XML Information Retrieval in Verteilten Systemen

Judith Winter

Institut für Informatik
J. W. Goethe-Universität
Robert-Mayer-Str. 11-15
60325 Frankfurt
winter@tm.informatik.uni-frankfurt.de

Abstract: Um die stetig wachsende Menge an Informationen unserer modernen Wissensgesellschaft zu durchsuchen, verwenden moderne Suchmaschinen erfolgreich Techniken des Information Retrievals (IR), um die zu einer Anfrage *relevanten* Dokumente ausfindig zu machen. Allerdings häuft sich die Kritik an Systemen wie Google, deren zentrale Architektur Möglichkeiten zur Auswertung privater Informationen über die Suchenden oder zu Zensur bietet. Zunehmende Verbreitung finden daher Suchsysteme, die auf selbstorganisierenden Peer-to-Peer (P2P) Architekturen basieren und gänzlich ohne zentrale Kontrollinstanz oder zentralen Index auskommen. Bisherigen Ansätzen mangelt es jedoch häufig an Effektivität (Suchqualität) oder Effizienz (sparsamer Ressourcenverbrauch). In der vorliegenden Arbeit wird erstmalig eine verteilte Suchmaschine entworfen, die diese Problematik durch den Einsatz von XML-Retrieval Techniken angeht und Strukturinformationen zu durchsuchender XML-Dokumentkollektionen ausnutzt. Das entwickelte System basiert auf einem Zusammenschluss gleichberechtigter und autonomer Peers und ist die weltweit erste Lösung für verteiltes XML-Retrieval. Untersucht wird, wie die verteilte Suche nach strukturierten Dokumenten in einem P2P-Netz ohne zentralen Index sowohl effizient als auch effektiv durchgeführt werden kann. Dabei kann bei skalierendem Kommunikationsaufwand und Ressourcenverbrauch eine Suchqualität nachgewiesen werden, die im Bereich der besten 10 XML-Retrieval Systeme weltweit liegt und somit dem internationalen Vergleich mit zentralisierten Lösungen standhält. Die Arbeit ist damit ein Beitrag dazu, P2P-Suchmaschinen zukünftig so performant zu gestalten, dass sie in der Praxis von den Benutzern tatsächlich eingesetzt werden können.

1 Einleitung – Motivation und Zielsetzung

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 [St07]. 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. 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. Seit einigen Jahren werden verstärkt Kollektionen mit strukturierten Dokumenten durchsucht, insbesondere seit Durchsetzung der eXtensible Markup Language (XML) als offizieller Standard des W3Cs [St07]. 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 – stark vorangetrieben seit 2002 durch die internationale Forschungsinitiative INEX zur Evaluierung von XML-Suchmaschinen – eine 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. Die gängigsten Ansätze sind die Fokussierung auf besonders relevante Dokumentteile (*Element-Retrieval*), die Benutzung von Strukturhinweisen der Benutzer– sogenannte *Content-And-Structure (CAS) Anfragen* – oder die unterschiedliche Gewichtung der einzelnen Textpassagen, also beispielsweise die Diskriminierung eines Fußnoten- gegenüber eines Titel-Elements [TGK09].

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 [WK10]. Techniken für verteiltes XML-IR werden in der Praxis auch dort benötigt, wo das zu durchsuchende System aus einer Vielzahl lokaler, heterogener 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. 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. 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 [Su04].

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. 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 [SW05].

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 [SW05]. 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.

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 üblicherweise 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. 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 [LLH03]. Techniken, die XML-Strukturinformationen zur Auswahl adäquater Informationen verwenden und somit zu Effizienzsteigerung beitragen können, wurden bisher noch nicht entwickelt.

In der vorliegenden Arbeit wird am Beispiel von P2P-Netzen erstmalig untersucht, inwiefern XML-Retrieval in verteilten Systemen effektiv und effizient möglich ist. Dazu wird eine P2P-Suchmaschine entwickelt, um zu zeigen, inwiefern XML-Struktur dazu genutzt werden kann, die verteilte Suche nach XML-Dokumenten effizient in Bezug auf Ressourcen- und Bandbreitenverbrauch zu gestalten und dabei effektiv umzusetzen, d.h. möglichst viele hochrelevante Dokumente aufzufinden. 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. Dies wäre ein wesentlicher Schritt dazu, P2P-Suchmaschinen zukünftig so performant gestalten zu können, dass sie in der Praxis von den Benutzern tatsächlich effektiv und effizient eingesetzt werden.

2 Architektur einer P2P-Suchmaschine für XML-Retrieval

Im Rahmen der Arbeit wurde eine konkrete P2P-Suchmaschine für XML-Retrieval entwickelt. Diese umfasst verschiedene Techniken des verteilten XML-Retrievals für Indizierung, Routing und Ranking. Insbesondere die aufwendige Beantwortung von aus mehreren Termen bestehenden Multitermanfragen sowie Verteilungsaspekte werden berücksichtigt. Neben zu erzielender Suchqualität steht der notwendige Kommunikationsaufwand im Vordergrund, alle Methoden wurden so entworfen, diesen möglichst gering zu halten.

Die Architektur jedes Peers des entwickelten Systems ist in Abbildung 1 dargestellt. Basis aller Relevanzberechnung sind die während der Indizierung der Kollektion

gesammelten Informationen, von denen jeder Peer einen im zugewiesenen Anteil verwaltet. Der *Dokumentindex* enthält dabei Dokumentstatistiken wie Termfrequenzen, also die Häufigkeiten der in den Dokumenten vorkommenden Terme. Der *Elementindex* speichert Statistiken über Elemente, die somit analog zu Dokumenten als Ergebnis dienen können. Und der *invertierte Index* beinhaltet für alle in der Gesamtkollektion vorkommenden Terme eine sogenannte *Postingliste*; jedes *Posting* dieser Liste referenziert ein Dokument, in denen der entsprechende Indexterm vorkommt, und enthält statistische Zusatzinformationen über die Bedeutung des Terms innerhalb des referenzierten Dokuments. Je Term wird auch seine XML-Struktur gespeichert, so dass zu jedem Term diverse Postinglisten verfügbar sind, nämlich je unterschiedlicher XML-Struktur eine eigene Liste. Dies ermöglicht beim Retrieval schnellen und effizienten Zugriff auf genau diejenigen Postings, die zu einer gewünschten Struktur passen. Die Kombination aus Indexterm und seiner XML-Struktur sei als *XTerm* bezeichnet.

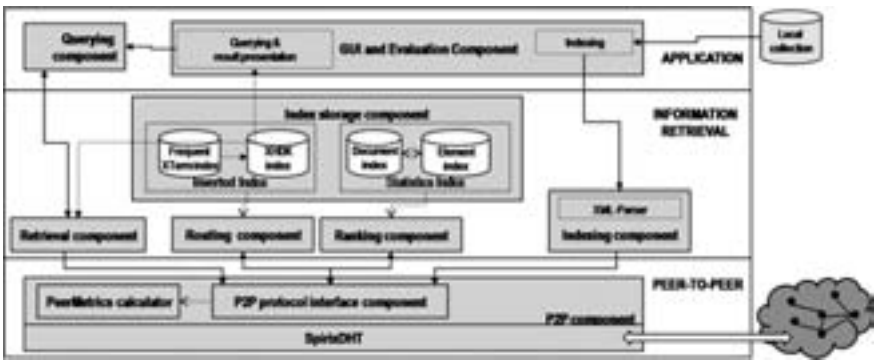

Abbildung 1: Architektur eines Peers

Alle drei Indizes sind über die Menge dem System teilnehmender Peers verteilt, wobei der jeweils für eine Informationseinheit zuständige Peer durch Anwendung einer Hash-abbildung auf den Schlüssel der Informationseinheit bestimmt wird. Für zu verteilende Postinglisten ist der entsprechende Schlüssel beispielsweise der Indexterm der Posting-liste sowie dessen XML-Struktur.

Lokalisiert werden die Daten über die Komponenten der P2P-Schicht, die auf dem P2P-Protokoll Chord [SML03] basiert und speziell für den Einsatz von XML-Retrieval angepasst wurde.

3 Routing strukturierter Anfragen

Die Beantwortung einer Anfrage auf Basis der beschriebenen Architektur besteht aus den nachfolgenden Routingschritten und dem Ranking selbst. Das Routing geht in drei Schritte vonstatten, für die jeweils *Routingnachrichten* zwischen den Peers ausgetauscht werden müssen.

Zunächst wird die Anfrage vom anfragenden Peer an diejenigen Peers weitergeleitet werden, die für die Postingliste der einzelnen Anfrageterme zuständig sind. Dazu müssen die passenden Postinglisten im Netz lokalisiert werden. Mit Chord-basierten Methoden kann dies effizient in *log(n)* Schritten durchgeführt werden [SML03].

Im zweiten Schritt des Routings werden aus den lokalisierten Postinglisten eines Anfrageterms passende Postings ausgewählt und mit den ausgewählten Postings der übrigen Anfrageterme abgeglichen, indem Teile der Postinglisten (z.B. die jeweils am höchsten gewichteten 50 Postings) von Peer zu Peer geschickt werden. Es können hierbei aus Effizienzgründen nicht alle Postings ausgewählt werden, da das System ansonsten bei ansteigender Anzahl von Dokumenten und entsprechendem Anstieg der Postinglistenlängen nicht skaliert – der Aufwand für das Übertragen kompletter Postinglisten zwischen den Peers zum Abgleichen der Postings wäre zu hoch. Es ist daher essentiell, dass geeignete Postings ausgewählt werden. Hier können XML-Retrieval Techniken zum Einsatz kommen.

Für die endgültig ausgewählten Postings müssen in einem dritten Schritt diejenigen Peers identifiziert und lokalisiert werden, die die Statistiken der Dokumente und der Elemente speichern, die von den Postings referenziert werden. Diese Statistiken sind nicht direkt in den Postinglisten abgelegt, da zu jedem Dokument eine Vielzahl von Elementen gehört, deren Statistiken nicht redundant in jeder Postingliste eines Dokuments gespeichert werden sollen. Die Anfrage wird daher zu allen Peers weitergeleitet, die die entsprechenden Statistiken speichern. Im worst case, wenn alle Statistiken auf verschiedenen Peers abgelegt sind, fällt je ausgewähltem Posting eine Weiterleitung an, die als *Rankingnachricht* bezeichnet sei.

Jeder Peer, der eine solche Rankingnachricht erhält, führt anhand seiner lokal gespeicherten Statistiken eine Relevanzberechnung für die ausgewählten Dokumente und Elemente durch, wobei gängige XML-Retrieval Techniken verwendet werden. Dies geschieht für jedes ausgewählte Posting, auf allen benachrichtigen Peers gleichzeitig.

Abbildung 2 stellt die Anfragebeantwortung einer einfachen Single-Term-Anfrage *q=(apple, \\p)* dar, wobei *q* ein einzelner XTerm ist. Dabei werde *q* durch den anfragenden Peer p_0 in einem System gestellt, das aus 11 Peers besteht, die auf einem Chord-Ring angeordnet seien. Die Peers bilden einen logischen Ring, der effizientes Lookup von Informationen mit Chord-Methoden in *log(n)* Schritten ermöglicht [SML03].

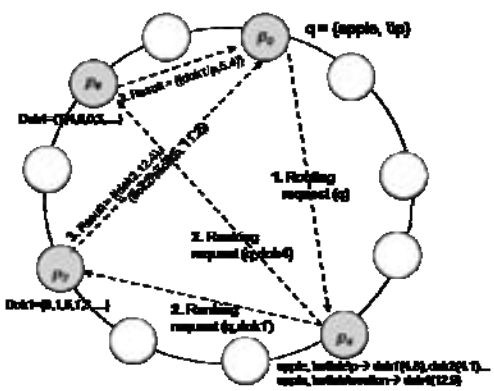

Abbildung 2: Anfragebeantwortung der Anfrage q = {apple,\\p}

Gesucht wird nach Absätzen (*p* steht für Paragraph) in Artikeln, deren Inhalt relevant zu dem Suchterm *apple* ist. Die Strukturangabe *p* wird dabei, analog zur inhaltsbasierten Suche mit IR-Methoden, als *vager* Hinweis des Benutzers gesehen, welche Strukturen von besonderem Interesse sind. Als relevante Ergebnisse werden daher auch solche angesehen, deren Struktur nicht exakt mit der angegebenen übereinstimmt.

Im Retrievalprozess wird q zunächst zu Peer p_4 weitergeleitet, der dem Hashwert von *apple* zugeordnet ist und somit alle Postinglisten von XTermen speichert, die den Indexterm *apple* enthalten. Peer p_4 wählt bei Eintreffen von q passende invertierte Listen aus, diese seien die für die XTerme (*apple, \article\p*) und (*apple, \article\section*), da eine starke Ähnlichkeit zwischen Absätzen (*\p*) und Sektionen (*\section*) angenommen wird. Aus den selektierten Postinglisten werden geeignete Postings ausgewählt, indem die Gewichte der Postings mit der Ähnlichkeit zwischen der Struktur des XTerms und dem Strukturhinweis der CAS-Anfrage q, nämlich *\\p*, multipliziert werden. Dies begünstigt alle Postings aus Listen von XTermen, deren Struktur ähnlich zu *\\p* ist. Im vorliegenden Fall werden die Postings mit Dokumentreferenzen auf *dok1* und auf *dok4* ausgewählt. Daraufhin wird die Anfrage weitergeleitet zu Peer p_7 und p_9, die die entsprechenden Dokument- und Elementstatistiken der ausgewählten Postings gespeichert haben. Beide Peers erhalten die Anfrage q, berechnen relevante Ergebnisse und senden diese zurück an den anfragenden Peer p_0. Als Ergebnis wurden dabei auch Elemente aus den beiden ausgewählten Dokumenten errechnet.

Somit umfasst die Liste der Ergebnisse sowohl das Dokument *dok2* als auch eine Sektion aus diesem Dokument (*dok2/section*) sowie einen Absatz aus Dokument1 (*dok1/p*).

Eine ausführliche Darlegung der entwickelten Algorithmen zur Auswahl adäquater Postings erfolgt in [Wi09]. Die wesentlichen Aspekte bei sind nachfolgend aufgeführt:

- Einem XTerm ist im invertierten Index für jede seiner Strukturen genau eine Postingliste zugeordnet. Jedes Posting wird dabei mit einem Gewichtungsfaktor bewertet.
- Zur Unterstützung von Multitermanfragen können Kombinationen von XTermen, sogenannte XHDKs, gebildet werden. Dies kann zu einer erheblichen Reduktion der Netzlast führen, wenn das Abgleichen von Postinglisten zwischen verschiedenen Peers entfällt [WD08].
- Alle Postings werden mit IR-Gewichtungsformeln bewertet, dabei kommen insbesondere Adaptionen der Gewichtungsformel BM25 oder BM25E zum Einsatz [RZT04]. Diese Formeln wurden an die Verwendung von XTermen adaptiert, so dass beispielsweise unterschiedliche Elemente verschieden gewichtet werden, also die Terme innerhalb eines Titel-Elements stärker berücksichtigt werden können als die eines Fußnoten-Elements.
- Bei der Auswahl der Postings wird die Struktur der Indexterme mit den Strukturhinweisen verglichen, die der Benutzer beim Stellen von CAS-Anfragen geben kann. Indexterme, deren Struktur eine Mindestähnlichkeit zur vom Benutzer gewünschten Struktur aufweist, werden mit einer Erhöhung ihres Gewichts belohnt.
- In die Bewertung fließen nicht nur Statistiken ein, die aus den durch Postings repräsentierten Dokumenten stammen, sondern auch Statistiken über die Elemente der jeweiligen Dokumente. Die Bewertung eines Postings erhöht sich, wenn das im Posting referenzierte Dokument viele hochgewichtige Elemente erhält. Dieses Vorgehen reflektiert den Einsatz von Element-Retrieval beim Ranking, also die Möglichkeit, auch einzelne XML-Elemente als Ergebnis zu verwenden.
- Auch der Peer, der für die Relevanzberechnung des referenzierten Dokuments zuständig ist, erhält eine Bewertung. Dieser *PeerScore* setzt sich zusammen aus Bewertungen seiner Zuverlässigkeit (Online-Verhalten), der Qualität seiner Kollektion, seines Erfolgs bei der Beantwortung früherer Anfragen sowie seiner technischen Merkmale (z.B. CPU, RAM, Antwortzeiten, Firewall, Bandbreite).

Die entwickelten Routingalgorithmen basieren also auf einer Kombination von Techniken und Berechnungen aus dem klassischen IR, aus dem XML-Retrieval zur Ausnutzung von Strukturinformation sowie aus dem P2P-Bereich.

4 Implementierung der Suchmaschine SPIRIX

Um das entworfene Konzept einer P2P-Suchmaschine für verteilte XML-Retrieval in Bezug auf Effektivität und Effizienz evaluieren zu können, wurde es in Form der Suchmaschine SPIRIX implementiert [WD09]. Das Akronym SPIRIX steht für: **S**uchmaschine für **P**eer-to-Peer **I**nformation **R**etri**e**val in **X**ML-Dokumenten. SPIRIX ist eine voll-funktionsfähige Suchmaschine, die XML-Dokumente in einem P2P-Netz indizieren, finden und ihre Relevanz berechnen kann. SPIRIX verfügt sowohl über Schnittstellen zur Durchführung einer großen Anzahl automatischer Evaluierungsläufe, als auch über eine komfortable graphische Benutzungsoberfläche, mit der ungeübte Benutzer sich von der Funktionalität der entwickelten Methoden überzeugen können. Für die Kommunikation zwischen Peers wurde im Rahmen der Arbeit ein P2P-Protokoll namens SpirixDHT entworfen, das auf Basis von Chord arbeitet und an XML-Retrieval angepasst wurde. Die Anpassung betrifft beispielsweise die Auswahl der Schlüssel, auf die die Hashabbildung angewandt wird: statt zufälliger Verteilung der Daten über das Netz werden diejenigen Informationen zusammen auf dem gleichen Peer abgelegt, die in einer Anfrage mit einer gewissen Wahrscheinlichkeit zusammen benötigt werden. Weiterer Kommunikationsaufwand wird durch Bündelung von Nachrichten eingespart. SPIRIX besteht insgesamt aus fast 40.000 Zeilen Java-Code.

5 Evaluation von SPIRIX

SPIRIX wurde durch eine Serie von umfangreichen Experimentenreihen evaluiert. Dabei wurden die Testkollektion, die Evaluierungsmethoden und die Evaluierungsmaße der Forschungsinitiative INEX verwendet, da sich INEX als Plattform zur Evaluierung von Verfahren zum XML-Retrieval international durchgesetzt hat.

Als weltweit erste Lösung für verteiltes XML-Retrieval musste zunächst die Funktionsfähigkeit von SPIRIX nachgewiesen werden, insbesondere die Effektivität der Suchmaschine war zu evaluieren. Dies geschah im Schwerpunkt durch Teilnahme an INEX 2008, wo eine Präzision der Ergebnisse erreicht wurde, die im Bereich der besten 10 XML-Retrieval Systeme weltweit liegt – und dies im Vergleich mit zentralisierten Lösungen! Damit wurde nicht nur der Beweis erbracht, dass SPIRIX eine funktionierende XML-Suchmaschine ist, sondern sogar, dass eine P2P-Suchmaschine durchaus mit zentralisierten Systemen mithalten kann.

Auf der hohen erreichten Suchqualität basierend wurden die einzelnen entworfenen Methoden für verteiltes XML-Retrieval mit INEX-Werkzeugen evaluiert; zunächst waren dies verschiedene Ranking-Gewichtungsmethoden, Element-Retrieval und Berücksichtigung von CAS-Anfragen. Es wurden jeweils erzielte Suchqualität und Aufwand gegenübergestellt und eine signifikante Verbesserung der Präzision nachgewiesen.

Die gewonnenen Erkenntnisse wurden auf den Routingprozess angewendet; hier war speziell die Fragestellung interessant, wie XML-Struktur zur Performanzverbesserung in Bezug auf die Effizienz eines verteilten Systems genutzt werden kann. Die untersuchten Methoden konnten zu einer signifikanten Reduzierung der Anzahl versendeter Nachrichten und somit der Netzlast verwendet werden, wobei gleichzeitig eine Steigerung der Suchqualität erreicht wurde. Abbildung 3 zeigt beispielsweise, dass die Suchqualität beim Einsatz der entwickelten XML-Retrieval basierten Auswahltechniken (AdvancedRouting) bei bis zu 500 Nachrichten deutlich über der Baseline liegt. Angegeben ist jeweils die erzielte Präzision der 1% besten Ergebnisse, gemessen mit dem offiziellen INEX-Maß iP[0.01], in Relation zu der gemessenen Anzahl Routingnachrichten.

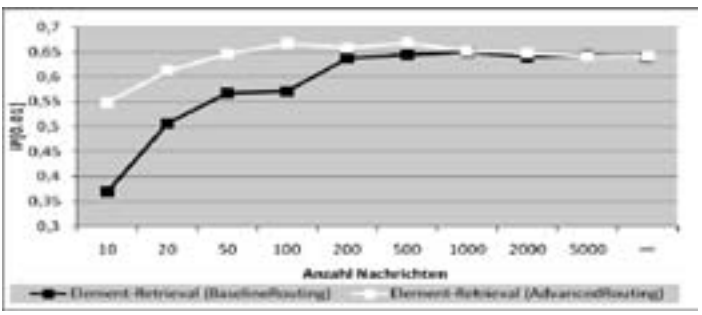

Abbildung 3: Präzision bei unterschiedlicher Anzahl Routingnachrichten

Das Routing-Verfahren wurde durch Methoden zur Unterstützung von in verteilten Systemen besonders problematischen Multitermanfragen erweitert, so dass der Effekt des Einsatzes von XTerm-basierten Schlüsselkombinationen (XHDKs) analysiert werden konnte. Es konnte demonstriert werden, wie die Anzahl kontaktierter Peers und der Kommunikationsaufwand erheblich reduziert werden können.

Grundlage der durchgeführten Experimente war jeweils ein kleines reales P2P-Netz mit bis zu 50 Peers oder ein simuliertes Netz mit bis zu 5000 kontaktierten Peers pro Anfrage. Der Nachrichtentransport in diesen Netzen (sowohl simuliert als auch tatsächlich stattfindend) geschah mittels des entwickelten P2P-Protokolls SpirixDHT. Für dessen Analyse wurden Evaluierungen durchgeführt, die die Einsparung von Kommunikationsaufwand durch Erkenntnisse aus den vorherigen Experimenten zum verteilten XML-Retrieval zeigen. Beispielsweise konnten Nachrichten, die in großer Anzahl während der Indizierung und dem Routing anfallen, gebündelt werden, um so die Anzahl der Nachrichten zu reduzieren.

6 Zusammenfassung und Ausblick

Im Rahmen der vorliegenden Arbeit wurde erstmals ein Forschungsansatz für verteiltes XML-Retrieval entwickelt. Dazu wurden verschiedene Techniken des verteilten XML-Retrieval entworfen, die sich auf Indizierung, Routing und Ranking beziehen.

Das Konzept wurde in Form von SPIRIX umgesetzt, einer P2P-Suchmaschine für verteiltes XML-Retrieval, die aus fast 40000 Zeilen Java-Code besteht. SPIRIX kann voll-funktionsfähig nach XML-Dokumenten in einem P2P-Netz suchen sowie deren Relevanz inhaltsbasiert bewerten.

Für die Evaluierung der entworfenen Techniken wurde die Suchqualität von SPIRIX durch Teilnahme an INEX 2008 offiziell ausgewertet. Dabei konnte eine Präzision festgestellt werden, die vergleichbar mit der Qualität der 10 besten XML-Retrieval Suchmaschinen weltweit ist. In weiteren Experimenten wurden die entworfenen Methoden für verteiltes XML-Retrieval evaluiert; dabei wurden jeweils die erzielte Suchqualität und der notwendige Aufwand gegenübergestellt und sichergestellt, dass insbesondere der Aufwand für die Kommunikation zwischen Peers so effizient ist, dass das System auch zu einer großen Anzahl von teilnehmenden Peers skalieren kann.

Als Fazit lässt sich festhalten, dass verteiltes XML-Retrieval prinzipiell möglich ist, und zwar sowohl effektiv als auch effizient. Des Weiteren zeigte die Evaluierung der konzipierten Routingtechniken, dass durch Ausnutzung von XML-Strukturinformationen eine signifikante Reduzierung der Anzahl versendeter Nachrichten, ihrer Größe und somit der Netzlast bei gleichzeitiger Steigerung der Suchqualität erreicht werden kann.

Auch nach Abschluss der vorliegenden Arbeit wurden die Evaluierungen zum Thema verteiltes XML-Retrieval fortgeführt. Bei INEX 2009 konnte SPIRIX eine hervorragende Suchqualität erzielen. Beim Ad Track wurde Platz 7, beim Efficiency Track sogar zweimal Platz 1 (in Bezug auf die Suchqualität) erreicht. Dabei kam erstmalig eine neue Testkollektion zum Einsatz, die aus über 50 GB besteht, aus mehr als 2.660.000 XML-Artikeln und aus einer großen Anzahl semantischer Tags, die für die Suche genutzt werden können. Bei der Indizierung und Verarbeitung dieser umfangreichen Kollektion hatten viele der zentralisierten INEX-Ansätze Schwierigkeiten während die verteilte Architektur von SPIRIX jedem Peer nur einen Teil der Gesamtlast bei der Suche zuweist, so dass der Indexplatzverbrauch verteilt und das Ranking parallel erfolgen kann [WK10]. Die Vorbereitungen zu INEX 2010 laufen bereits, hier wird untersucht werden, wie die neue, sehr viel strukturreichere Kollektion noch effizienter ausgenutzt werden kann durch Anwendung eines erweiterten Satzes von Strukturähnlichkeitsfunktionen.

Die Bewältigung der Informationsflut des Internets wird weiterhin eine der vorherrschenden Herausforderungen unseres Zeitalters bleiben. Mit dem starken Wachstum von Web-basierten Gemeinschaften wie beispielsweise soziale Netzwerke oder Wikis, die auf Zusammenarbeit und Austausch zwischen den einzelnen Teilnehmern beruhen, werden P2P-Architekturen voraussichtlich eine immer größere Rolle spielen. Ihr Anteil an der verursachten Netzlast im Internet übersteigt schon jetzt das Nachrichtenvolumen von C/S-Architekturen [SW05], wobei viele zentralisierte Lösungen inzwischen P2P-Ansätze integrieren. Durch die zunehmende Verwendung von XML für den Austausch strukturierter Informationen wird es daher immer wichtiger werden, mit adäquaten Methoden effizient und effektiv in solchen Systemen suchen zu können. Die in der vorliegenden Arbeit entwickelten Techniken, insbesondere die zur Reduktion des Kommunikationsaufwands, können dazu beitragen, den Übergang von P2P-Suchmaschinen als Forschungsthema zu realem, performantem Einsatz in der Praxis voranzutreiben.

Literaturverzeichnis

[LLH03] Li, J.; Loo, B.; Hellerstein, J.; Kaashoek, F.; Karger, D.; Morris, R.: On the Feasibility of Peer-to-Peer Web Indexing and Search. In: Proc. of the Second Int. Workshop on Peer-to-Peer Systems, 2003.

[RZT04] Robertson, Stephen E.; Zaragoza, Hugo; Taylor, Mary: Simple BM25 extension to multiple weighted fields. In: Proc. of Int. ACM Conference on Information and Knowledge Management (CIKM'04), ACM Press, New York, USA, 2004.

[SML03] Stoica, I.; Morris, R.; Liben-Nowell, D.; Karger, D.; Kaashoek, F.; Dabek, F.; Balakrishnan, H.: Chord - A Scalable Peer-to-peer Lookup Protocol for Internet Applications. In: IEEE/ACM Transactions on Networking, Vol. 11, No. 1, 2003.

[St07] Stock, Wolfgang: Information Retrieval. Oldenbourg Wissenschaftsverlag, München, 2007.

[Su04] SuMa-eV: Gemeinnütziger Verein zur Förderung der Suchmaschinen-Technologie und des freien Wissenszugangs e.V. (SuMa-eV). Satzung, Hannover, Deutschland, 2004.

[SW05] Steinmetz, R.; Wehrle, K. (eds.): Peer-to-Peer Systems and Applications. Lecture Notes in Computer Science No. 3485, Springer-Verlag, Berlin-Heidelberg, 2005.

[TGK09] Trotman, Andrew; Geva, Shlomo, Kamps, Jaap (Eds.): Advances in Focused Retrieval. Proc. of the 7th Int. Workshop of the INitiative for the Evaluation of XML Retrieval (INEX 2008), LNCS, Springer-Verlag, Berlin-Heidelberg, 2009.

[WD08] Winter, Judith; Drobnik, Oswald: A Distributed Indexing Strategy for Efficient XML Retrieval. In: Efficiency Issues in Information Retrieval Workshop (EIIR2008) at European Conference on Information Retrieval (ECIR2008), Glasgow, UK, 2008.

[WD09] Winter, Judith; Drobnik, Oswald: SPIRIX – A Peer-to-Peer Search Engine for XML-Retrieval. In: Advances in Focused Retrieval (INEX 2008), Lecture Notes in Computer Science, Springer-Verlag, Berlin-Heidelberg, 2009.

[Wi09] Winter, Judith: XML Information Retrieval. Dissertation, Johann Wolfgang Goethe Universität, Frankfurt, Deutschland, Sept. 2009.

[WK10] Winter, Judith; Kühne, Gerold: Achieving High Precisions With Peer-to-Peer Is Possible! In: Focused Retrieval and Evaluation. Lecture Notes in Computer Science, Springer-Verlag, Berlin-Heidelberg, 2010.

Judith Winter wurde 1974 in Frankfurt am Main geboren. 1994-1997 absolvierte sie eine Ausbildung als Industrie-Informatikerin (Mathematisch-technische Assistentin) beim Rhein-Main Rechenzentrum in Frankfurt. Anschließend (1997-2002) war sie als Datenbankspezialistin bei der Deutschen Bundesbank tätig. Von 2000-2006 studierte sie Diplom-Informatik an der J.W.Goethe-Universität Frankfurt und an der Queensland University of Technologie, Australien. Ihre Diplomarbeit zum Thema Geschäftsprozess-management entstand dabei in Kooperation mit der Technischen Universität Karlsruhe. Als Wissenschaftliche Mitarbeiterin an der J. W. Goethe-Universität arbeitete sie 2006-2009 in der Gruppe für Architektur und Betrieb Verteilter Systeme. In dieser Zeit entstand auch ihre Dissertation mit dem Titel „XML Information Retrieval in Verteilten Systemen". Sowohl das Studium als auch die Endphase der Promotion wurden durch mehrere Stipendien gefördert. Die Forschungsergebnisse der Dissertation wurden in mittlerweile 13 Publikationen veröffentlicht und auf Konferenzen und eingeladenen Forschungsaufenthalten international vorgestellt. Sie bilden die Basis des Forschungsprojekts „SPIRIX – Suche in Verteilten Systemen", das Judith Winter mittlerweile an der Fachhochschule Frankfurt leitet, wo sie seit 2009 eine Vertretungsprofessur für Wirtschaftsinformatik innehat.