Ein Verfahren zur thematisch spezialisierten Suche im Web und seine Realisierung im Prototypen HomePageSearch

Dissertation

Gerd Hoff
Universität Trier – Informatik

Kurzfassung

Insbesondere im wissenschaftlichen Bereich ist es zur Selbstverständlichkeit geworden, zur Befriedigung eines Informationsbedürfnisses zunächst das **World Wide Web** (Web) zu benutzen. Der komplexe Aufbau des Webs und die dort herrschende Unordnung machen das Auffinden bestimmter **Dokumente** zu einem diffizilen Vorgang. Suchmaschinen und Web–Verzeichnisse unterstützen den Benutzer hierbei; ihre effektive Nutzung ist jedoch nicht einfach. Es ist daher hilfreich, Suchdienste zu entwickeln, welche Kontextwissen mitbringen.

In der vorliegenden Arbeit wird das Verfahren der **thematisch spezialisierten Suche** präsentiert und anhand der Menge der persönlicher Home Pages von Informatikern exemplifiziert. Der Bibliographie–Server DBLP.uni-trier.de liefert u.a. zu über 4.000 Wissenschaftlern manuell gesammelte persönliche Home Pages. Durch verschiedene Analysen werden in dieser Menge häufig vorkommende, also **relevante Eigenschaften** ermittelt. Dies ist ein kreativer Prozess, in den sowohl Suchwissen — Wissen über Techniken des Suchens im Web — als auch Fachwissen — Wissen über die Beschaffenheit persönlicher Home Pages— einfließen. Nach der Überprüfung dieser Eigenschaften auf ihre **Signifikanz** in der Gesamtheit der Web–Dokumente ist ein **thematisch spezialisiertes Wissen** erschaffen worden.

Zur Durchführung der Suche würde man idealerweise alle Web–Dokumente mit Hilfe des erarbeiteten Wissens und des Anfrageterms — des vollständigen Namens eines Informatikers — untersuchen. Da das Web aus mehreren Milliarden Dokumenten besteht, ist dies praktisch nicht durchführbar: Es ist notwendig, eine geeignete **Vorauswahl** zu treffen. Suchmaschinen haben Daten zu vielen Web–Dokumenten erfasst. Deshalb wird mit ihrer Hilfe unter Verwendung des Anfrageterms eine Vorauswahl getroffen.

Um nun die Dokumente der Vorauswahlmenge gemäß der Anfrage zu bewerten wird das thematisch spezialisierte Wissen adäquat in einer **Bewertungsfunktion** repräsentiert, welche jedem Dokument einen Wert zuweist und somit eine Rangfolge erstellt. Das Design der Bewertungsfunktion ist eine schöpferische Tätigkeit: Die sachgerechte **Gewichtung** der verschiedenen Eigenschaften bestimmt maßgeblich das Gelingen des Verfahrens.

Der vornehmlich in Java implementierte, vollautomatisierte **Web–Dienst** HomePageSearch (`hpsearch.uni-trier.de`) führt regelmäßig für über 100.000 Namen diesen Suchvorgang durch. Der Benutzer kann auf das in einer DB2–Datenbank gespeicherte Ergebnis zurückgreifen. Durch die periodische Neuberechnung des thematisch spezialisierten Wissens wird der Dynamik des Webs Rechnung getragen. Der modulare und stark parametrisierte Aufbau des Systems erleichtert die Portierung des Verfahrens auf andere Mengen von Web–Dokumenten. Nicht zuletzt aufgrund der starken Verlinkung mit DBLP und ResearchIndex wird HomePageSearch von vielen Wissenschaftlern intensiv genutzt.