Fully Decentralised Search Engines: just a dream?

….no, because Google isn‘t future. Really not!

Herwig Unger
FernUniversität in Hagen, Faculty of Mathematics and Computer Science
Phone: +49 2331 9871155, +66 9797922070; Fax: +49 2331 987353
eMail: Herwig.Unger@gmail.com

State of the art…..

- Google became the all dominating empire.
- and: we even dont know much 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

…. the basis of the success of

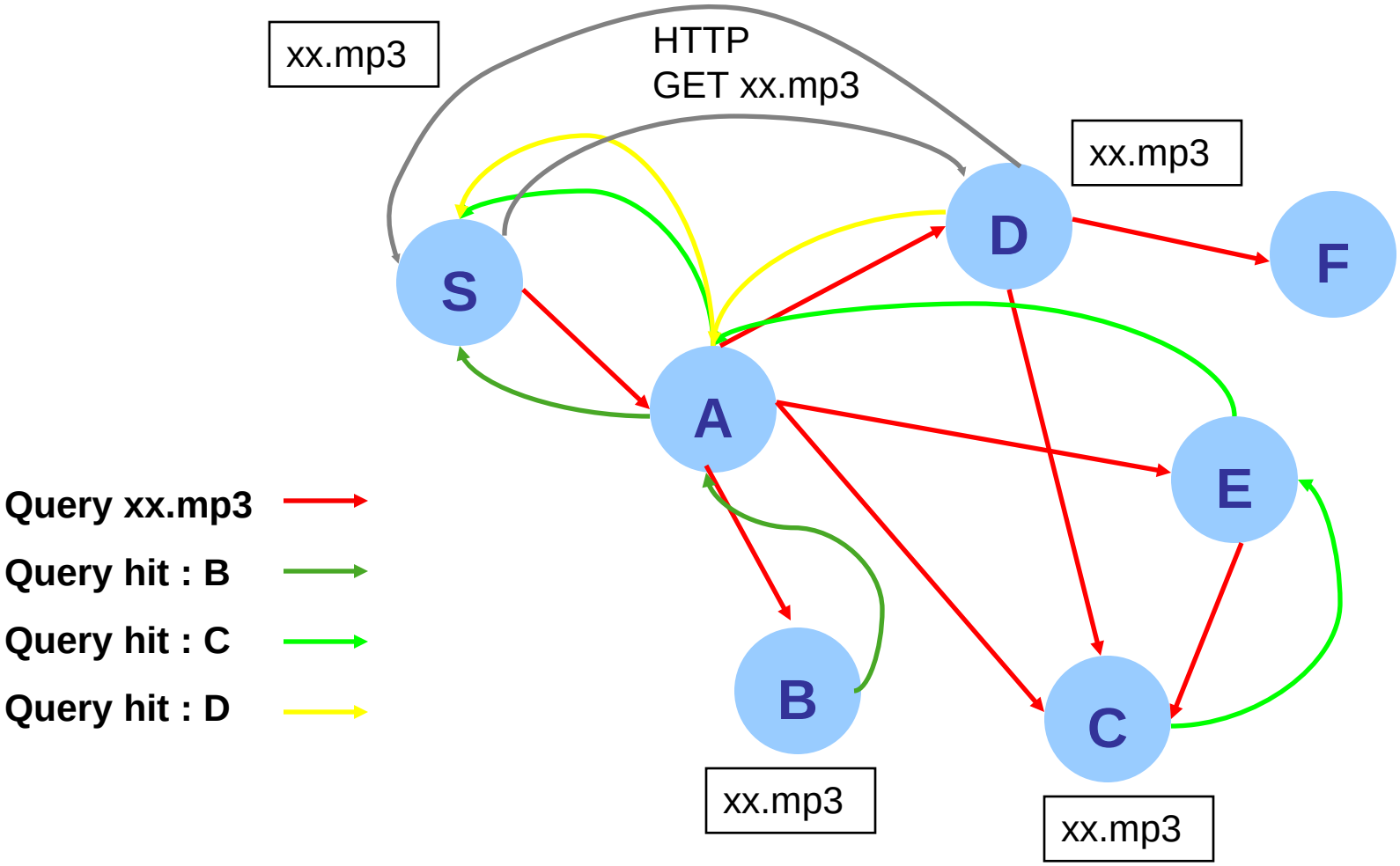
- **Parameters:**

u	a node in the web graph
d_i^+	out degree of a node i
$w_1, w_2, \ldots, w_k$	nodes pointing to u
η	normalization constant, <1
$PR(u)$	page rank of page u

- PageRank is given by

$$PR(u) = (1-\eta) + \eta \cdot \left(\frac{PR(w_1)}{d_1} + \frac{PR(w_2)}{d_2} » \frac{PR(w_k)}{d_k} \right)$$

Alternatives: GNUTELLA-Query/QueryHit/GET



The mistake of Jordan Ritter

About Jordan Ritter

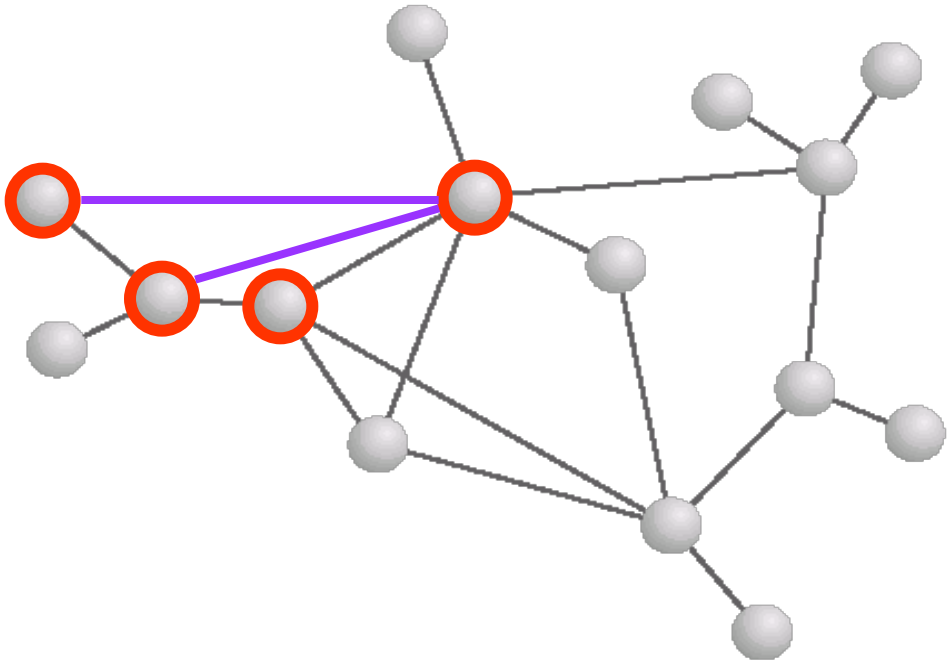
I've had a deep hand in building some of the most popular software you know. I don't blog much though, usually because I'm too busy building something.

Reachable Users								
	T=1	*T=2*	*T=3*	*T=4*	*T=5*	*T=6*	*T=7*	*T=8*
N=2	2	4	6	8	10	12	14	16
N=3	3	9	21	45	93	189	381	765
N=4	4	16	52	160	484	1,456	4,372	**13,120**
N=5	5	25	105	425	1,705	6,825	**27,305**	**109,225**
N=6	6	36	186	936	4,686	**23,436**	**117,186**	**585,936**
N=7	7	49	301	1,813	**10,885**	**65,317**	**391,909**	**2,351,461**
N=8	8	64	456	3,200	**22,408**	**156,864**	**1,098,056**	**7,686,400**

Source:
Why Gnutella Can't Scale. Really Not.

- Ritter (one of the Napster founder) claimed that there is an exponentially growing number of messages in Gnutella. **But it is a fairy tale.**
- In fact, every query has an ID. Every query is posted from every node to all of its neighbours until a TTL=0. **But each query (ID) only once.**
  So this is an additional bouncing mechanism in the protocol and so - there is a maximum of n^2 messages - if the graph would be a complete graph.
- Every node has a maximum of 4 open connections as Ritter realised. So there are even not n^2 **but only 4n messages**.

Alternatives: Freenet and its Search [Hong01]

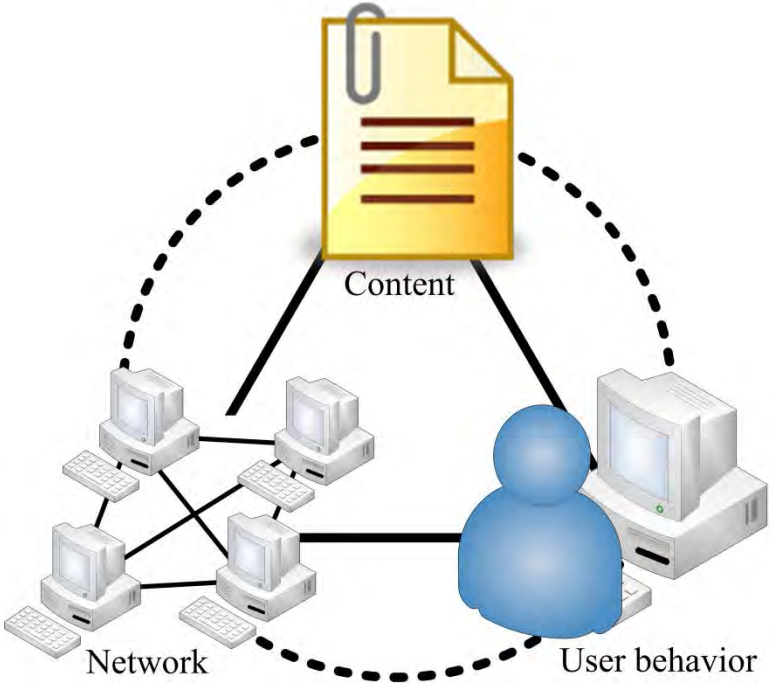


- a graph structure actively evolves over time
  - new links form between nodes
  - files migrate through the network
    - ⇒adaptive routing
    - ⇒most requested content is found fast

[Abere02]

Research Focus

- **Our Approach to look at communication networks:**



→ Consider the mutual influence between content, users and user activities as well as network with its parameters and configuration

So what we can do?

Motivation

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

→ big challenge to filter relevant information

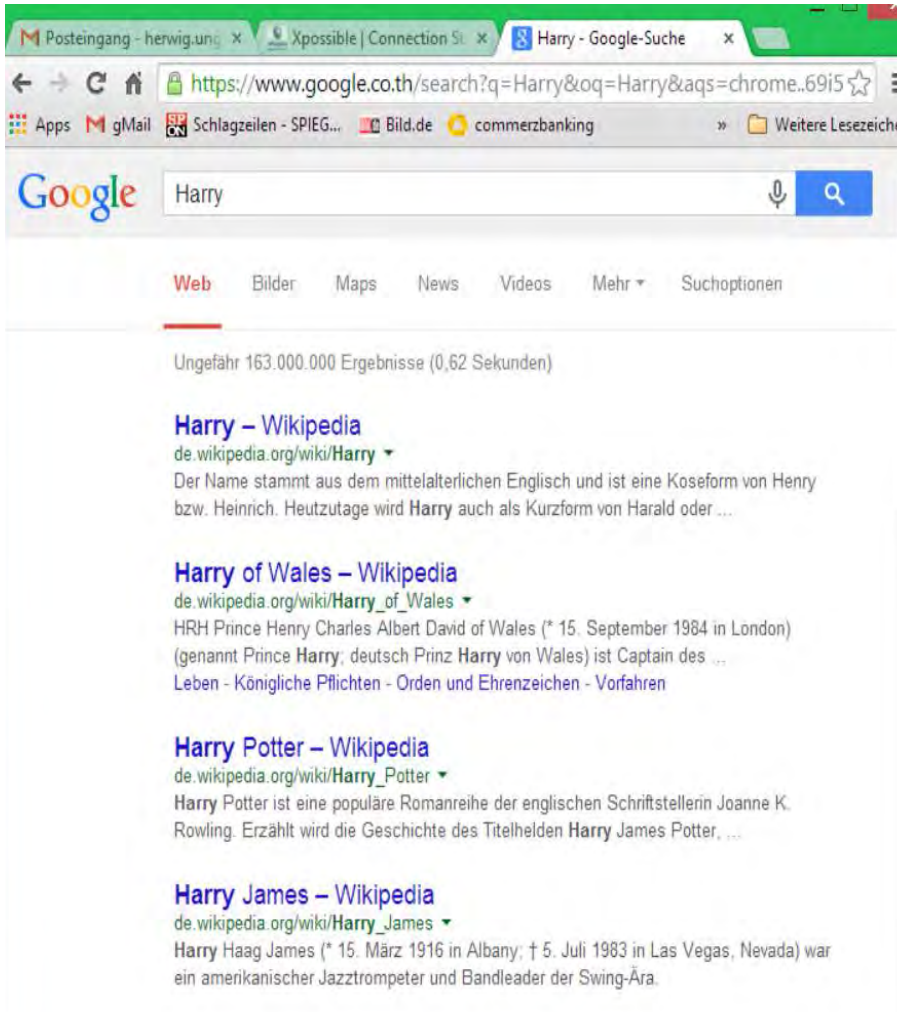
→ usually 2-3 keywords are a weak description of what the users are looking for

→ the typically received 10000+ search result overload the user and normally only the first 30 results are considered

→ 3 til 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.

An Example: Harry.



We get:

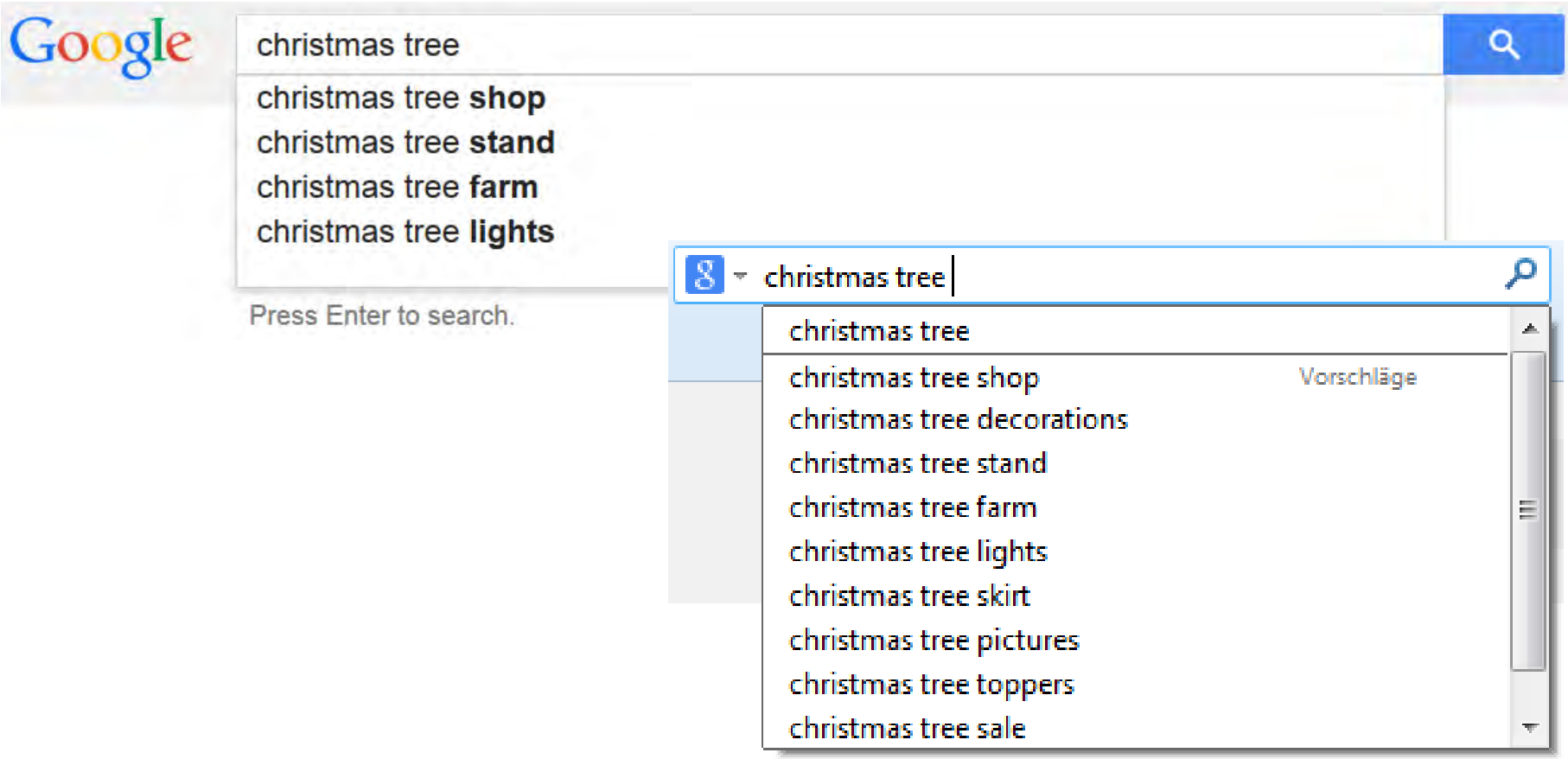
163 million results
on 16,3 million pages!!!!!

- → no structured presentation
- → no independend ranking
- → no evaluation of trostworthiness
- → no support for the user
- → no use of user evaluation

!! Google only offer search for text or pictures or similar pictures !!

Another Example: “christmas tree“

- What does Google offer?



So whats 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).

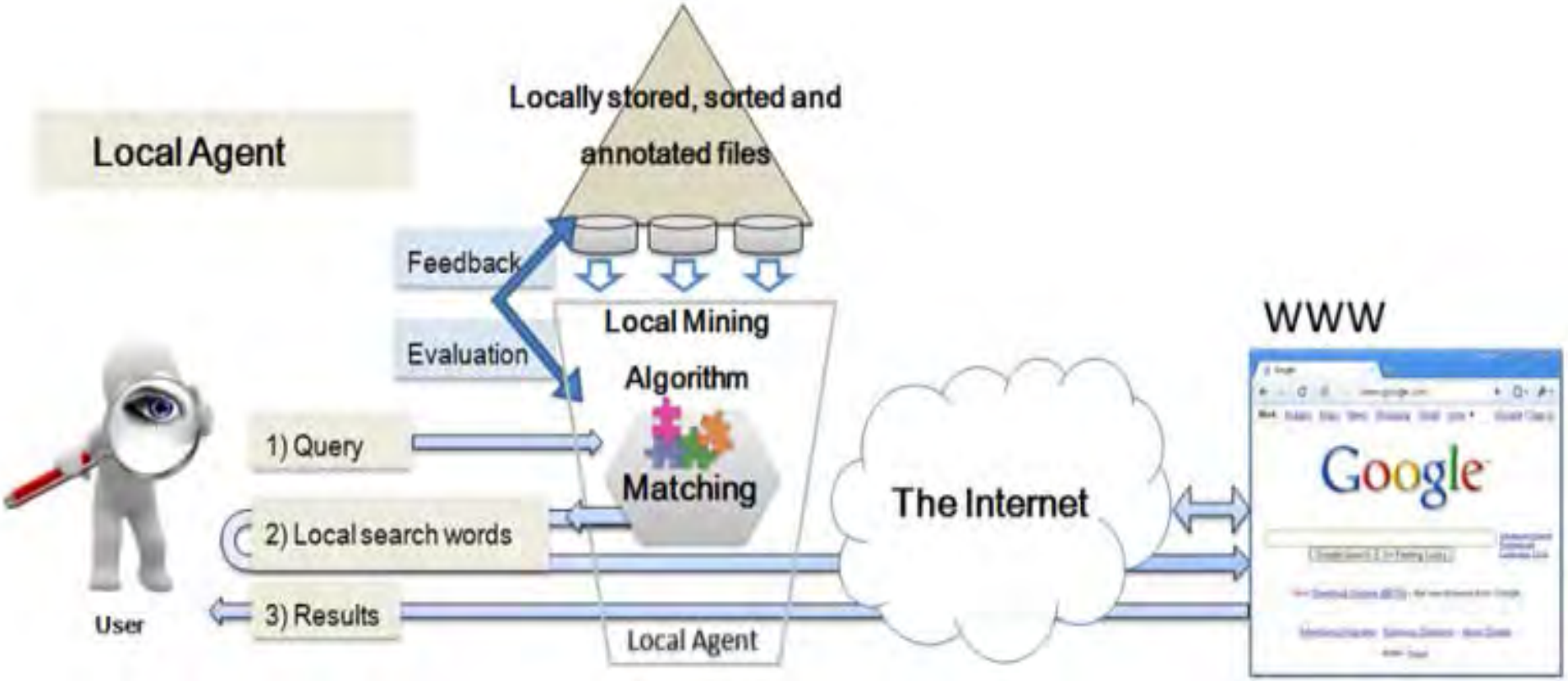
Motivation / Problem Statement 1: Disambiguation

- 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.) cube (maths, car)
christmas tree (biol, oil), Harry (sev.names)

Idea 1: Locality



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 Submit

Image Results for: Harry

Select/deselect all images for analysis

Search for more similar images Use keyword translation

Search with selected images

Result of Textanalysis

Select extracted keywords from your document (?):

- [x] Harry
- [x] Potter
- [x] book
- [x] film
- [ ] review
- [ ] series
- [ ] world
- [ ] cover
- [ ] movie
- [ ] school
- [ ] wand
- [ ] fan
- [ ] deathly
- [ ] Hermione
- [ ] friend
- [ ] Ron
- [ ] magic
- [ ] video
- [ ] wizardry
- [ ] Voldemort

Selected search words:

Harry Potter book film

Search Google™ Custom Search

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

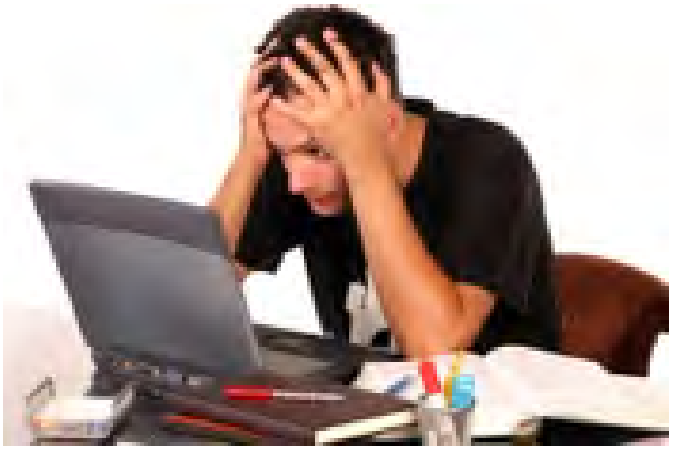
'Harry Potter'-Inspired **Film Series**
Sep 12, 2013 ... Expanding their longterm, lucrative partnership on the **Harry Potter** franchise, Warner Bros and author J.K. Rowling are putting a new **film series** ...
www.deadline.com/.../warner-bros-j-k-rowling-team-for-new-**harry-potter**- inspired-**film-series**/

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)**

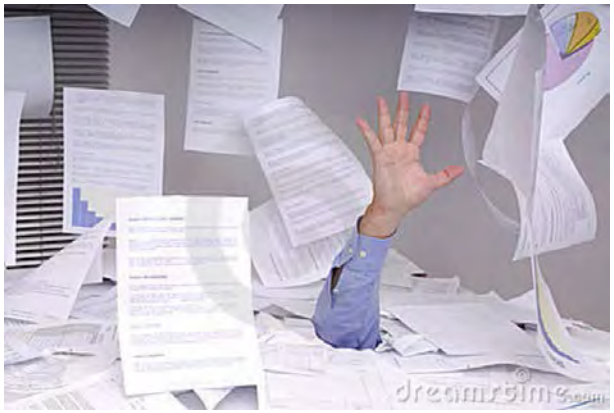
Motivation / Problem Statement 2

- 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 3: 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

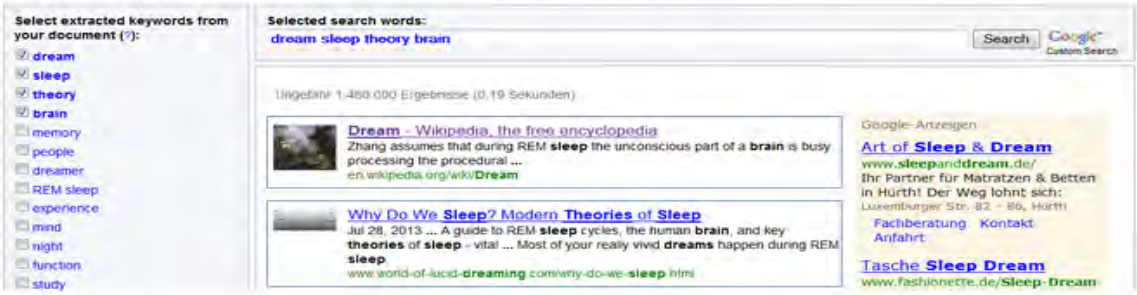
Try out DocAnalyser for Yourself at www.docanalyser.de!

DocAnalyser - Find Similar and Related Web Documents

What is DocAnalyser?

DocAnalyser is a new service that offers you 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.

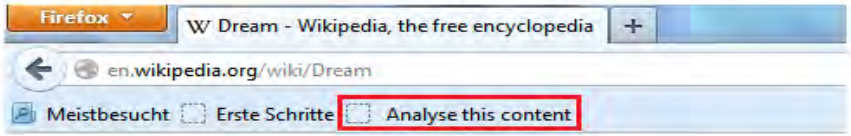
DocAnalyser (Alpha)



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)

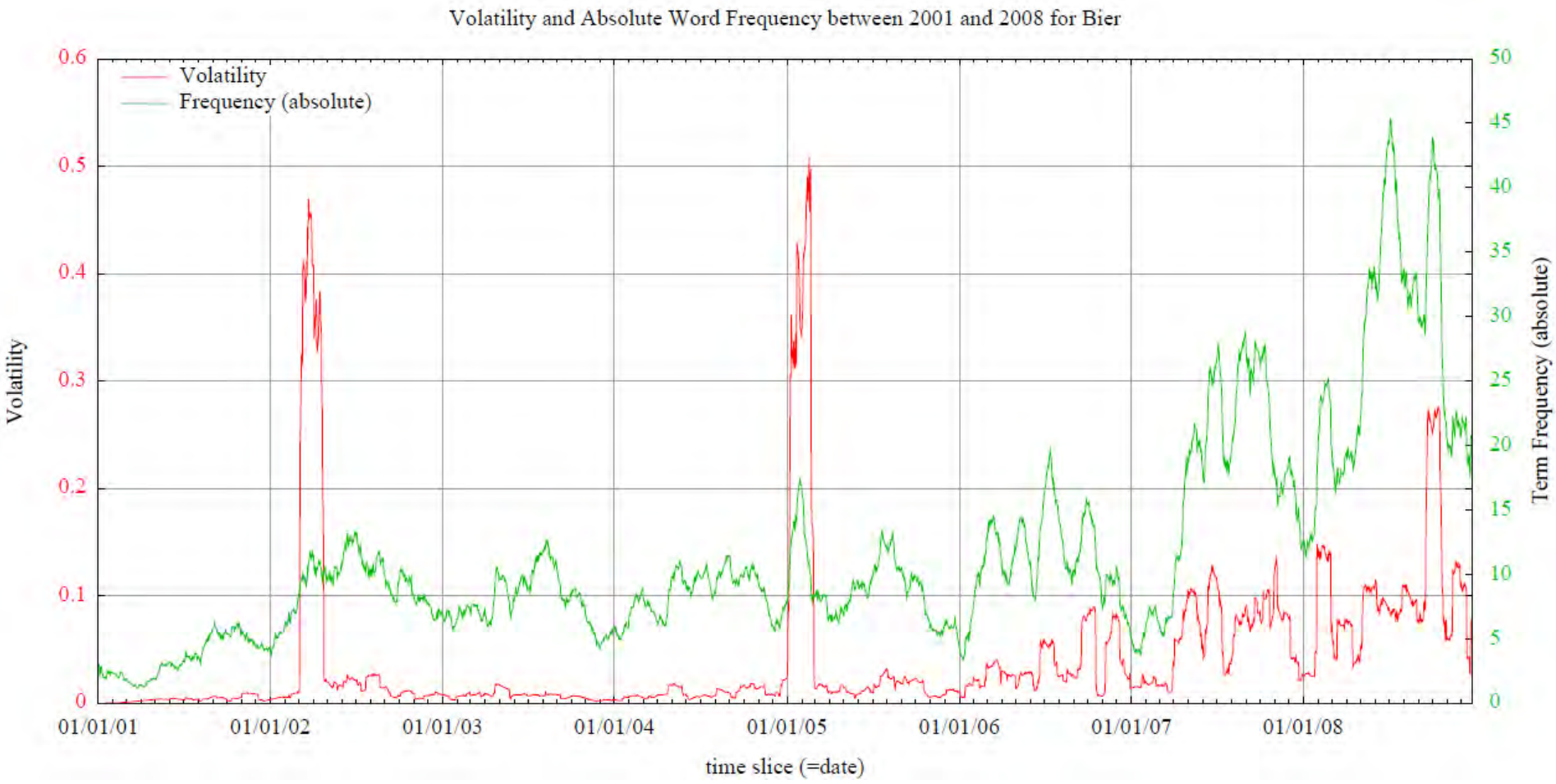
Remark: The following bookmarklet should only be used when analysis errors occur using the first bookmarklet: Analyse this content

Idea 4: Detection of topic changes

- Interesting for
  - Marketing
  - Trend detection
  - Determination of hotly discussed topics
- Approach
  - Topics do not only characterise events, they indicate an author's or society's view on these events, too.
  - This view can change over time and is therefore dynamic.
  - The detection of change is problematic with term frequencies
    - low term frequency for new topics
    - no indication of semantically related topics exists

→ Use term volatility !

The hotly discussed topic „beer“



Idea 5: Webserver and P2P (see also YaCy and Faroo)

WWW
Pages

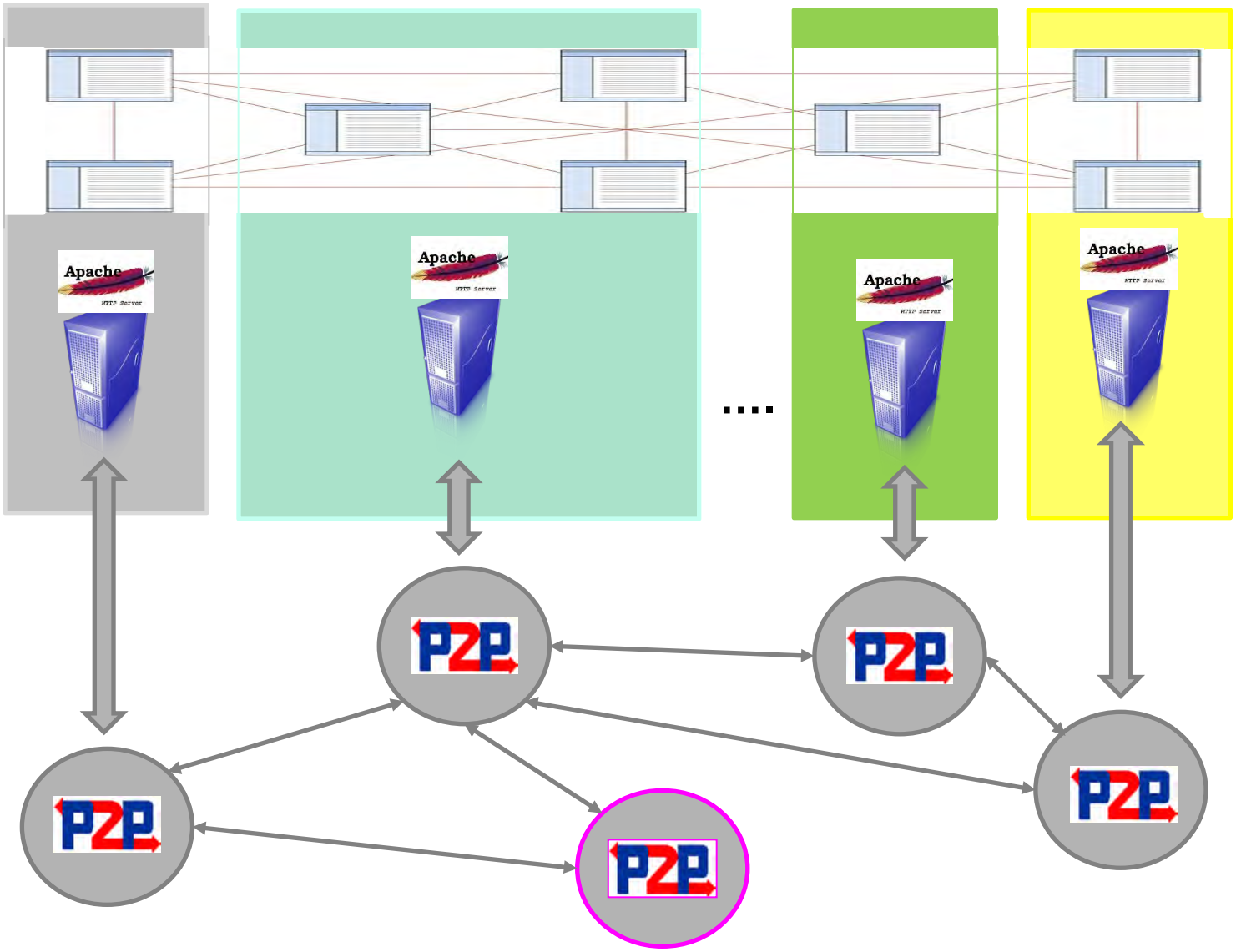
+

Webserver
Im Internet

+
P2P-Level

===

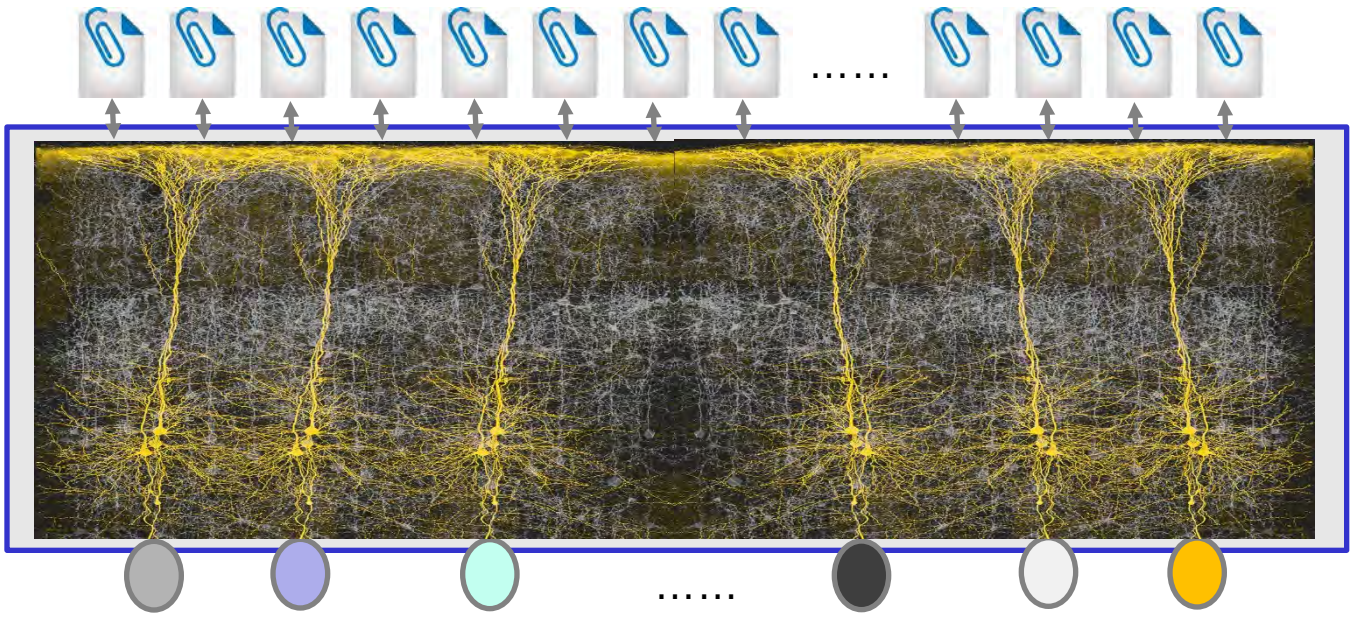
NEW
SYSTEM



A dimension problem...

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

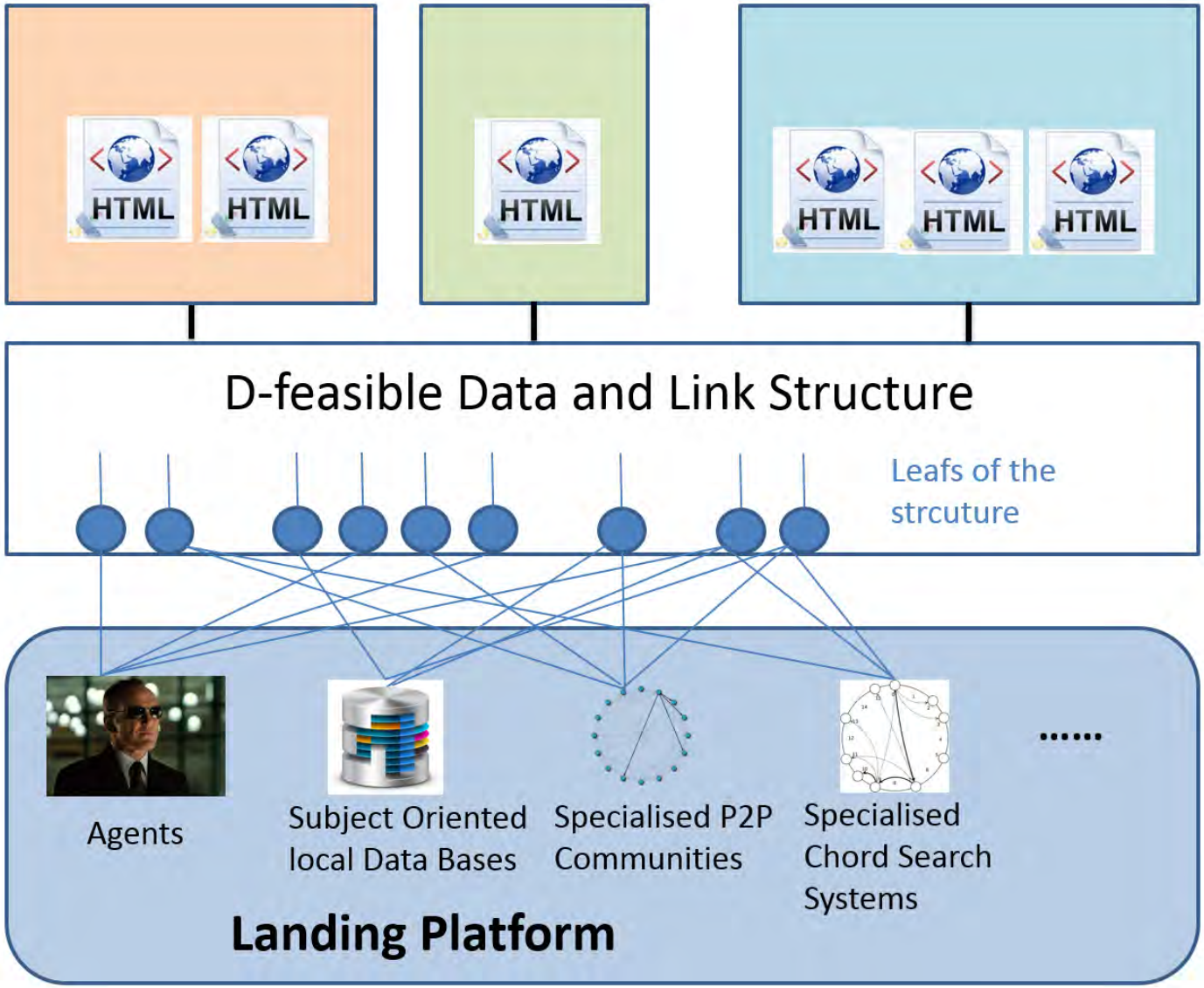
Billions of Web documents d



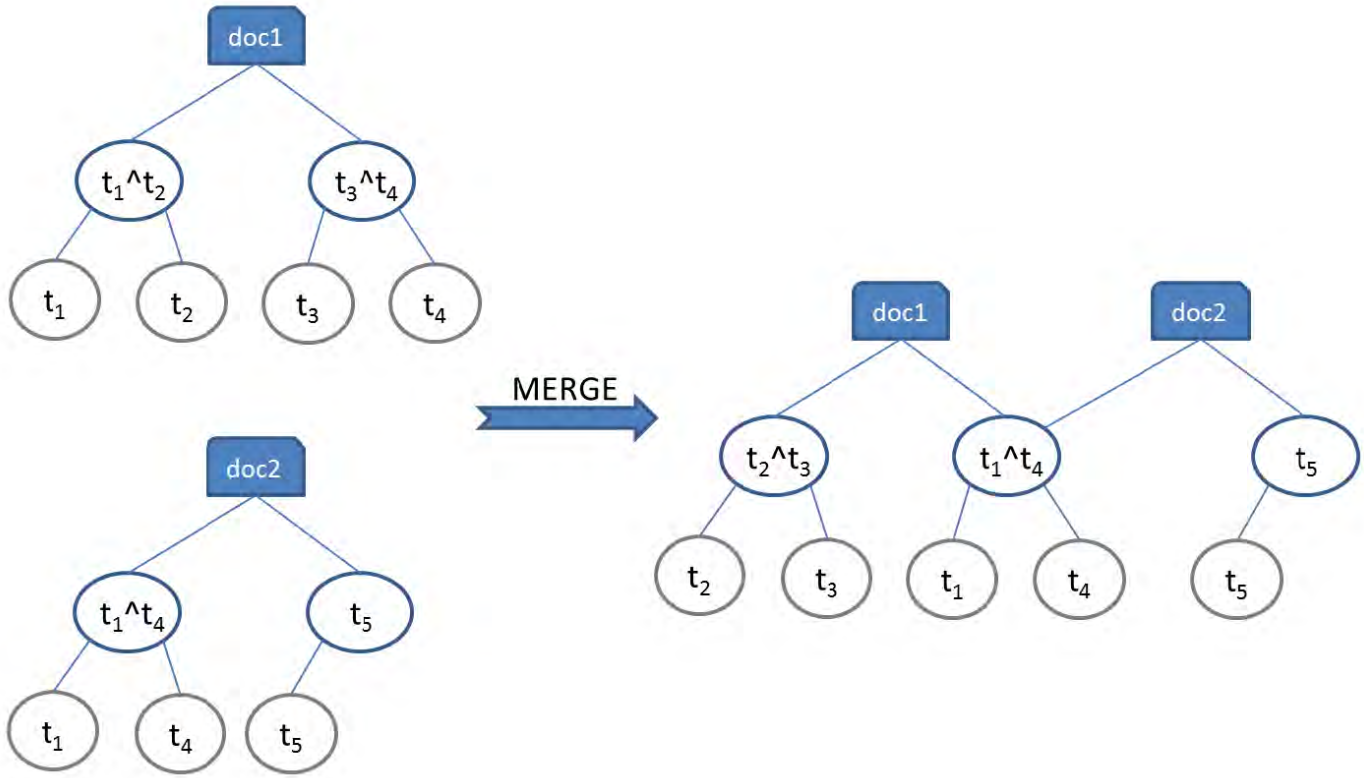
constant to ln(d) distance for scalability

Approx. 200,000 terms t

Solutions


HTML
HTML
HTML
HTML
HTML
HTML
D-feasible Data and Link Structure
Leafs of the strcuture
Agents
Subject Oriented local Data Bases
Specialised P2P Communities
Specialised Chord Search Systems
.....
Landing Platform

Solutions 2

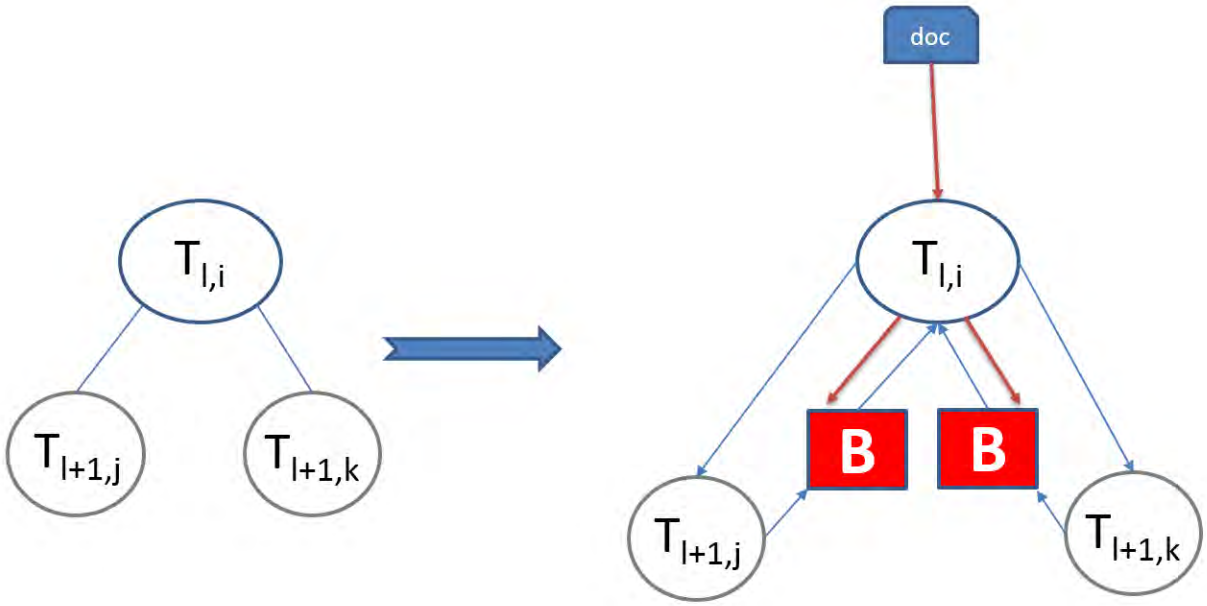


a) Document trees of Doc1 and Doc2

b) Possible, merged tree structure

Solutions 3: Bloomfilter

- to solve the problem of multi-keyword search
- requests pass only fitting edges
- solution made in the „middle“ part of the structure



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