Intelligentes Recherchesystem

LuMriX: Finden statt nur suchen

Gießener Forscherteam um Dr. Schneider und Dr. Hölzer ebnet den Weg zum »Semantischen Web«

Das Internet, hilfreiche, scheinbar unendliche Informationen als Wissenquell für Kuchenrezepte, Autokauf und hochkomplexe wissenschaftliche Zusammenhänge liefernd, ist in Wirklichkeit ein morastiger Dschungel. Dort hat der Suchende vor lauter Gewächs Mühe, das Wesentliche zu finden. Die vermeintlichen Lianen, Google und Co., die einen Überblick geben und helfen sollen, entsprechend schnell zum Ziel zu kommen, leisten das nicht in dem Maße, in dem es wünschenswert ist. Der Grund: Sie sind einfach zu dumm. Die üblichen Recherchesysteme, Internet-Suchmaschinen und Volltextsuchhilfe, stoßen schnell an ihre Grenzen, weil die Suchbegriffe inhaltlich nicht verknüpft werden.

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

Der Einbau dieser »Intelligenz«, d. h. das Erahnen des Rechercheziels einer Anfrage (Worauf will der Nutzer hinaus?), ist Inhalt von Forschungsarbeiten an der Justus-Liebig-Universität Gießen: Das Team um Privatdozent Dr. Simon Hölzer hat mit LuMriX eine Suchtechnologie entwickelt, der es gelingt, mehrere Suchbegriffe (Begriffsmoleküle) zu sinnvollen Themen zu kombinieren. Dabei wird die Struktur und die inhaltliche Bedeutung der Dokumente sowie Informationen zu deren thematischer Verknüpfung (Semantik) genutzt. LuMriX durchsucht so genannte Themen-Netze, die mit dem ISO-Standard »Tropic Maps« repräsentiert und aus beliebigen Dokumenten (PDF, HTML, XML, RDF etc.) aufgebaut werden.

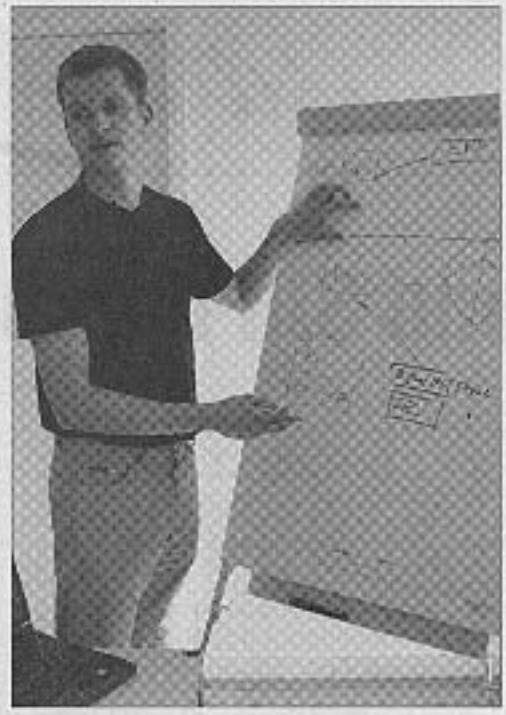
Ein Beispiel: Ein Nutzer gibt die Suchbegriffe »Autokauf Volkswagen Golf« ein und definiert damit ein für jedermann verständliches Suchziel. Für die Maschine bleibt jedoch die sinnvolle thematische Verknüpfung (Interesse an dem Kauf eines Golf der Marke VW) primär verborgen. Elektronisch wird nach dem Vorkommen von Einzelbegriffen und deren Kombination in verfügbaren Texten gesucht. Google findet mehr als 100 000 zumeist irrelevante Seiten. Dagegen erfolgt beim Ansatz von LuMriX eine sprachliche (Kauf und Auto) und thematische Auflösung der Suchanfrage, die von folgendem Themen-Netz ausgeht: Volkswagen = Automarke, Auto = PKW und Golf = Produkt der Firma Volkswagen. Dieses Themen-Netz definiert die Nähe und Art der Beziehung einzelner Begriffe. Der Begriff »Golf« steht damit in Beziehung zum Begriff »Auto«, »Auto« wird synonym zu »PKW« gebraucht, und gleichzeitig ist festgelegt, dass VW nicht direkt etwas mit Golfsport zu tun hat. Auf diese Weise erfolgt eine komplett andere Auswahl, Gewichtung und Sortierreihenfolge der Suchtreffer, ohne Einschränkungen bei der Vollständigkeit zu erleiden. Die Suche bleibt trotzdem einfach und intuitiv zu bedienen, weil der Nutzer selbst gewählte Stichwörter eingeben kann und trotz Tippfehler, Umlauten, zusammengesetzten Begriffen, Abkürzungen und anderer Schreibweisen fündig wird.

Grundlage von LuMriX ist die so genannte eXtensible Markup Language (XML), eine standardisierte Sprache zur Beschreibung und Strukturierung von Dokumenten für Datenhaltung und Datenaustausch im Internet. Mit XML können in bisher freien und unstrukturierten Texten einzelne Themen und Inhalte ausgezeichnet werden.

Die LuMriX-Suchmaschine, deren Name sich aus den Internet-Standards XML und URI (»Uniform Resource Identifier« entspricht der Internetadressinformation) zusammensetzt, wurde am Gießener Institut für Medizinische Informatik entwickelt. Als Ergebnis einer fünfjährige Entwicklungs- und Testphase konnte das Team um die Forschungsleiter Diplominformatiker Dr. Ralf Schweiger und Privatdozent Dr. Simon Hölzer zeigen, dass mittels XML mehrere Suchparameter wie die Präzision, die Vollständigkeit, die Toleranz und die Geschwindigkeit der Suche gleichzeitig optimiert werden können. Die Einsatzgebiete erstrecken sich derzeit schwerpunktmäßig auf die Informationsrecherche im Bereich der Medizin, der Rechtswissenschaften und des Bibliothekswesen.

Viele Themen-Netze sind bereits für spezifische Anwendungsbereiche definiert bzw. können halbautomatisch erstellt und gepflegt werden. Gleichwohl seien noch einige Anstrengungen notwendig, um eine nachhaltige Verbesserung der Suche in elektronischen Medien zu schaffen, berichtet Dr. Hölzer.

Auf diesem Weg zu einem »Semantischen Web« kommt der Strukturierung, Verschlagwortung und themenbezogenen Verknüpfung eine große Bedeutung zu. Dies erfordert eine erweiterte »Kultur« im Umgang mit elektronischen Medien, die die Unterstützung des gesamten Lebenszyklus eines elektronischen Dokumentes mit einbezieht (Information Lifecycle Management). Die Erfahrungen an der Universität Gießen zeigen, dass insbesondere die interdisziplinäre Zusammenarbeit zwischen Informatikern und fachgebietsspezifischen Experten (z. B. Mediziner, Dokumentare, Apotheker und Bibliothekare für den medizinischen Bereich) den zusätzlichen Mehrwert dieser Anwendungen ausmachen. Gleichzeitig findet hier ein erfolgreicher Wissenstransfer zwischen Universität, Public Domain und industrieller Anwendung statt. (-/pm)



Komplizierte Geschichte, ein vereinfachendes Recherchesystem zu entwickeln: Dr. Simon Hölzer und seinen Kollegen von der Uni Gießen ist dies mit LuMriX gelungen.

Wer das Recherchieren mit LuMriX ausprobieren möchte, kann dies in der elektronischen Bibliothek der JLU tun:
http/geb.uni-giessen.de/geb/
Weitere Infos: http://www.lumrix.net