Artikel erschienen in der Datenschleuder 086, http://www.ccc.de/updates/2005/ds86

YaCy – Peer-to-Peer Web-Suchmaschine

von Michael Christen

Information ist im Web eine stark kontrollierte Resource – Portale, Suchmaschinen und das DNS sind zentral in de-facto-Monopolen organisiert und bestimmen welche Daten verfügbar sind. Mit YaCy wird die Kontrolle wieder an die Nutzer zurückgegeben.

Das YaCy - Projekt wurde Ende 2003 mit dem Ziel gestartet, eine freie, unabhängige und nicht zensierbare P2P - basierende Web-Suchmaschine zu erstellen. Zu diesem Zeitpunkt existierten viele gut funktionierende P2P - File-Sharing - Techniken, und auch mehrere freie Implementationen von Crawlern/Indexierern, aber keine Technik die P2P mit Suchmaschinentechnik verband. Wir stellen hier die über die Funktion einer Suchmaschine hinausgehenden Eigenschaften und Architektur von YaCy vor.

Suchmaschine mit Mehrwert

YaCy ist nicht nur eine Suchmaschine mit Crawler und Suchfunktion, sondern auch ein Web-Server, ein caching http Proxy mit optionalem pre-fetch, eine DNS-Erweiterung, ein Messaging-System und ein Wiki. Warum das ganze? Es gibt einen gemeinsamen Schlüssel für all diese Funktionen: Synergien zwischen Suchtechnik und der Loslösung von zentral-gesteuerten Diensten im Internet. Im Detail:

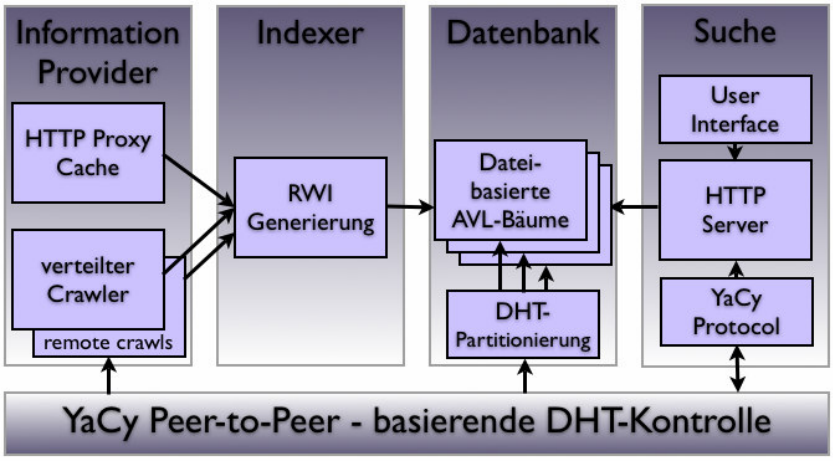
- **P2P-Suchmaschine und caching http Proxy**: Für die Existenz des Proxy in YaCy gibt es drei Gründe:
  1. Eine P2P-basierenden Software ist auf lange Onlinezeit angewiesen. YaCy soll nicht nur laufen während der User eine Suche absendet. *Synergie*: Durch die Nutzung des Proxy als Mehrwert wird eine lange Online-Zeit erreicht.
  2. Der Proxy ist ein 'Information Provider' für den Indexer. *Synergie*: die Suchmaschine erhält als Option 'kostenlos' Web-Seiten zum Indexieren ohne Crawling.
  3. Der Proxy bietet mit seinen eingebauten Filtern einen persönlichen Schutz vor ungewollten Webinhalten: das ist der notwendige 'Gegenpol' zur Zensurfreiheit durch die Suchmaschine. Die Idee ist, das jeder sich (seine Familie, sein Unternehmen, etc.) wieder soweit selbst zensieren kann, so wie er es ggf. von einem Suchmaschinenbetreiber erwarten würde (wenn er denn diese Erwartung hätte). *Synergie*: populäre Filter (bsp. Banner-Blocker) können von Peer zu Peer importiert werden.
- **Crawlen und Prefetching**: während Crawling eine typische Aufgabe einer Indexierungssoftware ist, liefert ein Proxy-Prefetch ggf. schnellere Zugriffszeiten für den Proxy-User. *Synergie*: beide Funktionen benutzen prinzipiell den gleichen Algorithmus.
- **Eingebauter Web-Server, File-Share, Wiki & Messaging**: wer durch die dezentrale Struktur der YaCy-Suche Informationsfreiheit sicherstellen möchte, will ggf. auch ein Publikationsmedium nutzen das in gleichem Maße keiner Zensur unterliegt. Web-Inhalte können zwar durch YaCy ad-hoc erfasst werden, aber selbst-erstellte Daten könnten weiterhin auf fremden Servern gesperrt oder entfernt werden. *Synergie*: unzensierte Suchergebnisse sind erst dann sinnvoll, wenn deren Resource auch geladen werden kann. YaCy gibt dazu einige Basiswerkzeuge an die Hand. Da der dazugehörige Server dann dem Nutzer gehört, kann er nicht zensiert werden.
- **Web-Server, Suchinterface und Proxy**: die natürlichste Umgebung für eine Web-Suche ist eine Web-Seite. Daher besteht YaCy's GUI aus einem integrierten Web-Server mit Servlet-Engine. *Synergie*: der Proxy, das GUI und die eigenen Webinhalte (siehe oben) können den gleichen eingebauten httpd-Server benutzen.
- **DNS-Umgehung und Erweiterung um die Top-Level-Domain ‚yacy'**: Das DNS-System ist mit seiner zentral-hierarchischen Struktur ein einfachster Angriffspunkt für Web-Zensur. YaCy bietet jedem Peer-Betreiber seine eigene ‚<peer-name>.yacy'-Domain, die automatisch durch den Proxy zum YaCy-Webserver des Peer-Betreibers aufgelöst wird. *Synergie*: die Nutzung des Proxies macht den Eingriff in die DNS-Auflösung möglich und die YaCy P2P-Verwaltung stellt ganz selbstverständlich eine Peer-zu-IP-Datenbank dar, was auch mit dynamisch zugewiesenen IP's funktioniert. Es existiert außerdem ein schlüssiges Konzept um ohne zentrale Datenbank einen Namens-Diebstahl der yacy-Domains unterbinden zu können

Zwar realisiert der eingebaute Proxy YaCy's zentrale Konzepte, aber die Software kann auch betrieben werden ohne den Proxy nutzen zu müssen. Dann dient YaCy ‚nur' als Suchinterface, Crawler, Web-Server etc.

Indexierung und Peer-Architektur

YaCy besteht aus einem Crawler, einem Indexierer, einer Datenbank, einem Suchinterface und der P2P-Organisation:

- Wir haben vom Crawler abstrahiert und sehen konzeptionell einen ‚**Information Provider**' vor. Als solcher kann ein Crawler eingesetzt werden, oder es ist möglich Seiten aus dem integrierten Proxy Cache als Input für den Indexers zu benutzen.
- Der **Indexer** erzeugt konzeptionell einen Reverse Word Index (RWI), d.h. zu jedem Wort eine Liste der URL's plus Ranking-Informationen. In unserer Implementation des RWI werden jedoch keine Wörter im Klartext gespeichert sondern lediglich Wort-Hashes. Die Wort-Hashes können allerdings nicht wieder zurück in Wörter übersetzt werden, das ist nicht notwendig und so können keine Klartextfragmente der ursprünglichen Webseiten auf den Rechnern der Peer-Betreiber gefunden werden. Diese können daher auch nicht zur Verantwortung für die bei ihm/ihr gelagerten Wörtern gezogen werden.
- Die **Datenbank** der RWI's ist eine hochspezialisierte Datenstruktur: eine AVL-Baumstrukur lässt ein geordnetes Aufzählen der URL's in RWI's zu und unterstützt damit effizientes Tabellen-Join, das für Wort-Kombinationssuche benötigt wird.
- Die Datenbank kann vom http Web-Interface des eingebauten **http-Servers** mit Servlet-Engine aus durchsucht werden. Der Web-Server dient auch zur Administration vom YaCy.



Das YaCy-P2P-Protokoll kontrolliert die Kooperation der Peers:

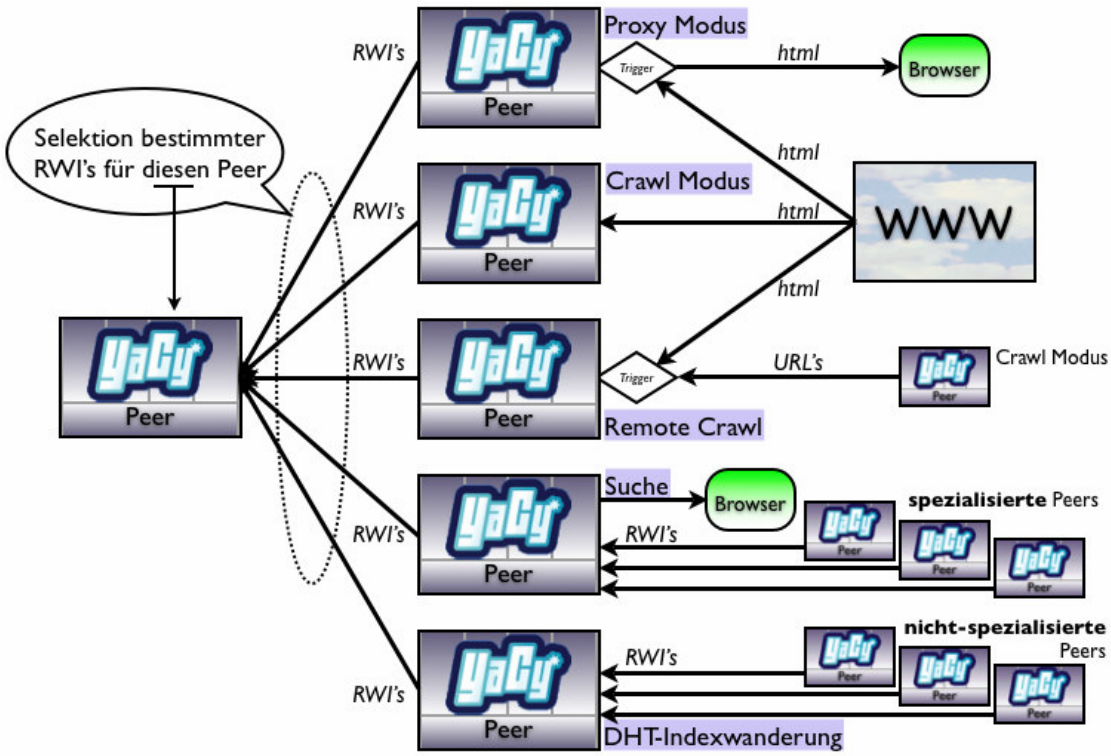
- RWI's werden in einer Distributed Hash Table (DHT) organisiert. Hierzu existiert ein komplexes Index-Wanderungskonzept zur Verteilung der Indexe zwischen den Peers.
- Peers können Teile eines Crawls an andere Peers abgeben. Es kann sichergestellt werden, dass URL's dabei nicht doppelt von verschiedenen Peers geladen werden.
- Jeder Peer publiziert seine eigenen Kontaktdaten regelmäßig an einen anderen Peer und pflegt eine Komplettliste aller Peers. Die Peer-Namen – zu – IP Zuordnung wird unter anderem zur Auflösung von <peer-name>.yacy – Domains im Proxy genutzt.

Index-Verteilung im YaCy-Netz

Es existieren verschiedene Möglichkeiten, wie ein YaCy-Peer seine Index-Datenbank erweitern kann. Im Überblick in der folgenden Grafik rechte Seite von oben nach unten:

1. der Peer ist im Proxy-Modus und indexiert Seiten aus dem Proxy-Cache.
2. der Peer führt einen lokal gestarteten Crawl aus.
3. der Peer erhält durch einen anderen Peer (der lokal einen Crawl ausführt) eine URL zum Indexieren zugewiesen und führt damit einen Remote Crawl aus.
4. der Peer bearbeitet eine lokal angestoßene Suchanfrage und fordert selektiv von anderen Peers RWI-Fragmente ein, die anschließend permanent in der lokalen Datenbank bleiben.
5. der Peer erhält von anderen Peers RWI-Fragmente zugewiesen, weil er für diese RWI's eine bessere Position entsprechend der DHT-Organisation hat.

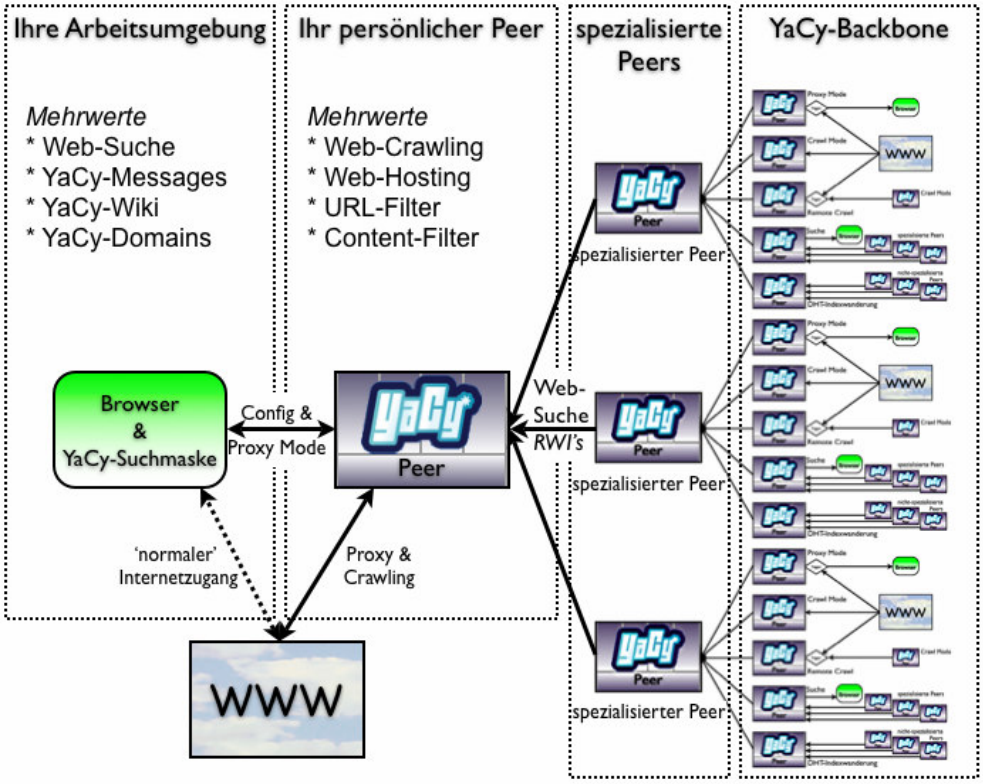
Alle Varianten der Index-Generierung können simultan ablaufen.



Spezialisierung von Peers auf bestimmte RWI-Bereiche (linke Seite der Grafik): simultan zur RWI-Gewinnung partitioniert jeder Peer seine RWI-Datenbanken in Teilmengen, die wiederum per DHT-Wanderung an einen anderen Peer zur permanenten Speicherung abgegeben werden. Dieser Zielpeer partitioniert ebenfalls seine RWI's, speichert die soeben zugewiesenen Indexe aber so lange bis ein neuer Peer auftaucht, der eine ggf. noch bessere Position in der DHT besitzt. Dies geschieht beispielsweise wenn das YaCy-Netz wächst und neue Peers hinzukommen.
Die permanent ablaufende Index-Wanderung sorgt für eine gute Durchmischung der RWI's, so das recht schnell nicht mehr feststellbar ist ob ein Index, der an einem bestimmten Peer anzutreffen ist, auch von diesem erzeugt wurde, und auch nicht durch welche Methodik (Crawlen oder Proxy-Use). Dies ist sehr vorteilhaft für jeden Peer-Betreiber da dieser konzeptionell ausschließen kann dass er für die Erzeugung einzelner Indexe verantwortlich ist.
Da die Index-Verteilung ohne zentralen Server organisiert ist, gibt es keine technische Möglichkeit zur Zensur.

Verteiltes Crawlen und Suchen

Es ist eine Prämisse für das Projekt Suchzeiten möglichst kurz zu halten. Suchanfragen werden nur zu Peers gesendet, die aufgrund der DHT-Konstruktion den entsprechenden RWI speichern sollen. Das funktioniert weil RWI's ja schon im Vorfeld einer Suche bereits zu dem Peer, wo sie bei einer Suche erwartet werden, gewandert sind.
Mit steigender Peer-Zahl kann sich die DHT in dem wachsenden Netz weiter spezialisieren. Die Suchzeit für ein Wort steigert sich aber theoretisch nicht mit der Größe des Netzes, da der RWI ja immer nur an einer bestimmten Position erwartet wird. Praktikabel ist aber ein gewisser Anstieg des Redundanzfaktors für die Anzahl der gleichzeitig abzusuchenden Peers, und eine Mischung mit Peers die einen besonders großen Index sowie Durchsatz besitzen.



Besondere Vorteile dieser dezentralen Suchtechnik:

- Der Benutzer kann selbst und ad-hoc bestimmen, welche Webseiten in den Index aufgenommen werden - Es können Seiten Indexiert werden, auf die kein Link existiert und an die somit kein Crawler heranreicht.
- Es bietet sich an ein Ranking durch die Kooperation der Benutzer zu gestalten. Die Aufnahme in den Index geschieht ja durch die Aufmerksamkeit der YaCy-User auf bestimmte Web-Seiten. Dadurch wird die Indexierung auf populäre und für die Nutzer interessante Seiten fokussiert.
- Bereits gefundene Web-Seiten werden vom suchenden Peer aus wieder per DHT-Wanderung verteilt. Dadurch verstärkt sich die Präsenz interessanter Suchergebnisse. Dies kann als implizite Moderation des globalen Index durch die Peer-User angesehen werden.
- Inhalte können nicht global zensiert werden.

Konzepte um die Peer-Anzahl stark zu vergrößern

YaCy könnte für folgende Bereiche interessant werden:

- Als Betriebssoftware für Internet-Cafes: der caching Proxy ist nutzbringend zur Bandbreitenbegrenzung, und eine geplante Abrechungs-Funktion mit Client-Kontrolle bringt einen zusätzlichen hohen Nutzwert. Einsatzgebiet wäre weniger Deutschland, eher weltweit. Die Internetabdeckung geschieht in vielen Regionen fast ausschließlich über Internet-Cafes, und die Abrechung wird oft nur auf Papier ohne Software gemacht.
- Als Browser-Erweiterung: YaCy könnte auch einen Browser Cache auslesen, die Proxy-Funktion ist dann hinfällig aber alle anderen Vorteile wie beispielsweise die Nutzung von .yacy-Domains bleiben erhalten. Open-source - Projekt wie Konqueror und Firefox könnten YaCy als Option in ihre Distribution aufnehmen. Dann hätte jeder Browser-Nutzer automatisch die komplette Kontrolle über Web-Suche und Inhalte des gemeinsamen Suchindex.

YaCy ist open-source (GPL-Lizenz), kompakt (rund 1 MB Download), portabel (Java), leicht zu installieren (nur auspacken, keine DB aufsetzten) und einfach zu betreiben. Wer mitmacht, der arbeitet aktiv an der Sicherstellung der Informationsfreiheit mit.

Links:
YaCy-Homepage: http://www.yacy.net
deutsche Dokumentation: http://www.suma-lab.de/yacy