Symposium „Wissenschaft und Praxis im Austausch über aktuelle Herausforderungen 2015“

https://cuvillier.de/de/shop/publications/7013

Prof. Dr. Judith Winter*

Forschungsprojekt SPIRIX –
Suche in Verteilten Systemen

1. Motivation und Hintergrund

Individuelles und vor allem kollektives Wissen ist Grundlage unserer heutigen Wissensgesellschaft, in der die Menge digital verfügbarer Informationen immer noch exponentiell steigt. Allerdings stellt ihre Größe und Komplexität den Zugriff und die Verarbeitung von Informationen häufig vor erhebliche Schwierigkeiten. Das Ende dieser Informationsflut nicht abzusehen, Experten sagen im Gegenteil eine Zunahme sowohl ihres Umfangs als auch ihrer Vielfalt voraus. Es gibt wohl kaum eine Information, die nicht irgendwo vorhanden wäre – die Kunst liegt nun darin, diese auch zu finden. Umso wichtiger sind adäquate Methoden, um sehr große Dokumentkollektionen durchsuchen zu können [6]. Im Gegensatz zur exakten Suche, bei der nach Dokumenten mit bekannten Dateinamen gesucht wird, werden Techniken des Information Retrieval (IR) dazu eingesetzt, relevante Ergebnisse zu einer Anfrage ausfindig zu machen [4]. Seit einiger Zeit werden verstärkt Kollektionen mit strukturierten Dokumenten durchsucht, insbesondere seit Durchsetzung der eXtensible Markup Language (XML) als offizieller Standard des World Wide Web Consortiums (W3C) [17]. Diese Strukturierung kann wertvolle Hinweise für die Suche geben, da sie den eigentlichen Textinhalt um Metainformationen wie die Gliederung des Textes, Diskriminierung der verschiedenen Textpassagen bis hin zu semantischen Informationen über den Text anreichert. Mittlerweile gibt es – initial vorangetrieben durch die internationale Forschungsinitiative INEX zur Evaluierung von XML-Suchmaschinen – eine ganze Reihe von Forschungsansätzen, bei denen IR-Methoden auf XML-Dokumente angewendet werden. XML Information Retrieval (XML-Retrieval) nutzt dabei die Struktur der Dokumente, um die Qualität der Suchergebnisse zu steigern [18].

2. State-of-the-art und Grundlagen

Die bisherigen Lösungen für XML-Retrieval beziehen sich jedoch alle auf zentralisierte Stand-Alone Suchmaschinen zu Forschungszwecken. Sehr große, über eine Vielzahl von Rechnern verteilte Datenkollektionen lassen sich damit nicht durchsuchen, was aber gerade für die Praxisrelevanz von Suchsystemen eine große Rolle spielt. Techniken für verteiltes XML-IR werden in der Praxis auch dort benötigt, wo das zu durchsuchende System aus einer Vielzahl lokaler, hete-

* Die Verfasserin ist Professorin für Wirtschaftsinformatik am FB3

rogener XML-Kollektionen besteht, deren Benutzer ihre Dokumente nicht auf einem zentralen Server speichern wollen oder können; solche Benutzer schließen sich häufig in Form eines dezentralen Peer-to-Peer (P2P) Netzes zusammen [1]. Dies trifft beispielsweise auf Benutzer zu, die viel Wert auf Privatsphäre legen oder die die Möglichkeit von Zensur und Manipulation fürchten. Datenschützer sehen insbesondere die große Anzahl Informationen sehr kritisch, die ein Suchender unfreiwillig über sich selbst preisgibt und die von zentralisierten Suchmaschinen gesammelt und ausgewertet werden. Außerdem ist die Möglichkeit, zentral gespeicherte Informationen zu zensieren, immer wieder ein umstrittenes Thema und führt zu einer verstärkten Anwendung von P2P-Architekturen für Dokumentkollektionen [10]. Auch im Zuge des NSA-Skandals, als der Öffentlichkeit bekannt wurde, in welchem Ausmaß Behörden wie die US-amerikanische NSA Daten sammeln, zu denen auch im großen Stil die Suchanfragen gehören, die über zentrale Server von Google & Co laufen, war der Ruf nach alternativen Suchmaschinen in den Medien immer wieder Thema, P2P-Suchmaschinen erleben seit 2013 einen neuen Aufschwung [5; 9].

P2P-Suchmaschinen bestehen aus dem Zusammenschluss einer teilweise sehr großen Menge gleichberechtigter und autonomer Rechner, den sogenannten Peers. Diese teilen Ressourcen wie Rechen- und Speicherkapazität miteinander, und zwar selbstorganisiert ohne zentrale Kontrolle oder zentralen Index [8]. Diese Selbstorganisation ermöglicht ein dynamisches Anpassen an die jeweils teilnehmenden Peers, so dass ein hoher Grad an Fehlertoleranz und Robustheit ohne Eingriff von außen realisiert werden kann. Ein P2P-Netz kann zu einer – zumindest theoretisch – unbegrenzten Anzahl von teilnehmenden Peers skalieren, deren andernfalls u.U. ungenutzten Ressourcen zu einem leistungsstarken System zusammengefasst werden können [7].

Während in der Praxis bisher lediglich die exakte Suche in P2P-Systemen unterstützt wird, gibt es viele Forschungsansätzen, bei denen P2P-IR-Methoden verwendet werden. Diese umfassen Lösungen, die sich auf das Retrieval von Textdokumenten spezialisieren, sowie Ansätze für das Auffinden von Bildern, Videos und Musikdateien. Bisher nutzt jedoch keine der entwickelten P2P-Techniken die Möglichkeit, speziell nach XML-Dokumenten zu suchen. Das Potential von XML-Retrieval Techniken, das in zentralisierten XML-Suchmaschinen zur Steigerung der Suchqualität bereits erfolgreich verwendet werden kann, bleibt bei P2P-Ansätzen bislang unberücksichtigt [16].

Unerforscht ist auch die Möglichkeit, XML-Struktur zur Steigerung der Effizienz von P2P-Suchmaschinen auszunutzen, obwohl eine der Hauptschwierigkeiten bei der verteilten Suche gerade die Skalierbarkeit des Systems ist, also die Gewährleistung einer effizienten Anfragebeantwortung auch bei zunehmender Größe des Systems. In der Praxis scheitern P2P-Suchmaschinen üb-

licherweise daran, dass sie im Vergleich mit zentralisierten Lösungen nicht performant genug sind. Dies liegt in dem hohen Kommunikationsaufwand begründet, der bei der Lokalisierung und dem Zugriff auf die zur Anfragebeantwortung notwendigen, aber verteilten Informationen anfällt [3]. Um die Netzlast zu reduzieren und Skalierbarkeit zu garantieren, benutzen P2P-Suchmaschinen daher i.A. nur eine begrenzte, ausgewählte Menge an Informationen. Dies allerdings wirkt sich negativ auf die Suchqualität aus. Techniken, die XML-Strukturinformationen zur Auswahl adäquater Informationen verwenden und somit zu Effizienzsteigerung beitragen können, wurden bisher noch nicht entwickelt.

3. Zielsetzung des Forschungsprojekts SPIRIX

Es wird daher im Rahmen des Forschungsprojektes SPIRIX am Beispiel von P2P-Netzen u.a. untersucht, inwiefern XML-Retrieval in verteilten Systemen effektiv und effizient möglich ist, d.h. inwiefern Dokumentstrukturen dazu genutzt werden können, die verteilte Suche nach XML-Dokumenten effizient in Bezug auf Ressourcen- und Bandbreitenverbrauch zu gestalten und dabei effektiv umzusetzen, also möglichst viele hochrelevante Dokumente aufzu-finden.

Ziel ist dabei die Ausnutzung von Strukturinformationen für eine signifikante Steigerung der Präzision der Suchergebnisse sowie für eine Reduktion des dazu nötigen Kommunikationsaufwands, so dass das System auch zu einer großen Anzahl von teilnehmenden Peers skaliert. Gerade der Fokus auf die Ressourcenökonomisierung von Suchsystemen soll ein Beitrag zur Performanz von dezentral verteilten Suchsystemen sein, so dass diese in der Praxis von den Benutzern tatsächlich effektiv und effizient eingesetzt werden können.

4. Die Suchmaschine SPIRIX

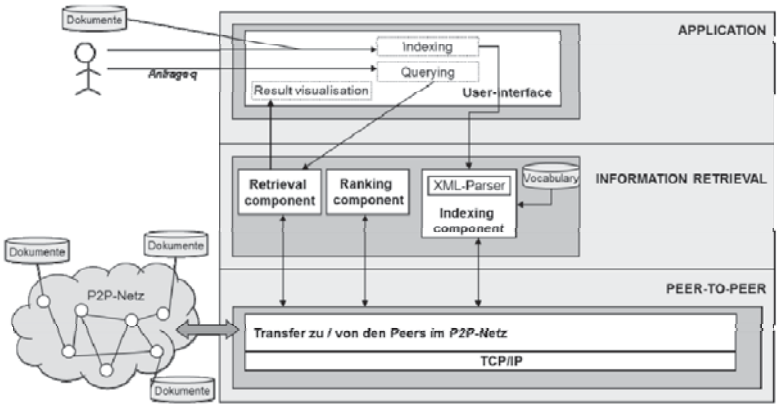

Abbildung 1: Architektur der Suchmaschine SPIRIX

Grundlage der Forschungsarbeiten ist die Entwicklung und Java-Implementierung einer dezentralen Suchmaschine, in die diverse XML-Retrieval Techniken integriert sind. Diese Suchmaschine namens SPIRIX „Search Engine for P2P Information Retrieval of XML-documents“ ist Namensgeberin des beschriebenen Forschungsprojekts [16].

Die Architektur (s. Abb.1 und Abb.2) von SPIRIX stellt sich wie folgt dar: Abfrageerstellung sowie Ergebnisdarstellung erfolgen über die Applikationsschicht, während das Bewerten von Suchergebnissen (Ranking und Retrieval) durch die Information Retrieval Schicht bewerkstelligt werden, wo auch die Indizierung verfügbarer Dokumentkollektionen stattfindet. Die Kommunikation zwischen teilnehmenden Rechnern (Peers) sowie das Auffinden nützlicher Informationen auf diesen findet auf der Peer-to-Peer Ebene statt.

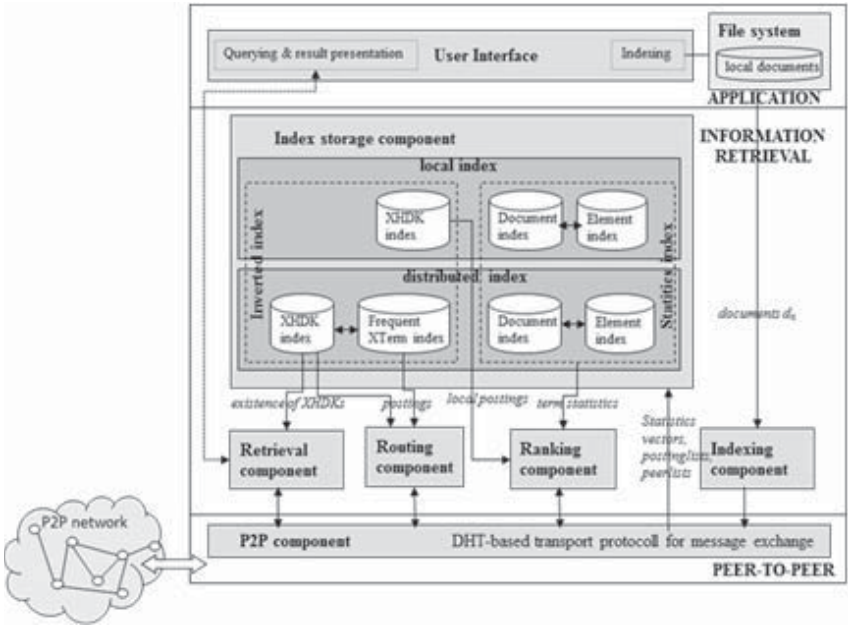

Abbildung 2: Informationsfluss innerhalb der Suchmaschine

5. Forschungsuntersuchungen und -ergebnisse

Im Rahmen des Projekts SPIRIX werden Strategien zur Ressourcen-schonenden Bewältigung von XML-Anfragen untersucht, wobei insbesondere diejenigen XML-Retrieval Techniken in SPIRIX evaluiert werden, die der Steigerung der Effizienz (Bandbreite) sowie der Effektivität (Suchqualität) des Systems dienen. Einer der Schwerpunkte ist des Weiteren die Suche nach XML-Dokumenten in sehr großen Netzen. Evaluiert werden dabei diverse Strategien auf Basis der INEX-Kollektion, einer Dokumenten-Sammlung im XML-Format der internationalen Forschungsgemeinde im Bereich XML-Retrieval. SPIRIX ist weltweit eines der wenigen XML-Retrieval-Systeme, die für die Suche in P2P-

Systemen konzipiert wurden und trotzdem mit der Suchqualität zentralisierter Suchsysteme mithalten kann, wie mehrere Teilnahmen am jährlichen INEX-Vergleich zeigten [11].

Auf Basis der Suchmaschine SPIRIX konnten bereits essenzielle Erkenntnisse im Bereich des dezentralen XML-Retrievals erarbeitet werden. Insbesondere der Einfluss diverser XML-Retrieval Maßnahmen zur Steigerung der Suchqualität auf die Effizienz des Systems (z.B. Bandbreiten- und Ressourcenverbrauch) u.u. konnte quantifiziert und durch Verwendung der INEX-Kollektion und international anerkannter Evaluierungsverfahren nachgewiesen werden. Hinsichtlich einer Ressourcenökonomisierung wurden beispielsweise Verfahren entwickelt und erfolgreich angewendet, um die Anzahl notwendiger Nachrichten zwischen verteilten Rechnern zu reduzieren, ohne die Suchqualität wesentlich zu beeinträchtigen. Eine Steigerung der Suchqualität, u.a. durch Einbindung von Strukturvergleichsfunktionen, wird angestrebt und wurde auch schon in einzelnen Strategien erzielt. (vgl. [10]-[16]).

Da das Thema Datenschutz und Datensicherheit, nicht zuletzt durch die jüngsten NSA-Skandale, sehr aktuell ist, werden im Rahmen von SPIRIX auch die gesellschaftlichen Aspekte betrachtet, die zur Verwendung von Peer-to-Peer Systemen führen und beispielsweise durch Wertschätzung der eigenen Privatsphäre motiviert sind oder durch Benutzer erfolgt, die die Möglichkeit von Zensur und Manipulation, beispielsweise in repressiven Staaten, fürchten. Datenschützer sehen beispielsweise insbesondere die große Anzahl Informationen sehr kritisch, die ein Suchender unfreiwillig über sich selbst preisgibt und die von zentralisierten Suchmaschinen gesammelt und ausgewertet werden. Auch die Möglichkeit, zentral gespeicherte Informationen zu zensieren, ist immer wieder ein umstrittenes Thema und führt zu einer verstärkten Anwendung von P2P-Architekturen für Dokumentkollektionen. Im Zuge des NSA-Skandal ist außerdem das Thema Wirtschaftsspionage im Zusammenhang mit Suchmaschinen aufgekommen – nicht für Dritte bestimmte Daten gibt ein Suchender u.U. nicht nur über sich als Person, sondern auch über sein Unternehmen preis; vielen Mitarbeitern mangelt es jedoch an entsprechender Sensibilität in Hinsicht auf die Auswertbarkeit zentraler Suchmaschinenlogs. Im Rahmen des Projekts SPIRIX wurden bereits einige Betrachtungen zu den Risiken und Chancen von Big Data in Bezug auf Suchmaschinen durchgeführt ([10;14]), weitere werden angestrebt.

Literaturverzeichnis

[1] ABERER, K. / HAUSWIRTH, M. 2002: Peer-to-peer information systems: concepts and models, state-of-the-art, and future systems. *In: IEEE 18th Int. Conference on Data Engineering (ICDE'02).*

[2] AMER-YAHIA, S. / LALMAS, M. 2006: XML Search: Languages, INEX and Scoring. *In: SIGMOD RecVol. 35, No. 4*.

[3] BAEZA-YATES, R.; CASTILLO, C.; JUNQUEIRA, F.; PLACHOURAS, VASSILLIS; SILVESTRI, F. 2007: Challenges on Distributed Web Retrieval. *IEEE Int. Conf. on Data Engineering (ICDE'07), Istanbul, Turkey.*

[4] BAEZA-YATES, R. / RIBEIRO-NETO, B. 2010: Modern Information Retrieval: The Concepts and Technology Behind Search. *Addison-Wesley Educational Publishers Inc, New Jersey, USA*.

[5] FRANKFURTER RUNDSCHAU (AFP) 2013. Sichere Suchmaschinen gefragt – Seit Prism haben Alternativen zu Google und Yahoo Hochkonjunktur. *FR Nr. 143, S. 3, 24.06.2013.*

[6] MANNING, C.; RAGHAVAN, P.; SCHÜTZE, H. 2008: Introduction to Information Retrieval. *Cambridge University Press.*

[7] STEINMETZ, R. / WEHRLE, K. (EDS.) 2005: Peer-to-Peer Systems and Applications. *Lecture Notes in Computer Science No. 3485, Springer-Verlag, Berlin-Heidelberg.*

[8] RISSON, J. / MOORS, T. 2004: Survey of research towards robust peer-to-peer networks – search methods. *Technical Report UNSW-EE-P2P-1-1, University of New South Wales, Australia.*

[9] WESTDEUTSCHE ZEITUNG / NAGEL, T. 2013: "DuckDuckGo": Nutzeransturm auf anonyme Suchmaschine dank "Prism". *http://www.wz-newsline.de/home/multimedia/specials/teststrecke-der-technik-ratgeber/duckduckgo-nutzeransturm-auf-anonyme-suchmaschine-dank-prism-1.1349474, 20.06.2013.*

[10] WINTER, JUDITH / CHRISTEN, MICHAEL 2014: Zentrale Suchmaschinen - Die neuen Bottlenecks des Internetzeitalters?. *In: Krah, H.; Müller-Terpitz, R.: "Suchmaschinen"; Passauer Schriften zur interdisziplinären Medienforschung, Bd. 4, Passau*

[11] WINTER, JUDITH / KÜHNE, GEROLD 2010: Achieving High Precision With Peer-to-Peer Is Possible! *In: Focused Retrieval. Lecture Notes in Computer Science (LNSC), Springer-Verlag, Berlin-Heidelberg*.

[12] WINTER, JUDITH 2011: Democratisation of Digital Search by Decentralisation. *In: Proc. of ACM 3rd International Conference on Web Science (WebSci'11), Koblenz, GERMANY.*

[13] WINTER, JUDITH 2011: Supporting Search Engines by Exploitation of Semi-Structured Online Sources. *In: Proc. of Workshop on Knowledge Extraction and Exploitation from semi-Structured Online Sources (KEESOS 2011) at XIV Conference of the Spanish Association for Artificial Intelligence (CAEPIA 2011), St. Cristobal, Spain*.

[14] WINTER, J. / CHRISTEN, M. 2013: Zentrale Suchmaschinen: notwendige Werkzeuge für effektiven Informationszugriff oder effektive Torwächter zwischen Benutzer und freier Wissensgesellschaft?, *In: "Suchmaschinen - die neuen Bottlenecks des Internetzeitalters?", IIM, Passau.*

[15] WINTER, J.; SAILER, J. 2012: Dezentrale Suche als Beitrag zur demokratischen Wissensgesellschaft?, *in Proc. of IR-2012, Workshop on Information Retrieval 2012, at LWA'12 (Lernen, Wissen, Adaption), Dortmund.*

[16] WINTER, JUDITH 2011: An Approach to XML Information Retrieval in Distributed Systems. *it-Information Technology, Jahrgang 53 (2011) Heft 4, DOI 10.1524/itit.2011.0645, Oldenbourg Wissenschaftsverlag, München.*

[17] W3C 2008: Extensible Markup Language (XML) 1.0 (Fifth Edition). *W3C Recommendation.*

[18] WALSH, D. ET AL: Overview of INEX 2014. *In: Information Access Evaluation. Multilinguality, Multimodality, and Interaction, 5th International Conference of the CLEF Initiative, CLEF 2014, Sheffield, UK.*