

(Akademische) Suchmaschinen und Empfehlungsdienste

Übersicht

Bitte beachten Sie: Alle Themenvorschläge hier sind lediglich Vorschläge. Wenn Sie sich für Suchmaschinen oder Recommendation Engines sowie Text Mining, Data Mining, Web 2.0, Mind Mapping, GUIs, Ähnlichkeitsanalysen, Collaborative Filtering, Neuronale Netze, oder Referenzanalysen interessieren, dann haben wir mit Sicherheit ein interessantes Thema für Sie.

Alle Themen können im Rahmen einer Diplom- oder Masterarbeit bearbeitet werden. Mit reduziertem Umfang auch als Studien- bzw. Bachelorarbeit und mit verstärktem Fokus auf Programmierung auch als Labor- bzw. Softwarepraktikum.

Wir – das [SciPlore](#)-Team – untersuchen, wie Informationstechnologie Forscher und Wissenschaftler bei ihrer Arbeit am besten unterstützen kann. "Informationstechnologie" ist hierbei weit gefasst und schließt unter anderem Computer und das Internet im Allgemeinen mit ein und Web 2.0, Grid Computing, Data & Text Mining sowie Künstliche Intelligenz im Speziellen. In den letzten Jahren haben diese Technologien Forschung und Wissenschaft revolutioniert: Digitale Bibliotheken, akademische Suchmaschinen und Online Referenzmanager sind nur einige der neuen Möglichkeiten die Wissenschaftlern und Forschern das Leben erleichtern und ganz neue Arbeitsweisen ermöglichen.

Unser Forschungsbereich ist ein junges Forschungsgebiet und bietet unglaublich viel Potential – das haben auch andere mittlerweile erkannt. Neben kleineren Projekten wie Bibsonomy [1] oder Scholarz [2] treiben große Firmen die Entwicklung voran. Dabei sind Google mit Google Scholar [3] und zukünftig Google Palimpsest [4], Microsoft Research mit eigenen Konferenzen [5] und Forschungsgruppen im Bereich eScience [6], SAP mit einer Software zum Managen ganzer Forschungseinrichtungen [7] und der Fachverlag Nature mit der Software Connotea zum Verwalten wissenschaftlicher Dokumente [8].

Wir möchten Sie herzlich Einladen unser Team zu verstärken und bei uns Ihre Abschlussarbeit zu schreiben. Für ihre Themenvorschläge sind wir offen – jede neue Idee wie IT (sei es nun das Web 2.0, Neuronale Netzwerke, Text Mining, innovative GUIs, oder, oder, oder...) Wissenschaftler unterstützen kann, ist willkommen. Im folgenden finden Sie einige konkrete Themenvorschläge.

Themenvorschlag: Zitationsbasierte Recommender Systeme

Stellen Sie sich vor, sie haben eine wissenschaftliche Publikation gelesen, die Ihnen gefällt. Nun möchten Sie ähnliche Dokumente finden. Mittels Zitationsanalysen ist dies leicht möglich: der einfachste Weg besteht darin, jene Dokumente zu lesen die in dem Ausgangsdokument referenziert sind oder die Dokumente die Ihr Ausgangsdokument referenzieren. Zusätzlich könnten Sie jene Dokumente lesen welche die gleichen Dokumente referenzieren wie ihr Ausgangsdokument. So und mit weiteren Methoden können Sie schnell eine lange Liste relevanter Literatur erhalten.

Bei zitationsbasierten Empfehlungsdiensten gibt es jedoch zwei Herausforderungen. Einer davon stellen Sie sich in Ihrer Abschlussarbeit.

Referenzanalyse in Wissenschaftlichen Dokumenten

Die erste Herausforderung liegt darin, Referenzen zuverlässig zu erkennen, zu extrahieren und zu strukturieren. Das heißt, zu analysieren wo im Text Referenzen stehen, und welcher Teil der Referenz den Autor darstellt, welche das Journal, etc. Was auf den ersten Blick trivial erscheinen mag ist es beileibe nicht. Hunderte verschiedene Zitierstile und Nachlässigkeit der Autoren erfordern intelligente und fehlertolerante Algorithmen. Bisherige Algorithmen sind dazu nicht immer in der Lage, wie Fehler bei Google Scholar und anderen Referenzdatenbanken zeigen.

Ihre Aufgabe besteht darin, die existierenden Algorithmen zu evaluieren und ihre Stärken und Schwächen zu analysieren. Danach entwickeln Sie einen neuen Algorithmus und testen ihn. Ihrer Kreativität sind dabei keine Grenzen gesetzt, denn es gibt keinen vorgeschriebenen und einzig richtigen Lösungsweg. Vielleicht lösen Sie das Problem mit einem einzigartigen Text Mining Algorithmus? Oder Sie entwickeln einen "Meta-Algorithmus", der sich der vorhandenen Algorithmen bedient und diese vorteilhaft kombiniert? Oder Sie haben eine geniale Idee im Web 2.0 Stil, der die wissenschaftliche Community dazu bringt, freiwillig Referenzen zu identifizieren? Oder... ? Die potentiellen Einsatzmöglichkeiten Ihres Algorithmus sind vielfältig. Zum einen wird er natürlich bei SciPlore eingesetzt. Doch vielleicht wird ihr Algorithmus auch die beste Open Source Lösung zum Extrahieren von Referenzen? Oder Sie schlagen mit Ihrer Lösung sogar den Algorithmus von Google Scholar?

Konnten wir Ihr Interesse wecken? Dann melden Sie sich bei uns!

Ähnlichkeitsbestimmung von Wissenschaftlichen Dokumenten

Die zweite Herausforderung besteht darin, relevante Dokumente zu ranken. Denn auch wenn sie viele Dokumente gefunden haben, die ähnlich ihrem Ausgangsdokument sind bleibt die Frage: Welches Dokument zu erst in der Ergebnisliste anzeigen bzw. zu erst lesen?

Ihr Ziel ist die Entwicklung von Rankingverfahren und Realisierung eines ersten Prototyps für einen Empfehlungsdienst. Hierbei soll der Anwender ein Dokument vorgeben das er mag und dann werden ihm weitere Dokumente angezeigt, die ihn interessieren könnten. Dafür untersuchen Sie die existierenden Rankingverfahren basierend auf Zitationen, verbessern diese und/oder entwickeln neue. Wir sind sicher, dass Ihnen nach einiger Einarbeitungszeit

dutzende Ideen kommen werden wie Ähnlichkeiten zwischen verschiedenen Dokumenten bestimmt werden können. In ihrer Arbeit konzentrieren Sie sich dann auf die besten zwei oder drei Ideen.

Falls Ihnen eigene Ideen fehlen können Sie auch eine unserer beiden Methoden weiter verbessern und untersuchen. Näheres zu den zwei Methoden erzählen wir Ihnen gerne bei einem persönlichen Gespräch.

Themenvorschlag: "Document Usage Mining" zur Unterstützung von Recommender Systemen

Zitationsanalysen stellen nur eine Möglichkeit dar um Empfehlungsdienste für wissenschaftliche Dokumente zu realisieren. Im Rahmen von SciPlore arbeiten wir an einem komplett neuen Ansatz. Beim von uns so genannten "Document Usage Mining" (DUM) analysiert eine Software im Hintergrund die Lesegewohnheiten von Wissenschaftlern. Also welche Dokumente ein Wissenschaftler liest, wie lange, welche er ausdrückt, in welchen Dokumenten er Textpassagen markiert, etc. Die Grundannahme dabei ist, dass je intensiver sich ein Wissenschaftler mit einem Dokument beschäftigt, desto relevanter ist es für ihn. Basierend auf diesen Daten werden Wissenschaftler in Gruppen eingeteilt. Jedem Wissenschaftler werden dann die Dokumente empfohlen, die die anderen Wissenschaftler aus der gleichen Gruppe gelesen haben. Das Vorgehen ist in etwa mit dem von Amazon vergleichbar, die ihren Kunden die Produkte empfehlen die ähnliche Kunden gekauft haben.

Bisher ist Document Usage Mining nur eine Idee zu der erste Gedanken existieren. Sie haben die Chance mit Ihrer Abschlussarbeit diese Idee mit Leben zu füllen. Überlegen Sie, welche Aktionen überwacht werden sollen (und können), wie der Einfluss einer jeden Aktion bewertet und Datenschutz gewährleistet wird. Entwickeln Sie einen ersten Prototyp und testen Sie ihn mit einer Gruppe von Wissenschaftlern.

Themenvorschlag: Reverse-Engineering Google Scholar

Um neue Such- bzw. Recommender Systeme entwickeln zu können, muss zuerst untersucht werden wie die vorhandenen System arbeiten. In ihrer Diplomarbeit analysieren Sie, wie Google Scholar die Ergebnisse rankt. Das heißt, sie "reverse-engineeren" den Ranking-Algorithmus von Google Scholar und geben Antwort auf die Frage welche Faktoren (z.B. Zitationsanzahl, Keywordhäufigkeit im Text, Reputation des Autors, etc.) dazu führen, dass ein Paper bei Google Scholar so gerankt wird, wie es gerankt wird. Zusätzlich überlegen Sie, wie der Algorithmus von Google Scholar verbessert werden könnte. Die ersten Schritte hierfür wurden bei uns im VLBA Lab bereits unternommen und erste Publikationen veröffentlicht (siehe http://www.sciplore.org/publications_en.php). Näheres erläutern wir bei einem persönlichen Gespräch.

Themenvorschlag: Dokumentenidentifikation zur Realisierung von Recommender Systemen

Grundlegend für SciPlore's Erfolg ist die zuverlässige Identifizierung von elektronischen Dokumenten, zumeist PDFs. Denn was nützt es beispielsweise beim Document Usage Mining zu wissen, dass ein Dokument intensiv bearbeitet wurde, aber nicht sagen zu können welche Publikation eigentlich hinter dem PDF steckt?

Mit ihrer Abschlussarbeit werden Sie somit quasi das Fundament für SciPlore legen. Sie entwickeln ein Konzept wie am zuverlässigsten gesagt werden kann, dass das PDF XYZ die Publikation mit dem Titel abc vom Autor 123 beinhaltet. Verschiedene Lösungsansätze sind denkbar, die Sie auch kombinieren können. Zum einen könnte via Text Mining der Titel und die Autoren extrahiert werden. Sie könnten aber auch versuchen mit den Referenzen im Dokument eine Art Fingerabdruck zu erstellen und so Dokumente eindeutig zu identifizieren. Die extrahierten Daten (Titel, Autor und/oder Referenzen) könnten Sie dann noch mit bestehenden Datenbanken (SciPlore, Google Scholar, ...) vergleichen. Denkbar ist auch der Aufbau einer Datenbank mit Hashwerten von PDF Dateien und den zugehörigen Metadaten. Also eine Art freedb für wissenschaftliche PDF-Dateien.

Themenvorschlag: Entwicklung eines Suchalgorithmus für Akademische Suchmaschinen

Im Bereich der Websuchmaschinen ist Google der unangefochtene Platzhirsch. Bei akademischen Suchmaschinen gibt es einen solchen Marktführer nicht und das hat einen Grund: keine der vorhandenen akademischen Suchmaschinen ist wirklich gut. In Ihrer Arbeit untersuchen Sie woran das liegen könnte und entwerfen einen eigenen Suchalgorithmus für wissenschaftliche Suchmaschinen. Das mag sich schwer anhören, ist es aber nicht. Kontaktieren Sie uns für mehr Informationen. Wir haben schon viele Ideen, die nur darauf warten von Ihnen aufgegriffen zu werden.

Themenvorschlag: Dokumentenklassifikation zur Realisierung von Recommender Systemen

In Ihrer Abschlussarbeit untersuchen Sie wie mittels Tagging, Text-Mining und/oder Zitationsanalyse wissenschaftliche Dokumente am besten einem oder mehreren Themengebieten zugeordnet werden können. Das Ziel ist die Entwicklung eines Prototypen der zu einem wissenschaftlichen Dokument anzeigt aus welchem Themenbereich es stammt. Für Ihre Arbeit lesen Sie sich als erstes in die oben genannten drei Themengebiete ein, um dann ein Konzept zu entwickeln welches Sie abschließend implementieren und testen. Insbesondere wenn Sie Tagging berücksichtigen muss beachtet werden, dass wissenschaftliche Dokumente spezielle Eigenheiten aufweisen die "normales" unstrukturiertes Tagging nicht optimal erscheinen lassen. Zudem werden vermutlich niemals alle wissenschaftlichen Dokumente von den Lesern getaggt werden. In diesem Zusammenhang könnten Sie nach Auswegen suchen. Erste Ideen haben wir bereits, und würden diese gerne gemeinsam mit Ihnen besprechen.

Organisation

Uns ist es wichtig, dass Ihre Arbeit einen starken Bezug zur Praxis hat und wissenschaftlich relevant ist. Wir befürworten eine Kooperation von Ihnen mit Partnern aus Wirtschaft und Wissenschaft und unterstützen Sie aktiv bei der Suche nach solchen Partnern. Auch eine Kooperation mit anderen Studenten ist denkbar. Deswegen stellen alle hier präsentierten Themen nur einen ersten Vorschlag dar. Entsprechend den Wünschen und Bedürfnissen von Ihnen und Ihren Partnern sind Variationen in der Themenstellung möglich. Bei sehr guter Qualität Ihrer Arbeit unterstützen wir Sie aktiv bei der Veröffentlichung als Buch oder in einem internationalen Journal. In jedem Fall soll der praktische Teil ihrer Arbeit später bei

SciPlore eingesetzt werden. Betreut werden Sie während Ihrer Arbeit von [Jöran Beel](#), [Béla Gipp](#) und einem Hochschulprofessor Ihrer Wahl. Beginnen mit der Abschlussarbeit können Sie jederzeit.

Wenn Sie noch Fragen haben oder wir Ihr Interesse wecken konnten, dann kontaktieren Sie Jöran Beel und Béla Gipp per Email (abschlussarbeit AT sciplore.org).

Referenzen

- [\[1\] Bibsonomy](#) 
- [\[2\] Scholarz](#) 
- [\[3\] Google Scholar](#) 
- [\[4\] Google Palimpsest](#) 
- [\[5\] Microsoft Research](#) 
- [\[6\] Forschungsgruppen im Bereich eScience](#) 
- [\[7\] SAP](#) 
- [\[8\] Software Connotea](#) 