Explicit Semantic Analysis

Christian Eiserloh

Seminar Information Retrieval im SoSe 2012
Dozent: Dr. Thomas Gottron

Abstract. Explicit Semantic Analysis ist eine neue Methode, um semantische Verwandtschaft zwischen natürlichsprachlichen Texten zu ermitteln. Der Text wird dabei als gewichteter Vektor in einem hochdimensionalen Raum von Konzepten repräsentiert. Diese Konzepte werden aus umfangreichen Quellen menschlichen Wissens wie beispielsweise Wikipedia abgeleitet. Im Vergleich zu herkömmlichen Verfahren sind hierbei erhebliche Verbesserungen bezüglich der Übereinstimmung von berechneter Verwandtschaft und menschlicher Beurteilung festzustellen.

Keywords: Concept-based, Information Retrieval, Wikipedia, Explicit Semantic Analysis

1 Grundprinzip von Explicit Semantic Analysis

Das Ermitteln semantischer Verwandtschaft zwischen natürlichsprachlichen Texten setzt umfangreiches Fach- und Allgemeinwissen voraus. Für den Menschen ist dies eine einfache Aufgabe, für Computer hingegen eine scheinbar unüberwindbare Hürde. Mit Explicit Semantic Analysis (ESA) wurde eine neue Methode zum Ermitteln semantischer Verwandtschaft entwickelt. Die grundlegende Idee von ESA ist es, semantische Verwandtschaft nicht wie die herkömmlichen Verfahren auf Wort-Ebene zu ermitteln, sondern auf der Ebene von Konzepten. Konzepte sind die Basiseinheiten der Bedeutung. Sie ermöglichen dem Menschen, sein Wissen zu organisieren und zu teilen. ESA repräsentiert dabei natürlichsprachliche Texte in einem hochdimensionalen Raum von Konzepten. Im Fall von Wikipedia-basiertem ESA werden die Konzepte von Wikipedia, der zurzeit größten Enzyklopädie, abgeleitet. Der Text wird dabei als gewichteter Vektor von Konzepten repräsentiert. Damit sind, wie bei anderen Vektor-basierten Verfahren, Vergleiche mit konventionellen Metriken, wie beispielsweise der Kosinus-Ähnlichkeit, möglich. Im Vergleich zu herkömmlichen Methoden ist bei ESA eine deutliche Verbesserung in der Übereinstimmung von berechneter Verwandtschaft und menschlicher Beurteilung festzustellen (vgl. [1]).

1.1 Wikipedia-basiertes Explicit Semantic Analysis

Der Grundgedanke von Wikipedia-basiertem ESA ist es, bei der Repräsentation von Texten auf eine riesige Menge von hochorganisiertem menschlichem Wissen zurück-

zugreifen. Wikipedia ist der zurzeit größte Wissensspeicher im Web und wächst kontinuierlich in Breite und Tiefe. Trotz des Ansatzes der freien Bearbeitung zeichnet sich die Enzyklopädie durch eine bemerkenswerte Qualität aus. Jeder Wikipedia-Artikel behandelt genau ein Thema, das detailliert beschrieben wird. Es ist naheliegend, dass das Wikipedia-basierte ESA diese Gegebenheit nutzt und entsprechend Konzepte durch Wikipedia-Artikel definiert. Es handelt sich also um natürliche, „explizite“ Konzepte, die von Menschen selbst definiert wurden (vgl. [1]).

1.2 Bestandteile und grundlegendes Vorgehen

Das grundlegende Vorgehen ist in Abbildung 1 veranschaulicht.

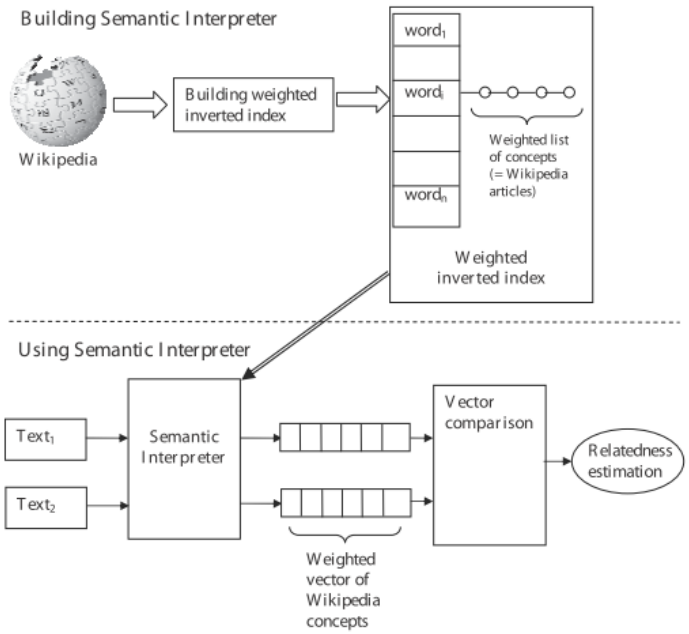

Fig. 1. Grundlegendes Vorgehen beim Explicit Semantic Analysis (Quelle: [1]).

Wie in Abbildung 1 zu sehen, werden Wikipedia-Artikel als Vektoren von darin enthaltenen Wörtern repräsentiert. Dabei findet eine Gewichtung mittels TF-IDF statt. Die Gewichtung drückt jeweils die Stärke der Assoziation zwischen dem Wikipedia-Artikel, also dem Konzept, und dem einzelnen Wort aus. Um die semantische Interpretation zu beschleunigen, wird ein Inverted Index verwendet. Dieser bildet jedes Wort auf eine Liste von Konzepten ab, in denen das Wort enthalten ist. Zur Generierung des Konzept-Vektors, der ein gegebenes Textfragment repräsentiert, wird ein Semantischer Interpreter verwendet. Dieser bildet das natürlichsprachliche Textfragment auf eine gewichtete Sequenz von Wikipedia-Konzepten ab. Dazu wird zunächst mittels TF-IDF ein gewichteter Vektor zum gegebenen Textfragment erzeugt. Anschließend iteriert der Semantische Interpreter über die Wörter im Vektor und sam-

melt die Konzepte, auf die die entsprechenden Wörter im Inverted Index verweisen. Aus diesen Konzepten wird schließlich der ebenfalls gewichtete Konzept-Vektor für das Textfragment zusammengesetzt. Die Bedeutung eines Textfragments wird also ausgedrückt durch seine Verwandtschaft zu Wikipedia-Konzepten. Der resultierende gewichtete Konzept-Vektor wird auch „Interpretationsvektor“ genannt. Möchte man nun zwei Textfragmente miteinander vergleichen, kann man einen Vektorvergleich, zum Beispiel anhand der Kosinus-Ähnlichkeit, durchführen.

2 Konzept-basiertes Information Retrieval mit ESA

Beim Information Retrieval geht es darum, die zu einer Nutzeranfrage relevantesten Dokumente aus einem Korpus zu ermitteln. Die heute dazu existierenden Verfahren erlauben Anfragen in natürlicher Sprache und werden unterstützt von automatischen Indexierungsmethoden. Die etablierten Verfahren greifen dabei auf die Bag-of-Words-Repräsentation („BOW“) zurück, die alle Terme als unabhängige Schlüsselwörter betrachtet und Textfragmente zum Beispiel als gewichtete Vektoren dieser Schlüsselwörter darstellt. Die Bag-of-Words-Repräsentation leidet allerdings unter dem sogenannten Vokabelproblem. Man unterscheidet dabei zwischen dem Synonymie-Problem und dem Polysemie-Problem. Das Synonymie-Problem entsteht dadurch, dass der Nutzer andere Schlüsselwörter verwendet als die Autoren des Textes. Die semantische Übereinstimmung von Synonymen wird also nicht beachtet. Dies hat einen geringeren Recall zur Auswirkung. Das Polysemie-Problem resultiert aus der kontextuellen Differenz von Schlüsselwörtern. Da der Kontext nicht berücksichtigt wird, werden Dokumente aufgrund von übereinstimmenden Termen als relevant zurückgeliefert, obwohl diese Terme in einem anderen Kontext auftreten und damit eine völlig andere Bedeutung haben. Es werden also vermehrt nicht-relevante Dokumente fälschlicherweise als relevant eingestuft, womit die Precision geringer ist (vgl. [2]).

Die grundlegende Idee beim Konzept-basierten Information Retrieval ist das Verwenden semantischer Konzepte, anstelle von oder zusätzlich zu Schlüsselwörtern. Damit ist die Repräsentation von Textfragmenten weniger abhängig von einzelnen Termen. Die genannten Probleme der Synonymie und Polysemie werden dabei deutlich gelindert. Konzept-basierte Verfahren finden auch relevante Dokumente, die keine übereinstimmenden Terme mit der Anfrage haben. Dies verringert das Synonymie-Problem und erhöht entsprechend den Recall. Gleichzeitig werden nicht-relevante Dokumente auch dann als solche erkannt, wenn gleiche Wörter wie in der Anfrage vorkommen. Damit ist auch das Polysemie-Problem verringert und entsprechend die Precision erhöht (vgl. [2]).

Explicit Semantic Analysis (ESA) kann für Konzept-basiertes Information Retrieval verwendet werden. ESA verwendet explizite Konzepte, die von Menschen definiert wurden. Die Methode greift existierende Verfahrensweisen des Information Retrieval auf und verwendet Konzepte sowohl bei der Indexierung als auch beim Retrieval. Des Weiteren greift ESA auf etablierte Datenstrukturen und Rankingme-

thoden zurück (vgl. [2]). In den folgenden Abschnitten wird ein in [2] entwickeltes ESA-Retrieval-System vorgestellt.

3 ESA-basiertes Retrieval

Im Folgenden wird ein erster Algorithmus zum Konzept-basierten Retrieval, basierend auf ESA, aus [2] vorgestellt. Dabei wird sowohl beim Indexieren, als auch beim Retrieval auf einen Wikipedia-basierten Konzept-Raum zurückgegriffen. Anschließend wird eine Evaluation dieses grundlegenden Verfahrens vorgestellt.

3.1 ESA-basierte Indexierung

Beim Explicit Semantic Analysis wird jedes Dokument als gewichteter Vektor von Konzepten repräsentiert. Da es unzählige Konzepte gibt, Dokumente aber in der Regel nur mit einer Teilmenge dieser Konzepte beschrieben werden, handelt es sich meist um einen nur spärlich gefüllten Vektor. Es tritt also wie bei BOW-Vektoren das Phänomen auf, dass der Vektor viele Null-Werte enthält. Gleichzeitig hat jedoch jedes Wort Beziehungen zu einer riesigen Menge von Konzepten. Effizientes Indexieren ist daher unter diesen Umständen nicht machbar. Eine Lösung hierfür ist die Beschränkung auf diejenigen Konzepte mit den höchsten Gewichten. Verwendet man einen nach Gewichten sortierten Vektor, findet man die resultierende Teilmenge der Konzepte entsprechend im Präfix des Vektors (vgl. [2]). Die Größe dieser Teilmenge wird durch einen sogenannten Cutoff festgelegt. Dieser gibt an, wie viele Konzepte beibehalten werden. Hierbei sollte man darauf achten, den Cutoff nicht zu groß zu wählen, da dies enorme Speicher- und Berechnungskosten zur Folge hätte.

Die Konzept-basierte Repräsentation von langen Dokumenten ist problematisch. Für eine gegebene Anfrage ist meist nur ein kleiner Abschnitt des Dokuments relevant. Dieser relevante Absatz kann im Konzept-Vektor des gesamten Dokuments allerdings unterrepräsentiert sein. Verfolgt man den genannten Ansatz des Beschränkens auf die Konzepte mit den höchsten Gewichten, so kann es sogar vorkommen, dass die relevanten Konzepte des kurzen Abschnitts wegen zu geringer Gewichte aus dem Index-Vektor entfernt wurden. Ein vergleichbares Problem existiert bei der BOW-Repräsentation von Dokumenten. Dabei besteht ein Lösungsansatz darin, die Term Frequency-Werte (TF) von Dokumenten unterschiedlicher Länge zu normalisieren. Ein ähnlicher Ansatz kann bei der Konzept-basierten Repräsentation angewendet werden. Danach werden lange Dokumente in kürzere Passagen aufgeteilt. Die einzelnen Passagen werden schließlich einem Ranking bezüglich ihrer Relevanz zum ursprünglichen Gesamt-Dokument unterzogen (vgl. [2]). Hierbei wurde beobachtet, dass eine Einteilung in Passagen fester Länge zu besseren Ergebnissen führt, als eine Trennung nach Syntax oder Semantik (siehe ebd.). Der letztlich in [2] realisierte Algorithmus verwendet überlappende Passagen fester Länge, die jeweils eine individuelle Repräsentation erhalten.

Zusammenfassend lässt sich festhalten, dass jedes Dokument in eine Menge von Passagen mit jeweils eigenem Konzept-Vektor aufgeteilt wird. Es wird der im Infor-

mation Retrieval übliche Inverted Index verwendet, wobei die Bezeichner der Konzepte als Token dienen. Äquivalent zum Term Frequency-Verfahren ergibt die Auswertung für jedes Konzept das Token-Gewicht (vgl. [2]).

3.2 ESA-basierter Retrieval Algorithmus

Das Verfahren konvertiert zunächst die gegebene Anfrage in einen ESA Konzept-Vektor. Die Anfrage wird also genauso repräsentiert wie die Dokumente im Korpus. Um eine sinnvolle Balance zwischen vollständigen Dokumenten und enthaltenen Passagen zu gewährleisten, werden die Relevanz-Auswertung des vollständigen Dokumentes und die der besten Passage summiert. Die Dokumente werden anschließend nach diesem kombinierten Wert sortiert. Als Ergebnis werden schließlich die Dokumente mit den besten Werten zurückgegeben. Das Verfahren lässt sich über einen Parameter s anpassen, der den Cutoff des Konzept-Vektors der Anfrage festlegt. Genau wie in der Indexierungsphase werden nur diejenigen Konzepte mit den höchsten Gewichten beibehalten. Wie bereits erwähnt, ist ein hoher Cutoff bei der Indexierung des Korpus äußerst ineffizient (siehe Abschnitt 3.1). Bei der Anfrage hingegen kann man zur präziseren Repräsentation einen höheren Cutoff verwenden, ohne einen bedeutenden Kostenanstieg in Kauf nehmen zu müssen (vgl. [2]).

3.3 Evaluation und Analyse

In [2] werden Experimente zur Auswertung der Nützlichkeit von ESA-basiertem Retrieval beschrieben und analysiert. Im Folgenden werden das Vorgehen und die Ergebnisse zusammengefasst.

Die Basis der Experimente ist Xapian, eine quelloffene, wahrscheinlichkeitesbasierte Information Retrieval Bibliothek. Es wurde der Xapian Inverted Index verwendet. Für einen Vergleich mit BOW-Retrieval wurde die entsprechende Xapian Implementation von Okapi BM25 Ranking verwendet. Es wurde auf den Datensätzen TREC-8 Adhoc und Robust-04 gearbeitet, welche dieselben 528.000 Dokumente sowie jeweils etwa 50 unterschiedliche Themengebiete behandeln. Um die Experimente möglichst realitätsnah zu gestalten, wurden nur die kurzen der in TREC-8 vorgegebenen Anfragen („title“-Anfragen) verwendet, die 1-3 Wörter umfassen. Als Evaluationsmaß wurde Mean Average Precision (MAP) verwendet. Beim ESA wurden die Dokumente in überlappende Passagen, die 50 Wörter umfassen, eingeteilt. Beim Indexieren wurde ein Vektor Cutoff von s=50 Konzepten gewählt (siehe [2]).

Ergebnisse der durchgeführten Experimente sind in Abbildung 2 veranschaulicht. Abbildung 2 zeigt, dass mit der Einteilung in Passagen deutlich bessere Ergebnisse erzielt werden als mit vollständigen Dokumenten. Die Kombination beider Ansätze liefert im Vergleich die besten Ergebnisse. Ebenfalls zu erkennen ist, dass die nachrangigen Konzepte im Vektor bei höherem Cutoff keine Verbesserung bringen, sondern Recall und Precision verringern. Dies lässt darauf schließen, dass sich einige Konzepte nachteilig auswirken, was auf die Notwendigkeit einer weitergehenden Auswahl der Anfrage-Konzepte hindeutet. Eine wichtige Erkenntnis dieses Experiments ist die Beobachtung, dass ESA mit einem MAP von 0.1760 schlechter ab-

schneidet als BOW (MAP von 0.2481). Dies wird insofern als überraschend beschrieben, als dass ESA in früheren Textanalyse-Anwendungen bessere Ergebnisse liefern konnte (siehe [2]).

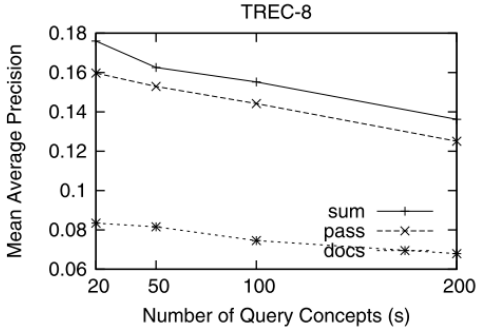

Fig. 2. ESA-basiertes Retrieval: Performance als Funktion des ESA Cutoffs (Quelle: [2]).

In [2] wird bei der Analyse der durchgeführten Experimente positiv festgehalten, dass ESA relevante Dokumente selbst dann erkannt hat, wenn keine Terme aus der Anfrage und auch keine einfachen Synonyme enthalten waren. Als Problem wurde hingegen ein Effekt erkannt, der dem der Anfrageverzerrung bei exzessiver Anfrageerweiterung ähnelt. Besteht die Anfrage beispielsweise aus zwei Termen, kann es vorkommen, dass Konzepte aufgrund eines hohen TF-IDF-Wertes bezüglich eines der Terme als relevant eingestuft werden, obwohl sie für die gesamte Anfrage eigentlich nicht relevant sind. Man muss also davon ausgehen dass der initiale Konzept-Vektor Störfaktoren und Ungenauigkeiten beinhaltet (vgl. [2]). Daher ist eine präzisere Auswahl der Konzepte notwendig. Der nächste Abschnitt behandelt eine entsprechende Erweiterung des Verfahrens.

4 Selektives ESA-basiertes Retrieval

Wie im vorherigen Abschnitt beschrieben, kann die grundlegende ESA-Repräsentation Störfaktoren und Ungenauigkeiten enthalten. Wie bereits angedeutet, ist daher eine Abstimmung der betrachteten Konzepte auf die Anfrage sinnvoll. Es wäre zwar auch möglich die Repräsentation der Dokumente anzupassen, allerdings wäre eine erneute Indexierung notwendig, was kostenintensiv wäre. Außerdem sind Dokumente aufgrund ihrer Länge in der Regel weniger anfällig für Störungen. Im Gegensatz dazu ist die meist sehr kurze Anfrage in dieser Hinsicht äußerst anfällig. Außerdem gibt die Anfrage den Kontext vor. Daher ist es sinnvoll, zur Beschreibung der Anfrage eine präzisere Auswahl von Konzepten zu treffen (vgl. [2]). Im Folgenden wird zunächst das Grundprinzip erläutert. Anschließend wird der angepasste Retrieval Algorithmus aus [2] vorgestellt. Abschließend werden die erzielten Verbesserungen beschrieben.

4.1 Feature Selection mit Pseudo-Relevance Feedback

Als Feature Selection bezeichnet man Verfahren, die anhand von benannten Trainingsdaten die Nützlichkeit von Features, die hier den Konzepten entsprechen, bewertet. Im Gegensatz zur Text-Kategorisierung besitzt das Information Retrieval allerdings keine solchen benannten Trainingsdaten. Daher ist hier ein alternatives Verfahren zur Bewertung der Nützlichkeit von Features notwendig. Man verwendet in diesem Fall Relevance Feedback. Beim Relevance Feedback erhält der Nutzer eine initiale Menge von Ergebnissen und hat die Aufgabe, diese bezüglich ihrer Relevanz zu bewerten. Das Feedback des Nutzers wird anschließend verwendet, um die Anfrage umzuformulieren, sodass bessere Ergebnisse geliefert werden können. Das Feedback übernimmt damit die Rolle der Trainingsdaten. Dieses Verfahren lässt sich auch automatisieren, in dem Sinne, dass der Nutzer nicht mehr involviert werden muss. Man spricht dabei von Pseudo-Relevance Feedback (PRF). Dabei werden die hochrangigen Ergebnisse der initialen Anfrage als relevant angenommen. Der in [2] schließlich verwendete Ansatz ist inspiriert durch PRF.

Beim verwendeten Ansatz werden die initialen Ergebnisse eines Schlüsselwort-basierten Retrievals als Grundlage für die Auswertung verwendet. Es handelt sich somit um eine zwei-phasige Retrieval Methode. Im ersten Schritt wird ein Schlüsselwort-basiertes Retrieval durchgeführt. Die Ergebnisse legen nun fest, wie die Konzepte der Anfrage angepasst werden. Im zweiten Schritt wird schließlich mit der angepassten Anfrage ein Konzept-basiertes Retrieval durchgeführt.

Es wird zwischen zwei Möglichkeiten unterschieden, anhand der initialen Ergebnisse, eine Teilmenge von Konzepten für die angepasste Anfrage auszuwählen. Zum einen kann man die hochrangigen, pseudo-relevanten Dokumente als positive Beispiele verwenden. Zum anderen kann man analog die nachrangigen, pseudo-nicht-relevanten Dokumente der initialen Ergebnismenge als negative Beispiele nutzen. Im vorgestellten Verfahren wird von beiden Möglichkeiten Gebrauch gemacht, da beides Vorteile mit sich bringt. Die positiven Beispiele liefern zusätzliche relevante Konzepte, die negativen Beispiele entfernen nicht-relevante Konzepte aus der initialen Anfrage (vgl. [2]). Auch hier beim Schlüsselwort-basierten Retrieval hat man die Beobachtung gemacht, dass Passagen deutlich bessere Ergebnisse liefern als vollständige Dokumente (siehe ebd.). Daher wurden bei der Auswahl von positiven und negativen Beispielen entsprechend hochrangige und nachrangige Passagen gewählt.

4.2 Selektiver ESA-basierter Retrieval Algorithmus

Wie im vorherigen Abschnitt herausgestellt, wird die Repräsentation der Anfrage modifiziert, indem der bereits vorgestellte Retrieval Algorithmus entsprechend angepasst wird. Der resultierende Prozess ist in Abbildung 3 dargestellt.

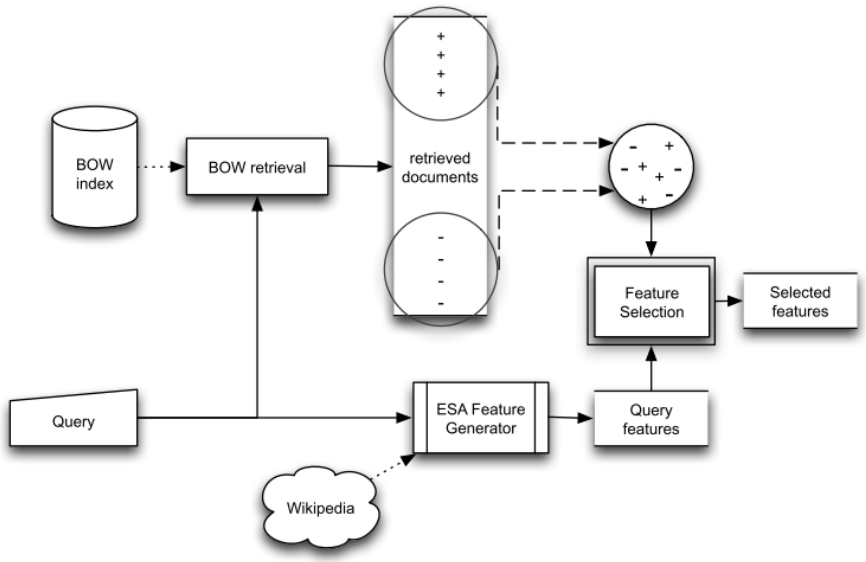

Fig. 3. PRF-basierter Feature Selection Prozess (Quelle: [2]).

Wie in Abbildung 3 dargestellt, wird zunächst der Konzept-Vektor zur gegebenen Anfrage erzeugt. Anschließend werden die ersten n Ergebnisse eines Schlüsselwort-basierten Retrievals festgehalten. Davon werden die oberen k Ergebnisse als pseudo-relevant bzw. positive Beispiele markiert, die unteren k Ergebnisse als pseudo-nicht-relevante bzw. negative Beispiele. Das Feature Selection Verfahren wählt nun anhand dieser Beispiele die besten Konzepte aus dem ursprünglichen Anfrage-Vektor und erzeugt daraus einen modifizierten Anfrage-Vektor. Auf diesem angepassten Vektor wird nun Konzept-basiertes Retrieval durchgeführt.

4.3 Evaluation und Analyse

In [2] wurden drei Feature Selection (FS) Methoden evaluiert: FS mit Information Gain (IG), FS mit inkrementellem Information Gain (IIG) sowie FS mit einem Rocchio Vektor (RV) (Näheres siehe [2]). Zum Vergleich wurde eine vierte Methode herangezogen, die die Teilmenge der Features zufallsbasiert wählt. Es wurde jeweils ein Cutoff von s=50 verwendet. Die Ergebnisse wurden in Abhängigkeit von zwei Parametern betrachtet. Der erste Parameter k ist die Größe der pseudo-relevanten Ergebnismenge. Der zweite Parameter θ beschreibt das Aggressivitätslevel der Feature Selection (siehe [2]). Die beiden Parameter können durch Trainingsdurchläufe abgestimmt werden. Dies ist auch erforderlich, da die Parameter erheblichen Einfluss auf die Ergebnisse haben (vgl. ebd.). Da in [2] ein relativ konsistentes Systemverhalten beobachtet werden konnte, wird davon ausgegangen, dass die erzielte Abstimmung auch für andere Testmengen verwendet werden kann.

Anhand der Experimente in [2] zur Feature Selection wurde bestätigt, dass diese Verfahren tatsächlich effektiv sind und zu besseren Retrieval Ergebnissen führen, die nicht nur mit der kleineren Menge von Konzepten zu erklären ist. Bei den Experimenten wurden durch das Hinzufügen von Feature Selection zum Retrieval Verbesserungen von bis zu 40% erreicht. Allerdings liegt man damit immer noch erst bei 85% des BOW-Retrievals (vgl. [2]). Wie man dennoch bessere Ergebnisse erreicht, wird im nächsten Abschnitt beschrieben.

5 Fusioniertes selektives ESA-basiertes Retrieval

Um die Vorteile des Konzept-basierten Ansatzes zu nutzen und gleichzeitig besser zu sein als Schlüsselwort-basiertes Retrieval, greift man auf ein Verfahren zurück, das im Information Retrieval häufiger zum Einsatz kommt: Die Fusion. Bei der Fusion handelt es sich um das Kombinieren verschiedener Retrieval Methoden. Genauer gesagt werden die Ergebnisse der verwendeten Methoden kombiniert, was zu besseren Endergebnissen führen kann. Der erzielte Effekt ist umso stärker, je deutlicher sich die kombinierten Verfahren unterscheiden. Da BOW und ESA schon vom Ansatz her einen signifikanten Unterschied aufweisen, ist eine deutliche Verbesserung durch Kombination zu erwarten (vgl. [2]).

5.1 Vorgehen

Die Auswertung der Dokumente wird als gewichtete Summe der Auswertungen der einzelnen Retrievalmethoden berechnet. Hier verwendet man das weit verbreitete Modell der Linearkombination. Dazu werden zunächst die Dokument-Auswertungen beider Retrievalverfahren normalisiert. Anschließend werden die resultierenden Werte gewichtet und aufsummiert. Den Gewichtungsfaktor erhält man durch Parameter-Abstimmung auf einem Datensatz mit relevanten Beurteilungen (vgl. [2]).

5.2 Das MORAG System

Das gesamte System mit allen vorgestellten Komponenten nennen die Entwickler MORAG (siehe [2]). Die Architektur ist in Abbildung 4 dargestellt.

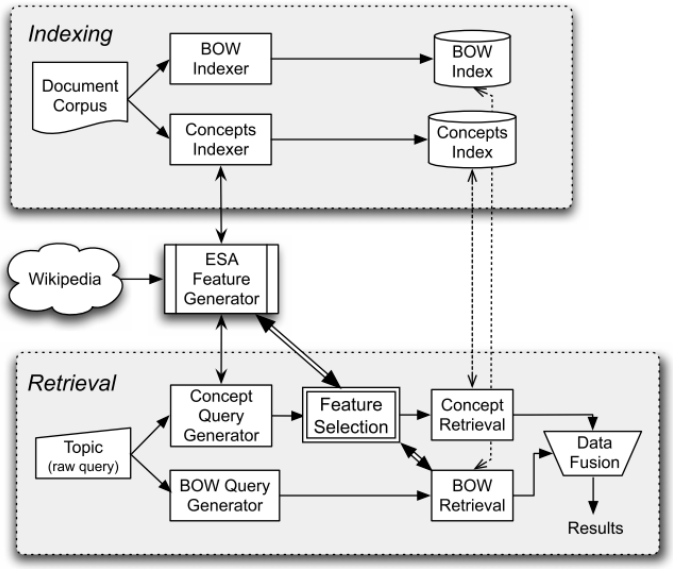

Fig. 4. Das MORAG System (Quelle: [2]).

Zunächst wird ein ESA-Modell aus Wikipedia oder einer anderen Quelle erstellt (siehe Abbildung 4). Anschließend wird die Indexierung des Korpus in BOW- und in ESA-Repräsentation durchgeführt. An die Indexierungsphase schließt sich die eigentliche Retrievalphase an. Dabei wird zuerst die BOW-Anfrage übermittelt. Die Ergebnisse werden für die Fusionsphase aufbewahrt. Außerdem werden die Ergebnisse zusammen mit der ESA-Anfrage in das Feature Selection Modul übergeben. An dieser Stelle wäre es auch möglich, zwei unterschiedliche BOW-Systems zu verwenden, eines für die Fusion, ein anderes für die Feature Selection (vgl. [2]). Die resultierenden Features werden schließlich für das Konzept-basierte Retrieval verwendet. Abschließend werden die Ergebnisse des Konzept-basierten Retrievals mit denen des Schlüsselwort-basierten Retrievals fusioniert. Das Resultat entspricht dem finalen MORAG Ergebnis.

5.3 Evaluation und Analyse

In [2] wurden umfangreiche Experimente zur Evaluation des MORAG Systems durchgeführt. Im Folgenden werden die wichtigsten Erkenntnisse zusammengefasst.

Es wurden beeindruckende Verbesserungen zur BOW Basislinie für alle drei Feature Selection Methoden festgestellt (siehe Abbildung 5). Die Ergebnisse zeigen, dass für große Werte von k (Größe der pseudo-relevanten Feature-Menge) die Rocchio Vektor (RV) Methode am besten funktioniert (Details siehe [2]). Des Weiteren wurde gezeigt, dass die Fusion mit ESA-basierten Ergebnissen zu einer Verbesserung führt, die quantitativ etwa der ESA Performance selbst ähnelt. Wichtig ist die Erkenntnis, dass ESA einzeln betrachtet zwar eher schwach ist, bei Fusion aber zu deutlicher Verbesserung führt (vgl. [2]).

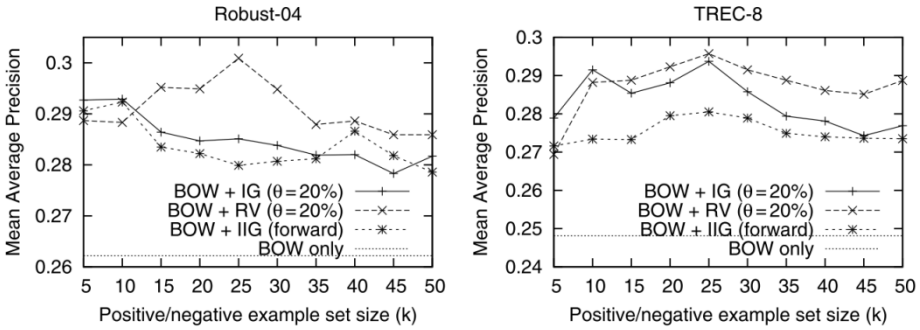

Fig. 5. MORAG Performance als Funktion der Anzahl von pseudo-relevanten Beispielen (k), alle FS Methoden, mit Fusion (Quelle: [2]).

Um den wahren Beitrag der ESA-Konzepte zu zeigen, wurde in [2] gezeigt, dass die Fusion verschiedener BOW-Systeme sehr viel schwächer ist als die Fusion von BOW und ESA bei MORAG. Damit wurde bestätigt, dass die Verbesserungen nicht alleine der Fusion angerechnet werden können, sondern mit dem hinzugefügten Wert der Konzept-basierten Retrievalmethode zu erklären sind. Als Grund für die besseren Ergebnisse der Fusion von ESA und BOW wurde die geringere Überlappung der Systeme ausgemacht. Diese resultiert in einer höheren Wahrscheinlichkeit dafür, dass beide Systeme jeweils neue relevante Dokumente hinzufügen. Der Recall fällt somit höher aus (vgl. [2]).

Mit den getesteten Feature Selection Verfahren, die auf Pseudo-Relevanz beruhen, sinkt in der Regel die Qualität der Ergebnisse mit steigender Größe der verwendeten Beispiele (vgl. ebd.). Ideal wäre Feature Selection dann, wenn vom Nutzer als relevant oder nicht-relevant markierte Dokumente benutzt werden würden. In [2] wurden Tests durchgeführt, bei denen nur die tatsächlich relevanten Dokumente unter den hochrangigen als positive Beispiele verwendet wurden. Die Beispiele sind dann zwar wegen gleichbleibender Anzahl aus einer größeren Teilmenge gewählt, aber garantiert relevant. Hierbei wurde festgestellt, dass dies zu Verbesserungen von 10-15% führen kann (siehe [2]). Dies ist ein Indiz dafür, dass raffiniertere Methoden zur Auswahl von pseudo-relevanten Dokumenten wertvoll sein können.

ESA-basierte Performance hängt direkt von der Auswahl der Features ab. Interessant ist die Frage, welche Verbesserungen bei MORAG mit besseren Feature Selection Methoden erwartet werden können. Dazu wird in [2] ein Experiment vorgestellt, das diese Frage klären soll. Bei dem Experiment wurden alle möglichen Teilmengen mit maximal 4 Elementen aus einer größeren Menge von Features ausgewählt und bezüglich ihrer Performance miteinander verglichen. Mit der Auswahl der besten Teilmengen konnten hierbei bessere Ergebnisse als mit Schlüsselwort-basierten Systemen und weit bessere Ergebnisse als mit reinem ESA erzielt werden. Durch Fusion mit BOW-Ergebnissen konnte sogar eine Verbesserung von fast 50% erreicht werden (siehe [2]). Im Rahmen dieses Experiments wurde zudem bestätigt, dass die Auswahl weniger, qualitativ hochwertiger Features zu deutlich besseren Ergebnissen führt als eine größere Auswahl von Features. Die optimale Anzahl hängt letztlich von der Anfrage und der Qualität der Features ab (vgl. ebd.). Zusammenfassend lässt sich fest-

halten, dass bessere Feature Selection Verfahren direkt zu besseren Ergebnissen führen.

Die in [2] vorgestellten Tests des MORAG Systems wurden auf einem handelsüblichen Rechner mit einem Vier-Kern-Prozessor und 12GB RAM durchgeführt. Die Ergebnisse zu einer Anfrage konnten damit in unter einer Sekunde geliefert werden. Die Indexe benötigten für die Datensätze TREC4-5 etwa 40GB Speicherplatz (siehe [2]). Damit kann das MORAG System auch ohne Optimierungen bezüglich der Rechenperformance als relativ effizient eingestuft werden.

6 Weitergehende Selektionsmethoden

Bei der Auswertung des MORAG Systems (siehe vorheriger Abschnitt) wurde festgehalten, dass eine bessere Feature Selection Komponente direkt zu besseren Retrieval Ergebnissen führen würde. Es ist also erstrebenswert, zu einer Anfrage oder einem Themengebiet diejenigen Features (Konzepte) zu identifizieren, die das entsprechende Themengebiet möglichst gut repräsentieren und es möglichst gut von anderen Themengebieten abgrenzen. Aus der Sicht der Text-Klassifikation geht es also darum, anhand von Trainingsdokumenten semantisch wertvolle Wikipedia-Artikel (Konzepte) auszuwählen. Damit sollen letztlich die Verbesserungen erreicht werden, die mit dem Verwenden einer besseren Feature Selection Komponente erwartet werden. Im Folgenden werden zum einen entsprechende Selektionsverfahren vorgestellt, die auf Text-Klassifikation basieren. Zum anderen soll geklärt werden, ob die Auswahl von Konzepten effektiver ist, wenn man die Verwandtschaft von Konzepten modelliert.

6.1 ESA für Text-Klassifikation

Bevor auf die einzelnen Verfahren eingegangen wird, wird an dieser Stelle zunächst beschrieben, wie ESA in der Text-Klassifikation eingesetzt werden kann. Die zur Klassifikation gegebenen Trainingsdokumente werden jeweils durch einen Konzept-Vektor repräsentiert. Um ein Testdokument nun in eine Klasse einzuordnen, erzeugt man zunächst auch für dieses Testdokument einen Konzept-Vektor. Dieser wird nun mit den Konzept-Vektoren der Trainingsdokumente verglichen. Abschließend wird das Testdokument anhand der besten Trainingsdokumente hinsichtlich der Kosinus-Ähnlichkeit einer Klasse zugewiesen.

6.2 Schwerpunkt-Strategie

Bei der Schwerpunkt-Strategie wird für jede Klasse anhand der Trainingsdokumente der Schwerpunkt dieser Klasse im Vektorraum aller Wörter ermittelt. Anschließend findet für den Schwerpunkt jeder Klasse ein Ranking der Wikipedia-Artikel nach ihrer Kosinus-Ähnlichkeit mit diesem Schwerpunkt statt. Die besten Artikel werden schließlich als Repräsentanten für diese Klasse gewählt. Das Vorgehen ist in Abbildung 6 veranschaulicht.

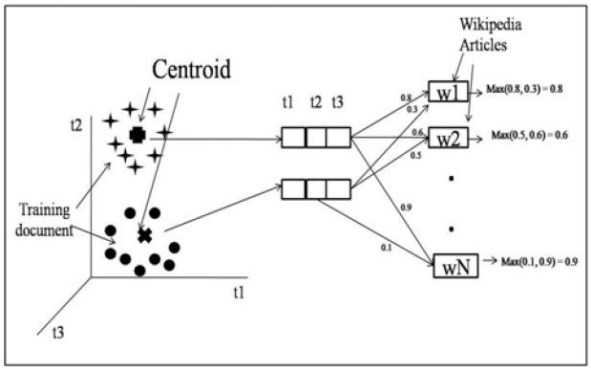

Fig. 6. Schwerpunkt-Strategie für die Selektion von Wikipedia-Artikeln (Quelle: [3]).

Die grundlegende Idee des Verfahrens ist es, Artikel auszuwählen, die prototypisch für die Kategorie sind (vgl. [3]). Allerdings sind hier mehrere Nachteile zu nennen. Zum einen ist es möglich, dass verschiedene Kategorien „verhungern", also keine angemessene Anzahl an Artikeln zugewiesen bekommen. Zum anderen können Artikel prototypisch für mehr als eine Klasse sein, wodurch sie nicht gut zwischen Klassen unterscheiden können. Außerdem sind die Schwerpunkte nicht unbedingt repräsentativ, da die Trainingsdokumente unter Umständen ungünstig im Vektorraum angeordnet sind.

6.3 Strategie der k-Nächsten Nachbarn (kNN)

Dieses Verfahren setzt auf lokale Nachbarschaft, anstatt auf Nähe zu Klassenschwerpunkten. Damit behebt es den entscheidenden Nachteil der Schwerpunkt-Strategie. Die Methode wählt zu jedem Wikipedia-Artikel die drei Trainingsdokumente, die die beste Kosinus-Ähnlichkeit aufweisen, und summiert diese Ähnlichkeitswerte. Jeder Klasse werden nun die Artikel mit den höchsten Werten zugeordnet. Dabei kann es sich allerdings um ähnliche Artikel handeln, was zu Redundanz führt (vgl. [3]).

6.4 Strategie mit Wahrscheinlichkeitsrate

Dieses wahrscheinlichkeitsbasierte Verfahren ermittelt die relative Wichtigkeit eines Artikels für eine Klasse anhand der Trainingsdokumente. Dazu wird die a-posteriori-Wahrscheinlichkeit jeder Klasse zu einem gegebenen Wikipedia-Artikel berechnet. Der Wikipedia-Artikel wird der Klasse mit dem höchsten Wahrscheinlichkeitswert zugeordnet. Für jede Klasse werden abschließend die in dieser Hinsicht besten Wikipedia-Artikel ausgewählt.

6.5 Erweitertes ESA

Erweitertes ESA basiert auf dem Ansatz, die Repräsentation aus einer Kombination von Wörtern und Konzepten zu gestalten. Dies soll bewirken, dass Wörter nicht ver-

loren gehen, die gut zwischen Klassen unterscheiden (vgl. [3]). Das Verfahren versucht entsprechend, die Unterscheidungsfähigkeit von Wikipedia-Seiten sowohl auf Wort-Ebene als auch auf Konzept-Ebene zu ermitteln.

6.6 Semantische Verwandtschaft

ESA geht davon aus, dass die Konzepte (Wikipedia-Artikel) nicht in Beziehung zu anderen Konzepten stehen. Tatsächlich ist dies aber nicht der Fall. Daher stellt sich die Frage, ob das Modellieren der semantischen Verwandtschaft verschiedener Konzepte bessere Retrieval Ergebnisse ermöglichen kann (vgl. [3]). Ein Ansatz dafür ist das Case Retrieval Network (CRN).

Ein Case Retrieval Network bietet die Möglichkeit, die paarweise Ähnlichkeit von Konzepten zu erfassen. Dabei wird ein Dokument als Vektor repräsentiert, der zunächst in jeder Komponente die Relevanz zu einem Konzept ausdrückt. Es werden die Werte 0 (nicht relevant) und 1 (relevant) verwendet. Das CRN verbindet Paare von Konzepten mit sogenannten Ähnlichkeitsbögen (siehe Abbildung 7). Relevante Konzepte dürfen ähnliche Konzepte „aktivieren“. Man spricht hierbei von Spreading Activation. Für jedes Konzept werden die eingehenden Aktivierungen aggregiert. Die aggregierten Werte bilden schließlich den erneuerten Vektor.

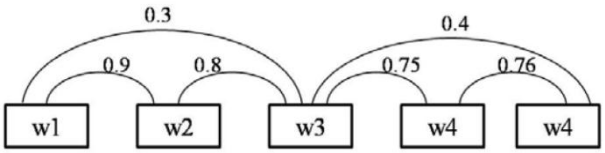

Fig. 7. Case Retrieval Network (Quelle: [3]).

Die Ähnlichkeit zwischen Wikipedia-Artikeln wird mittels Latent Semantic Analysis (LSA) ermittelt, genauer gesagt mit Sprinkled LSA (siehe [3]). Dabei wird die Repräsentation der Dokumente um Terme erweitert, die eine bestimmte Kategorie des Dokuments repräsentieren. Dies hat den Effekt, dass Dokumente einer Kategorie zusammengezogen und Dokumente unterschiedlicher Kategorien deutlicher getrennt werden (vgl. ebd.).

6.7 Auswertung

Um die vorgestellten Selektionsstrategien auszuwerten, hat man in [3] verschiedene Tests damit durchgeführt. Dabei wurde festgestellt, dass die Selektionsstrategien zu deutlich besseren Ergebnissen führen, insbesondere im Vergleich zum Schlüsselwort-basierten Vector Space Model. Die besten Ergebnisse konnte man mit der Schwerpunkt-Strategie und erweitertem ESA erzielen. Des Weiteren wurde festgehalten, dass lokale Strategien wie die der k-Nächsten Nachbarn mit komplexeren Themengebieten deutlich besser klar kommen. Die Modellierung von semantischer Verwandtschaft zwischen Konzepten wurde hingegen als nicht überzeugend bewertet (siehe [3]).

7 Zusammenfassung und Fazit

Zunächst wurde die Grundidee von Explicit Semantic Analysis (ESA) vorgestellt. ESA repräsentiert natürlichsprachliche Texte durch Konzepte, die menschliches Wissen umfassen. Diese Konzepte werden aus Quellen wie Wikipedia gewonnen. Anschließend wurde das Information Retrieval System MORAG behandelt, das auf ESA aufbaut. MORAG enthält eine Feature Selection Komponente, die auf Pseudo-Relevance Feedback basiert. Feature Selection dient dem Optimieren der Repräsentation der Anfrage. Dabei sollen Verzerrungen und Ungenauigkeiten entfernt werden. Des Weiteren nutzt MORAG Fusion, um noch besssere Ergebnisse zu erzielen.

Experimentelle Auswertungen haben die Verbesserungen bezüglich der Performance im Vergleich zu herkömmlichen Verfahren demonstriert. Gleichzeitig besteht aber noch Potenzial für Verbesserungen, insbesondere hinsichtlich der Feature Selection Komponente. Die im letzten Abschnitt vorgestellten Selektionsverfahren setzen an diesem Punkt an und bewirken somit weitere Verbesserungen.

Der Konzept-basierte Ansatz von ESA sowie das entwickelte System MORAG können als Fortschritt im Information Retrieval betrachtet werden. Die Entwickler von MORAG sehen ihr System zudem als mögliche Verschiebung im Information Retrieval Paradigma, von einem rein Schlüsselwort-basierenden zu einem nun eher auf menschlichen Konzepten basierenden. Es wird sicherlich interessant sein, zu beobachten, inwiefern der Konzept-basierte Ansatz in Zukunft das Information Retrieval beeinflussen wird.

Quellen

1. Gabrilovich, E. & Markovitch, S. Veloso, M. M. (Ed.) Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12, 2007, 2007, 1606-1611
2. Egozi, O.; Markovitch, S. & Gabrilovich, E. Concept-Based Information Retrieval Using Explicit Semantic Analysis ACM Trans. Inf. Syst., ACM, 2011, 29, 8:1-8:34
3. Patelia, A.; Chakraborti, S. & Wiratunga, N. Ram, A. & Wiratunga, N. (Eds.) Selective Integration of Background Knowledge in TCBR Systems Case-Based Reasoning Research and Development, Springer Berlin / Heidelberg, 2011, 6880, 196-210
4. Radinsky, K.; Agichtein, E.; Gabrilovich, E. & Markovitch, S. A word at a time: computing word relatedness using temporal semantic analysis Proceedings of the 20th international conference on World wide web, ACM, 2011, 337-346
5. Gottron, T.; Anderka, M. & Stein, B. Insights into Explicit Semantic Analysis CIKM'11: Proceedings of 20th ACM Conference on Information and Knowledge Management, 2011, 1961-1964