Real Time Suche

Dirk Lewandowski
Hochschule für Angewandte Wissenschaften Hamburg, Fakultät DMI, Department Information, Berliner Tor 5, 20099 Hamburg. dirk.lewandowski@haw-hamburg.de

Einleitung

Twitter, Facebook und Co. sind nun schon seit einiger Zeit in aller Munde. Die Nutzungszahlen steigen rasant an und durch die massive Erstellung von Inhalten auf diesen Plattformen entstehen Kollektionen jenseits des offenen, für die Suchmaschinen indizierbaren Webs. Insofern müssen Suchmaschinen diesen neuen Diensten auf zwei Ebenen begegnen: Einerseits müssen sie sicherstellen, dass sie Zugriff auf die dort angebotenen Inhalte bekommen, andererseits müssen sie im eigenen Interesse dafür sorgen, dass die Nutzer weiterhin auf ihren Seiten suchen und nicht zu den Sozialen Netzwerken „abwandern".
Die beiden großen Web-Suchmaschinen Google und Bing haben in den letzten Monaten Lösungen vorgestellt, die Real-Time-Informationen in die reguläre Websuche integrieren (Lewandowski, 2009). Es soll auch Thema dieses Aufsatzes sein, inwieweit diese Lösungen gelungen sind.
Doch zunächst einmal: Was versteht man überhaupt unter Real Time Suche? Und wie lassen sich darunter so unterschiedliche Dienste wie Facebook und Twitter fassen?
Dazu müssen wir zuerst einmal klären, was Real Time Informationen sind. Ich schlage folgende Definition vor:

> Real Time Informationen sind Informationen, die über laufende Kanäle verbreitet werden, im Moment ihrer Aktualisierung für Nutzer relevant werden und deren Informationswert zeitkritisch ist.

Der erste Teil dieser Definition bezieht sich darauf, dass die Informationen schon kurz nach der Erstellung über laufend aktualisierte Ticker (auch: *feeds* oder *streams*) verfügbar sind. Zweites ist es für Nutzer wichtig, diese Informationen sofort zur Kenntnis nehmen zu können, während der dritte Teil der Definition der Tatsache Rechnung trägt, dass diese Informationen (zumindest in der Regel) schnell veralten. Man kann also sagen, dass einem Nutzer von Real-Time-Informationen ein Informationsvorsprung erwächst, wenn er denn diese Informationen sofort wahrnimmt. Dies mag man nun für alle Informationsarten konstatieren, die in irgendeiner Weise zeitkritisch sind. Für die Real-Time-Informationen gilt es aber in besonderem Maße, da der kritische Zeitpunkt, an dem sie relevant werden und der Zeitpunkt, an dem sie irrelevant werden, in der Regel sehr nah beieinander liegen.

Bei all dem Hype um die Real-Time-Informationen stellt sich natürlich die Frage, ob wir es tatsächlich mit einem neuen Phänomen zu tun haben. Dies ist klar zu verneinen: Auch in der Vergangenheit wurden zeitkritische Informationen wie Börsenkurse oder Wetterdaten in Echtzeit aktualisiert und abrufbar gemacht. Neu in der Diskussion der Real-Time-Informationen ist allerdings, dass wir es nun mit einer massenhaften Erstellung solcher Informationen durch eine Masse von Autoren zu tun haben. Ähnlich wie bei der Veränderung der Welt der Online-Informationen durch das World Wide Web sehen wir nun, wie die Idee der Datenstreams aus der Welt der Profis zu einem Massenphänomen wird.

Hinzu kommt, dass wir damit einen Wandel im Informationsverhalten der Web-Nutzer erleben: Während der Informationszugang bislang entweder über den direkten Aufruf von Web-Angeboten oder aber über eine Suche erfolgte, werden nun die schon bei den Blogs populären Feeds abgefragt. Man kann also von einem Wandel hin zu einer Feed-basierten Informationsaufnahme sprechen. Inwieweit dadurch die Ad-Hoc-Recherche beeinflusst wird, ist allerdings noch unklar.

Wenn wir nun von Real-Time-Suche sprechen, so geht dies über die reine Zusammenstellung von Feeds hinaus. Wir können verschiedene Stufen des Informationszugangs zu Real-Time-Informationen unterscheiden:

1. Direkter Aufruf eines Feeds. Hier werden Real-Time-Informationen eines einzelnen Angebots angezeigt, ohne dass diese durch den Nutzer (oder Informationen, die über diesen Nutzer vorliegen) gefiltert wurden. Die Suche beschränkt sich also auf das Auffinden des geeigneten Feeds.
2. Aufruf eines individuell zusammengestellten Feeds. Hier werden Real-Time-Informationen bereits nach den Wünschen des Nutzers (oder aufgrund seines Profils) gebündelt. Wie bei Alertingdiensten steckt der Vorgang der Suche also in der Zusammenstellung der Quellen, was hier allerdings weit gefasst werden kann: Feeds können beispielsweise in sozialen Netzwerken aus den Statusmeldungen der mit einem Nutzer verbundenen Personen bestehen.
3. Suche in Real-Time-Informationen mittels einer Suchmaschine. Hier werden Real-Time-Informationen in einer Ad-Hoc-Recherche durchsucht. Ergebnisse werden dabei gerankt nach
    a. allgemeinen, für alle Nutzer gültigen Kriterien, oder
    b. aufgrund des Nutzerprofils (beispielsweise abgeleitet aus dem Kontaktnetzwerk des Suchenden).

In diesem Aufsatz soll vor allem der dritte Punkt betrachtet werden, wobei der Fokus auf den allgemeinen Suchmaschinen und ihrer Integration der Real-Time-Recherche liegt. Dabei werde ich zuerst die Aktualität der Datenbestände der Suchmaschinen generell diskutieren. Dann folgt ein Überblick über die Inhalte der Real-Time-Suchen, darauf wird die Frage diskutiert, inwieweit sich Real-Time-Informationen überhaupt als explizite Such-Inhalte eignen. Nach einem Überblick über die Recherchemöglichkeiten in diesem Bereich schließt der Artikel mit einem Fazit und einem Ausblick.

Aktualität der Datenbestände von Suchmaschinen

Dass Suchmaschinen mit Real-Time-Informationen nicht oder nur schwer umgehen können, wurde bereits 2001 von Chris Sherman und Gary Price in ihrem Buch über das Invisible Web festgestellt (Sherman & Price, 2001). So war es nur folgerichtig, dass sie eben auch diese Inhalte zum Invisible Web rechneten.
Seitdem sind allerdings viele der damals für die Suchmaschinen unsichtbaren Inhalte sichtbar geworden. Auf der einen Seite haben die Suchmaschinen sowohl technisch als auch hinsichtlich der Indexkapazitäten enorme Fortschritte gemacht, auf der anderen Seite haben aber auch die Inhalte-Anbieter die Notwendigkeit erkannt, ihre Informationen für die Suchmaschinen aufzubereiten und so deren Zugänglichkeit zu gewährleisten.
Bisher schienen allerdings die Real-Time-Informationen eine Ausnahme zu sein. Ein Notbehelf sind sog. Smart Answers oder One-Box Results, also von den Suchmaschinen eigens erstellte, besondere Hinweise auf einzelne Datenbanken (Höchstötter &

Lewandowski, 2009; Lewandowski & Höchstötter, 2009). Im Fall der Real-Time-Informationen wurden diese Datenbanken auch direkt abgefragt und das Ergebnis in der One Box angezeigt. Abbildung 1 zeigt zwei Beispiele aus der Google-Suche: Wetterinformationen und Börsenkurse.

Abb. 1: Einbindung von Real-Time-Informationen (links: Wetter bei Google, rechts: Börsenkurse bei Yahoo)

Der Weg zu den Real-Time-Informationen markiert allerdings nur den Endpunkt in einer immer weiteren Verbesserung der Suchmaschinen hinsichtlich der Aktualität der Datenbestände bzw. der Erweiterung der Suchmaschinen um weitere, immer aktuellere Datenbestände. Damit einher geht die Ergänzung des Crawlings um die Erfassung von Feeds.

Bereits zu einem Zeitpunkt, an dem die Web-Suchmaschinen tatsächlich nur einen einzigen Index abfragten (das, was wir heute als den Web-Index bezeichnen), wurde klar, dass die Aktualität der Datenerfassung von essentieller Bedeutung ist. Kursorische Tests machten deutlich, dass sich die Suchmaschinen in dieser Hinsicht wesentlich unterschieden (Notess, 2003), spätere systematische Tests bestätigten dies (Lewandowski, Wahlig, & Meyer-Bautor, 2006).

Die ersten ergänzenden Datenbestände, die von den Suchmaschinen aufgebaut wurden, waren die Nachrichtendatenbanken. Diese wurden zuerst als eigenständige Suchmaschinen (die natürlich von den Start- und Trefferseiten der Web-Suchmaschinen aus verlinkt wurden) aufgebaut. Da sie jedoch von den Nutzern häufig übersehen wurden (und als eigenständige Recherchequellen auch weiterhin übersehen werden), fand im Zuge der Hinwendung der Suchmaschinen zu einer sog. Universal Search (s. Quirmbach, 2009) eine Integration dieser Inhalte in die reguläre Web-Suche statt.

Nachrichteninhalte wurden nach dem gleichen Prinzip wie Web-Inhalte erfasst, nämlich mittels Crawling, d.h. die Suchmaschine erreicht die neuen Inhalte über das Verfolgen von Links. Dabei unterscheidet sich die Nachrichtensuche von der Websuche dadurch, dass nur in vorher festgelegten Quellen nach neuen Inhalten gesucht wird und durch die kürzeren Aktualisierungsintervalle (Machill, Lewandowski, & Karzauninkat, 2005).

Mit dem Aufkommen von Blogs erreichten Suchmaschinen eine neue Stufe der Geschwindigkeit. Auch hier wurden zuerst wieder separate Blog-Suchmaschinen aufgebaut, die dann wiederum in die Unversal Search integriert wurden. Neu waren aber nicht nur die wiederum größere Notwendigkeit der schnellen Aktualisierung des Datenbestands, sondern auch die Art der Erfassung. Gängige Blog-Software ist in der Lage, die Inhalte mittels RSS zu übermitteln. Suchmaschinen identifizieren automatisch Blogs, deren Inhalte in den Datenbestand aufgenommen werden sollen und fragen diese regelmäßig nach Veränderungen ab (Thelwall & Hasler, 2007, S. 469). Damit kann eine bessere Aktualität ohne übermäßige Belastung der Server des Anbieters erreicht werden.

Die letzte Stufe der Aktualität bei Suchmaschinen ist nun mit der Real-Time-Suche erreicht. Auch hier stellen sich wieder die Fragen nach der Erfassung und Erschließung der Inhalte sowie nach der Integration in die Universal Search. Eine Integration meint dabei grundsätzlich, dass die Inhalte aus der in die Web-Suche integrierten Kollektion nicht nur statisch an der immer gleichen Stelle innerhalb der Trefferpräsentation angezeigt werden, sondern dass das Ranking die Inhalte aus verschiedenen Datenbeständen im Ranking miteinander verbindet. Zwar mögen die Inhalte der unterschiedlichen Kollektionen immer noch in der Trefferliste für sich stehen, jedoch werden diese Container innerhalb der Trefferliste je nach Relevanz für die Anfrage an unterschiedlicher Stelle platziert. Abbildung 2 zeigt als Beispiel die Integration von Real-Time-Informationen in die Trefferliste von Google.

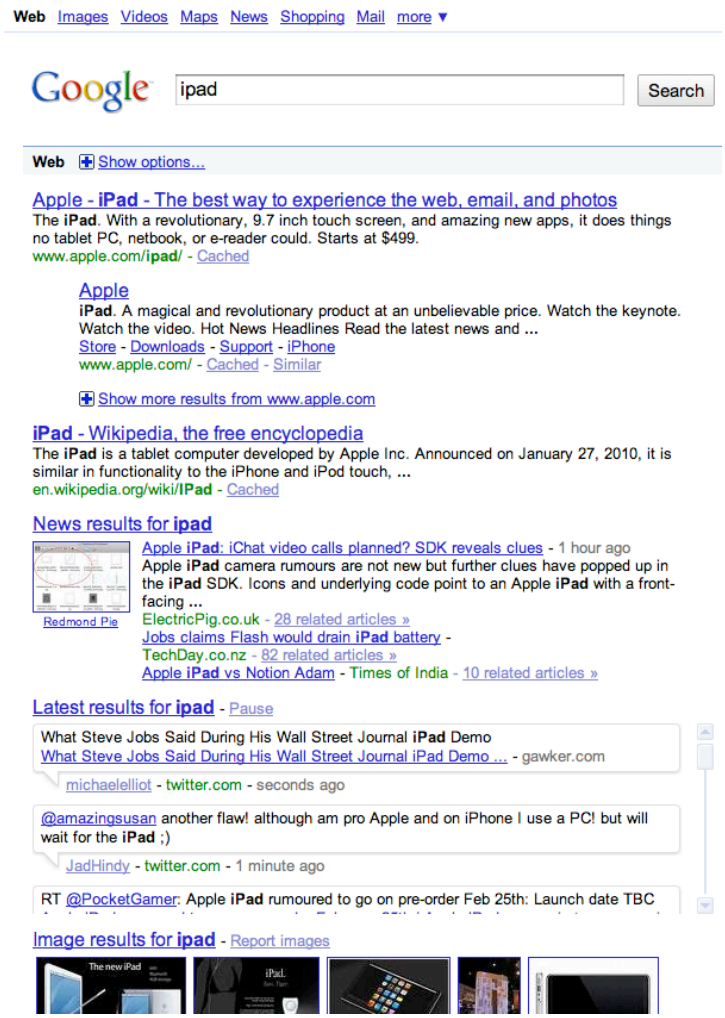

Abb. 2: Integration eines Containers mit Real-Time-Informationen bei Google

Inhalte der Real-Time-Suche

Bereits angesprochen wurden einige Arten von Real Time Informationen wie Börsenkurse und Wetterinformationen. Neben diesen „klassischen“ Arten hat sich in den letzten Jahren, vor allem durch das Aufkommen der sozialen Netzwerke, eine neue Form etabliert, die sog. Statusmeldungen. Diese treiben im wesentlichen die Entwicklung der Real-Time-Suche voran; man kann sogar sehen, dass andere Formen von Real-Time-Inhalten (bisher) für die Web-Suchmaschinen nur eine marginale Rolle spielen.

Unter Statusmeldungen versteht man kurze Postings von Nutzern in Sozialen Netzwerken, die vom Empfängerkreis dieses Nutzers gelesen werden können. Beispiele für Statusmeldung sind in Abb. 3 und Abb. 4 zu sehen.

Abb. 3: Beispiel für eine implizite Statusmeldung in einem Sozialen Netzwerk

Abb. 4: Beispiel für eine explizite Statusmeldung in einem Sozialen Netzwerk

Zu unterscheiden ist zwischen

- impliziten Statusmeldungen (die Statusmeldung wird vom System aufgrund einer Nutzeraktivität generiert, s. Abb. 3)
- expliziten Statusmeldungen (ein Nutzer gibt explizit etwas bekannt, d.h. er schreibt einen Text, s. Abb. 4).

Weiterhin ist zu unterscheiden zwischen Statusmeldungen, die nur aus einem Text bestehen und solchen, die (zusätzlich) einen Link enthalten, der auf ein externes Dokument hinweist.
Statusmeldungen fallen vor allem in den großen Sozialen Netzwerken (wie Facebook oder Xing) sowie in Twitter an. Daher ist es kaum verwunderlich, dass sich die bekannten Websuchmaschinen sich (zumindest bislang) auf diese beschränken[1].

Real-Time-Daten als *Such*inhalte

Real-Time-Informationen werden in den Sozialen Netzwerken oder in Twitter bislang kaum als Suchinhalte betrachtet. Vielmehr geht es um einen kontinuierlichen Datenstrom, den man entweder direkt beobachtet (meint: „nebenher laufen lässt“) oder aber in den man zu einem bestimmten Zeitpunkt einsteigt, um sich die Meldungen der letzten Zeit anzusehen. Eine Ad-hoc-Recherche ist dabei bislang nicht vorgesehen, auch wenn sich in den Angeboten wie selbstverständlich Suchboxen finden.
Möchte man der Frage nachgehen, inwieweit sich nun die Statusmeldungen als Suchinhalte eignen (und wie man dann ihre Relevanz als Ergebnis einer Suche bewerten soll), so muss man erst einmal nach der Intention des Nutzers fragen, der seinen Status bekannt gibt. An wen richtet er sich?
Man kann dabei nach den Adressaten der Kommunikation folgendermaßen unterscheiden:

- One-to-one: Der Sender möchte genau einen Empfänger erreichen.
  Man könnte hier also von einer privaten Nachricht sprechen. Diese in einer Suche (durch andere Nutzer) auffindbar zu machen, wäre wohl kaum wünschenswert. Man kann aber beobachten, dass solche One-to-one-Nachrichten in Sozialen Netzwerken zwar *primär* privat sein mögen, allerdings auch von anderen Nutzern eingesehen werden können. Dies ist beispielsweise bei den Kommentaren, die sich „Freunde“ bei

[1] Twitter wird hier nicht als Soziales Netzwerk gesehen, obwohl es einige wesentliche Merkmale eines solchen Netzwerks aufweist. Allerdings fehlt bei Twitter die sonst bei Sozialen Netzwerken übliche gegenseitige Kontaktbestätigung sowie das Verbergen der nutzergenerierten Inhalte vor externen Suchmaschinen.

Facebook gegenseitig an ihre Pinnwände schreiben können, der Fall. Weiterhin lassen sich Postings mit dem Namen eines Adressaten versehen (üblich durch eine Kennzeichnung mit dem @-Zeichen), wodurch diese Nachricht als für diese Person bestimmt markiert wird. Allerdings bedeutet dies nicht, dass sie nicht auch für andere Personen einsehbar ist. Insofern ist die One-to-One-Kommunikation mittels Statusmeldungen nicht klar von den anderen Formen zu unterscheiden.

- One-to-many: Der Sender möchte viele Nutzer erreichen.
  Hier besteht keine Beschränkung hinsichtlich des Adressatenkreises, oder zumindest ist die Grenze nur durch die Grenze des Gesamtsystems bestimmt, d.h. alle Nutzer des gleichen Netzwerks können die Nachricht lesen.
  Vor allem die Nachrichten bei Twitter gehören zu diesem Typ. Jede Nachricht dort kann von allen Nutzern des Dienstes eingesehen werden. Dabei ist es unerheblich, ob ein Nutzer bei Twitter angemeldet ist. Direkt zugestellt bekommen die Nachricht allerdings nur diejenigen, die „Follower" des Senders sind, d.h. sie haben die Meldungen dieses Senders abonniert.
- One-to-some: Der Sender möchte eine bestimmte Zielgruppe erreichen.
  Der Adressatenkreis der Nachricht ist beschränkt auf die Kontakte innerhalb des Sozialen Netzwerks. In der Regel erfolgt die Beschränkung auf die direkten Kontakte (Kontakte ersten Grads), teils werden aber auch andere Kreise adressiert (dann vor allem die Kontakte zweiten Grads).

An dieser Form der Kommunikation kann man sehen, dass die Inhalte den Interessenten nicht primär über eine Suche, sondern über einen (abonnierten) Datenstrom erreichen. Diejenigen, die „Freunde" oder „Follower" des Senders sind, bekommen dessen Nachrichten automatisch in ihren individuellen Datenstrom eingespeist. Interessant dabei ist, dass wir es hier mit Diensten zu tun haben, bei denen nicht (wie sonst üblich) die Möglichkeit zur Ab-Hoc-Recherche durch Alerting-Dienste ergänzt wird, sondern umgekehrt.

Wie sieht nun diese Ergänzung um eine Suche aus? Auf der einen Seite bieten Dienste wie Facebook und Twitter eigene Suchfunktionen, auf der anderen Seite bemühen sich externe Suchmaschinen, die Inhalte aus verschiedenen Diensten zu aggregieren und durchsuchbar zu machen. Probleme entstehen hier vor allem aus den sehr kurzen Texten der Statusmeldungen und natürlich durch die beständige Aktualisierung.
Suchmaschinen haben im Lauf der Jahre ein relativ beständiges Set an Rankingfaktoren entwickelt, mit denen sich textuelle Inhalte ranken lassen. Dabei mag es sich im einzelnen um hunderte Faktoren (oder besser: Signale) handeln, diese lassen sich jedoch auf vier Gruppen von Rankingfaktoren reduzieren:

- Textstatistik: Hier wird die Suchanfrage mit dem Inhalt des Dokuments verglichen, wobei die Häufigkeit und die Position der Suchbegriffe im Dokument berücksichtigt werden. Es handelt sich dabei um Standardverfahren des Information Retrieval.
- Popularität: Die Popularität eines Dokuments, gemessen beispielsweise anhand der Nutzerzugriffe und der Verweildauer auf den Dokumenten sowie – bestimmend für das Ranking von Web-Dokumenten – anhand der Verlinkung des Dokuments innerhalb des Web-Graphen, wird zur Bewertung der Qualität des Dokuments herangezogen. Dabei wird nicht einfach auf die Masse der Klicks bzw. Links gesetzt, sondern es kommen gewichtete Modelle zum Einsatz, die eine differenzierte Bewertung ermöglichen. Diese Modelle sind in der Literatur gut dokumentiert (unter

anderem (Culliss, 2003; Dean, Gomes, Bharat, Harik, & Henzinger, 2002; Kleinberg, 1999; Page, Brin, Motwani, & Winograd, 1998).

- Aktualität: Die Bewertung der Aktualität spielt bei Web-Suchmaschinen auf zwei Ebenen eine Rolle. Einerseits geht es um die Ermittlung des tatsächlichen bzw. relativen Erstellungs- bzw. Aktualisierungsdatums der Dokumente (Acharya et al., 2005), andererseits um die Frage, in welchen Fällen es sinnvoll ist, bevorzugt aktuelle Dokumente anzuzeigen.
- Lokalität: Für Suchmaschinen essentiell ist auch die Bewertung von Dokumenten nach ihrer Nähe zum Nutzer.

Betrachten wir nun die Suche nach Real-Time-Informationen, so wird deutlich, dass die im Web-Kontext verwendeten Rankingfaktoren auch auf diesen Bereich übertragbar sind, allerdings die Gewichtungen verändert werden müssen. Dies ist nicht ungewöhnlich und gilt für andere Datenbestände ebenso (z.B. für die News-Suche, s. (Machill et al., 2005).
Bei Real-Time-Informationen ist die Verwendung von textspezifischen Faktoren natürlich notwendig, um Suchanfrage und Dokumente miteinander zu vergleichen. Allerdings entsteht durch die in der Regel sehr kurzen Texte (bei Twitter gibt es beispielsweise eine Beschränkung auf 170 Zeichen pro Nachricht) das Problem, dass Gewichtungen kaum anwendbar sind.
Ebenso problematisch ist die Verwendung von linkbasierten Rankingfakoren: Real-Time-Nachrichten können schlicht noch keine Links auf sich gezogen haben, wenn man davon ausgeht, dass die Suchanfrage bereits kurz nach der Erstellung der Nachricht erfolgt. Anderes sieht es bei Recherchen aus, die zu einem späteren Zeitpunkt durchgeführt werden.
Allerdings sind nicht alle Popularitätsfaktoren unbrauchbar: Die Anzahl der Follower eines Nutzers kann ein Indikator für die Qualität von dessen Nachrichten sein[2], ebenso die Anzahl von Links auf seine Nachrichten in der Vergangenheit.
Während Lokalität als Rankingfaktor bei Real-Time-Informationen ähnlich zu sehen ist wie bei den Web-Inhalten auch, ist die Aktualität natürlich einer der bestimmenden Faktoren bei dieser Art von Inhalten.
Zusätzlich kann noch die oben beschriebene Intention des Senders als ein Kriterium für die Relevanz der Nachricht für den Suchenden herangezogen werden, wenn man davon ausgeht, dass man in der Suche überhaupt Zugriff auf alle Nachrichtentypen hat. So könnte eine Nachricht, die sich an alle richtet, höher gewichtet werden als eine, die nur an einen beschränkten Adressatenkreis gerichtet ist. Ein Problem hierbei könnte allerdings sein, dass sich die Sender wohl meist nicht bewusst sind, an welchen Adressatenkreis sie eine Nachricht gerade senden.

Recherche nach Real-Time-Informationen

Ähnlich wie bei anderen Spezialsuchen auch ist bei einer Recherche nach Real-Time-Informationen die geeignete Suchmaschine unter anderem nach den Kriterien der Abdeckung und der Aggregations auszuwählen. Sullivan (2009) unterscheidet zwischen

1. Suchmaschinen, die einen einzigen Datenbestand (z.B. Twitter) abfragen, deren Mehrwert aber in einer Verbesserung der Suchergebnisse liegt.

[2] Wobei allerdings zu bedenken ist, dass Accounts mit vielen Followern inzwischen auch verkauft werden, damit sie anderweitig genutzt werden. Man fühlt sich an das Problem der abgelaufenen Domains erinnert, das den Suchmaschinen vor einigen Jahren Probleme bereitete.

2. Suchmaschinen, die ihren Datenbestand aus den im Real-Time-Web gesendeten Links aufbauen und so eine alternative Suche zu den konventionellen Web-Suchmaschinen darstellen.
3. Websites, die selbst den Großteil der Inhalte des Real-Time-Web erstellen (wie Facebook und Twitter) und eine eigene Suchfunktion anbieten.

Ergänzt werden muss diese Aufstellung noch um

4. Allgemeine Web-Suchmaschinen (wie Google und Bing), die Meldungen des Real Time Web in ihre Suchergebnisse integrieren.

Bei der Recherche stellt sich also die Frage, welche dieser Suchmöglichkeiten genutzt werden sollten. Ich werde im folgenden einige Empfehlungen für die Recherche nach Real-Time-Informationen geben.

1. *Die Recherche direkt in den einzelnen Angeboten ist erstens anzuraten, wenn Inhalte nicht (oder nicht vollständig) an die Aggregatoren weitergegeben werden.*
Wie aus dem Vorangegangenen schon deutlich wurde, ergeben sich bei der Erschließung von Real-Time-Informationen Probleme, die sich unmittelbar auf die Recherche auswirken. So stellt sich die sonst übliche Frage, ob man besser bei den einzelnen Diensten direkt oder bei einem Aggregator recherchiert, im Fall der Real-Time-Informationen auf noch eine andere Weise. Während es sonst vor allem um den Umfang der Suchfunktionen und den Komfort der Recherche geht, haben wir es nun auch damit zu tun, dass ein Teil der vorhandenen Inhalte erst gar nicht bei den Aggregatoren verfügbar ist, da er nur in geschlossenen Systemen vorhanden ist. Ein Beispiel dafür ist Facebook. Dort sind weite Teile der von den Nutzern verfassten Meldungen nicht für Suchmaschinen zugänglich, auch wenn sich Facebook in dieser Hinsicht (durch Veränderungen der Nutzungsbedingungen) zunehmend öffnet. Allerdings wird auch bereits davon gesprochen, dass Facebook aufgrund seiner Masse an genuinen Inhalten eine ernsthafte Konkurrenz für Google darstellt, welches auf das offene Web angewiesen und beschränkt ist.[3]

2. *Die Recherche direkt in den einzelnen Angeboten ist zweitens anzuraten, wenn es um Vollständigkeit und „absolute Aktualität" geht.*
Aggregatoren beziehen die Inhalte aus den Real-Time-Angeboten zwar als Feeds (und damit ohne großen Zeitverzug), durch die Aufbereitung der Inhalte kommt es aber zu Verzögerungen. Außerdem versuchen die Aggregatoren (wie bei den Web-Inhalten auch), Spam-Meldungen zu filtern. Durch die (zusätzlichen) Relevanzbewertungen der Meldungen werden nicht unbedingt die aktuellsten Meldungen zuerst angezeigt.

3. *Um eine alternative Sicht auf das aktuelle Web zu bekommen, lohnt sich die Recherche in Suchmaschinen, die sich auf das Real Time Web beschränken.*
Diese Suchmaschinen[4] erfassen die in Meldungen aus dem Real Time Web enthaltenen Links und bilden daraus einen durchsuchbaren Index. Sie bieten sich als Suchalternative zur Recherche in den konventionellen Web-Suchmaschinen an, da sie einen anderen, stark auf Aktualität fokussierten Index durchsuchen. Zwar mögen die von ihnen indexierten Dokumente auch in den Web-Suchmaschinen vorhanden sein, die Beschränkung auf die aktuell diskutierten Seiten[5] macht eine aktualitätsbezogene Recherche hier jedoch komfortabler.

---

[3] http://www.wired.com/techbiz/it/magazine/17-07/ff_facebookwall?currentPage=all

[4] Beispiel: www.oneriot.com

[5] Es ist wichtig, hier zwischen einer einfachen Beschränkung auf aktuelle Seiten (nach dem Aktualisierungsdatum) und der Beschränkung auf aktuell diskutierte Seiten zu unterscheiden. In

4. *Die Recherche in Web-Suchmaschinen lohnt sich, wenn primär Web-Inhalte gefunden werden sollen, eine Ergänzung um Meldungen aus dem Real Time Web aber erwünscht ist.* Suchmaschinen wie Bing bieten zwar auch Spezialsuchen für Tweets[6], die einen Mehrwert gegenüber der Twitter-eigenen Suche bieten, indem sie nicht nur die aktuellsten Tweets anzeigen, sondern auch die meistgelinkten Seiten inklusive der Kommentare (vgl. Abb. 5). Allerdings liegt ihre besondere Leistung in der Integration von Real-Time-Informationen in die Trefferlisten der Websuche. Abbildung 2 zeigt die Integration von Real Time Informationen in Google. Diese Suchmaschine zeigt einen aggregierten Datenstrom nach Aktualität an (der laufend aktualisiert wird). Es ist allerdings damit zu rechnen, dass Google (wie auch Bing) diese Ergebnisse stärker vorverarbeiten und in die Ergebnislisten integrieren wird.

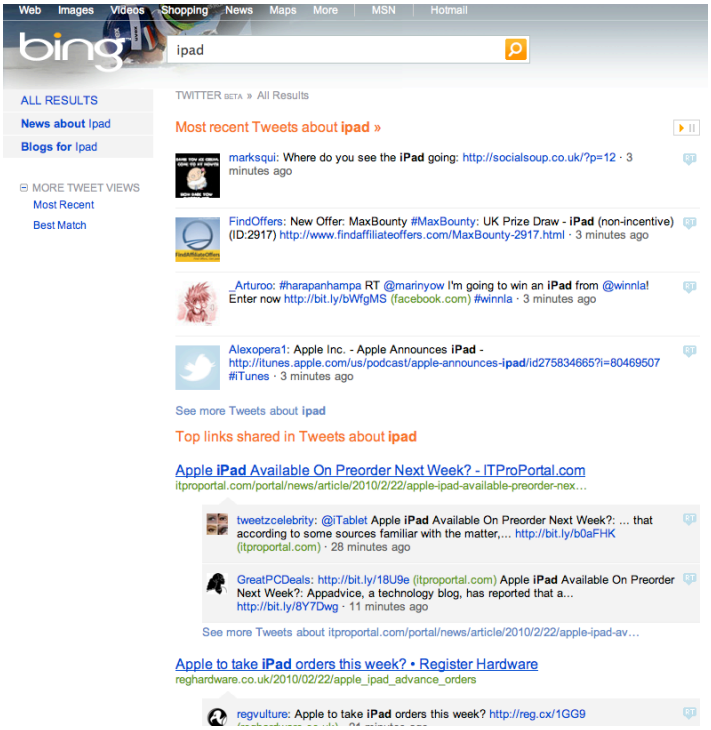

Abb. 5: Ausschnitt aus einer Trefferseite einer Spezialsuche bei Bing

Fazit

Die Beschäftigung mit dem Thema Real Time Suche macht deutlich, dass die Suchverfahren in diesem Bereich noch ganz am Anfang stehen. Dabei wurde klar, dass sich zwar die Suchverfahren verändern werden, das Real Time Web jedoch bereits in einem großen Umfang besteht und nicht wieder verschwinden wird. Allerdings dürfte sich der Hype um dieses Thema legen, und das Thema wird selbstverständlicher Bestandteil des Webs und damit auch der Web-Recherche werden.

Bislang indexieren die Real-Time-Suchmaschinen (und auch die Suchfunktionen der Real-Time-Angebote selbst) entweder nur die Inhalte eines einzigen Anbieters oder sie aggregieren Inhalte verschiedener Anbieter, wobei sie jedoch weit von einer vollständigen Abdeckung des Real Time Web entfernt sind. Dies ist vor allem als eine

letzterem Fall können durchaus auch ältere Dokumente auftauchen, denen aber aktuell eine Bedeutung beigemessen wird.

[6] http://www.bing.com/twitter

Zukunftsaufgabe der allgemeinen Websuchmaschinen zu sehen, die jedoch an den Punkten auf Probleme stoßen, an denen ihre Crawler durch die Anbieter ausgeschlossen werden. Eine Lösung besteht entweder im Aufbau eigener Real-Time-Plattformen (was allerdings wiederum nicht zu einer vollständigen Abdeckung aller Real-Time-Informationen führt) oder in Partnerschaften mit den Anbietern (wie Microsofts Partnerschaft mit Facebook).
Bei der Recherche darf nicht vergessen werden, dass das, was wir hier unter Real-Time-Web verstehen, nur ein Teil des Ganzen ist. Denn auch Audio- und Videostreams sind diesem Teil des Web zuzuordnen, werden jedoch von den auf Text fokussierten Suchmaschinen nicht erfasst.
In diesem Artikel wurde auch deutlich, dass es sich bei Real-Time-Informationen nur zum Teil um ein Thema der *Suche* handelt. Vielmehr geht es auch um die kontinuierliche Beobachtung. Zwar war auch in der Vergangenheit schon eine Beobachtung im Web möglich und notwendig (siehe dazu Calishain, 2007), diese wird mit dem Monitoring des Real Time Web aber auf eine neue Stufe gehoben.

Literatur

Acharya, A., Cutts, M., Dean, J., Haahr, P., Henzinger, M., Hoelzle, U., et al. (2005). Information retrieval based on historical data. US Patent Application No. 20050071741.

Calishain, T. (2007). Information Trapping: Real-time Research on the Web. Berkeley, CA: New Riders.

Culliss, G.A. (2003). Personalized search methods. USA: Ask Jeeves, Inc. US Patent No. US 6,539,377 B1.

Dean, J.A., Gomes, B., Bharat, K., Harik, G., & Henzinger, M.H. (2002). Methods and apparatus for employing usage statistics in document retrieval. USA: Google, Inc. US Patent Application No. 09/797,754.

Höchstötter, N., & Lewandowski, D. (2009). What Users See – Structures in Search Engine Results Pages. Information Sciences, 179(12), 1796-1812.

Kleinberg, J.M. (1999). Authoritative sources in a hyperlinked environment. Journal of the ACM, 46(5), 604-632.

Lewandowski, D. (2009) Microsoft Bing: Einstieg in Real Time Search. Password (11), 4-5.

Lewandowski, D., & Höchstötter, N. (2009). Standards der Ergebnispräsentation. In D. Lewandowski (Hrsg.), Handbuch Internet-Suchmaschinen (S. 204-219). Heidelberg: Akademische Verlagsgesellschaft Aka GmbH.

Lewandowski, D., Wahlig, H., & Meyer-Bautor, G. (2006). The Freshness of Web search engine databases. Journal of Information Science, 32(2), 133-150.

Machill, M., Lewandowski, D., & Karzauninkat, S. (2005). Journalistische Aktualität im Internet. Ein Experiment mit den "News-Suchfunktionen" von Suchmaschinen. In M. Machill & N. Schneider (Hrsg.), Suchmaschinen: Herausforderungen für die Medienpolitik (S. 105-164). Berlin: Vistas.

Notess, G.R. (2003). Search Engine Statistics: Freshness Showdown. http://www.searchengineshowdown.com/statistics/freshness.shtml [22.2.2010]

Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). The PageRank citation ranking: Bringing order to the Web. http://dbpubs.stanford.edu:8090/pub/1999-66 [22.2.2010]

Quirmbach, S. (2009). Universal Search - Kontextuelle Einbindung von unterschiedlicher Quellen und Auswirkungen auf das User Interface. In D. Lewandowski (Hrsg.), Handbuch Internet-Suchmaschinen. Heidelberg: Akademische Verlagsgesellschaft Aka GmbH.

Sherman, C., & Price, G. (2001). The Invisible Web: Uncovering Information Sources Search Engines Can't See. Medford, NJ: Information Today.

Sullivan, D. (2009). A behind-the-scenes look at the real-time search circus. Advertising Age, 80(37), 40.

Thelwall, M., & Hasler, L. (2007). Blog search engines. Online Information Review, 31(4), 467-479.