Kontextbasiertes Retrieval unter Verwendung verknüpfter Annotationen

Ingo Frommholz, Holger Brocks, Ulrich Thiel und Adelheit Stein

Fraunhofer IPSI, Dolivostr. 15, D-64293 Darmstadt, Germany
{frommholz, brocks, thiel, stein}@ipsi.fraunhofer.de

Abstract: Kollaborative Arbeitsumgebungen im Web können Mechanismen enthalten, mit denen neben dem Erstellen von zum Dokument gehörenden Metadaten auch ein wissenschaftlicher Diskurs über das eigentliche Dokument geführt werden kann (z.B. über freie Annotationen). Dieser Diskurs kann wertvolle Informationen über das Dokument enthalten, die aus den Metadaten nicht ersichtlich sind. Es wird gezeigt, wie sich ein solcher wissenschaftlicher Diskurs mittels Annotationen und Diskursstrukturrelationen modellieren läßt und wie man die daraus gewonnenen Informationen beim Retrieval ausnutzen kann.

1 Einführung und Motivation

Das COLLATE-Projekt[1] hat sich zum Ziel gesetzt, ein webbasiertes Kollaboratorium[2] zu entwickeln, mit dessen Hilfe Archivare, Wissenschaftler und Endbenutzer gemeinsam am und mit digitalisiertem kulturellen Material arbeiten können. In diesem Projekt geht es um historisches Filmmaterial, konkret digitalisiertes Material über europäische Filme des frühen zwanzigsten Jahrhunderts (für detailliertere Informationen siehe auch [BTSDW01]). Beispielsweise kann es sich um Zensurdokumente handeln, denen sich der Zensurgrund und die im Film zensierten Stellen entnehmen lassen. Zu jedem digitalisierten Dokument werden in COLLATE Metadaten gesammelt, die z.B. die zensierende Behörde, die zensierende Person, verschiedene Namen des dazugehörigen Films, Art des Dokumentes (wie oben erwähnt kann es ein Zensurdokument sein, aber auch ein Zeitungsartikel, Foto, etc) enthalten. Zusätzlich zu diesen definierten Metadaten besteht in COLLATE noch die Möglichkeit, Dokumente zu annotieren bzw. auch bestehende Annotationen zu annotieren. Mit Hilfe dieser Annotationen können *wissenschaftliche Diskurse* verwirklicht werden. Aus diesen Diskursen lassen sich u.U. wertvolle Informationen über den Gesprächsgegenstand (das digitalisierte Dokument) entnehmen, die während des Retrievalprozesses ausgenutzt werden können. Die Diskurse lassen sich also für Anfragen wie "gib mir alle Zensurdokumente über ein bestimmtes Zensurthema" ausnutzen, wie ein klei-

[1] Collaboratory for annotation, indexing and retrieval of digitized historical archive material IST-1999-20882, http://www.collate.de/

[2] Ein *Kollaboratorium* (engl.: Collaboratory) ist ein aus den Begriffen *Collaboration* und *Laboratory* zusammen gesetzter Begriff

nes Beispielszenario verdeutlichen soll. Angenommen, der Suchende interessiert sich für alle Zensurentscheide aus politischen Gründen, wobei dieses Informationsbedürfnis sich in der Anfrage q manifestieren würde. Das System berechnet nun für jedes digitalisierte Dokument d ein Retrievalgewicht $P(R|d, q)$, welches die Wahrscheinlichkeit darstellt, dass d bzgl. der Anfrage q relevant (R) ist. Von den Informationen, die wir aus einem Zensurdokument selbst bekommen (in Form von daraus extrahierten Metadaten) können wir einen Zensurgrund ablesen, z.B. Unsittlichkeit. Ein solches Dokument wäre auf Grund der vorliegenden Informationslage nicht relevant zur Frage nach politischen Zensurgründen, würde also ein entsprechend niedriges Retrievalgewicht bekommen. Es könnte aber sein, dass ein Filmwissenschaftler den auf dem Dokument angegebenen Zensurgrund anzweifelt und statt dessen mutmaßt, dass doch eher politische Gründe ausschlaggebend waren, auch wenn dies (aus welchen Gründen auch immer) auf dem Dokument selbst nicht vermerkt ist. Der Informationssuchende könnte daher das Dokument doch noch interessant finden, was sich in einer Erhöhung des Retrievalgewichts niederschlägt. Um das Beispiel ein wenig weiter zu führen, könnte ein zweiter Filmwissenschaftler wiederum die Aussage des ersten Wissenschaftlers anzweifeln, was im Extremfall dazu führen könnte, dass der Anfrager in dem Dokument doch nicht das findet, wonach er sucht (da z.B. die Aussage des ersten Wissenschaftlers als zu abwegig dargestellt wurde). Das Retrievalgewicht zu d könnte wieder gesenkt werden.

Dieses einfache Beispiel soll zeigen, dass es sich durchaus lohnen kann, beim Retrieval ein Augenmerk auf die über ein Dokument entbrannte wissenschaftliche Diskussion zu werfen. Um zu erkennen, in welchem Verhältnis eine gemachte Aussage zu dem Objekt der Aussage steht, stellen wir im nächsten Abschnitt *Diskursstrukturrelationen* vor. Im darauf folgenden Abschnitt wird das Konzept des *kontextbasierten Retrievals* erläutert, welches auf Annotationen und deren Beziehungen zueinander (eben o.g. Diskursstrukturrelationen) zurückgreift. Zuletzt wird eine Zusammenfassung und ein Ausblick gegeben.

2 Diskursstrukturrelationen

Wie eingangs erwähnt, stellen Annotationen das Hauptkonzept dar, um kollaboratives Arbeiten mittels wissenschaftlicher Diskurse zu ermöglichen. Dabei können Annotationen zu den digitalen Dokumenten selbst, aber auch zu anderen Annotationen erstellt werden. Auf diese Weise entsteht ein gerichteter, azyklischer Graph, der als Quellknoten das digitalisierte Dokument hat, und bei dem die Annotationen die restlichen Knoten bilden. Dieser Graph wird, in Anlehnung an Diskussionsthreads, wie sie z.B. aus Newsgroups bekannt sind, *Annotationsthread* genannt.

Zwischen den Annotationen existieren bestimmte typisierte Links, die die Beziehung zwischen zwei Annotationen definieren. Unsere Definition dieser Beziehungen basiert auf Konzepten aus der Diskurstheorie; die folgenden *Diskursstrukturrelationen* (siehe auch [MH93]) finden in COLLATE Anwendung: *Ausarbeitung* (*Elaboration*) ist das Bereitstellen zusätzlicher, detaillierterer Information, beispielsweise “Frankfurt a.d. Oder, nicht am Main”; *Vergleich* (*Comparison*) ist unterteilt in *Analogie* (*Analogy*) und *Unterschied* (*Difference*) und soll semantische Ähnlichkeiten oder Kontraste beleuchten; *Ursache* (*Cause*)

Abbildung 1: Digitalisiertes Dokument, Annotationen und Metadaten

meint Angaben über eine bestimmte Ursache eines Ereignisses, gewisser Umstände, etc. *Hintergrundinformation* (*Background Information*) ist die Information über den Hintergrund eines Autors, z.B. “der Autor ist Laie auf dem Gebiet und berücksichtigt nicht die historischen Aspekte”; *Interpretation* ist die subjektive Deutung einer Aussage (z.B. “was der Autor hiermit sagen möchte...”), *Argumentation* ist aufgeteilt in *unterstützendes Argument* (*Support Argument*) und *Gegenargument* (*Counterargument*); eine Aussage wird entweder bestärkt oder aber es wird dagegen argumentiert.

3 Kontextbasiertes Retrieval

Kontextbasiertes Retrieval kann sich auf verschiedene Arten von Kontexten beziehen (beispielsweise ein semantischer Kontext, gegeben durch eine Kategorie, in der sich ein Dokument befindet). In COLLATE arbeiten wir mit dem so genannten *Diskurskontext*. Dies bedeutet, dass wir für jede Aussage berücksichtigen, an welcher Stelle im Diskurs diese gemacht wurde, auf welche andere Aussage sie sich bezieht und welcher Art die Aussage ist (eine der in Abschnitt 2 definierten Diskursstrukturrelationen).

Zu jedem digitalisierten Dokument existiert ein Annotationsthread, mit dem der wissenschaftliche Diskurs über dieses Dokument abgebildet wird. Zusätzlich dazu können wir noch auf verschiedene Arten von Metadaten zurückgreifen. Filmwissenschaftler können Schlüsselwörter vergeben, die entweder kontrolliert (d.h. aus einer Ontologie entnommen) oder frei sein können. Ferner können noch Katalogisierungsinformationen gegeben werden, in denen z.B. die zum Dokument zugehörigen Filmtitel, die zensierende Institution, der Dokumenttyp (z.B. Zensurdokument, Artikel), usw. angegeben werden kann. Abbildung 1 zeigt ein Beispieldokument mit dazugehörigen Metadaten und Annotationen.

Wie erwähnt, berechnen wir beim Retrieval in COLLATE für jedes digitalisierte Dokument d die Wahrscheinlichkeit $P(R|d,q)$, dass das Dokument relevant bezüglich der Anfrage q ist. Auf diese Weise erhalten wir ein Ranking von Dokumenten. Zur Bestimmung von $P(R|d,q)$ können einerseits die zum Dokument zugehörigen Metadaten zur Geltung kommen, andererseits kann aber auch die Information, die wir aus dem Annotationsthread erhalten, ausgenutzt werden (wie in Abschnitt 1 motiviert).

Sei nun $P_{meta}(R|d,q)$ die Wahrscheinlichkeit, dass d bzgl. q relevant ist, wobei hier nur die Metadaten berücksichtigt werden. Analog sei $P_{dis}(R|d,q)$ die Relevanzwahrscheinlichkeit, bei der nur der im Annotationsthread abgebildete Diskurs über d betrachtet wird. Dann ist

$$P(R|d,q) = \alpha \cdot P_{dis}(R|d,q) + (1-\alpha) \cdot P_{meta}(R|d,q).$$

Mittels des Parameters $\alpha \in [0,\ldots,1]$ kann der Suchende den Grad bestimmen, inwieweit er die Ergebnisse des Diskurses oder nur die Metadaten berücksichtigen möchte. Auf die Berechnung von $P_{meta}(R|d,q)$ soll in diesem Artikel nicht weiter eingegangen werden. Statt dessen konzentrieren wir uns nun auf die Abschätzung von $P_{dis}(R|d,q)$.

Zu einem Dokument d haben wir eine Menge $\{A_1,\ldots,A_n\}$ von zum Dokument zugehörigen Annotationen. Ferner kann zwischen zwei Annotationen $A_i, A_j \in \{A_1,\ldots,A_n\}$ die Beziehung $rel(X,A_i,A_j)$ existieren, wobei X eine der in Abschnitt 2 eingeführter Diskursstrukturrelationen darstellt. Analog sind auch Beziehungen zwischen d und dessen direkten Annotationen mittels rel definiert. Die Annotationen, deren Beziehungen und das digitalisierte Dokument d als Wurzel bilden den Annotationsthread.

Zu jeder Annotation A ist nun die Wahrscheinlichkeit $P(R|A,q)$ zu berechnen, dass A relevant zur Anfrage ist. Da wir auch bei Annotationen den Kontext berücksichtigen wollen, d.h. alle zu dieser Annotation gemachten direkten Annotationen (seine Söhne im Annotationsthread) und deren Relationstypen, ergibt sich ähnlich wie bei den Dokumenten

$$P(R|A,q) = \beta \cdot P_{dis}(R|A,q) + (1-\beta) \cdot P_{ann}(R|A,q),$$

wobei $P_{ann}(R|A,q)$ die Relevanzwahrscheinlichkeit angibt, bei der nur die Annotation selbst betrachtet wurde, während bei $P_{dis}(R|A,q)$ wiederum ausschließlich die Söhne im Annotationsthread berücksichtigt werden. Da Annotationen aus freiem Text bestehen, kann $P_{ann}(R|A,q)$ mittels wohlbekannter Retrievaltechniken abgeschätzt werden. β gibt an, wie stark der Kontext der Annotation gewichtet werden soll.

Zu Berechnung von $P_{dis}(R|A,q)$ werden die direkten Söhne $A' \in succ(A)$ betrachtet (für die wiederum $P(R|A',q)$ berechnet werden muss). Für jedes Paar (A,A') wird $P(R|A,A',rel(X,A,A'),q)$, die Wahrscheinlichkeit, dass A relevant zur Anfrage q ist, wenn die Diskursrelation X zwischen A und A' existiert, abgeschätzt. Dabei sind für die verschiedenen Diskursrelationstypen entsprechende Berechnungen zu definieren.

Hat man $P(R|A,A',rel(X,A,A'),q)$ für alle Söhne von A berechnet, ergibt sich

$$P_{dis}(R|A,q) = \frac{1}{|succ(A)|} \sum_{\substack{A' \in \\ succ(A)}} P(R|A,A',rel(X,A,A'),q).$$

Sollte A ein Blatt im Annotationsthread sein, gilt $P(R|A,q) = P_{ann}(R|A,q)$.

Der nachfolgende Algorithmus p_annotation leistet die Berechnung von p = $P(R|A,q)$. Dabei ist $\mathcal{A}$ die Menge der Annotationen im Annotationsthread zu d, p_ann ein Algorithmus zur Berechnung von $P_{ann}(R|A,q)$, und p_rel dient der Berechnung von

$P(R|A, A', rel(X, A, A'), q)$ (in diesem Algorithmus erfolgt ein erneuter Aufruf von p_annotation mit der Annotation A' als Parameter).

```
    if A ist Blatt in 𝒜 then
2:    p = p_ann(A,q)
    else
4:    s = Anzahl direkter Nachfolger von A in 𝒜
      p_dis = 0
6:    for all direkte Nachfolger A' von A in 𝒜 do
        p_dis = p_dis + (p_rel(A,A',rel(X,A,A'),q) / s)
8:    end for
      p = beta * p_dis + (1 - beta) * p_ann(A,q))
10: end if
```

Hat man $P(R|A, q)$ mit obigem Algorithmus für alle direkten Nachfolger A von d im Annotationsthread berechnet, so ergibt sich

$$P_{dis}(R|d, q) = \frac{1}{|succ(d)|} \sum_{\substack{A \in \\ succ(d)}} P(R|d, A, rel(X, d, A), q),$$

wobei für die Abschätzung von $P(R|d, A, rel(X, d, A), q)$ wieder eine von der Diskursrelation X abhängige Funktion zu finden ist, in der $P(R|A, q)$ mit einfließt.

4 Zusammenfassung und Ausblick

In diesem Beitrag haben wir gezeigt, wie Annotationen (zugehörig zu einem Dokument d), zwischen denen eine typisierte Verbindung besteht (basierend auf Diskursstrukturrelationen) ausgenutzt werden können, um zusätzliche Informationen für das Retrieval von d zu gewinnen. Die Annotationen und deren Beziehungen modellieren dabei einen wissenschaftlichen Diskurs, der das Dokument d als Thema hat. Für die Berechnung von $P(R|d, A, rel(X, d, A), q)$ bzw. $P(R|A, A', rel(X, A, A'), q)$ sind entsprechende Funktionen zu definieren, die die Diskursrelation X zwischen Dokument und Annotation (bzw. zwischen zwei Annotationen) geeignet reflektiert.

Literaturverzeichnis

[BTSDW01] H. Brocks, U. Thiel, A. Stein, and A. Dirsch-Weigand. Customizable Retrieval Functions Based on User Tasks in the Cultural Heritage Domain. In Panos Constantopoulos and Ingeborg Sølvberg, editors, *Proceedings of the ECDL 2001, Darmstadt, Germany, Sept. 2001*, Lecture Notes in Computer Science, pages 37–48, Berlin et al., 2001. Springer.

[MH93] E. Maier and E.H. Hovy. Organising Discourse Structure Relations Using Metafunctions. In *New Concepts in Natural Language Processing*, pages 69–86. Pinter, London, 1993.