

Suchmaschinen der Zukunft: Was kommt nach Google?

Zeit und Ort: 25.04.2006, EEAR

Referent: Prof. Dr. Gerhard Weikum, Max-Planck-Institut für Informatik, Saarbrücken

Protokoll: Ass. jur. Iris Speiser

Ausgehend vom Thema der Veranstaltung ging Prof. Weikum zunächst auf die Leistungsfähigkeit von Google ein.

Google sei ein sehr gutes Werkzeug für Suchanfragen nach sehr konkret beschreibbaren Informationen, z.B. die Suche nach Informationen über wichtige Personen. Ein großer Vorteil hierbei sei die Toleranz gegenüber Tippfehlern.

Bei komplexen Abfragen erreiche Google jedoch sehr schnell seine Leistungsgrenzen. Google suche nur nach einzelnen Webseiten; bei komplexen Abfragen nach Informationen, die über mehrere Unterseiten verteilt sind, versage der Algorithmus hingegen. Wenn man nach Informationen über einen Professor aus Saarbrücken suche, der sich mit Datenbanken oder Information Retrieval befasse und Forschungsprojekte zu XML habe, werde Google vermutlich keine guten Ergebnisse liefern, da die Forschungs- und Lehrtätigkeiten eines Professors in der Regel auf verschiedenen Seiten seiner Homepage dokumentiert werde. Ebenfalls problematisch seien beschreibende Suchabfragen, die die in der Seite enthaltenden Suchbegriffe nicht enthielten, sondern lediglich verwandte Begriffe. Als weiteres Beispiel nannte Weikum die Suche nach der Person, die an einer Veranstaltung teilgenommen habe, jedoch nicht als Referent aufgetreten sei.

Für Suchmaschinenteknologie jenseits von Google gebe es weite Anwendungsgebiete. Wichtig sei die Einbeziehung von Hintergrundwissen und semantischen Daten als ausdrucksstärkerer Repräsentation vom Wissen. Der Mensch müsse, z.B. durch gezielte Rückfragen, stärker in die Suche einbezogen werden. Auch der Kontext müsse stärker berücksichtigt werden.

Die semantische Suche als Versuch einer logischen Repräsentation von Weltwissen sei bereits Gegenstand verschiedener Forschungsprojekte. So würden Thesauri

entwickelt, die eine Suche nach verwandten Begriffen erlaube (hierzu verwies Weikum auf die Thesaurus-Projekte [WordNet](#) von der Princeton University und das deutsche [GermaNet](#)). Suchbegriffe könnten gewichtet werden; Beziehungen zwischen Begriffen könnten durch XML-Strukturen abgebildet werden. Die Datenbanktechnologie mit ihren logischen Funktionen könne mit strukturiertem XML kombiniert werden. Vielversprechend seien auch Forschungsarbeiten zur automatisierten Erkennung von Eigenschaften in Texten (z.B. [GATE/ANNIE](#)).

Ein weiteres Konzept sei die personalisierte Suche, bei der versucht werde, ein individuelles, persönliches Ergebnisranking zu bilden. Dies könne durch eine Frageauswertung anhand von persönlichen Interessen erfolgen. Ermitteln könne man derartige Interessenprofile durch das Erstellen von Userprofilen.

Google enthalte bereits Ansätze zum Personal Information Management, den Google PageRank. Google bewerte den PageRank im Wesentlichen nach folgenden Kriterien: Ausgehend von der Prämisse, dass man Links eher auf gute, thematisch verwandte Seiten setzt, würden Links zu einer Seite als Empfehlung gewertet. Links von "wichtigen" Seiten werde ein höherer Wert zugewiesen als solchen von "unwichtigen". Dieses System lasse sich zu einem personalisierten PageRank weiterentwickeln. Der Nutzer springe nämlich bevorzugt zu für ihn relevanten Seiten. Die Einbeziehung persönlicher Präferenzen, die sich z.B. aus Bookmarks ermitteln ließen, könnte zur Ermittlung eines solchen personalisierten PageRank benutzt werden.

Einen weniger individuellen Ansatz verfolge die Soziale Suche, die menschliches Verhalten und die so genannte "wisdom of crowds" ausnutze. Hierbei könne auf Sammlungen persönlicher Bookmarks, Suchlogs, Clickstreams und Communities wie z.B. Blogs zurückgegriffen werden. Bei der Bewertung der Suchergebnisse könnten dann Suchergebnisse bevorzugt werden, die von ähnlichen Nutzern präferiert würden. Problematisch hierbei sei allerdings die damit einhergehende Tendenz zum "gläsernen Nutzer", die insbesondere von Datenschützern sehr kritisch gesehen werde. So seien in jüngerer Vergangenheit Dienste wie "Yahoo! MyWeb" in die Kritik geraten, da sie für SPAM und ähnlichen Missbrauch anfällig seien.

Andererseits hätten derartige Communities schon heute beachtliche Fähigkeiten. Das Tagging in Flickr erlaube z.B. die Suche nach Bildern, die Stimmungen ausdrücken. Auf andere Weise ist eine solche Suche nicht automatisierbar.

Als abschließenden Ausblick gab Prof. Weikum folgendes Fazit:

Das Web habe ein enormes Potential als Wissensbasis, das noch nicht genutzt werde.

Das Wissen im Web lasse sich erschließen mit semantischer Suche, personalisierter Suche sowie einer in P2P-Netzwerke eingebetteten sozialen Suche.

Die Datenqualität müsse durch geeignete Maßnahmen sichergestellt werden.

In der abschließenden Diskussion wurde die Frage diskutiert, inwieweit neue Technologien zum Information Retrieval bereits das Stadium der Anwendungsreife erreicht haben. Funktionen wie Amazons "andere Käufer dieses Buches kauften auch..." stellen sicher nicht die abschließende Umsetzung einer sozialen Suche dar.