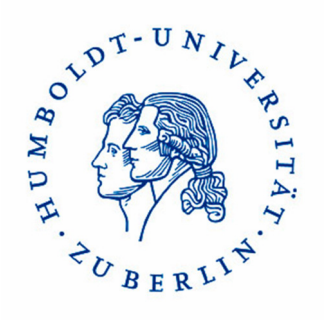

Exposé zur Studienarbeit

Relevanzranking in Lucene im biomedizinischen Kontext

Christoph Jacob

Betreuer: Phillipe Thomas, Prof. Dr. Ulf Leser

04. August 2010

1. Motivation

„Sucht und ihr werdet finden“ – dieses wohlbekannte Bibelzitat hat kaum an Relevanz für die heutige Informationsgesellschaft verloren. Wie schon vor 2000 Jahren ist bei der Suche jedoch entscheidend, *was* man findet. Aus dem manuellen Vorgang des Suchens in einem überschaubaren Bestand an Dokumenten ist im vergangenen Jahrhundert ein computergestützer Prozess geworden, was Fluch und Segen gleichzeitig bedeutet. Segen deshalb, weil man in Sekundenbruchteilen mehrere Millionen Zeilen Text durchsuchen kann, Fluch weil die potenziell große Menge an Suchergebnissen oft schwer zu überblicken ist.

In den Naturwissenschaften hat sich mittlerweile das digitale Publizieren durchgesetzt und spielt insbesondere in der Biomedizin und Genforschung eine zentrale Rolle. Dies führt dazu, dass der Bestand an Volltexten in diesen Bereichen rapide zunimmt. Diese Volltexte sind in aller Regel sehr gehaltvoll an Informationen, sind aber im Gegensatz zu strukturierten Datenbeständen, wie man sie beispielsweise in Datenbanken oder xml-Dateien vorfindet, schwer zu durchsuchen und stellen besondere Anforderungen an eine effiziente und zielführende Suche.

So lassen sich bei schwach strukturierten Datenbeständen wie Volltexten selten klare Suchkriterien definieren, welche die Suchergebnisse genügend einschränken. Während bei einer diskreten Suche in einem strukturierten Kontext gut ermittelbar ist ob ein Datensatz relevant für eine Query ist, ist dies bei der Volltextsuche schwer zu bestimmen – ein Ergebnis kann hier mehr oder weniger relevant sein. Beispielsweise ist die Suche in einer relationalen Datenbank mit Informationen über Flugbuchungen nach einem Flug mit einer bestimmten Nummer recht trivial, da die Informationen gut strukturiert vorliegen und ein Flug in aller Regel über seine Nummer eindeutig identifiziert werden kann. Hingegen ist die Suche in einem Korpus von Volltexten, wie man sie im biomedizinischen Kontext vorfindet, schwieriger bezüglich der Ermittlung der Relevanz eines Dokuments für eine Query.

Typisch für eine solche Suche könnte eine Anfrage nach Termen wie Krankheiten oder Genen sein. Sucht der Nutzer beispielsweise nach dem Gen mit den Namen *FMR1*, so erwartet er in den Suchergebnissen Dokumente, die eine Relevanz zum gesuchten Gen haben. Diese Relevanz kann sich im Auftreten des Gennamens oder dessen Bezeichnung äußern, beispielsweise aber auch darin, dass die im Dokument enthaltenen Informationen einen direkt oder indirekten Bezug zum Gen haben. Da das Gen FMR1 für die Codierung des Proteins *FMRP* verantwortlich ist, könnten auch Dokumente relevant sein die dieses Protein behandeln oder dessen Funktion beschreiben.

All diese Kriterien sind wichtig, um ein Ranking der Suchergebnisse durchzuführen, damit diese dem Nutzer in einer Reihenfolge präsentiert werden können, welche die relevantesten Dokumente höher listet als die weniger relevanten. Ziel dieses Relevanzranking ist es, dem Nutzer das mühsame Durchsuchen der Ergebnisse nach den gewünschten Informationen zu erleichtern.

Eine Suche und das entsprechende Relevanzranking im biomedizinischen Kontext wird weiterhin dadurch erschwert, das bei der Formulierung der Query selten ein einheitliches Vokabular genutzt wird. Die Suchterme „FMR1“, „fragile X mental retardation 1“ oder die Entrez Gene [1] GeneID „2332“ sollte nach Möglichkeit dasselbe Ergebnis liefern. Ein weiteres Problem stellt die Doppeldeutigkeit von Suchtermen dar. So kann der Suchterm „fragile X mental retardation 1“ das menschliche Gen FMR1 (GeneID 2332) meinen oder das Gen FMR1 der Wanderratte (GeneID 24948). Ein Relevanzranking im biomedizinischen Kontext muss sowohl die Synonymie als auch die Ambiguität beachten.

2. Ziele

Im Zuge dieser Studienarbeit sollen Kriterien für ein adäquates Relevanzranking im biomedizinischen Kontext ermittelt und Hilfe der Open-Source Volltextsuchmaschine *Lucene* [2] in ein bestehendes System implementiert werden. Im Vordergrund stehen dabei Queries, welche ein oder mehrere Gene als Suchterme enthalten. Gene können dabei sowohl durch ihre Symbole, als auch durch einen eindeutigen Schlüssel z.B. der GeneID bezeichnet werden. Die Relevanz der Antworten wird daher in erster Linie durch das Vorkommen der Gene im Dokument bestimmt.

2.1. GeneView

Im Rahmen des Wettbewerbs BioCreative III IAT: Interactive Demostration Task for Gene Indexing and Retrieval (IAT 2010) [3] wurde am Lehrstuhl für Wissensmanagement in der Bioinformatik das Informationssystem *GeneView* angepasst, welches im Rahmen von *ColoNet* [4] entwickelt wurde. GeneView indiziert wissenschaftliche Artikel und stellt diese in einer Weboberfläche für eine Schlüsselwortsuche zur Verfügung. GeneView soll dahingehend erweitert werden, dass die Suchergebnisse mittels eines Relevanzranking in eine Sortierung gebracht werden, die dem Nutzer die für seine Suchanfrage relevantesten Artikel oben anzeigt. Da der GeneView-Index mittels Lucene erstellt wurde ist zu prüfen, inwiefern das Standard-Lucene Ranking verbessert werden kann. Dabei könnten auch kontextabhängige Kriterien eine Rolle spielen.

2.2. Kriterien und Methoden für das Relevanzranking

Für das Relevanzranking sollen zunächst folgende Standardmodelle eingesetzt sowie erweiterte Methoden implementiert werden:

tf-idf: term frequency - inverse document frequency
Diese im Information Retrieval häufig eingesetzte Gewichtung für Terme und Dokumente setzt sich zusammen aus der *Termfrequenz*, welche eine Aussage über die Bedeutung eines Terms für ein einzelnes Dokument trifft, sowie der *inversen Dokumenthäufigkeit*, welches eine Aussage über die Bedeutung eines Terms für die Gesamtheit aller Dokumente trifft. [5]

tf-idf für einzelne Sektionen innerhalb eines Dokuments
Trotz der schwachen Strukturierung von Volltexten lassen sich diese in einzelne *Sektionen* wie z.B. Title, Abstract, Summary unterteilen. Eine Aussage über das Vorkommen von Suchtermen innerhalb bestimmter Sektionen könnte nützlich für das Ranking sein. Beispielsweise könnte das Auftreten des Suchterms im Titel eines Dokuments ein Hinweis darauf sein, dass das Dokument relevanter für die Suche ist. Hingegen könnte das Auftreten eines Suchterms innerhalb der Diskussion als weniger relevant bewertet werden, da hier die Wahrscheinlichkeit für das Auftreten von Termen die nur am Rande mit dem eigentlichen Dokument zu tun haben größer ist.

Zu prüfen wäre weiterhin für beide Rankings, ob ggf. durchgeführte *Normierungen* bezüglich der Dokumentlänge und der Häufigkeit von Auftreten von Termen im biomedizinischen Kontext sinnvoll sind oder nicht.

GeneOntology Integration
Beim Information Retrieval im biomedizinischen Kontext spielt die *GeneOntology* [6] eine zentrale Rolle. Diese stellt ein einheitliches Vokabular sowie Annotationen zu Genen und Proteinen zur Verfügung. Diese Annotationen könnten genutzt werden, um die Suchterme des Nutzers zu erweitern und eine „breitere“ Suche durchzuführen. Diese Suche würde allerdings auch ein modifiziertes Ranking erfordern, da die Gewichtung des ursprünglichen Suchterms größer sein muss als die Gewichtung des Annotationsterme. Daher ist zu prüfen, ob die Integration der GeneOntology praktikabel ist und sich positiv auf das Relevanzranking auswirkt.

JournalRanking
Eine weitere Möglichkeit die Relevanz eines Dokuments zu bestimmen könnte sein, ein Ranking unter den *Journals* zu erstellen, in denen die wissenschaftlichen Artikel publiziert wurden. Im Hintergrund steht hier die Frage ob es im biomedizinischen Bereich Journals gibt, die ein hohes Ansehen genießen und in denen konstant „gute“ Artikel publiziert werden. So könnte auch der Nutzer daran interessiert sein, Artikel aus rennomierten Journals ein besseres Ranking zu geben.

2.3. Benchmarking

Um herauszufinden, welche eingesetzten und implementierten Methoden das Relevanzranking verbessern, ist es notwendig die unterschiedlichen Verfahren gegeneinander zu testen. In aller Regel werden für solche Zwecke *Experten* herangezogen, die für gegebene Suchterme ein Ranking der Suchergebnisse händisch vornehmen. An diesem „Gold-Standard“ lassen sich dann die Ergebnisse der jeweiligen Ranker messen [7]. Da es schwierig werden könnte einen solchen Experten zu Rate zu ziehen, dieses Verfahren aber bei verschiedenen Wettbewerben im Bereich des Information Retrieval angewandt

wird (so auch im BioCreative-Wettbewerb), ist zu prüfen, ob stattdessen auf bereits vorhandene Datenbestände zurückgegriffen werden kann.

In diesem Zusammenhang ist zu prüfen, wie andere wissenschaftliche Arbeiten auf dem Gebiet des Relevanzrankings ihre Ranking-Verfahren austesten. Eventuell können hier Methoden übernommen und angepasst werden. In jedem Fall soll ein Vergleich der eigenen Ranking-Methoden zu anderen gezogen werden.

3. Vorgehensweise

Wie bereits erwähnt erfolgt die Implementierung des Relevanzrankings innerhalb des bestehenden Informationssystems GeneView. Da die verwendeten Indexstrukturen mittels Lucene erstellt wurden, wird das Ranking ebenfalls in Lucene implementiert werden. Lucene bietet dazu eine eigene API mit welcher man zunächst das tf-idf-Ranking einsetzen und modifizieren kann: über die Klassen *Similarity* und *Scoring* lassen sich eigene Rankingfunktionen und Normierungen erstellen [8]. Diese können dann über die Suche auf einen bestehenden Index getestet werden.

Für die Integration der GeneOntology müsste man vor der eigentlichen Suche im Index eine *QueryExpansion* des gesuchten Gens mithilfe der Annotationen in der GeneOntology durchführen. Anschließend ist zu prüfen, ob man dem ursprünglichen Suchterm (das Gen) über einen *QueryBoost* eine höhere Gewichtung als den Annotationen gibt, um das Ranking nicht zu stark auf den Annotationen basieren zu lassen. Sowohl die QueryExpansion als auch der QueryBoost können im Gegensatz zu den folgenen Methoden auf einen bereits bestehenden Index implementiert werden.

Die Gewichtung der Sektionen lässt sich über einen *FieldBoost* auf die im Index vorhandenen Felder für die Sektionen implementieren. Dieser FieldBoost muss allerdings bei der Indexerstellung gesetzt werden und kann nicht bei der Suche auf einem bestehenden Index dynamisch gesetzt werden. Dies erschwert das Testen und macht eine spätere Interaktion des Benutzers mit den Boost-Werten so gut wie unmöglich.

Das JournalRanking kann über die dritte Art von Boost implementiert werden kann die Lucene zur Verfügung stellt: den *DocumentBoost*. Dieser Boost kann ebenfalls nur bei der Indexerstellung gesetzt werden und würde ausgehend von einem Ranking der Journals Artikeln aus höher gerankten Journals einen Boost geben. Ein dynamisches Boosten ist hier zunächst auch nicht praktikabel, wäre aber sicherlich aus Nutzersicht wünschenswert da dieser evtl. unterschiedliche Präferenzen bezüglich des Rankings der Journals hat.

Schließlich bleibt das Benchmarking der implementierten Ranker, um herauszufinden welche Methoden eine Verbesserung des Relevanzrankings bewirken. Als *Baseline* soll hier das Standard tf-idf-Ranking von Lucene ohne jegliche Modifikation dienen. Der Gold-Standard müsste für einige wenige Suchterme von einem Experten erstellt werden. In Bezug auf andere wissenschaftliche Arbeiten auf dem Gebiet des Relevanzrankings kann alternativ oder zusätzlich eine Analyse der Ranker-Performance auch auf Grundlage gefundener Methoden erstellt werden.

Literatur

[1] Maglott, D., J. Ostell, et al. (2005). "Entrez Gene: gene-centered information at NCBI." Nucleic Acids Res **33**(Database issue): D54-58.

[2] Apache Lucene – Overview. http://lucene.apache.org/java/docs/index.html

[3] BioCreative III – IAT Task. http://www.biocreative.org/tasks/biocreative-iii/iat/index.html

[4] ColoNet. http://www.colonet.de/

[5] Manning, C. D., P. Raghavan, et al. (2009). Introduction to Information Retrieval. Cambridge [u.a.], Cambridge Univ. Press. Kap 6.2.

[6] (2010). "The Gene Ontology in 2010: extensions and refinements." Nucleic Acids Res 38(Database issue): D331-335.

[7] Roberts, P. M., Cohen, A. M., et al. (2008) Tasks, topics and relevance judging for the TREC Genomics Track.

[8] Gospodnetić, O. and E. Hatcher (2005). Lucene in Action : [a guide to the Java search engine]. Greenwich, Conn., Manning.