The WWW: A library without librarian

...Why Google isn’t future. Really not!

Herwig Unger & Mario Kubek
FernUniversität in Hagen, Faculty of Mathematics and Computer Science
Phone: +49 2331 9871155, +66 9797922070; Fax: +49 2331 987353
eMail: Herwig.Unger@gmail.com
Google became the all dominating empire.

and: we even don‘t know much of how it works!
PageRank

- Page, Brin, 1998
- evaluation regardless of the contents of the web page
- based solely on its location in the web graph

- **Parameters:**
  - $u$: a node in the web graph
  - $d_i^+$: out degree of a node $i$
  - $w_1, w_2, ..., w_k$: nodes pointing to $u$
  - $\eta$: normalization constant, $<1$
  - $\text{PR}(u)$: page rank of page $u$

- PageRank is given by

$$
\text{PR}(u) = (1 - \eta) + \eta \cdot \left( \frac{\text{PR}(w_1)}{d_1} + \frac{\text{PR}(w_2)}{d_2} + \ldots + \frac{\text{PR}(w_k)}{d_k} \right)
$$
An example: Harry.

Harry !?

Just Harry ???

Crazy ---
Research focus

- Our Approach to look at communication networks:
  - Consider the mutual influences between content, users and user activities as well as network with its parameters and configuration.
Alternatives: GNUTELLA-Query/QueryHit/GET

Query xx.mp3
Query hit : B
Query hit : C
Query hit : D
Alternatives: Freenet and its Search [Hong01]

- A graph structure actively evolves over time
  - New links form between nodes
  - Files migrate through the network
    ⇒ Adaptive routing
    ⇒ Most requested content is found fast

[Abere02]
A dimension problem...

Billions of Web documents $d$

constant to $\ln(d)$ distance for scalability

Approx. 200,000 terms $t$

1  11  12  2  8  7  6  3  10  4  9  5  

Ranking
80% of all information in the WWW is given in a textual form!

→ big challenge to filter relevant information
→ usually 2-3 keywords are a weak description of what the users are looking for
→ the typically received 10000+ search result overload the user and normally only the first 30 results are considered
→ 3 till 6 words may return more precise results but it is hard to find words with high selection rate

The task is to find out fastly what the user is looking for and support him in this process.
Librarians: refining the task determination

... are active intermediaries between users and resources

- Provision, archiving and maintenance of information in many formats
- Referring patron
- Researcher for special topics
- Managing access in an efficient manner → pathfinder, bibliographies, suggestions
- **information literacy** i.e. "... the ability to know when there is a need for information, to be able to identify, locate, evaluate, and effectively use that information for the issue or problem at hand."
What we can do: The local librarian…
Idea 1: Locality
Another example: “christmas tree“

- What does Google offer?
So what's about “christmas tree“

This of course:

But also this:

An assembly of control valves, fittings, pressure gauges and pipes at the top of a well to control the flow of oil and gas after the well has been drilled and completed.

→ Looks like a decorated christmas tree (with some imagination).
Motivation / Problem statement 1: Disambiguation

Disambiguation (also called word sense disambiguation or text disambiguation) is the act of interpreting an author's intended use of a word that has multiple meanings or spellings.

Word sense disambiguation (WSD) is the task of selecting the appropriate senses of a word in a given context.

→ e.g. mouse (animal, comp.)
   christmas tree (biol, oil),
   cube (maths, car)
   Harry (sev.names)
Idea 2: Pictures

- Already in 1911 the expression "Use a picture. It's worth a thousand words." appears in a newspaper article by Arthur Brisbane discussing journalism and publicity.
- The roots of that phrase are even older and have been expressed by earlier writers.
- The Russian writer Ivan Turgenev wrote (in *Fathers and Sons* in 1862), "A picture shows me at a glance what it takes dozens of pages of a book to expound."
Example: Harry

Search for Images: Harry

Image Results for: Harry

Select/deselect all images for analysis

Search for more similar images  Use keyword translation

Search with selected images
Result of text analysis

Select extracted keywords from your document (*):
- Harry
- Potter
- book
- film
- review
- series
- world
- cover
- movie
- school
- wand
- fan
- deathly
- Hermoine
- friend
- Ron
- magic
- video
- wizardry
- Voldemort

Selected search words:
Harry Potter book film

Ungefähr 24.400.000 Ergebnisse (0,58 Sekunden)

Harry Potter Film Wizardry, Brian Sibley, 9780061997815.
Product Description: Immerse yourself in the world of the spectacular Harry Potter film series. Learn why Yule Ball ice sculptures never melt, where Galleons, ...
www.amazon.com/Harry-Potter-Wizardry-Brian.../0061997811

Harry Potter-Inspired Film Series
Sep 12, 2013 ... Expanding their long-term, lucrative partnership on the Harry Potter franchise, Warner Bros and author J.K. Rowling are putting a new film series ...

Harry Potter (film series) - Wikipedia, the free encyclopedia
The Harry Potter film series is a British–American feature film series based on the Harry Potter novels by author J.K. Rowling. The series is distributed by Warner ...
en.wikipedia.org/wiki/Harry_Potter_(film_series)

Filme schauen
www.watchever.de/Filme
Die Online-Flatrate für Filme.
Jetzt 30 Tage kostenlos testen!

WATCHEVER® auf dem iPad
WATCHEVER® auf Apple TV
Filme sofort angucken
Serien sofort gucken
WATCHEVER® auf deinem PC
Breaking Bad – Das Finale

Harry Potter Film Book
www.amazon.de/
Niedrige Preise, Riesen-Auswahl und kostenlose Lieferung ab nur 20 EUR.
***** 1.074 Bewertungen für amazon.de
Idea 3: Systematic presentation of search results

→ From D. Weiss, S. Osinski. Carot Search
http://project.carrot2.org/release-3.5.0-notes.html
Idea 4: User based evaluation of search results
Motivation / Problem statement

- Searching the WWW...
  - Manual query formulation is a tedious and error-prone task

- Evaluating large result sets is time-consuming

So why not let the computer read and find useful web documents for you?
Idea 5: Documents as queries

**Concept:**

- Use documents as the only initial search parameter while browsing

- Technically:
  - extract web (DocAnalyser) or local document’s (FxResearcher) main topics
  - search for topical sources (important inherent, influential aspects / basics)
  - use them as search words (query terms)

- Find similar and related content or track topics in real time (on-line) or when the user is off-line
Try out DocAnalyser for yourself at www.docanalyser.de!

DocAnalyser - Find Similar and Related Web Documents

What is DocAnalyser?
DocAnalyser is a new service that offers you a novel way to **search for similar and related web documents** and to **track topics** without the need to enter search queries manually. You just need to provide a web content to be analysed. DocAnalyser then extracts its main topics and their sources (important inherent, influential aspects / basics) and uses them as search words.

Installing DocAnalyser
In order to be able to use DocAnalyser, please **drag and drop one or both of the following bookmarklets to your bookmarks toolbar** of your favourite web browser:

**Bookmarklet 1:** Analyse this content (analyse currently shown/selected web content)

**Bookmarklet 2:** Analyse a web content (analyse another web content)

**Remark:** The following bookmarklet should only be used when analysis errors occur using the first bookmarklet: Analyse this content
A new concept…
Idea 6: Decentralised search engines (see also YaCy and Faroo)
Idea 7: The librarian of the web

Empty bookshelf

...growth process...

...full shelf 😞

Classify & Sort 😊 !

Catalogue or Order algorithm
The物理analogon:
→ the centre of mass

- words = mass point
- distance vector = distance in co-occ. graph

→ e.g. school is the centroid of a document containing classroom, students, teacher but also computer

→ The centroid of a document is the term with the minimal average distance to all words of the respective document in the co-occ. graph.
Properties of centroids

<table>
<thead>
<tr>
<th>Title of Wikipedia Article</th>
<th>Centroid Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tay-Sachs disease</td>
<td>mutation</td>
</tr>
<tr>
<td>Pythagoras</td>
<td>Pythagoras</td>
</tr>
<tr>
<td>Canberra</td>
<td>Canberra</td>
</tr>
<tr>
<td>Eye (cyclone)</td>
<td>storm</td>
</tr>
<tr>
<td>Blade Runner</td>
<td>Ridley Scott</td>
</tr>
<tr>
<td>CPU cache</td>
<td>cache miss</td>
</tr>
<tr>
<td>Rembrandt</td>
<td>Louvre</td>
</tr>
<tr>
<td>Common Unix Printing System</td>
<td>filter</td>
</tr>
<tr>
<td>Psychology</td>
<td>psychology</td>
</tr>
<tr>
<td>Universe</td>
<td>shape</td>
</tr>
<tr>
<td>Mass media</td>
<td>database</td>
</tr>
<tr>
<td>Stroke</td>
<td>blood</td>
</tr>
<tr>
<td>Mark Twain</td>
<td>tale</td>
</tr>
<tr>
<td>Ludwig van Beethoven</td>
<td>violin</td>
</tr>
<tr>
<td>Oxyrhynchus</td>
<td>papyrus</td>
</tr>
<tr>
<td>Fermi paradox</td>
<td>civilization</td>
</tr>
<tr>
<td>Milk</td>
<td>dairy</td>
</tr>
<tr>
<td>Health</td>
<td>fitness</td>
</tr>
<tr>
<td>Tourette syndrome</td>
<td>tic</td>
</tr>
<tr>
<td>Agriculture</td>
<td>crop</td>
</tr>
<tr>
<td>Malaria</td>
<td>disease</td>
</tr>
<tr>
<td>Fiberglass</td>
<td>fiber</td>
</tr>
<tr>
<td>Continent</td>
<td>continent</td>
</tr>
<tr>
<td>United States Congress</td>
<td>Senate</td>
</tr>
<tr>
<td>Turquoise</td>
<td>turquoise</td>
</tr>
</tbody>
</table>

- The centroid can be a word, which is not contained in any of the documents.
- Often, generalising terms will be found.
- Theoretically, a document may have more than one centroid.
- The distance of two document centroids in the co-occurrence graph can be used to define the similarity of the documents.
- Even to short queries may a centroid term may be assigned.
Properties of centroids

- The centroid can be a word, which is not contained in any of the documents.
- Often, generalising terms will be found.
- Theoretically, a document may have more than one centroid.
- The distance of two document centroids in the co-occurrence graph can be used to define the similarity of the documents.
- Even to short queries may a centroid term may be assigned.

Reference Article: "Abgas-Skandal - Schummel-Motor steckt auch in Audi A4 und A6"
Building a self-specialising hierarchy

Rules of the game

- If a level is full, the local co-occ. graph is partitioned.
- Document links are given to one node of the lower level depending on the location of its centroids. (some words of a document may be in the other partition, however)
- The upper levels remain as a chunky classification of new arriving documents or queries which are later refined
- The co-occ. graph in the lower level will be refined by documents assigned to the respective node
- In case the next node is full, the game is repeated in a successive manner.
Idea 8: Document evaluation

The physical analogon:

→ bouyancy vs. weight

→ User access add air to a thought swim bladder attached to every document
→ Time let air leak from the bladder
Traffic Balancing

(sub-) root nodes overloaded

Rules of the game

✓ Parts of the tree-like structure may be on demand dynamical copied and share load
✓ Entry points on different levels may be established from different stakeholders
✓ Traffic may be adapted to load
✓ Possibly established copies of instances must be kept consistent
Summary and Outlook. A first idea…

... of a fully decentralised search engine.

- No more copying of the whole WWW
- 100% actual information
- As fast as google
- New services
- New interfaces
- No more NSA

Thank you for your time! Q&A.

Prof. Dr.-Ing. habil. Herwig Unger
Herwig.Unger@gmail.com
LINE: hu2106
+49 176 8183 2106 / +66 979 722 070