Benutzerorientiertes Dokumenten-Clustering durch die Verwendung einer Anfragemenge

Marc Lechtenfeld
Informationssysteme
Universität Duisburg-Essen

Abstract

Der Einsatz einer Anfragemenge ermöglicht die Berücksichtigung des Informationsbedürfnisses des Benutzers beim Strukturierungsprozess von Dokumenten. Der vorgestellte benutzerorientierte Ansatz kann den Benutzer damit bei mehreren Suchaktivitäten unterstützen. Geplante Benutzerexperimente sollen die Effektivität, Effizienz und Benutzerzufriedenheit prüfen.

1 Einführung

Ein Benutzer kann mithilfe zahlreicher Information-Retrieval-Verfahren nach wohlspezifizierten Informationen suchen, indem er sein Informationsbedürfnis in Form einer Anfrage an sie richtet. Er ist jedoch nicht immer in der Lage sein Informationsbedürfnis genau zu spezifizieren und in einer speziellen Systemanfrage auszudrücken. Um seine Aufgabe in diesem Fall trotzdem erfüllen zu können, ist er dann auf die Unterstützung weiterer Suchaktivitäten durch das System angewiesen [Belkin, 1993].

Explorative Suchaktivitäten

Das Durchstöbern einer von einem Clustering-Verfahren strukturierten Dokumentmenge stellt für den Benutzer eine Möglichkeit dar, nach relevanten Dokumenten zu suchen, ohne dabei sein Informationsbedürfnis explizit ausdrücken zu müssen. Diese Dokumentmenge könnte beispielsweise eine große Kollektion von Dokumenten sein, die vom Benutzer über die in ihr enthaltenen Themen durchsucht werden kann oder ein Suchergebnis zu einer mehrdeutigen Anfrage, das durch die Gruppierung der Dokumente mit gleicher Bedeutung für den Benutzer verständlicher wird.

Benutzerorientierung

Klassische Dokumenten-Clustering-Verfahren strukturieren Dokumentmengen allerdings häufig nur nach thematischen Gesichtspunkten. Das konkrete Informationsbedürfnis des Benutzers wird nicht direkt in den Strukturierungsprozess einbezogen. Ein Benutzer kann sich jedoch für unterschiedliche Aspekte eines Dokuments interessieren, beispielsweise für die Textsorte, die Verständlichkeit oder die inhaltliche Qualität eines Dokuments. So möchte er für eine Sammlung von Zeitungsartikeln vielleicht herausfinden, über welche Ereignisse in den unterschiedlichen Ressorts negativ berichtet wurde. Dies erfordert beispielsweise sowohl eine thematische Gruppierung der Dokumente nach Ressorts (Politik, Sport, usw.) als auch eine Gruppierung nach Sentiment [Pang *et al.*, 2002] (positive, neutrale und negative Äußerungen).

2 Benutzerorientiertes Dokumenten-Clustering

Um ein für den Benutzer nützliches Clustering zu erzeugen, sollte sein Informationsbedürfnis den Strukturierungsprozess auf eine grundlegende Art und Weise beeinflussen.

2.1 Optimum Clustering Framework

Die theoretische Basis dazu bildet das *Optimum Clustering Framework* (OCF) [Fuhr *et al.*, 2010]. Zur Bestimmung der Dokumentähnlichkeit führt es eine Anfragemenge ein, die in Verbindung mit einem Retrievalmodell dazu verwendet wird, die Relevanz eines Dokuments für den Benutzer zu schätzen. Die Ähnlichkeit zweier Dokumente ergibt sich dann aus dem Skalarprodukt der zwei Dokumentrepräsentationen. Diese beiden Vektoren bestehen aus den Relevanzwahrscheinlichkeiten bezüglich aller Anfragen der Anfragemenge. Dokumente werden demnach als ähnlich definiert, wenn sie bezüglich möglichst vieler Anfragen gleichzeitig potenziell relevant sind.

Klassische Dokumenten-Clustering-Verfahren verwenden dabei eine Anfragemenge, die aus allen in der Dokumentmenge vorkommenden Termen besteht. Experimentelle Ergebnisse zeigen jedoch, das Clusterings besser werden, wenn die verwendeten Anfragen stärker den tatsächlichen Anfragen entsprechen, die der Benutzer für eine anfrageorientierte Suche verwenden würde.

2.2 Berücksichtigung einer Anfragemenge

Um eine Strukturierung der Dokumentmenge zu erzeugen, die sich stärker am Informationsbedürfnis des Benutzers orientiert und damit nützlicher für ihn ist, kann daher das Optimum Clustering Framework auf eine Anfragemenge angewandt werden, die an das Informationsbedürfnis des Benutzers stärker angepasst ist.

Die Anfragemenge sollte dabei so gestaltet sein, dass sie die Anfragen, die das Informationsbedürfnis des Benutzers am besten ausdrücken, enthält, aber nicht wesentlich größer ist. Beim Clustering eines Suchergebnisses kann diese Anfragemenge beispielsweise auch alle potenziellen Anfrageerweiterungen der Suchanfrage enthalten.

2.3 Bestimmung der Anfragemenge

Liegen über das Informationsbedürfnis des Benutzers keinerlei Informationen vor, so erscheint es sinnvoll als Anfragemenge die Menge aller in der Dokumentmenge vorkommenden Terme als Eintermanfragen zu verwenden. Dies entspricht dann dem klassischen Clustering nach dem dominantesten Aspekt, der in der Regel das Thema ist. Da eine spezialisierte Anfragemenge jedoch zu besseren Strukturierungen führt, sollte, wenn weitere Hinweise

über das konkrete Informationsbedürfnis des Benutzers verfügbar sind, eine Anfragemenge verwendet werden, die an diese Informationen angepasst ist.

Anfragemengen für bestimmte Aspekte
Durch die Interaktion mit dem Benutzer können besser an sein Informationsbedürfnis angepasste Anfragemengen gefunden werden. Eine Möglichkeit dazu besteht darin, den Benutzer explizit nach den Aspekten zu fragen, für die er sich interessiert. Beispielsweise durch Vorlage einer Liste von Aspekten, die allgemein für viele Benutzer interessant sind und die in der vorliegenden Dokumentmenge vorkommen. So könnte eine Kollektion von Büchern z. B. nach den Themengebieten, nach den Leserbewertungen oder nach der Verständlichkeit gruppiert werden. Zur Strukturierung wird dann jeweils die vorbereitete Anfragemenge eingesetzt, die den Aspekt am besten beschreibt, den der Benutzer ausgewählt hat. Durch die Möglichkeit zur Auswahl verschiedener Aspekte wird so ein mehrdimensionales Clustering aufgrund unterschiedlicher Aspekte möglich.

Individuelle Anfragemengen
Das System könnte während der Interaktion mit dem Benutzer jedoch auch Informationen über den Benutzer indirekt sammeln. Diese könnten dabei helfen die anfänglich große Anfragemenge, die ein allgemeines Informationsbedürfnis ausdrückt, sukzessive zu spezialisieren und so an das individuelle Informationsbedürfnis des Benutzers anzupassen. Auf diese Weise wäre es vielleicht möglich, automatisch zu bestimmen, für welche Aspekte sich der Benutzer wahrscheinlich interessiert.

Mögliche Quellen für die Eingrenzung der potenziellen Anfragen sind neben dem Interaktionsverhalten des Benutzers mit dem System beispielsweise auch Charakteristika der Kollektion, Adaptierungen der Benutzerschnittstelle, die verwendeten Retrievalmodelle oder externe Quellen wie z. B. die Enzyklopädie *Wikipedia*.

2.4 Unterstützte Suchaktivitäten

Die Kombination von Verfahren des Information-Retrieval und Dokumenten-Clustering ermöglicht die Unterstützung verschiedener Suchaktivitäten.

Interpretation einer Dokumentmenge
Einen ersten Überblick über eine unbekannte Kollektion oder über ein Suchergebnis kann sich der Benutzer mithilfe einer Strukturierung dieser Dokumentmengen verschaffen. So könnte man durch eine Gruppierung der Ergebnisse einer Buchsuche beispielsweise einen Überblick über aktuelle Programmiersprachen erhalten.

Die Möglichkeit die Strukturierung aufgrund unterschiedlicher Aspekte durchführen zu lassen, kann dem Benutzer dabei helfen die Beschaffenheit der Dokumentmenge oder sein Informationsbedürfnis besser zu verstehen. Die Bücher über Programmiersprachen lassen sich beispielsweise auch nach den Vorkenntnissen z. B. in Bücher für Anfänger und Fortgeschrittene oder nach der Leserzufriedenheit über die Rezensionen gruppieren.

Über die jeder Gruppierung zugrundeliegende Anfragemenge ist für den Benutzer möglicherweise erkennbar, für welche Anfragen oder Aspekte die Dokumente der gerade betrachteten Gruppierung relevante Informationen liefern und wie sich die Dokumente der unterschiedlichen Cluster voneinander unterscheiden. Auf diese Weise kann er lernen, über welche Anfragen die gewünschten Dokumente vom Retrievalsystem zurückgeliefert werden bzw. wie er eine Anfrage mit weiteren Termen ergänzen könnte, um die gewünschten Dokumente zu erhalten. Beim Clustering eines Suchergebnisses kann man die Fragemenge beispielsweise als Menge von möglichen Frageerweiterungen betrachten.

Finden durch Erkennen
Durch das Durchstöbern der gruppierten Dokumentmenge kann der Benutzer Dokumente finden, ohne sein Informationsbedürfnis explizit ausdrücken zu müssen. Er kann relevante Dokumente entdecken, indem er die Dokumente – nach der Terminologie von Belkin [Cool and Belkin, 2002] – *scannt* und die relevanten Dokumente in der Dokumentmenge als für ihn relevant *erkennt*. So könnte der Benutzer bei einer Suche nach einem Einsteigerbuch für eine neue Programmiersprache z. B. auf eine Untergruppe von Umsteigerbüchern stoßen, die für ihn nützlicher sind.

Durch die Interaktion mit dem System etwa durch die Auswahl einzelner Cluster oder durch das manuelle Gruppieren einzelner Dokumente, kann das System die Fragemenge automatisch Schritt für Schritt an das Informationsbedürfnis des Benutzers anpassen. Auf diese Weise kann ein auf den Benutzer zugeschnittenes Clustering der Kollektion oder der Ergebnismenge erzeugt werden.

3 Benutzerorientierte Evaluation

Es sind Benutzerexperimente geplant, die die Nützlichkeit dieses Ansatzes für den Benutzer bestimmen sollen.

Es sollen dabei insbesondere auch Informationsbedürfnisse untersucht werden, die sich nur schwer in Form einer spezifischen Anfrage ausdrücken lassen. Dabei ist zu berücksichtigen, dass sich der Benutzer für verschiedene Aspekte interessieren kann, die zu unterschiedlichen Strukturierungen führen können. Ein Vergleich der erzeugten Strukturierung mit einer manuell erstellten Klassifikation, die nur eine Sichtweise abbildet, reicht daher nicht.

Mithilfe von *Simulated Work Tasks* [Borlund, 2000] soll geprüft werden, ob der Benutzer die ihm gestellten Aufgaben lösen kann (Effektivität), wie viel Zeit er dafür benötigt (Effizienz) und wie zufrieden er ist (Zufriedenheit), also ob er damit insgesamt bei der Suche profitiert.

Literatur

[Belkin, 1993] Nicholas J. Belkin. Interaction with texts: Information retrieval as information-seeking behavior. In *Information Retrieval*, pages 55–66, 1993.

[Borlund, 2000] Pia Borlund. Experimental components for the evaluation of interactive information retrieval systems. *Journal of Documentation*, 56:71–90, 2000.

[Cool and Belkin, 2002] Colleen Cool and Nicholas J. Belkin. A classification of interactions with information. In H. Bruce, R. Fidel, P. Ingwersen, and P. Vakkari, editors, *Emerging frameworks and methods. Proceedings of the 4th COLIS*, pages 1–15, Greenwood Village, 2002. Libraries Unlimited.

[Fuhr *et al.*, 2010] Norbert Fuhr, Marc Lechtenfeld, Benno Stein, and Tim Gollub. The Optimum Clustering Framework: Implementing the Cluster Hypothesis. 2010. Submitted.

[Pang *et al.*, 2002] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–86, 2002.