Einsatz von Internet-Suchmaschinen bei der beruflichen Recherche

Chancen und Risiken

TuTech Innovation GmbH, 23.11.2006, Hamburg
Michael Nebel

Teil I - Informationsgewinnung aus dem Internet

Teil II - Suchmaschinenrealität

Teil III - Suchmaschinen-Wirtschaft

Fazit

Informationsgewinnung

- Suchmaschinen im Internet
- Klassische Recherche
- Internetsuche
- Veränderung der Information
- Eine perfekte Suchmaschine

Suchmaschinen im Internet

- ⇨ Das Internet gewinnt als Informationsquelle immer mehr an Bedeutung.
- ⇨ Suchmaschinen bestimmen unser Verhalten und unsere Erwartungshaltung bei der Informationsgewinnung. (Gatekeeper-Funktion)
- ⇨ Informationsentdeckung/-wiederentdeckung wird immer schwieriger.

Klassische Recherche

- Lexikon, Bibliothek, Dienstleister, Wartungsverträge
- Hochwertiger und relativ vollständiger Datenbestand
- Strukturierte Fragen

=> Qualitativ hochwertige Antworten

=> Antwortzeit Minuten bis Tage

Internetsuche

- "googlen" / Wikipedia
- Vom Nutzer autodidaktisch gelernt
- Umfangreicher Datenbestand von sehr heterogener Qualität
- Ausgehend von wenigen Worten wird der Ergebnisraum schrittweise unterteilt.
- Beachtung nur der ersten Ergebnisse.

=> Scheinbar preisgünstig

=> Reaktionszeit Sekunden.

Veränderung der Information

- Die Bedeutung der Informationen im Internet wächst:
  - Unternehmensdarstellung
  - Vertriebskanal
  - Werbekanal
  - Marketingwerkzeug
  - Kriminalität
  - Selbstdarstellung

- „Internet“ ist keine Frage des „kann man“ mehr, sondern ein kritischer Geschäftsprozess.

Eine perfekte Suchmaschine

- *Vollständige Erfassung der verfügbaren Informationen*
  Eine Suchmaschine im Internet soll "das (ganze) Internet" durchsuchen. Je nach der Erfahrung des Nutzers mit dem Medium vielleicht auch nur "den wichtigsten Teil des Internet".
- *Hochwertige Ergebnisse*
  Die Qualität der Ergebnisse soll einen Bezug zur Frage besitzen und objektive Antworten liefern.
- *Einfachheit bei der Benutzung*
  Die Suchmaschine soll bereits mit nur wenigen Stichworten das erwartete Ergebnis liefern. Die Fragen werden selten ausformuliert sondern durch eine Liste von Stichworten umschrieben.

Eine perfekte Suchmaschine

- *Hohe Aktualität der Informationen*
  Die der Suche zu Grunde liegenden Informationen sollen dem aktuellen Abbild der Quelle entsprechen. Änderungen an der Quellseite gehen ohne Verzögerung in die Suche mit ein.
- *Objektive Behandlung der Inhalte*
  Eine Suche wird in der Regel mehr als ein Ergebnis liefern. Diese sollen so gewichtet werden, dass das "beste Ergebnis" als erstes geliefert wird.

Suchmaschinenrealität

- Komponenten einer Suchmaschinen
- Informationsquellen
- Zusätzliche Informationsquellen
- Entstehende Probleme
- Grenzen von Suchmaschinen
- Deep-Web
- Problematische Datenquellen

Komponenten einer Suchmaschinen

Komponente	Beschreibung
Crawler/Fetcher	Sammeln der Webseiten
Parser/Indexer	Analyse der Dokumente
Datenspeicher	Indices und gecachte Dokumente
Suchfrontend	Interface für die Benutzer
Monitoring	Überwachung der Komponenten
Wartung/Betrieb	Hardware, Software

Informationsquellen

Woher kommt das "Wissen" des Internets?

Strukturierte Quellen

- Redaktionelle Inhalte, Medien
- Hersteller
- Forschungseinrichtungen
- Öffentliche Einrichtungen

Merkmale:

- bekannte Quellen
- häufig für Suchmaschinen optimiert
- zu weiten Teilen von Suchmaschinen erfasst

Dynamische Quellen

- Newsgroups, Mailinglisten
- Private Homepages
- Vereine
- Foren, Blogs

Merkmale:

- entstehen und vergehen ungeordnet
- häufig "subjektive" aber auch hochwertige Inhalte
- schwierig zu erfassen

Zusätzliche Informationsquellen

- Analyse der Linkstruktur (Stichwort: Page-Rank)
- Zusätzliche Felder durch die Autoren (Meta-Tags, Dublin-Core-Tags)
- Zuordnung der Website zu Kategorien in Internet-Verzeichnisse (Dmoz, Yahoo)
- Manuelle Kennzeichnung
- Kennzeichnung durch Communities (z.B. Yahoo-Groups)
- Nutzung von Thesauri, Lexika
- Grammatikalische Analyse der Sprache

Entstehende Probleme

Technisch

- Datenvolumen
- Datendoubletten
- nicht eindeutige Datenquellen
- „Müll“
- Monopolisierung

Politisch/Gesellschaftlich

- Monopolisierung
- Bewertung und Filterung von Inhalten, ...
- Einflussnahme/Lobbyismus

Grenzen von Suchmaschinen

- *Datenmenge*
  Der Umfang der Ausgangsdaten ist kaum abschätzbar. Kontinuierlich werden neue Informationen erzeugt und alte verschwinden. Die Qualität der Ausgangsdaten wird durch Werbemüll und Webspam gezielt verringert.
- *Aktualität der Datenbasis*
  Eine Suchmaschine muss die Ausgangsdaten präventiv erfassen und vorverdichten. Zwischen der Erfassung und der Anfrage verstreicht immer eine gewisse Zeitspanne, so dass die Ergebnisse der Suchmaschinen in der Regel veraltet sein müssen.
- *Verfügbare IT-Technik*
  Die zu verarbeitenden Datenmengen und Anforderungen an Antwortzeiten und Verfügbarkeit stellen die aktuelle Hard- und Software vor große Herausforderungen.
- *Doubletten/Herkunft*
  Identische oder stark ähnelnde Inhalte sind oft an verschiedenen Stellen im Internet verfügbar. Die Originalquelle ist nur noch selten erkennbar. Eine Bewertung der Information an Hand des Autors ist nicht mehr möglich.

Grenzen von Suchmaschinen

- *Mehrdeutigkeit der Suchanfragen*
  Eine Recherche "per Stichwort" ist ein iterativer Prozess, bei dem sich die Benutzer schrittweise an die erwartete Antwort heran tasten. Der Informationsgehalt der Frage ist oftmals für eine korrekte Antwort zu gering.
- *Bewertung der Relevanz*
  Die Relevanz einer Antwort auf eine Frage hängt häufig auch von der Umwelt des Fragenden ab. Die Definition einer objektiven Relevanz ist auf Grund der fehlenden Rahmendaten nur schwer möglich.
- *Juristische Forderungen*
  Suchmaschinen sind Teil einer Gesellschaft und müssen sich den Regeln/Gesetzen der Gesellschaft ihres Heimatlandes unterwerfen. Im Rahmen des Internets gelten diese Regeln aber oft nur für einen Teil der Benutzer. Der andere Teil kennt die ursprünglichen Regeln und ihren Einfluss auf die Suchmaschine und ihre Ergebnisse nicht.

Deep-Web

"deep-", "dark-" oder "invisible web"

- kurz was Suchmaschinen nicht sehen.

- Suchmaschinen folgen Links von Seite zu Seite
- oftmals nur die "obersten" Ebenen indiziert (Surface Web)

- Erfasste deutschspachige Dokumente in den großen Suchmaschinen ~100 Mio. Seiten (geschätzt!)
- Umfang des deutschsprachigen Netzes: über 1 Mrd. Seiten (geschätzt)
- Inhalte oftmals sehr hochwertig

Der Begriff wurde 1994 von Dr. Jill Ellsworth geprägt (siehe Brightplanet).

Problematische Datenquellen

- Fehlerhafte Webseiten (Link-Loops, ungültige Links, Crawler-Traps, ...)
- Gesperrte Webseiten
- zu große Anzahl an Dokumenten auf einer Website (ohne Priorisierung)
- zu langsames Antwortverhalten des Webservers
- unbekannte Websites
- Datenbanken (Eingabefelder)

=> Breitencrawler oder Tiefencrawler ?

Suchmaschinen-Wirtschaft

- Wie werden Websites gefunden?
- Suchmaschinenoptimierung (SEO)
- Ergebnismanipulation
- Werbung im Internet
- Von der Werbung zum Webspam
- Beispiele für Webspam
- Erkennen relevanter Informationen
- Geschäftsmodelle von Suchmaschinen
- Marktanteile

Wie werden Websites gefunden?

Es gibt nicht „das“ Internet. Suchmaschinen müssen selber neue Websites finden und in ihren Index aufnehmen.

Initiiert vom Websitebetreiber:

- Eintragung von URLs in Verzeichnissen (DMOZ, ...)
- Direkte Anmeldung von URLs bei den Suchmaschinen

Automatisch von der Suchmaschine:

- Suche nach ausgehenden Links
- Domainlisten, IP-Nummernlisten
- Manuelle Einbindung wichtiger Websites

In der Regel ist es ein Problem der **Betreiber**, Inhalte für Suchmaschinen zugänglich zu machen.

Suchmaschinenoptimierung (SEO)

- Das Internet ist oft ein weiterer wichtiger Vertriebskanal.
- Die Positionierung auf den Ergebnissseiten der großen Suchmaschinen hat **direkten** Einfluss auf den Geschäftserfolg.

=> Die Präsentation und Position in den Suchmaschinen ist für die Websitebetreiber wichtig.

- Suchmaschinen finden nicht immer von sich aus die "richtigen" Inhalte
- Eine direkte Einflussnahme auf die Suchmaschinen und die Positionierung in den Ergebnislisten ist nur bedingt möglich.

=> neues Beraterfeld: "Suchmaschinenoptimierung"

Ergebnismanipulation

Aktuelle (dubiosen) Tricks zur Manipulation der großen Suchmaschinen:

- Cloaking:
  Unterschiedliche Inhalte für Besucher und Suchmaschinen
- Link-Farm:
  ermogeln eines Page-Rank
- Guestbook-Spamming:
  ermogeln eines Page-Rank durch automatische Eintragung
- Doorway Pages:
  Aufbau vorgelagerter Webseiten zum "Einfangen" von Besuchern
- Hidden-Text:
  weisse Schrift auf weissem Grund. Fremde Schlüsselworte.
- Meta-Keywords:
  Nutzung fremder Schlüsselworte in den Meta-Tags.

Werbung im Internet

Der Betrieb eines Internetangebots kostet Geld. Viele Betreiber versuchen sich über Werbung zu finanzieren.

- Früher wurde "Tausender-Kontakt-Preise" bezahlt, d.h. für jede Darstellung einer Werbung erhielt der Websitebetreiber Geld. Dies ist nur noch selten der Fall.

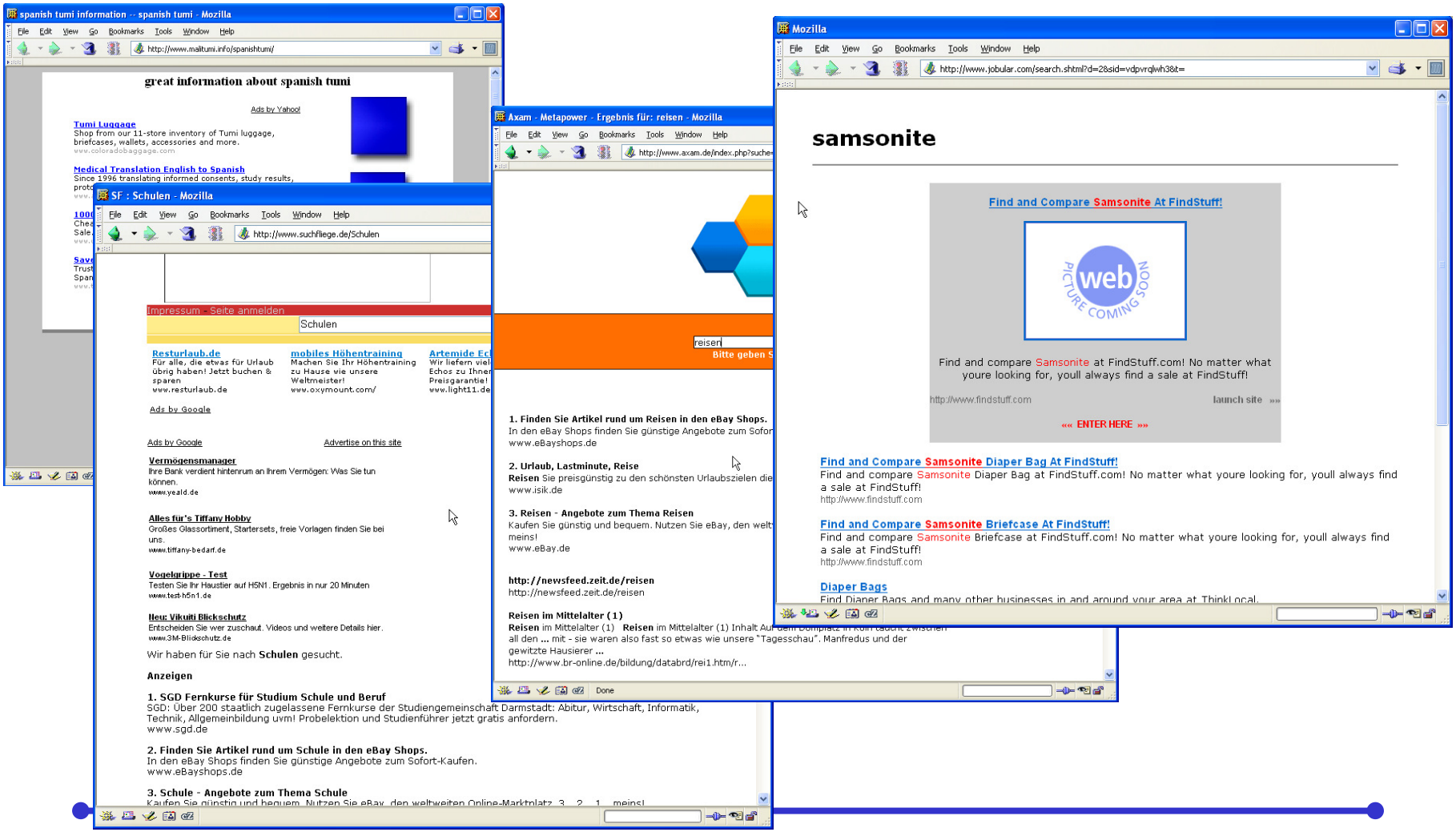
- Aktuell: Die Vermarkter (Amazon. Ebay, Google Adsense, Overture u.v.a.m) zahlen dem Betreiber der Website je Klick auf eine Anzeige einige Eurocent

Von der Werbung zum Webspam

- Webspammer bauen automatisiert Internetseiten (Webspam-Seiten)
- Die Webspam-Seiten sind mit Anzeigen versehen
- Üblicherweise werden alte aber etablierte Domains aufgekauft
- Suchmaschinen erfassen diese Seiten und listen sie als Ergebnisse
- Ein kleiner Prozentsatz der Suchenden wird so zu den Spam-Sites geleitet (je nach Position in der Ergebnisliste)
- Ein noch kleinerer Prozentsatz der Besucher klickt dann auf eine Anzeige

=> in der Menge verdienen große Spammer darüber mehrere Tausend Euro je Monat

Beispiele für Webspam



Erkennen relevanter Informationen

- Bezug zur Suchanfrage?

- Art und Menge der Werbung auf der Seite
- Standardkonformes Design (W3C, ...)
- Sonstiger Inhalt der Website, Thematisierung, Navigation
- Syntaktisch korrekte Sprache (Schreibfehler)
- Semantisch konsistenter Schreibstil/Formulierungen
- Aktualität
- Autor/Impressum

- Reverse-Suche über alternative Suchmaschinen (link:)

Geschäftsmodelle von Suchmaschinen

Der Betrieb einer Suchmaschine kostet Geld. Wie können die Ausgaben finanziert werden?

- Banner-Werbung
- Einbindung von contextbezogener Werbung
- Plazierung von "Sponsored Links" in den Ergebnislisten
- Individualisierte und zielgruppenbezogene Werbung / Ergebnisseiten
- Beratung und Projekte
- Lizenzen, Produkte

Daneben gibt es verschiedene Suchmaschinen, die von Forschungseinrichtungen, Vereinen oder im Rahmen des Unternehmensimages betrieben werden.

Marktanteile

	03.07.2003	07.06.2004	11.10.2005	12.04.2006	Umsatz 2004	Umsatz 2005
Google	67.1%	75.1%	83.1%	84%	3,2 Mrd USD	6,2 Mrd USD
Yahoo	7.1%	5.8%	4.2%	4,00%	3,5 Mrd USD (inkl. Overture)	5,3 Mrd USD
MSN Web-Suche	7.7%	5.1%	4.1%	4,30%	(2,2 Mrd USD)	(2,2 Mrd USD)
AOL Suche	1.5%	2.7%	2.5%	2,30%	-	-
T-Online	3.4%	2.4%	1.6%	0,50%	-	-
Altavista	1.4%	0.9%	0.4%	0,50%	(Yahoo)	(Yahoo)
WEB.DE	1.3%	1.5%	0.5%	0,30%	43 Mio. Euro	(United Internet: 420 Mio. Euro)
MetaGer	1.5%	1.2%	0.4%	0,40%	-	-
Lycos (Europa)	2.1%	1.0%	0.4%	0,70%	103 Mio. Euro	-

zum Vergleich:

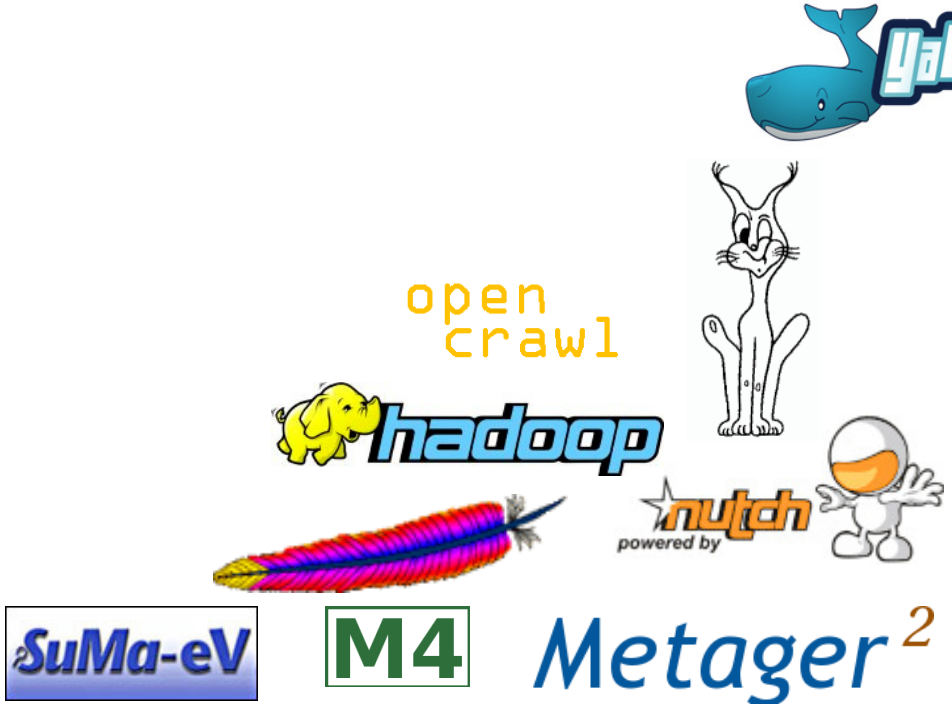
- Gruner + Jahr AG & Co. KG: Umsatz 2004: 2,4 Mrd. Euro
- Axel Springer AG: Umsatz 2005: 2,4 Mrd. Euro

Quelle:http://www.webhits.de/deutsch/index.shtml?/deutsch/webstats.html, Pflichtveröffentlichungen

Fazit

- Alternativen
- Zusammenfassung
- Was tun?
- Kontakt

Alternativen



Zusammenfassung

- Es gibt ein Suchmaschinenmonopol
- Es gibt ein Vermarktungsoligopol
- Qualität der Suchergebnisse wird gezielt untergraben und weiter sinken.
  => Risiko durch falsche Informationen
- Qualitätssicherung und Bewertung der Ergebnisse muss durch den Benutzer erfolgen.
  => versteckte Kosten der Informationsgewinnung
- Treiber der aktuellen Entwickung sind die Marktführer im Suchmaschinenmarkt
  => Ausbau der Monopolstellung
- Die Markt-Monopole verhindern die Erfassung des "deep"-Web, Spam-Bekämpfung, ...
  (Synergie von Suchmaschinenbetreibern und -optimierer)
- Der freie Zugang zu den Informationen des Internets wird schwieriger.
  (Möglichkeit der Zensur, nur wenige Suchmaschinen sind das Ziel der Spammer)

Was tun?

Was für Alternativen gibt es für Unternehmen / Interessensgruppen?

- Schaffung/Identifikation hochwertiger Informationsquellen
- Aufbau eigener spezialisierter Suchmaschinen

Was kann der Einzelne tun?

- Alternativen ausprobieren und Feedback geben.
- Recherchieren ist nicht nur "googlen" oder "wikipedia'n"!
  Fördern Sie den kritischen Umgang mit den Suchergebnissen!

Kontakt

Dipl.-Ing. Michael Nebel

eMail: michael@nebel.de
Telefon: +49 40 851 581 45

Freier Berater für Internet-Dienste und Suchmaschinentechnologie

Mitglied und technischer Beirat des SuMa-eV
Gemeinnütziger Verein zur Förderung der Suchmaschinen-Technologie und des freien Wissenszugangs

Mitarbeiter des Suchmaschinenlabors am RRZN / Universität Hannover (metager.de)

Betreiber der Suchplattform opencrawl.de

Betreiber der Suchmaschine netluchs.de

Links

welche Suchmaschine füttert wen:

- http://www.ihelpyou.com/search-engine-chart.html,
- http://www.zeromillion.com/webmarketing/SERP-distribution.html

Alternative Suchmaschinen:

- http://www.metager.de/
- http://www.metager2.de/
- http://www.netluchs.de/
- http://www.yacy.net/

Simulator einer Suchmaschinensicht:

- http://playground.nebel.de/sumaview/sumaview/
- http://www.gritechnologies.com/tools/spider.go(Poodle Predictor)

Backlink Auswertung:

- http://www.webuildpages.com/neat-o/
- http://www.seo-consulting.de/online-tools/backlink-link-check.php