Google für Fortgeschrittene

Prof. Edy Portmann: http://human-ist.unifr.ch/ , edy.portmann@unifr.ch.

Was ist Information Retrieval (IR)?

- Ein **Werkzeug** mit dem **Informationen** ausgewählt werden
- **Interaktion** mit Informationen
  - Informationsbedarf des Benutzers muss dem System übermittelt werden
  - Die gefundenen Informationen bzw. eine geeignete Darstellung wird dem Benutzern präsentiert

- Zentrale **Probleme:**
  - Umsetzung des menschlichen Informationsbedarfs in eine für die Maschine verständliche Form
  - Darstellung der maschinengerecht vorliegenden Informationen in eine für Menschen geeignete Form

IR Basiskonzepte

- Wer befasst sich damit?
  - **Informatik**: Technologie, Computerwissenschaften, Datenstrukturen
  - **Linguistik**: Sprache, Konzepte
  - **Informationswissenschaft**: Systeme, Werkzeuge und Strategien des Suchens, Digitale Bibliotheken

- Retrievalsoftware: Volltext-Suchmaschinen, kommerzielle Datenbanksysteme, Suchmaschinen im Web

Linguistik und IR

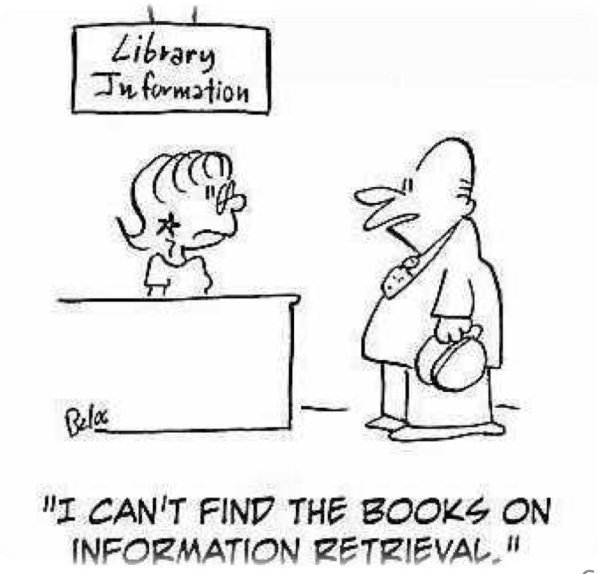
- Sprache
    - Strings
    - Worte
    - Terme
    - Phrasen

- Konzepte
    - Textsammlungen, Kollektionen (Typen, Charakteristiken, statistische Analysen, erkennen von Konkordanzen)
    - Indexierung und Klassifikation
    - Such-
        - Hilfen
        - Prozesse

Geschichte des IR

- Bei der Entwicklung von Informationssystemen spielten folgende Faktoren eine grosse Rolle:
    - Informationsbedarf
    - Technische Machbarkeit
- Computer waren teuer und gross
- Erste Systeme um den Bestand an wissenschaftlicher Literatur zu verwalten
- Bibliothekskontext, Katalogsysteme
    - Klassische Aufgabenstellung
    - Inhaltliche Beschreibungen
- **Information Retrieval**

Websuche ist eine Form von IR



Suche im WWW


Technorati


Google


Ask
.com


YAHOO!


bing


Wolfram
Alpha


LYCOS


Baidu 百度


see & find
quintura

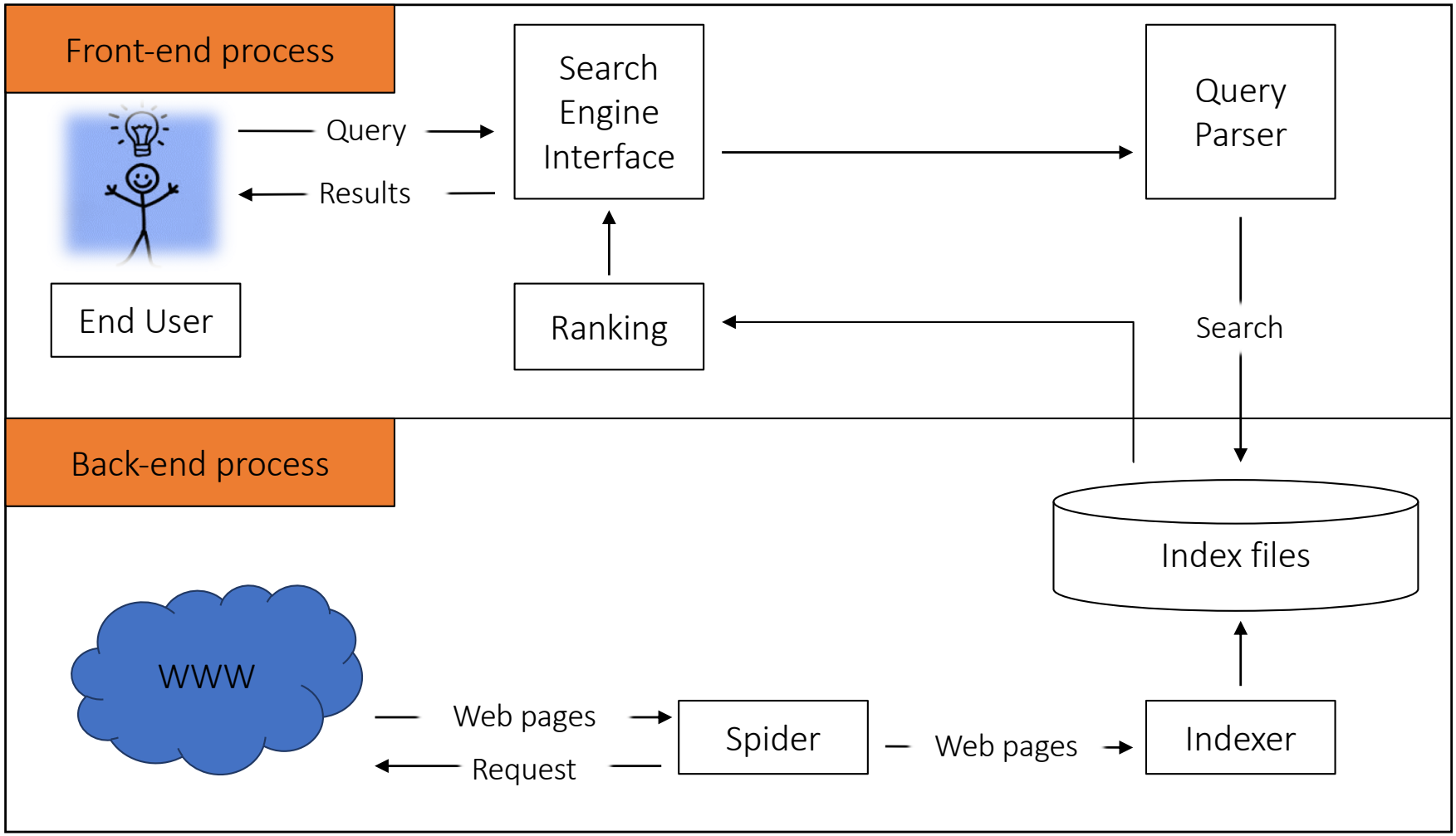

altavista


dmoz

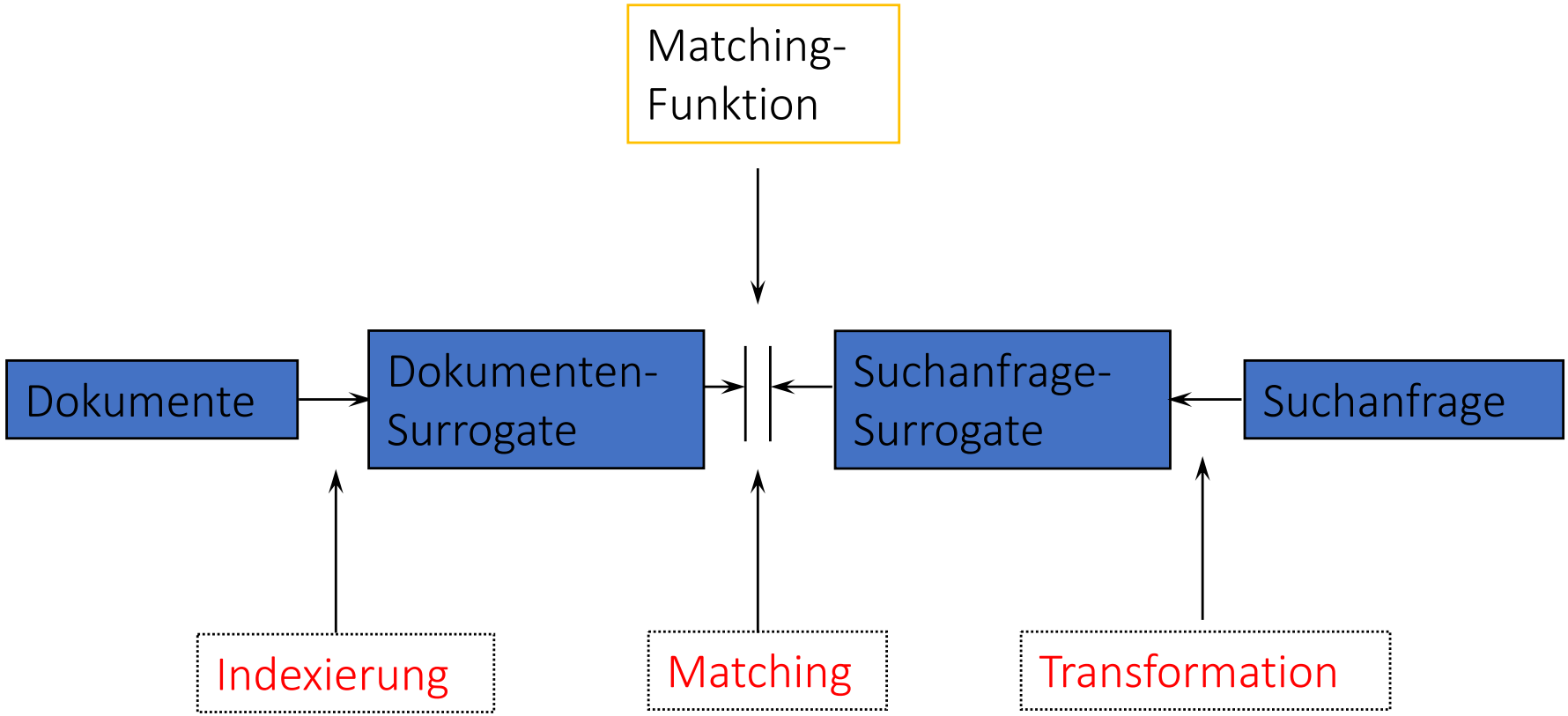
WWW-Suchdienste

- **Suchmaschinen**
    - Zentrale Systemarchitektur
    - Verteilte Systemarchitektur
- **Meta Suchmaschinen**
    - MetaGer (http://meta.rrzn.unihannover.de)
- **Spezialisierte Suchmaschinen** und Datenbanken
    - Gefilterte Informationen zu speziellen Gebieten
    - Strukturiert durch Portale
    - Informationsflut des Internets wird eingegrenzt
    - Oft breites Zusatzangebot an Materialien und Tools
    - Themenspezifische Suchdienste oder Datenbanken
    - Email-Adressen
    - News

Zentrale Architektur von Suchdiensten



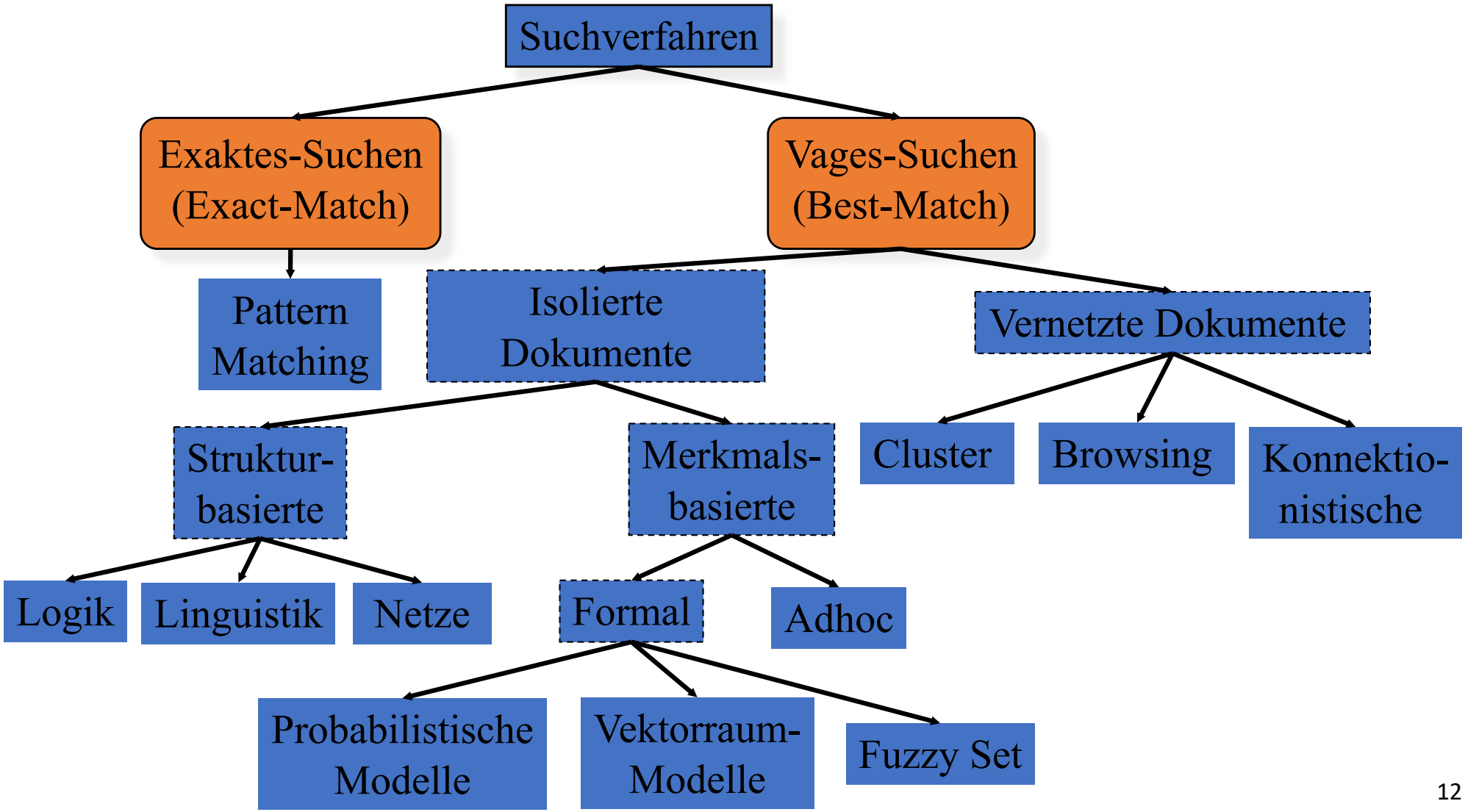
Einfaches Information Retrieval Modell



Arbeitsweise von Suchdiensten

- Strategien von Spidern
    - Parallele Spider
    - Aufteilung der Web nach Domänen, Adressbereichen
    - Startliste von URLs
    - Besuchsintervalle bei der Indexierung
    - Schnelle Indexierung bei ‚bezahlten' URL
- Heuristiken
    - Breadth First versus Depth First
    - Volldurchlauf?
- Probleme
    - Last auf Netz und Server
    - Ausschluss der Indexierung durch Robot Exclusion
    - Keine Indexierung dynamisch generierten Seiten, Datenbanken, etc.

Suchverfahren der Maschinen



Hauptaufgaben des Web Information Management

- **Zugang zur Information**
    - Suche (Suchmaschinen, z.B. Google, Altavista)
    - Navigation (Browsers, z.B. IE, Firefox)
    - Filtering (Empfehlungssystem, z.B. Amazon)
- **Organisation der Information**
    - Kategorisierung (Web Directories, z.B. Yahoo!, Open Directory Project, DMOZ)
    - Clustering (Hitliste wird kategorisiert, z.B. Vivisimo)
- **Information Discovery** (Information Mining, z.B. Enterprise Information Portal (EIP) von IBM)

Möglichkeiten und Probleme

- Übersichtsseiten im WWW sind meistens von einzelnen Personen ehrenamtlich zusammengestellt (ähnliche Probleme wie beim persönlichen Nachfragen).
- Suchmaschinen beschränken sich auf die Suche nach einzelnen Wörtern.
- Die meisten Dokumente im WWW sind nicht systematisch strukturiert.
    - Nur HTML-XML-Dokumente, keine animierte Seiten
    - Auf Angebote in Datenbanken und speziellen Infomationsseiten können Suchmaschinen nicht zugreifen.
- Im WWW ist es nur bedingt möglich, die Richtigkeit zu überprüfen.

Eigenschaften von Information im WWW

- **“Unendliche” Informationsmenge** (Surface vs. Deep Web)
  - Surface = statische Webseiten (HTML)
  - Deep = dynamisch generierte Webseiten (DB’s)
- **Semistrukturierte Dokumente** (kein festes Schema)
  - Strukturierte = HTML tags, hyperlinks, etc.
  - Unstrukturierte = Text
- **Unterschiedliche Formate** (PDF, MS-Word, PS, ...)
- **Multimedia** (Textuelle, Audio, Bilder, ...)
- Grosse Qualitätsunterschiede („Viel Mist“)
- Alles ist publizierbar!
- **Herausforderung** ist: Verteilte Daten, Fluktuation, unterschiedliche Qualität, heterogene Daten

Herausforderungen

- **Verteilte Daten**
  - Auf verschiedenen Plattformen, welche mit unterschiedlichen Bandbreiten vernetzt sind
  - Was ist ein Dokument, was kann indexiert werden?
- **Grosse Fluktuation** der Daten
  - Daten können leicht verändert werden
- **Unterschiedliche Qualität** der Daten
  - Falsche Informationen
  - Veraltet / keine Aktualisierungsautomatik
  - Schlecht geschrieben
  - ‚Unstrukturierte' und redundante Daten
- **Heterogene Daten**
  - Format, Multimedia
  - Verschiedene Sprachen, Zeichensätze

How do I do it?

Bewusst sein, was effektiv gesucht wird und in welchem Kontext

Wo und wie könnte die gewünschte Information aufgeführt sein

Überlegungen vor der Suche können viel Arbeit ersparen

Stichwort: VORBEREITUNG

Wie Suchen?

- Welche Stichwörter beschreiben das Problem besonders gut?
- Wörter, die **spezifisch für die Fragestellung** sind, aber **so allgemein, dass sie in jedem "wichtigen" Artikel vorkommen**.
- Allenfalls erste grob Information über Wikipedia, um anschliessend gezielter zu suchen

- **Beispiel**:
  - Eine Agentur erhält von einem Institut für Rechtsmedizin den Auftrag, eine Informationsbroschüre für Laien zum Thema DNA-Vaterschaftsanalyse zu erstellen. Das Grundlagenmaterial des Kunden für die Arbeit besteht aus einigen wissenschaftlichen Artikeln, den Informationen auf der Website des Instituts und einigen Passagen aus Jahresberichten vergangener Jahre. Aufgabe des zuständigen Redaktors.

Mögliches Vorgehen

1. Verständnis der Thematik fördern – Erst Information
   - Beispielsweise über Wikipedia
     - Die verschiedenen Quellenartikel des Wikipedia-Eintrages durchgehen und anschauen
     - Eröffnet neue Ideen zur Weitersuche oder zu effizienten Keywords

Weblinks [Bearbeiten]

Wikinews: Bundesverfassungsgericht: Heimliche Vaterschaftstests sind unzulässig – Nachricht

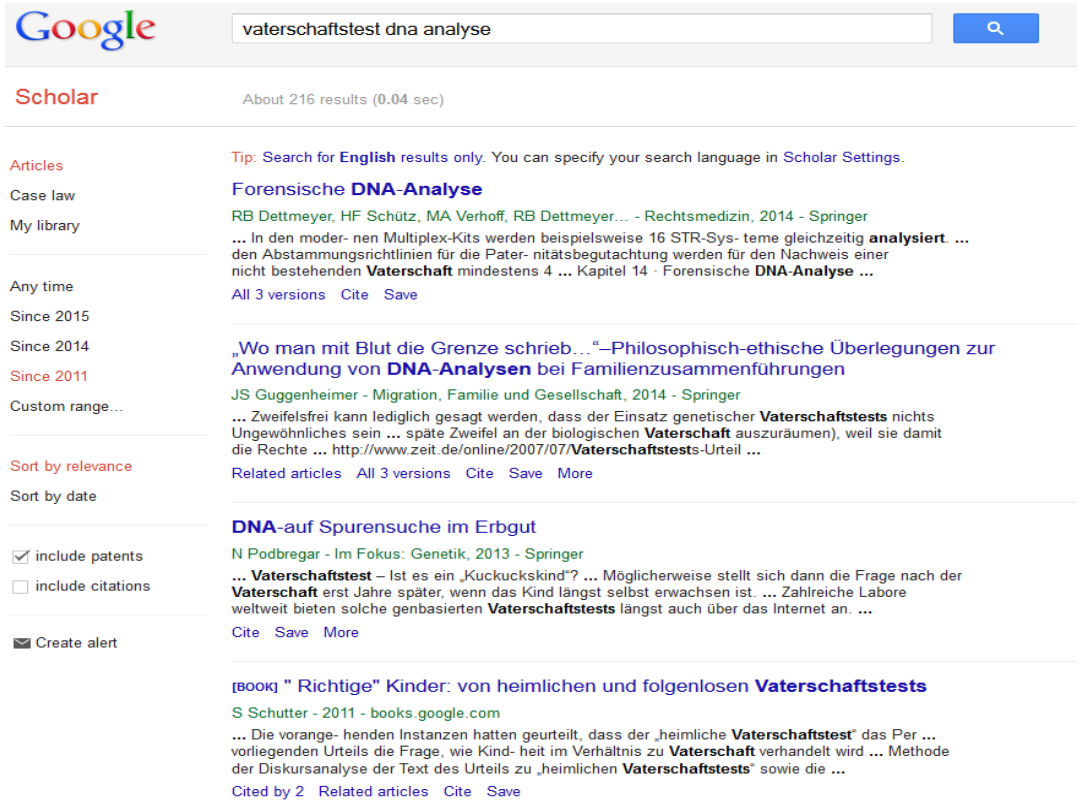
- Rechtsmedizin Uni-Mainz (Memento vom 18. Mai 2006 im *Internet Archive*) Methoden der DNA-Analyse bei Abstammungsbegutachtung und forensischer Spurenkunde (im Internetarchiv)
- Beschlossener Gesetzentwurf des Gesetzes zur Klärung der Vaterschaft unabhängig vom Anfechtungsverfahren (PDF; 260 kB)
- Neue Regeln für Vaterschaftstests Bundestag stimmt Gesetzentwurf zu (Memento vom 10. März 2009 im *Internet Archive*) von den Seiten des Deutschen Bundestages
- Bundestag verabschiedet Gesetz zur Vaterschaftsfeststellung (22. Februar 2008) (Memento vom 4. März 2008 im *Internet Archive*)
- Gesetz zur Klärung der Vaterschaft unabhängig vom Anfechtungsverfahren im Bundesgesetzblatt (BGBl. 2008 I S. 441) (PDF; 53 kB)

Einzelnachweise [Bearbeiten]

1. ↑ BVerfG, Beschluss vom 18. August 2010, Az. 1 BvR 811/09, Volltext.
2. ↑ OLG Stuttgart, Beschluss vom 10. August 2009, Az. 17 WF 181/09, Volltext; Abs. 5 bis 7.
3. ↑ § 17 Abs. 3 Nr. 2 Gendiagnostikgesetz.
4. ↑ BGH, Urteile vom 12. Januar 2005 (PDF; 34 kB), Az. XII ZR 60/03, Volltext und Az. XII ZR 227/03 (PDF; 42 kB), Volltext.
5. ↑ BVerfG, Urteil vom 13. Februar 2007, Az. 1 BvR 421/05, Volltext.
6. ↑ § 26 Abs. 2 Gendiagnostikgesetz.
7. ↑ Mitteilung der Gendiagnostik-Kommission (GEKO) vom 10. September 2010 (Memento vom 19. September 2010 im *Internet Archive*), abgerufen am 10. Mai 2015.
8. ↑ Petra Gehring: *Heimliche Vaterschaftstests: Biowaffen im Geschlechterkampf*. In: thema forschung *(TU Darmstadt) 2/2005, S. 31–33.*
9. ↑ aerztezeitung.de: Männer sind für heimliche Vaterschaftstests
10. ↑ Dokumentation: Gesetz der Bundesregierung-Entwurf eines Gesetzes zur Klärung der Vaterschaft unabhängig vom Anfechtungsverfahren, in: Familie, Partnerschaft, Recht (Fachzeitschrift) 2007, Seite 403
11. ↑ *Zypries will heimliche Vaterschaftstests verbieten lassen.* In: *Spiegel Online*. 3. Januar 2005, abgerufen am 21. Januar 2011.
12. ↑ Dietmar Hipp: *Kuckucksei im Nest*. In: *Der Spiegel*. Nr. 7, 2007 (online).
13. ↑ OLG Stuttgart, Beschluss vom 11. Juli 2008, Az. 8 WF 102/08, Volltext.
14. ↑ Anlage 2 (zu § 10 Abs. 1 JVEG).

Mögliches Vorgehen

2. Google Scholar zur Hand nehmen zur Überprüfung von neueren Artikel und der Relevanz der in Wikipedia aufgeführten Artikel
   - Anzahl Zitierungen kann ein Anhaltspunkt sein
     - Suchen sowohl auf Englisch wie auch auf Deutsch durchführen
       - Querlesen der Artikel



Mögliches Vorgehen

3. Mit dem neuen Wissen verfeinerte (Google-) Suchen (nutzen von Operatoren und der Erweiterten Suche) durchführen
   - Bspw. Land einschränken oder ähnliches
4. Herkunft der Infos kritisch hinterfragen und Querchecken
   - Auf was für einer Seite befinde ich mich, habe ich bereits ähnliche Informationen gesammelt , etc.
5. Gesammeltes Sichten und Verarbeiten

Probleme

- Homographen
  - Gleich klingendes Wort
  - Lehre/Leere
- Polyseme
  - Worte mit mehreren Bedeutungen (historisch bedingte Gleichheit für dieselbe Bedeutung)
  - Bank: Sitzgelegenheit oder Geldinstitut
- Flexionsformen
- Derivationsformen
- Komposita
- Vagheit – Unschärfe – Unsicherheit
  - Sowohl in der Anfrage wie auch in der gegebenen Information

Kompositazerlegung/Motivation

- Im Deutschen treten Komposita sehr häufig auf; werden “spontan” gebildet.
- Vollständige Erfassung sämtlicher Komposita quasi nicht möglich.
- aber: Rückführung auf Grundformen möglich, endliche Menge von Regeln zur Kompositazerlegung
    - z.B. Einfügung eines -s: Recht + Streit = Rechtsstreit
    - z.B. Einfügung eines -n: Bauer + Kind = Bauernkind
- Dies ermöglicht:
    - Das Auffinden von Dokumenten, die nur einen Teil eines Kompositas enthalten, das in der Anfrage spezifiziert wurde.
    - Das Auffinden von Dokumenten, die Komposita enthalten, in welchen ein Term der Suchanfrage vorkommt.
- Verarbeitungsverfahren analog zur morphologischen Verarbeitung

Grundsätzliche Fragestellungen

- Welche Möglichkeiten zur Spezifikation einer Anfrage gibt es?
- Wie werden diese Anfragen bearbeitet?
- Welche standardisierten "Sprachen" (Protokolle) zur Abfrage von Datenbeständen (Online-Datenbanken, Suchmaschinen) gibt es?

Anfragesprache

- Informationssuche ist die Suche nach Information mittels einer Sprache. Diese Sprache wird als Anfragesprache (Retrievalsprache) bezeichnet.
    - Formale Sprachen
        - Abstrakte Sprache wie z.B. Programmiersprachen
    - Anfragesprachen können wie folgt eingeteilt werden:
        - nach Einsatz und Benutzungsgrad
        - nach Dokumententyp
- Beispiele:
    - Suchmaschinen
        - Anfragen mit Stichworten
    - Faktenretrieval SQL (Structured Query Langage)
        - Anfragen auf Tabellen, Datenbanken
    - Anfragen auf Dokumentenstrukturen XML (XQuery)

Anfragen mit Stichworten (1)

- Grundlegende Annahme:
    - Die kleinste Einheit in einem Dokument sind Worte.
- Einfachste Anfrage: ein Wort
    - Je nach System eine Folge von Buchstaben, Zahlen, Sonderzeichen
- Je nach IR-Model muss mindestens eines der Anfrageworte im Dokument enthalten sein.
- Ranking:
    - Gewichtung mittels *term frequency* und *inverse document frequency.*

Anfragen mit Stichworten (2)

- Boolesche Anfragen
  - Aussagenlogische Formeln
  - Operatoren: AND, OR, NOT (oftmals nur AND NOT)
  - Kombination mit regulären Ausdrücken
    - Sing* AND (Beatles OR "Pink Floyd")
- Natürlichsprachliche Anfragen
  - Keine Beschränkung bzgl. der Anfragenformulierung
  - Schwierigkeiten bei der Verarbeitung der natürlichen Sprache

Anfrageverfeinerung durch Operatoren (1/2)

Symbol	Verwendung
+	Suche nach Google+ Seiten oder Blutgruppen Beispiele: `+Chrome` oder `AB+`
@	Suche nach sozialen Tags Beispiel: `@agoogler`
$	Suche nach Preisen Beispiel: `nikon 400 $`
#	Suche nach beliebten Hashtags zu Trendthemen Beispiel: `#throwbackthursday`
–	Wenn Sie einen Bindestrich vor einem Wort oder einer Website platzieren, werden Ergebnisse entfernt, die dieses Wort oder diese Website enthalten. Dies kann bei Wörtern mit verschiedenen Bedeutungen nützlich sein, etwa bei "Jaguar", bei dem Ergebnisse zur Automarke und zum Tier ausgegeben werden. Beispiele: `jaguar geschwindigkeit -auto` oder `panda -site:wikipedia.org`

- Mit dem Operator ~ werden ebenfalls nach Synonymen des nachfolgenden Wortes gesucht
- Mit | können Begriffe separiert werden (entweder das eine oder das andere) Suchmenge wird grösser.

Quelle: https://support.google.com/websearch/answer/2466433?hl=dec

Anfrageverfeinerung durch Operatoren (2/2)

"	Wenn Sie ein Wort oder eine Wortgruppe in Anführungszeichen setzen, enthalten die Ergebnisse nur solche Seiten, auf denen die Wörter in der gleichen Form und in der gleichen Reihenfolge wie innerhalb der Anführungszeichen vorkommen. Verwenden Sie Anführungszeichen nur dann, wenn Sie nach einem ganz speziellen Wort oder einer exakten Wortgruppe suchen. Ansonsten könnten Sie viele hilfreiche Ergebnisse versehentlich ausschließen. Beispiel: `"imagine all the people"`
*	Fügen Sie ein Sternchen als Platzhalter für alle unbekannten Begriffe hinzu. Beispiel: `"wer den * nicht ehrt, ist des * nicht wert"`
..	Trennen Sie Zahlen durch zwei Punkte ohne Leerzeichen, um Ergebnisse mit einem bestimmten Zahlenbereich zu erhalten. Beispiel: kamera `50..100 $`

> Mit mit **filetyp:** pdf oder csv usw. kann die Suche auf bestimmte Dokumenttypen eingeschränkt werden.

Quelle: https://support.google.com/websearch/answer/2466433?hl=de

Weitere hilfreiche Möglichkeiten

Suchoperator	Verwendung
`site:`	Gibt Ergebnisse für bestimmte Websites oder Domains aus Beispiele: `olympiade site:zeit.de` und `olympiade site:.de`
`link:`	Mit diesem Operator können Sie nach Webseiten suchen, die auf eine bestimmte andere Seite verweisen. Beispiel: `link:youtube.com`
`related:`	Mit diesem Operator finden Sie Websites, die einer bestimmten Webadresse ähneln. Beispiel: `related:spiegel.de`
`OR`	Mit diesem Operator finden Sie Webseiten, die irgendeinen von mehreren Begriffen enthalten. Beispiel: `marathon OR lauf`
`info:`	Dieser Suchoperator liefert Informationen zu einer Webadresse. Dazu zählen unter anderem die im Cache gespeicherte Version der Seite, ähnliche Seiten und Seiten, die auf die angegebene Website verweisen. Beispiel: `info:google.de`
`cache:`	Mit diesem Suchoperator können Sie die Seitenversion abrufen, die Google beim letzten Besuch der angegebene Website zwischengespeichert hat. Beispiel: `cache:bundesregierung.de`

Quelle: https://support.google.com/websearch/answer/2466433?hl=de

Anfragen auf Dokumentstrukturen

- Integration von inhaltlichen und strukturellen Aspekten in Anfragen

- Auswertung: inhaltlicher Anteil der Anfrage liefert eine Menge von potentiellen Ergebnissen, auf der dann die strukturellen Constraints überprüft werden.

- Unterschiedliche Arten von Text-Struktur erlauben unterschiedliche Anfragen.
    - Festgelegte Struktur
    - Hypertext-Struktur
    - Hierarchische Struktur

Hypertext-Struktur / Beispiele

- "Finde Informationen über Bern, eingeschränkt auf Domäne Schweiz."
    - *SELECT d.url, d.title*
      FROM Dokument d SUCH THAT d ***MENTIONS*** *“Bern"*
      WHERE d.url ***CONTAINS*** *".ch";*

- "Finde Dokumente über Informatik, die auf die Universität Bern verweisen."
    - *SELECT d.url, d.title*
      FROM Dokument d SUCH THAT d ***MENTIONS*** *"Informatik",*
      Verweis v SUCH THAT basis = d
      WHERE v.anker ***CONTAINS*** *"unibe.ch“*

    - MENTIONS: Wird durch Suchanfrage an eine globale Suchmaschine (z.B. Google) realisiert; Zwischenergebnis wird weiterverarbeitet
    - CONTAINS: Lokale Zeichenkettensuche innerhalb eines Strings

Operationen auf Anfragen

- Initiale Spezifikation des Informationsbedürfnisses oft unzureichend, dementsprechend schlechte Qualität der Ergebnisse nach dem ersten Retrievalschritt
- Reformulierung notwendig, aber durch den Benutzer oft schwierig, da Kenntnisse über Dokumentensammlung kaum vorhanden
- Ziel: automatische Reformulierung
- Basis zur automatischen Reformulierung:
    - Rückmeldungen vom Benutzer über die Relevanz einzelner Dokumente aus dem Ergebnis (Relevance Feedback)
    - Analyse der initial gefundenen Dokumente im Ergebnis (lokal)
    - Analyse des gesamten Datenbestandes (global)

Erweiterung/Modifikation der Anfrage der Termgewichte im VSM

- (Cluster-)Annahmen
    - Gewichtsvektoren der als relevant markierten Dokumente sind einander ähnlich
    - Gewichtsvektoren von irrelevanten Dokumenten sind den obigen Vektoren nicht ähnlich

- Idee: Anfragevektor derart modifizieren, dass er "stärker" in Richtung der relevanten Dokumente zeigt

Stemming und Lemmatisierung

- Motivation:
- Simultane Suche nach allen morphologischen oder orthographischen Varianten.
- Verbesserung des Recalls ohne Verschlechterung der Precision.
- "Einfache Lösung": Benutzer muss durch Verwendung von Trunkationsoperatoren (*) oder Disjunktion über alle Formen selbst dafür Sorge tragen.
- Problem bei Trunkierung: Es werden ungewollte Fortsetzungen erzeugt: auto* findet Auto und Autos, aber auch automatisch, Autor oder Automorphismus usw.

Lemmatisierung

- Reduktion der Wortformen auf ihre Grundform (und weitere Information) durch Nachschlagen in einem elektronischen Wörterbuch.
- Vollformenlexikon: Jede Wortform kann direkt im Lexikon nachgeschlagen werden.
- Grundformenlexikon
    - Wortform wird durch morphologische Regeln auf eine potentielle Grundform reduziert, die dann im Lexikon nachgeschlagen wird.
- Vollformenlexikon
    - Ist aufwendiger hinsichtlich Speicherplatz aber effizienter bei der Verarbeitung

Nachteile von Lemmatisierung

- Erfordert umfangreiches elektronisches Wörterbuch, aufwendig in der Erstellung und Wartung.
- Relativ hohe Anforderungen an Verarbeitungszeit oder Speicherplatz.
- Was passiert mit Wortformen, die nicht im Lexikon gefunden werden?
    - Eigennamen
    - Komposita im Deutschen

Probleme bei der Kompositazerlegung im Deutschen

- Verschiedene korrekte Zerlegungen
    - Wachstube in Wachs + Tube oder wach + Stube

- Simplizia können irrtümlich zerlegt werden
    - Pomade -> po + made
    - Proletarier -> prolet + arier
    - Tangente -> tang + ente

- Grosse Anzahl von Zerlegungsambiguitäten, die nicht korrekt sind:
    - Aluminiumherstellung kann auf 12 versch. Arten zerlegt werden, z.B. alu+mini+umher+stellung
    - Alleinerziehende -> all+ein+erzieh+ende

Suche im WWW


Technorati
Google
Ask.com
YAHOO!
bing
LYCOS
Wolfram Alpha
Baidu 百度
see & find
quintura
altavista
dmoz

Alternative Suchmaschinen

- Andere grosse Suchmaschinen *(Bing/Ask/AOL/Blekko)*
- Duckduckgo *(!google dem Begriff voransetzen, wenn trotzdem über Google gesucht werden soll, die Suchanfrage bleibt aber anonymisiert)*
- Wikipedia
- Baidu *(für asiatische Inhalte)*
- Yandex
- Qwant *(Personensuche in Sozialen Medien)*
- Swisscows *(für lokales)*
- Wolfram Alpha *(äusserst stark im wissenschaftlichen Bereichen)*
- CC Search *(für Inhalte unter der Creative Common Lizenz)*
- Topsy *(für das Auffinden von Tweets – durchsucht Twitter)*
- Metasuchmaschinen benutzen *(je nach Art, kann thematisch gruppierter gesucht werden)*

Limitationen heutiger Suchmaschinen (1)

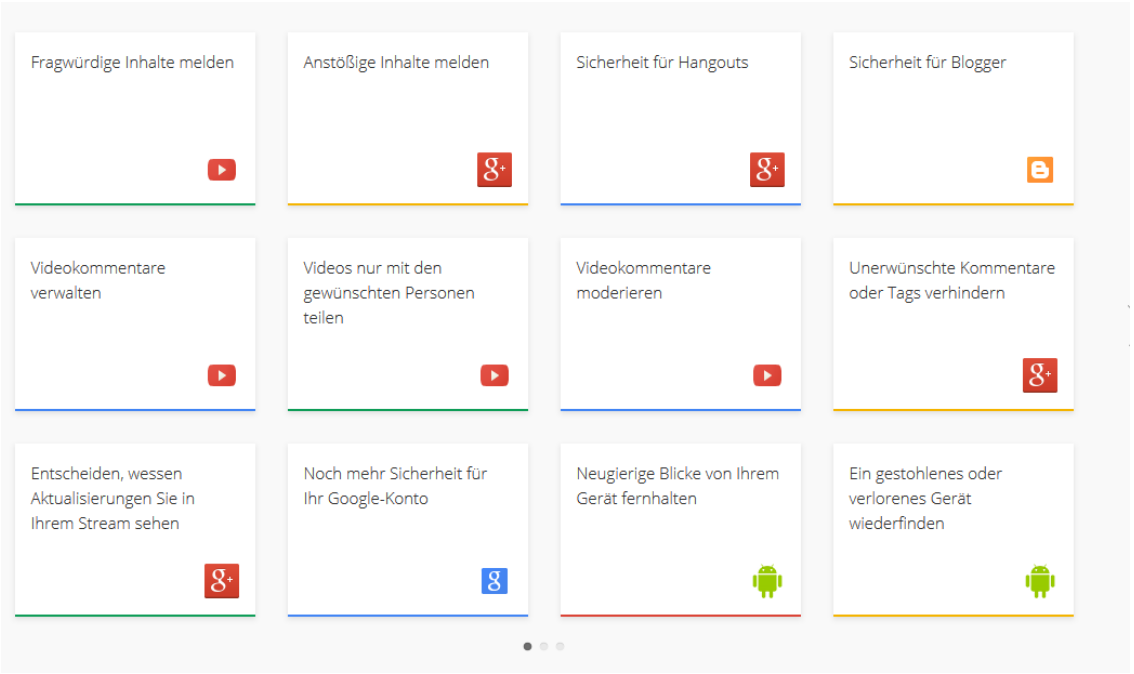
- Suchanfragesprache:
  - Syntax, Lesbarkeit, Verständlichkeit
- Eingeschränktes Verständnis des Dokumenteninhalts (Semantik)
- Ein Sack von Abfrageworte (Zusammenhangslos, Disambiguation) muss verarbeitet werden
- Statistische Gewichtung beim Matching von Suchanfrage zu Dokument
  - Keine Garantie für gute Suchresultate
  - Keine Modellierung des Benutzerkontextes
  - Unterschiedliche Resultate für die gleiche Suchanfrage
  - Ein Benutzer erhält unterschiedliche Resultate zu unterschiedlichen Zeiten und Orten

Limitationen heutiger Suchmaschinen (2)

- Wenig Support für den Suchenden
    - Keine eigentlichen Interaktionen möglich
    - Passive Unterstützung der Suche, keine Empfehlungen
    - Statische Navigation, keine dynamisch generierte Links
    - Schlechte Integration des Retrievals, Empfehlungen und Navigation

Relevanz und Glaubwürdigkeit

Quelle: https://www.google.ch/intl/de_ch/safetycenter/tools/



- Häufig durch User-Community unterstützt, keine Garantie
- Google pflegt einen Relevanz-Algorithmus, der gewährleisten soll das sucherfreundlicher Inhalt höher gewertet wird
- Algorithmus Updates (wie Google Panda, Penguin, Humingbird etc.) bemühen sich darum Qualität des Inhalts zu werten

Sicherheitsangebot von Google

Sicherheitstools von Google

Unsere Produkte bieten Ihnen Tools und Einstellungsmöglichkeiten, mit denen Sie die Internetnutzung Ihrer Kinder steuern können. Die folgenden Abschnitte enthalten Informationen zu SafeSearch, zum sicheren Modus bei YouTube und zum Filtern von Inhalten in Android.

Google SafeSearch

Mit SafeSearch werden Websites gefiltert, die sexuell eindeutige Inhalte enthalten, und aus Ihren Suchergebnissen entfernt. Kein Filter ist absolut sicher, SafeSearch hilft Ihnen jedoch dabei, Inhalte auszublenden, die Sie lieber nicht sehen möchten oder über die Ihre Kinder nicht stolpern sollten.

Standardmäßig ist der moderate SafeSearch-Filter aktiviert, mit dem unangemessene Bilder aus Ihren Suchergebnissen verbannt werden. Wenn Sie Ihre Einstellung ändern und stattdessen die strikte Filterung verwenden, werden sowohl unangemessene Textinhalte als auch Bilder herausgefiltert.

So ändern Sie die SafeSearch-Einstellungen auf Ihrem Computer

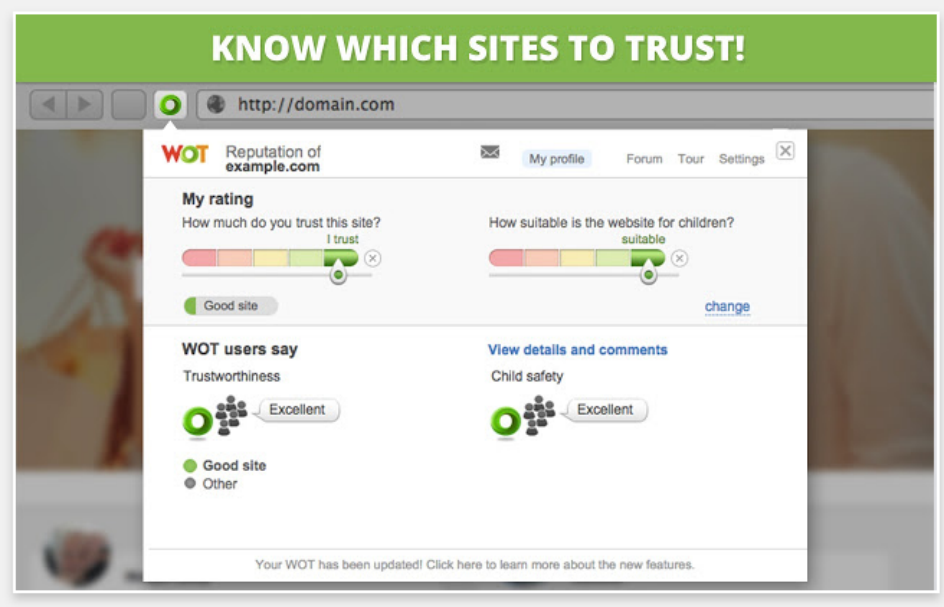
SafeSearch-Sperre

Wenn Sie befürchten, dass andere die strikte SafeSearch-Filterung ohne Ihr Wissen ändern, können Sie die Einstellung über die SafeSearch-Sperre mit einem Passwort schützen. Sobald die Sperre aktiviert ist, ändert sich das Aussehen der Google-Suchergebnisseite, damit der Nutzer weiß, dass SafeSearch aktiviert ist.

Quelle: http://www.google.ch/intl/de/goodtoknow/familysafety/tools/

User hat Möglichkeiten sich zu schützen

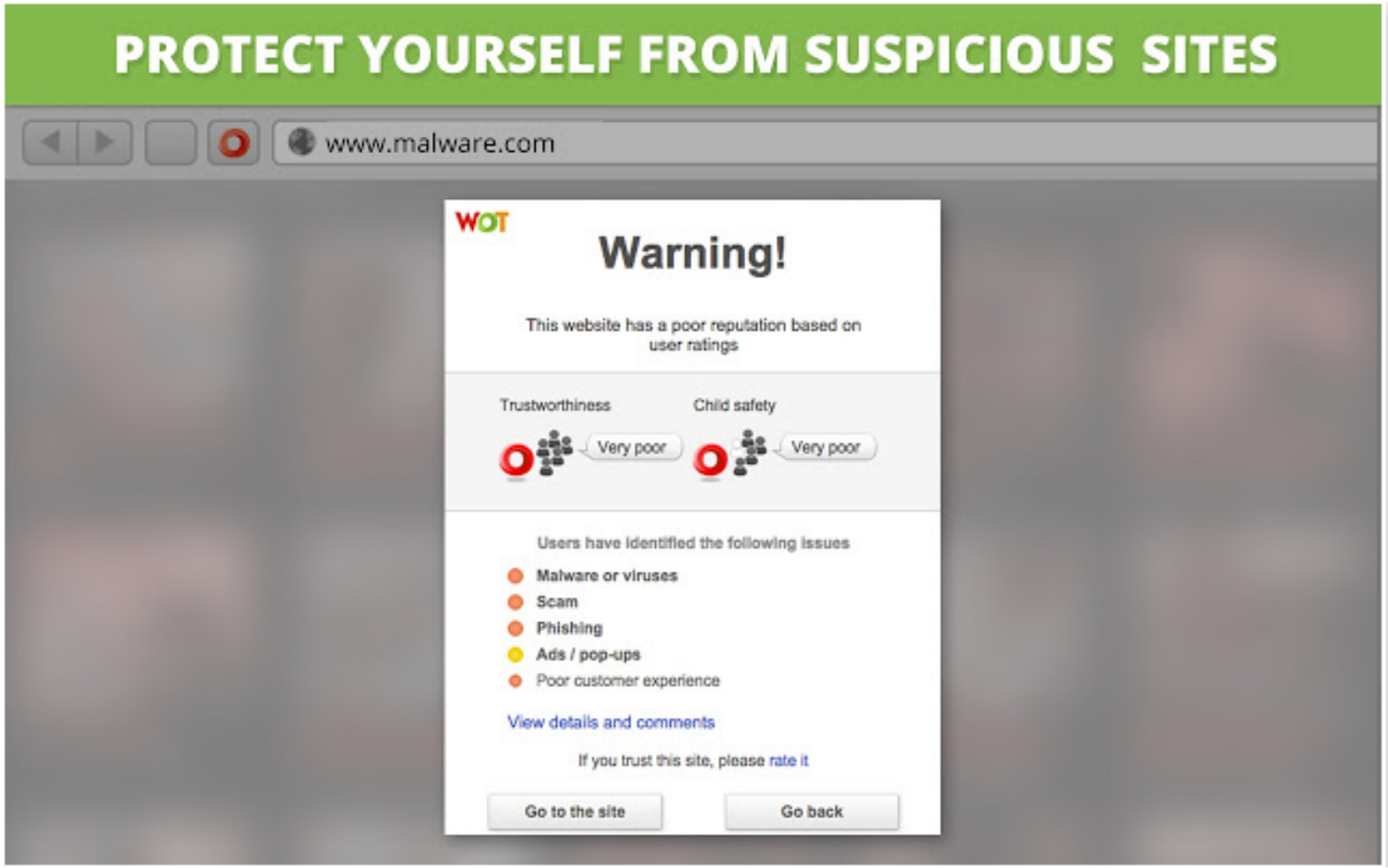
- Tools für Firefox
  - Https Everywhere
  - WoT (Web of Trust)
  - NoScript
  - BetterPrivacy
  - RequestPolicy



- Ähnliche Tools für andere Browser vorhanden
- Web of Trust auch für Chrome erhältlich

Quelle: https:chrome.google.com/webstore/detail/wot-web-of-trust-website/

Web of Trust



Quelle: https:chrome.google.com/webstore/detail/wot-web-of-trust-website/bhmmomiinigofkjcapegjjndpbikblnp?hl=de//

Information: Ein soziales Phänomen

- Fundamentale **Verschiebung**: progressive Verbreitung von sozialen Aktivitäten,
    - Konversationen, Labeling, Rating und Informationsorganisation in Online Netzwerken

- **Warum** gehen Benutzer Online?
    - Um zu kommunizieren
    - Um informiert zu sein
    - Um sich zu unterhalten
    - Um ein teil eines grossen sozialen Netzes zu sein

Zukunft und Herausforderungen

- Soziale Medien beinhalten **Strukturen** wie Benutzer miteinander interagieren
- Viele Daten warten nur auf ihre Nutzbarmachung
    - “Mining” und “Modellierung”
- Sehr heterogene Daten
- Enormer Umfang von Daten
- Lernen aus Daten und Modelle
    - Lösung von schwierigen Problemen oder für neue Anwendungen
- Websuche ist eine relativ junge wissenschaftliche Disziplin

Trends in „Next Generation Search Engines“

- Support für **Reformulierungen** von Suchanfragen
    - „Query by examples“
    - Automatische Generierung von Suchanfragen (Empfehlungen)
- Bessere **Modellierung des Kontextes**
- Bessere **Personalisierung**
- Bessere Dokumentenanalysen („ sentiment analysis“, **Semantik**, etc.)
- **Kontext-sensitives Ranking**

Trends in „Next Generation Search Engines“

- Spezialisierter / **Angepasst an Benutzerbedürfnisse**
    - Spezielle Gruppierungen (Community Suche)
        - *Beispiel*: CiteSeer (CiteSeerx) Scientific Literature Digital Library (Digitale Bibliothek wissenschaftlicher Literatur) ist eine Suchmaschine und Zitationsdatenbank, schwerpunktmässig Informatik
    - Personalisierte Suche
    - Suche innerhalb einer Domäne
- Integration von Suche, **Navigation**, Filtering
- Verbesserung der **Präsentation** der Suchresultate
    - Generierung von Summaries
- **Interaktive** Suche


Besten Dank!