Latent Semantic Indexing zum Aufschlüsseln philosophischer Werke

Exposé

Nils Alberti

Betreuer: Prof. Ulf Leser

Humboldt-Universität zu Berlin
Mathematisch-Naturwissenschaftliche Fakultät II
Institut für Informatik
Wissensmanagement in der Bioinformatik

1 Einführung

Seit den Tagen der Bibliothek von Alexandria hat sich die Aufbereitung geisteswissenschaftlicher Erkenntnisse nicht grundlegend verändert: Gelehrte strukturieren und verschriftlichen Ideen, um sie in Textform der Öffentlichkeit zugänglich zu machen. Eine solche Form der Darstellung induziert das lineare Nachvollziehen der Gedankengänge. Für die wissenschaftliche Forschung spielt bloßes Rekapitulieren gedruckter Gedanken eine untergeordnete Rolle: Sie versucht, über bestehende Theorien hinauszugehen, indem sie Teilaspekte beleuchtet, andere Schwerpunkte setzt, neue Fragen stellt und Gemeinsamkeiten wie Unterschieden in den Sichtweisen verschiedener Denker nachgeht. Hier hemmt die Aufbereitungsform des Wissens dessen Weiterentwicklung. Bei vielen Forschungsvorhaben bedarf es nicht nur eines theoretischen Verständnisses des philosophischen Bereichs, sondern auch umfassender Kenntnis der entsprechenden Werke, um die relevanten Textstellen zu finden.

Behelfen kann man sich mit Querlesen, Registern und seit wenigen Jahren mit Google Books. Diese Suchmaschine findet zumindest Stellen, wo ein gesuchter Begriff auftritt, sozusagen ein Register für alle Begriffe. Die Erwähnung passender Worte bedeutet jedoch nicht zwingend eine inhaltliche Nähe zur Fragestellung. Hinzu kommt, dass die Suche Kenntnis der verwendeten Terminologie voraussetzt. Man erhält folglich oft zu viele Dokumente und übersieht gleichzeitig andere.

Gäbe es die Möglichkeit, auf Knopfdruck sämtliche jemals veröffentlichten Textpassagen zu einer Forschungsfrage zu erhalten, wäre dies ein Quantensprung für die Geisteswissenschaft.

2 Zielstellung

Auf dieses Ziel möchte ich einen kleinen Schritt zu gehen und untersuchen, ob sich Methoden der computergestützten Textanalyse auch für die Unterstützung der Suche in philosophischen Texten eignen. Lassen sich zu einer textuell beschriebenen Forschungsthematik inhaltlich passende Textstellen aus dem Œuvre eines Autors finden?

3 Vorgehen

3.1 Auswahl der Werke

Kriterien für die Auswahl der Werke sind ihre digitale Verfügbarkeit sowie die einheitliche Aufbereitung. Auch gilt es, diese Methode auf zwei Werke anzuwenden, um die Reproduzierbarkeit zu belegen. Die Wahl fiel auf die CD-Sammlung „Digitale Bibliothek“ und die darin veröffentlichten Ausgewählte Werke von Karl Marx und Friedrich Engels [Marx und Engels, 2004], sowie die Gesammelten Schriften von Theodor W. Adorno [Adorno, 2004]. Die beiden Korpora sind aufgrund ihrer Verschiedenheit ideal, um das Verfahren zu verifizieren. Karl Marx changiert zwischen der nüchternen Sprache seiner ökonomischen Analysen [vgl. Marx, 1987] und der bildreichen Ausdrucksweise der philosophischen Frühschriften [vgl. Marx und Engels, 1970]. Durch Hinzunahme von Friedrich Engels' Schriften konfrontieren wir das Verfahren mit dem Sprachstil unterschiedlicher Autoren.

Das Werk Theodor W. Adornos unterscheidet sich grundlegend von den komplexen, aber verständlichen Texten von Marx und Engels. Bis heute gilt sein Schreibstil als Inbegriff des hochgestochenen „Professorendeutschs“. Der Frankfurter Philosoph hat ein uneinheitliches Gesamtwerk geschaffen. Philosophische Abhandlungen wechseln sich mit musiktheoretischen ab; die Aphorismensammlung Minima Moralia steht neben dem philosophischen Magnum Opus Negative Dialektik. Adorno, 20 Jahre nach Marx' Tod geboren, beschäftigte sich zeitlebens mit dessen Theorie [Braunstein, 2011]. Für uns ist dies interessant, weil sich somit Themen überschneiden und wir hierzu Ergebnisse aus beiden Korpora erhalten.

3.2 Verwendete Algorithmen

Computer sind noch weit davon entfernt, Bedeutung zu „verstehen“. Sie sind aber hervorragend geeignet, Ähnlichkeit von Texten anhand übereinstimmender Begriffe zu bewerten. Mehrere Untersuchungen zeigen, dass ihre Ergebnisse qualitativ denen von Menschen entsprechen [Burstein und Chodorow, 1999]. Im Forschungsgebiet des Information Retrievals wurden eine Vielzahl von Algorithmen zur Ähnlichkeitsmessung von Dokumenten entwickelt. Als Standard hat sich das Vektorraummodell etabliert, bei dem von der Reihenfolge der Worte abstrahiert wird. Dokumente werden dabei in eine Term-Dokument-Matrix überführt, wobei Spalten die Dokumente und Zeilen die Wörter repräsentieren. Kommt ein Wort in einem Dokument vor, so

wird eine 1 notiert, sonst eine 0.

Texte werden als umso ähnlicher betrachtet, je mehr Worte sie gemeinsam haben. Möchte man wissen, ob ein Text für eine Fragestellung relevant ist, so vergleicht man den Text wortweise mit der Fragestellung, der sogenannten Query. Mathematisch kann dies durch den Kosinus des Winkels zwischen beiden Vektoren, der sogenannten Kosinus-Ähnlichkeit, ausgedrückt werden. Deutlich verbessert wird das Verfahren, wenn die Aussagekraft von Wörtern berücksichtigt wird: Sie werden in diesem Fall umso stärker gewichtet, je seltener sie in allen und je häufiger sie in dem betrachteten Dokument vorkommen.

Ein Manko hat das Vektorraummodell, das es für unser Vorhaben problematisch erscheinen lässt: Zwei Dokumente können dasselbe Thema behandeln, ohne (aussagekräftige) Begriffe gemein zu haben. Gerade bei philosophischen Texten ist eine „Vocabulary Gap“ zu erwarten, weil, wie Jean Paul Sartre bemerkte, „die meisten Philosophen [...] Sätze literarischer Art“ [Sartre, 1977, S. 184] verwenden.

3.2.1 Latent Semantic Indexing

Um dieses Problem zu überwinden, entwickelten Scott Deerwester, Susan Dumais, George Furnas, Thomas Landauer und Richard Harshman Latent Semantic Indexing (LSI) [Deerwester et al., 1990]. Die Idee besteht in der Überführung der Term-Dokument-Matrix in einen deutlich kleineren, „semantischen“ Raum. Häufig gemeinsam auftretende Wörter werden dabei zu einem „Konzept“ komprimiert: In einem Korpus aus Zeitungsartikeln würden Wörter wie Audi, Mercedes und Porsche, genauso wie Deutschland, Frankreich und Italien annähernd zusammengefasst. Nicht weil der Algorithmus „versteht“, dass es sich um Automarken und Ländernamen handelt, sondern weil diese Wörter häufig gemeinsam in vergleichbaren Artikeln auftreten. Schriftstücke werden bei diesem Verfahren auch dann als ähnlich klassifiziert, wenn sie Wörter aus dem gleichen Kontext, nicht aber dieselben Wörter enthalten.

Mathematisch wird dies durch eine Singulärwertzerlegung erreicht, bei der die Term-Dokument-Matrix in drei kleinere Matrizen, deren Produkt die ursprüngliche Matrix ergibt, aufgeteilt wird. Eine Matrix enthält dann die absteigend sortierten Quadratwurzeln der Eigenwerte, die Singulärwerte, in der Diagonalen. Verkleinert man diese Matrix und beschränkt sie auf die k größten Werte, so ergibt das Matrixprodukt eine Approximation der Term-Dokument-Matrix, die durch die Hauptkomponenten, die Konzepte, bestimmt wird.

Wie in der ursprünglichen Matrix kann die thematische Nähe von Texten anhand

ihrer Kosinus-Ähnlichkeit ermittelt werden: Ein Text oder eine Query werden dabei in den semantischen Raum projiziert. Ein einzelnes Wort wird hierbei genauso wie ein dicker Wälzer durch einen k-dimensionalen Vektor abgebildet. In ihm haben die Vektorkomponenten hohe Werte, die ein Konzept des ursprünglichen Korpus repräsentieren und gleichzeitig eine gewichtige Rolle in der Anfrage spielen.

3.2.2 Segmentierung der Kapitel

Wir müssen überlegen, welches die kleinste Einheit sein soll, die zu separieren ist. Man könnte einzelne Worte, Sätze, Absätze oder ganze Kapitel wählen. Text an Worten zu separieren, ist plausibel, um einzelne Daten zu extrahieren. Kapitel hingegen sind eine grobe Einteilung, bei der Exkurse oder Teilaspekte keine Berücksichtigung finden.

So bleibt die Wahl zwischen der feinen Unterscheidung auf Satzebene und der gröberen auf der Ebene von Absätzen. Einzelne Sätze haben vielfach eine zu geringe Informationsdichte [Choi, 2000], weshalb es, durch eingeschobene Sätze, leicht zum „Zerhäckseln“ von Gedankengängen kommen kann. Vergegenwärtigen wir uns das Ziel dieser Arbeit, in komplexen philosophischen und geisteswissenschaftlichen Texten die für den Rezipienten passenden Auszüge zu finden, so erscheinen Absätze als geeignete Wahl.

3.2.3 Zusammenfassen von Absätzen

Wir sollten es vermeiden, thematisch ähnliche Absätze auseinanderzureißen. Sie sollten nur getrennt werden, wenn auch ein Themenwechsel erfolgt. Um diese Stellen zu erkennen, messen wir die Kosinus-Ähnlichkeit benachbarter Absätze und fassen sie so zu Abschnitten (Clustern) zusammen, dass Abstände zwischen getrennten Abschnitten maximiert und Clusterinnenabstände, also die Summe der Kosinus-Ähnlichkeiten zwischen einzelnen Absätzen und dem Clustermittelpunkt, minimiert werden. Mit anderen Worten: Abschnitte sollen sich von ihren Nachbarn inhaltlich deutlich unterscheiden. Darin enthaltene Absätze sollen sich hingegen möglichst ähnlich sein. Für jedes Unterkapitel berechnen wir mit dynamischer Programmierung ein solches Clustering [Hansen und Jaumard, 1997].

3.3 Verifizieren der Methode

Wir sind mit dem Problem konfrontiert, dass kein Maß für „inhaltliche Nähe“ existiert, mit dem wir unsere Ergebnisse bewerten können. Die Ähnlichkeit von Texten einschätzen zu können, setzt deren Verständnis voraus. Uns bleibt nichts anderes übrig, als auf die Bewertung von Menschen und hier auf die der Experten im jeweiligen Fachgebiet zu vertrauen.

Für die Evaluierung erstellen wir aus Sekundärliteratur zu mehreren Themengebieten einen „Goldstandard“: Wir erfassen die Referenzstellen, die die Exegeten zu einem Punkt herangezogen haben und testen, wie viele davon der Algorithmus findet.

Literatur

T. W. Adorno. Theodor W. Adorno: Gesammelte Schriften - Digitale Bibliothek Band 11. [CD-ROM], 2004.

D. Braunstein. *Adornos Kritik der politischen Ökonomie.* transcript Verlag, 2011.

J. Burstein und M. Chodorow. Automated essay scoring for nonnative English speakers. In: *Proceedings of a Symposium on Computer Mediated Language Assessment and Evaluation in Natural Language Processing*, S. 68–75. Association for Computational Linguistics, 1999.

F. Choi. Advances in domain independent linear text segmentation. In: *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, S. 26–33. Morgan Kaufmann Publishers Inc., 2000.

S. Deerwester, S. Dumais, G. Furnas, T. Landauer, und R. Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

P. Hansen und B. Jaumard. Cluster Analysis and Mathematical Programming. *Mathematical Programming*, 79:191–215, 1997.

K. Marx. *Das Kapital - Kritik der politischen Ökonomie, Erster Band.* Dietz Verlag, 1987.

K. Marx und F. Engels. *Das Kommunistische Manifest.* Dietz Verlag, 1970.

K. Marx und F. Engels. Karl Marx / Friedrich Engels: Ausgewählte Werke - Digitale Bibliothek Band 11. [CD-ROM], 2004.

J. P. Sartre. *Sartre über Sartre - Autobiographische Schriften Band 2.* Rowohlt, 1977.