiger Baustein für die in diesem Beitrag beschriebene Methode. Der Lemmatizer ermittelt für jedes analysierte Wort im Text die Grundform und die Wortart (Verb, Nomen usw.).

Wird das Wort »ging« in einem Text analysiert, erhält man die Information, dass das Wort ein Verb ist und die Grundform »gehen« lautet. Das folgende Beispiel zeigt die generierte Information der Wörter »ging« und »Häuser«.

Ging → gehen (*Cat V*)  
 Häuser → Haus (*Cat N*)

Diese Information wird bei der Analyse verwendet, um alle Elemente der gleichen Wortfamilie (Lexeme) zu gruppieren und diese dann für die Berechnung der Relevanz eines Wortes zu verwenden. Dadurch können die Relevanz und die Qualität des Indexierers wesentlich verbessert werden.

### Fazit

Die WMTrans-Technologie eignet sich durch maßgeschneiderte Produkte für den Einsatz in Sprachanwendungen. Durch die konsistente Datenverwaltung und die umfangreiche Datenbasis wird ein extrem hoher Qualitätsstandard der erzeugten Daten gewährleistet.

## 4 TF-IDF

Um Ähnlichkeiten zwischen Dokumenten zu untersuchen, wird häufig ein Verfahren namens *TF-IDF* eingesetzt. Da diese Methodik schon seit geraumer Zeit bekannt ist, wird TF-IDF (term frequency - inverse document frequency) oft verwendet bei *Information-Retrieval*- und *Text-Mining*-Verfahren (vgl. dazu [Salton & McGill 1983]). Die Hauptaufgabe von TF-IDF ist es, herauszufinden, wie *wichtig* einzelne Wörter in einem Dokument im Verhältnis zu anderen Dokumenten innerhalb einer Datenbank sind. Dazu wird zu jedem Wort innerhalb eines Dokumentes gezählt, wie häufig das Wort innerhalb dieses Dokumentes erscheint (tf = term frequency).

$$tf_{i,j} = \frac{freq_{i,j}}{\max_l(freq_{l,j})}$$

Diese Methode wird pro Term (Schlüsselwort)  $i$  abhängig vom Dokument  $j$  betrachtet.  $freq_{i,j}$  ist die Auftrittshäufigkeit des betrachteten Terms  $i$  im Dokument  $j$ . Im Nenner steht die Maximalhäufigkeit über alle  $k$  Terme im Dokument.

Danach wird festgestellt, wie häufig dieses Wort in allen anderen Dokumenten in jener Datenbank vorkommt, innerhalb der nach ähnlichen Dokumenten gesucht wird (idf = inverse document frequency).

$$idf_i = \log \frac{N}{n_i}$$

Hier ist  $N = |D|$  die Anzahl der Dokumente im Korpus und  $n_i$  die Anzahl der Dokumente, die Term  $i$  beinhalten.

Das Gewicht  $w$  eines Terms  $i$  im Dokument  $j$  ist dann nach *TF-IDF*:

$$w_{i,j} = tf_{i,j} \cdot idf_i = \frac{freq_{i,j}}{\max_l(freq_{l,j})} \cdot \log \frac{N}{n_i}$$

Allerdings sollte man sich vor dem Einsatz von TF-IDF Gedanken darüber machen, welche Wörter man für die Analyse einbezieht. Standardmäßig werden bei diesem Verfahren alle Wörter mit in die Untersuchung einbezogen. Ein einzelnes Wort wird dabei meistens erkannt, wenn es durch ein Leerzeichen von anderen getrennt ist. Hierbei muss man natürlich erst einmal alle »unwichtigen« und »zu häufigen« Wörter erkennen und ausschließen. Dazu kann man eine Stoppwortliste für die häufigsten Wörter definieren (z.B. *und, oder, mit, ...*) und dadurch viele häufige Wörter ausschließen, jedoch obliegt es hier einer aufwendigen manuellen Einstellung, welche Wörter zugelassen werden und welche nicht. Eine andere Problematik bekommt man jedoch durch einfaches Erstellen einer Stoppwortliste nicht in den Griff: Wortformen. Und diese führen oft zu starken Verfälschungen der Ergebnisse, da beispielsweise schon Pluralformen eines Wortes (etwa *Bank* oder *Banken*) und Konjugationen (*gehen, gehts, ging, ...*) als unterschiedliche Wörter behandelt werden. Auch Wortbildungen werden als unterschiedliche Wörter vom Rechner wahrgenommen, beispielsweise *Vorstandsvorsitzender, Vorstandschef, Chef, Vorstand*. Sowohl inhaltlich als auch syntaktisch sind sich diese Wörter ähnlich, das TF-IDF-Verfahren stuft sie jedoch alle als unterschiedliche Wörter ein und wertet sie entsprechend.

### Fazit

TF-IDF kann man als etabliertes und robustes Verfahren bezeichnen, um Ähnlichkeiten zwischen Dokumenten zu analysieren. Jedoch muss man, wenn man qualitativ hochwertige Ergebnisse bekommen möchte, viel an Vorarbeit investieren, um für die Analyse geeignete Schlüsselwörter herauszufinden.

## 5 Semantische Kategorisierung und themenbasierte Verschlagwortung von Dokumenten mit der Wikipedia

Generell kann man sagen, dass die drei beschriebenen Ansätze für sich jeweils starke Vorzüge besitzen.

So ermöglicht es die Struktur der Wikipedia (a), eine Unmenge an Sachverhalten digital thematisch nachzuschlagen. Durch die gut aufbereitete Struktur der Wikipedia und einer Lizenz, die selbst das kommerzielle Verwenden der Wikipedia erlaubt<sup>7</sup>, kann dieser Fundus zur Analyse von Dokumenten eingesetzt werden.

Allerdings kann man die Wikipedia zwar herunterladen<sup>8</sup>, jedoch ist damit noch nicht viel erreicht, da die Wikipedia nach dem Download nicht zum Gebrauch bzw. zur Analyse von Dokumenten zu verwenden ist. Um wirklich effizient damit arbeiten zu können, muss man sich eine entsprechende Architektur selbst überlegen und diese aufsetzen.

Die »Sprachwerkzeuge« der Canoo AG WMTrans (b) bieten die Möglichkeit, digitale Dokumente von starkem »Rauschen« zu befreien und beispielsweise eine Suche in einer Datenbank effizienter zu gestalten. Die WMTrans-Palette bietet hier eine Vielzahl von kleinen »Helferlein« an, um Texte besser und produktiver analysieren zu können.

Zuletzt wurde TF-IDF (c) besprochen und dabei aufgezeigt, dass der Algorithmus an sich zwar robust ist, man jedoch einiges an Vorarbeit investieren muss, um ordentliche Resultate zu bekommen.

In den beiden nun folgenden Punkten wird gezeigt, wie durch die Kombination der verschiedenen Techniken neue – und vor allem bessere – Ergebnisse aus der Analyse von Dokumenten entstehen.

### 5.1 Das Auffinden ähnlicher Dokumente

Wie in Abschnitt 4 angesprochen ist TF-IDF zwar ein robustes Verfahren, um Ähnlichkeiten zwischen Dokumenten herauszufinden, jedoch muss einiges in die Vorarbeit investiert werden, um nur die relevanten Wörter eines Textes für die Analyse zu nutzen. Es hat sich als produktiv erwiesen, hier eine Kombination der WMTrans-Produkte und TF-IDF einzusetzen.

Je nach Datenbasis ist es meistens wünschenswert, für die Analyse nur die Substantive eines Dokumentes zu verwenden, was man mit WMTrans einfach bewerkstelligen kann. Des Weiteren kann man verschiedene Wortformen vollständig erkennen und auf die Grundwortform konjugieren, was das – bereits mehrfach angesprochene – »Rauschen« bei Textanalysen deutlich senkt. Dieses »Rauschen« wird klarer, wenn man sich beispielsweise die verschiedenen Wortformen von *gehen* ansieht: gehe, gehst, gehen, gehet, ginge, gingest, gingen, ginget, gehend, gegangen. Ohne den Einsatz von WMTrans würde TF-IDF hier 11 unterschiedliche Wörter identifizieren und entsprechend gewichten anstelle von nur dem einen Wort: *gehen*. Dazu können noch – wie in Abschnitt 4 bereits angesprochen – Wortkombinationen erkannt und ebenfalls auf die jeweilige Grundform zerlegt werden.

Das Verfahren TF-IDF kann man jedoch nicht nur auf die Wörter eines Dokumentes verwenden, sondern auch auf die inhaltlichen Strukturen der Wikipedia.

Zu jedem Wort, das nach der Filterung durch WMTrans noch zur Analyse zur Verfügung steht, kann man feststellen, ob es dazu einen passenden Wikipedia-Artikel gibt.

Daraufhin kann man die dort vorhandenen semantischen Strukturen verwenden, um ein erneutes TF-IDF auf diese Informationen durchzuführen.

Die folgenden Beispiele zeigen den Umgang mit zwei in einem Text gefundenen Wörtern: »UBS« und »Credit Suisse«. Mit WMTrans

7. <http://de.wikipedia.org/wiki/Wikipedia:Lizenzbestimmungen>

8. <http://download.wikimedia.org/>

und TF-IDF alleine sind dies unterschiedliche Wörter und sie werden in der Analyse auch als solche behandelt. Beim Betrachten der semantischen Struktur innerhalb der Wikipedia erkennt man jedoch die Gemeinsamkeiten beider Wörter:

UBS:

Kategorien: Kreditinstitut (Schweiz) |

Unternehmen (Zürich) | Unternehmen (Basel)

Credit Suisse:

Kategorien: Kreditinstitut (Schweiz) |

Unternehmen (Zürich)

Beide Wörter besitzen die gemeinsamen Kategorien »Kreditinstitut (Schweiz)« und »Unternehmen (Zürich)«.

Bei einer doppelten Analyse von TF-IDF – einmal über die Wörter an sich und einmal über die zugehörigen semantischen Kategorien der gefundenen Wörter – ergeben sich mit einer Vorfilterung durch WMTrans erstaunlich gute Ergebnisse, da TF-IDF die mehrfach auftretenden Kategorien als wichtige Schlüsselwörter erkennen und entsprechend gewichtet.

Abbildung 4 zeigt einen Vergleich einer normalen Analyse (ohne Vorfilterung der Wörter) mit einer Analyse, die mit WMTrans und Wikipedia unterstützt durchgeführt wurde. Leider ist es schwer, diesen Vergleich qualitativ zu messen, da »Ähnlichkeiten« zwischen zwei Objekten immer auch zum Teil einen subjektiven Charakter besitzen. Allerdings kann man durch den Vergleich von Abbildung 4 zeigen, dass der gefundene Artikel (»Buchvorabdruck Jack Welch: Keine Krise ohne Blutbad«<sup>9</sup>) innerhalb des Testdatensatzes bei der normalen TF-IDF-Analyse eigentlich nur wenig inhaltlich mit dem Text »Diamantenkonzern De Beers spürt Wirtschaftskrise«<sup>10</sup> zu tun hat, während der Ar-

tikel »Manche verkaufen ihre liebsten Stücke«<sup>11</sup> sich auf die Wirtschaftskrise und den Verkauf von Schmuck bezieht (aufgrund der schlechten Zeiten).

## Fazit

Bei einer Vorfilterung der Wörter eines Dokumentes und einer weiteren Analyse der semantischen Kategorien sind die Ergebnisse eines TF-IDF-Verfahrens bedeutend besser, als wenn man keine Filterung vornimmt. Zudem werden die Ergebnisse noch verbessert, wenn man TF-IDF noch in einem zweiten Schritt auf die jeweiligen Kategorien der Artikel anwendet, die als Wörter in einem Dokument vorhanden sind.

## 5.2 Automatisches Kategorisieren von Dokumenten

Um automatisiert Dokumente semantischen Kategorien zuzuordnen, muss man einen ähnlichen Weg beschreiten wie bei der Analyse von Abschnitt 5.1.

Für die semantische Analyse wurden alle Substantive eines Textes extrahiert (durch WMTrans wurde im Vorfeld festgestellt, welche Wörter Substantive sind). Danach werden über alle gefundenen Wörter hinweg die semantischen Kategorien der entsprechenden Wörter gesammelt, und es wird im Anschluss daran untersucht, welches die häufigsten Kategorien innerhalb eines Textes sind. Die besten (am häufigsten gefundenen) Kategorien davon werden dann in der Datenbank – mit dem entsprechenden Dokument assoziiert – indexiert. Dadurch ist es möglich, die Vielzahl der Kategorien, die es in der Wikipedia gibt, auf wenige relevante zu beschränken und diese einem Text zuzuordnen. Wir haben dieses Verfahren »WikiTagging« genannt.

9. [www.bilanz.ch/edition/artikel.asp?Session=%3C&AssetID=2146](http://www.bilanz.ch/edition/artikel.asp?Session=%3C&AssetID=2146)

10. [www.tagblatt.ch/aktuell/wirtschaft/wirtschaft/Diamanten-Konzern-De-Beers-spuert-Wirtschaftskrise;art623,1266318](http://www.tagblatt.ch/aktuell/wirtschaft/wirtschaft/Diamanten-Konzern-De-Beers-spuert-Wirtschaftskrise;art623,1266318)

11. [www.fine-diamonds.ch/fileadmin/Dateien/Artikel/stocks\\_Manche\\_verkaufen.pdf](http://www.fine-diamonds.ch/fileadmin/Dateien/Artikel/stocks_Manche_verkaufen.pdf)

# Space separation

Start

Dokument:Webseite

Neuer Text

Diamanten-Konzern De Beers spürt Wirtschaftskrise

Die Wirtschaftskrise setzt auch der Edelstein-Industrie zu. Beim weltgrößten Diamanten-Konzern De Beers schossen die drei Aushilfskassen in den vergangenen Monaten um über 500 Mio. Dollar an. Unter dem weltweiten Konjunkturschwung ließe die Nachfrage nach Diamanten und die Liquidität des Konzerns, teilte De Beers mit. Nach einem Rekordumsatz in den ersten neun Monaten sei das Geschäft Ende 2008 eingeknickt. Auch für 2009 zeichnet De Beers ein trübes Bild. "Wir erwarten, dass die Bedingungen das ganze Jahr über schwierig bleiben", erkläre das Unternehmen, das rund 40 Prozent des Rohdiamantenmarkts kontrolliert. Für das Gesamtjahr konnte De Beers noch ein Umsatzplus von einem Prozent auf 6,9 Mrd. Dollar verbuchen. Der Gewinn vor Zinsen, Steuern und Abschreibungen legte um 0,5 Prozent auf 1,2 Mrd. Dollar zu. Die Produktion sank um 6 Prozent auf 48,1 Mio. Karat - vor allem weil der Konzern in Südafrika ein Bergwerk verkauft und ein weiteres geschlossen hat.

Zum Thema

BILANZ

Buchvorbestud Jack Welch: Winnow + Keine Krone ohne Blütezeit

STOCKS

Online-Shop ist einblitz

Handelszeitung

Stimme aus demn Legenden werden

Handelszeitung

Womittels Geld recht zu versichern

STOCKS

Angeln Sie die Top-Aktien

BILANZ

Who's who: Das Who's who der Leute

Dokument:Webseite

Neuer Text

Diamanten-Konzern De Beers spürt Wirtschaftskrise

Die Wirtschaftskrise setzt auch der Edelstein-Industrie zu. Beim weltgrößten Diamanten-Konzern De Beers schossen die drei Aushilfskassen in den vergangenen Monaten um über 500 Mio. Dollar an. Unter dem weltweiten Konjunkturschwung ließe die Nachfrage nach Diamanten und die Liquidität des Konzerns, teilte De Beers mit. Nach einem Rekordumsatz in den ersten neun Monaten sei das Geschäft Ende 2008 eingeknickt. Auch für 2009 zeichnet De Beers ein trübes Bild. "Wir erwarten, dass die Bedingungen das ganze Jahr über schwierig bleiben", erkläre das Unternehmen, das rund 40 Prozent des Rohdiamantenmarkts kontrolliert. Für das Gesamtjahr konnte De Beers noch ein Umsatzplus von einem Prozent auf 6,9 Mrd. Dollar verbuchen. Der Gewinn vor Zinsen, Steuern und Abschreibungen legte um 0,5 Prozent auf 1,2 Mrd. Dollar zu. Die Produktion sank um 6 Prozent auf 48,1 Mio. Karat - vor allem weil der Konzern in Südafrika ein Bergwerk verkauft und ein weiteres geschlossen hat.

Zum Thema

STOCKS

Manche verkaufen ihre letzten Stücke

STOCKS

Online-Shop ist einblitz

Handelszeitung

Stimme aus demn Legenden werden

BILANZ

Who's who: Das Who's who der Leute

STOCKS

Millarden für den Börsenaufschwung

Handelszeitung

http://www.handelszeitung.ch/arsen/Finanz\_20080721.html

wmtrans



Abb. 4: Vergleich normale TF-IDF-Analyse vs. Wikipedia- und WMTrans-unterstützte Analyse

Abbildung 5 zeigt das Analyseergebnis der gefundenen Kategorien durch WikiTagging von fünf verschiedenen Artikeln. Dort wird über verschiedene Versicherungsunternehmen berichtet: *Bâloise*, *Zurich*, *Axa Winterthur*, *Nationale Suisse* und *Swiss Life*. Durch alleiniges Verarbeiten der Wörter innerhalb der Texte ist es sehr schwierig, hier eine Gemeinsamkeit zwischen den Texten zu finden. Durch WikiTagging kann man die gemeinsame Kategorie: *Versicherungsunternehmen (Schweiz)* identifizieren und eine Beziehung zwischen den Texten herstellen, die sonst nur schwer möglich gewesen wäre. Dadurch kann man beispielsweise beim Suchen einem Benutzer – alternativ zu einer »normalen« Suche – die Möglichkeit anbieten, seine Ergebnisse durch Themen zu filtern.

## 6 Schlussbetrachtung und Ausblick

Diese Arbeit zeigt zwei Wege auf, mit der gebräuchliche Textanalysen qualitativ gewinnen können. Einmal ist es erforderlich, gute »Werkzeuge« zur Vorverarbeitung von Dokumenten

einzusetzen, um das »Rauschen« innerhalb der Analyse möglichst klein zu halten. WMTrans eignet sich hierzu in den angewandten Fällen gut und steht für Deutsch, Englisch und Italienisch zu Verfügung. Durch diese »einfachen« Filterungen werden die Ergebnisse bewährter Algorithmen deutlich verbessert.

Das Spezielle an diesem Ansatz ist jedoch die Kombination der Wikipedia mit gebräuchlichen Verfahren und das daraus entwickelte »WikiTagging«. Mit der Wikipedia steht eine mächtige und gut strukturierte Datenbank zur freien Verfügung<sup>12</sup>. Da die Betreiber der Wikipedia in regelmäßigen Abständen einen aktuellen Download anbieten, kann selbst zeitgemäßer Inhalt angemessen analysiert und verarbeitet werden. Vor allem darf man nicht außer Acht lassen, dass die Wikipedia jeden Tag weiter wächst und die Qualität der Einträge einer relativ guten Selbstkontrolle unterliegt. Zwar muss man selbst Hand anlegen, um mit der Wikipe-

12. <http://de.wikipedia.org/wiki/Wikipedia:Lizenzbestimmungen>

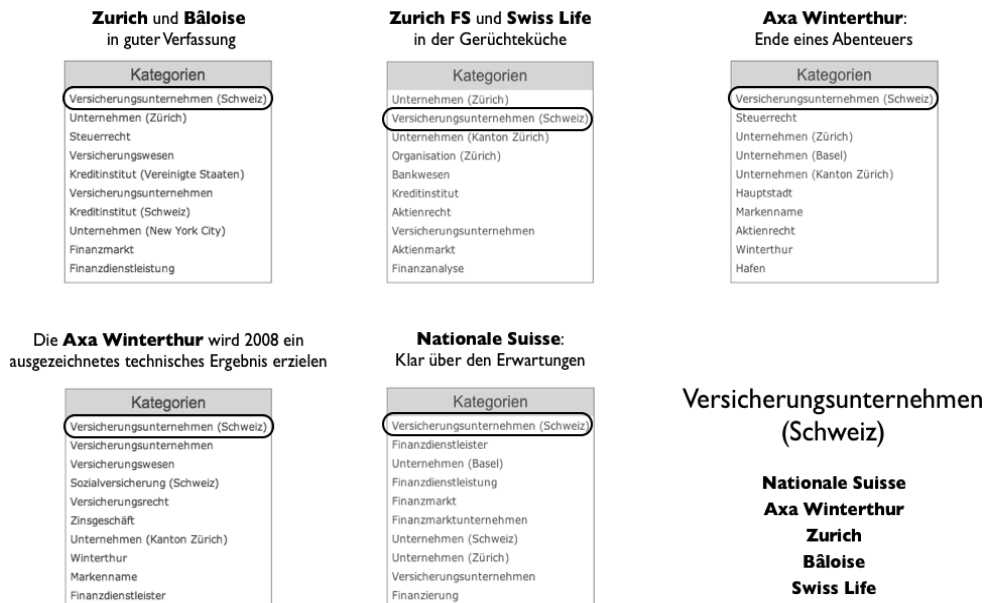


Abb. 5: Das Auffinden gemeinsamer Kategorien von scheinbar unterschiedlichen Artikeln

die Textanalysen durchführen zu können, aber diese Mühe lohnt sich.

In dieser Arbeit wurde nur rudimentär auf die weiteren Analyse- und Einsatzmöglichkeiten der Wikipedia eingegangen (Disambiguierungsmöglichkeiten in Abschnitt 2), es sollte aber erwähnt werden, dass die kausal vorhandenen Linkstrukturen innerhalb der Wikipedia noch in weiteren Fällen herangezogen werden können (Rankings, Auffinden weiterer Informationen etc.), dies aber den Rahmen dieser Arbeit sprengen würde.

Diese Arbeit hatte insgesamt zum Ziel, das Verständnis dafür zu schärfen, wie man mithilfe der Wikipedia »normale« Verfahren qualitativ stark bereichern kann, und zu zeigen, dass es dazu noch zahlreiche weitere Möglichkeiten gibt, bessere Strukturen in große Datenbanken zu bringen. Die Autoren sind sich sicher, dass in absehbarer Zeit noch weitere Einsatzmöglichkeiten mit der Wikipedia gefunden werden, um der digitalen Informationsflut Herr zu werden.

## 7 Literatur

- [Bunescu & Pasca 2006] *Bunescu, R.; Pasca, M.*: Using Encyclopedic Knowledge for Named Entity Disambiguation. In: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06), Trento, Italy, 2006, S 9-16.
- [Cucerzan 2007] *Cucerzan, S.*: Large-Scale Named Entity Disambiguation Based on Wikipedia Data. In: Proceedings of Empirical Methods in Natural Language Processing (EMNLP 2007), Prague, Czech Republic, 2007, S. 708-716.
- [Finkelstein et al. 2002] *Finkelstein, L.; Gabrilovich, Y. M.; Rivlin, E.; Solan, Z.; Wolfman, G.; Ruppin, E.*: Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20 (1), 2002, S. 116-131.
- [Gabrilovich & Markovitch 2006] *Gabrilovich, E.; Markovitch, S.*: Overcoming the brittleness bottleneck using Wikipedia: Enhancing text categorization with encyclopedic knowledge. In: Proceedings of the 21st National Conference on Artificial Intelligence, Boston, MA, 2006, S. 1301-1306.
- [Gabrilovich & Markovitch 2007] *Gabrilovich, E.; Markovitch, S.*: Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI'07), Hyderabad, India, 2007.
- [Hacken 2009] *Hacken, P. ten*: WordManager. In: State of the Art in Computational Morphology, Workshop on Systems and Frameworks for Computational Morphology (SFCM 2009), Zurich, Proceedings Series: Communications in Computer and Information Science, Vol. 41, Springer-Verlag, 2009.
- [Karttunen 1994] *Karttunen, L.*: Constructing Lexical Transducers. In: The Proceedings of the 15th International Conference on Computational Linguistics. Coling 94, 1, Kyoto, Japan, 1994, S. 406-411.
- [Koskenniemi 1983] *Koskenniemi, K.*: Two-level Morphology. A General Computational Model for Word-Form Recognition and Production. Department of General Linguistics, University of Helsinki, 1983.
- [Milne & Witten 2008] *Milne, D.; Witten, I. H.*: Learning to link with Wikipedia. In: Proceedings of the ACM Conference on Information and Knowledge Management (CIKM'2008), Napa Valley, California, 2008.
- [Salton & McGill 1983] *Salton, G.; McGill, M. J.*: Introduction to modern information retrieval. McGraw-Hill, 1983.

Dr. Stephan Gillmeier  
 Dr. Urs Hengartner  
 Dr. Sandro Pedrazzini  
 Canoo Engineering AG  
 Kirschgartenstr. 5  
 CH-4051 Basel  
 {stephan.gillmeier, urs.hengartner,  
 sandro.pedrazzini}@canoo.com  
 www.canoo.com