4. Übung zur Vorlesung "Internet-Suchmaschinen" im Sommersemester 2009

Prof. Dr. Gerd Stumme, Wi.Inf. Beate Krause

04. Juni 2008

1 Phrasensuche

Bis jetzt behandelt die in der Übung programmierte Suchmaschine einfache AND und OR Anfragen. Die Eigenschaften der Anfrage an sich spielen (noch) keine besondere Bedeutung.

1. Geben Sie Beispiele an, bei denen es Sinn machen würde, die Reihenfolge der Suchterme zu berücksichtigen.
2. Geben Sie Beispiele an, bei denen es sinnvoll wäre, Großbuchstaben von Kleinbuchstaben zu unterscheiden.
3. Skizzieren Sie einen Algorithmus, der anstelle von einfachen Anfragen im invertierten Index nach Phrasen suchen können soll, also z. B. nach "information retrieval" als anstatt bloß nach Dokumenten, die "information" und "retrieval" enthalten.

4. Müssen Sie dazu die Datenstruktur erweitern? Was wurde bisher nicht berücksichtigt?

2 Relevance Feedback, Anfrageerweiterung

1. Nehmen Sie an, ein Benutzer einer Suchmaschine hat die folgende Anfrage formuliert: *free music free downloads top free music*. Der Benutzer schaut sich zwei Dokumente, d_1 und d_2 genauer an. Er beurteilt d_1 mit dem Inhalt *music, free, software, free, music* relevant, und d_2 mit dem Inhalt *free, thrills, downloads* als nicht relevant. Nehmen Sie an, dass die Suchmaschine zur Anfrageerweiterung die Termhäufigkeit benutzt. Benutzen Sie Rocchio Relevance Feedback, wie in der Vorlesung auf Seite 7 angegeben, um einen neuen Anfragevektor zu bestimmen. Rechnen Sie mit $\alpha = 1$, $\beta = 0.75$ und $\gamma = 0.25$.
2. Wie würde man die Parameter α, β und γ gewichten, wenn nicht viel oder kein vertrauenswürdiges Feedback vorliegt?
3. Heutige Web-Suchmaschinen benutzen in der Regel kein klassisches Relevance Feedback oder Anfrageerweiterung, z. B. weil Anwender nicht bereit sind, aufwendige Formulare zur Anfrage auszufüllen.

   Welche Ansätze von Relevance Feedback oder Anfrageveränderungen können in Web-Suchmaschinen heutiger Machart umgesetzt werden, ohne ein kompliziertes Interface zu benutzen?
4. Anfrageerweiterung zielt darauf ab, durch Veränderung der Anfrage mehr relevante Dokumente aus einem gegebenen Korpus zu finden.

   Sehen Sie eine Parallele (oder Symmetrie) zwischen dieser Idee und der Arbeit von Suchmaschinenoptimierern?

3 Einfluss der verschiedenen Methoden auf Precision und Recall

Bestimmen Sie, wie die folgenden Techniken Precision und Recall beeinflussen. Entscheiden Sie, ob das jeweilige Maß erhöht, erniedrigt oder unbeeinflusst bleibt.

- Anfragen mit dem booleschen AND
- Anfragen mit dem booleschen OR
- Platzhalter in der Anfrage
- Anfrageerweiterung
- Beachtung der Groß- und Kleinschreibung
- Stemmen von Wörtern
- Entfernung von Stoppwörtern
- Phrasenberücksichtigung
- Einsatz eines Thesaurus

4 String-Suche: Reguläre Ausdrücke

Bei der String-Suche werden nicht mehr Terme als atomare Einheit betrachtet sondern einzelne Zeichen. Eine Möglichkeit, Muster herauszufinden, ist die Anwendung von regulären Ausdrücken. Können Sie die folgenden Informationen mit Hilfe eines regulären Ausdruckes wiedergeben?

- Eine Telefonnummer mit vierstelliger Vorwahl. Diese kann durch Klammern abgetrennt werden.

  Eine deutsche Postleitzahl mit Ortsnamen, evtl. mit Bindestrich, z. B. “34117 Kassel”, “D-34132 Kassel-Oberzwehren”
- Eine Kasseler Straßenadresse. Diese besteht aus einem Präfix “Königs”, “Kaiser”, “Wilhelms” oder “Karls” mit einem Suffix “platz”, “straße”, “tor”, “allee” und einer ein- bis dreistelligen Hausnummer. Eventuell steht vor dem Straßennamen noch eines der Adjektive “Obere”, “Unterer”, “Großes”, “Kleines” (in einem beliebigen Genus).
- Spezifizieren Sie die Abhängigkeit des Genus des Adjektivs vom Suffix im regulären Ausdruck, oder begründen Sie, warum das nicht möglich ist.

5 Praxisübung; Abgabe am 16.06.2009

Implementieren Sie die Phrasensuche (siehe Aufgabe 1). Zeigen Sie über Ihre Suchschnittstelle das Dokument, sowie einen kurzen Kontext aus dem gefunden Dokument. Der Kontext besteht aus (z.B.) fünf Wörtern vor und nach dem (ersten) Auffinden des Suchbegriffes / der Suchphrase.