Wikipedia Search

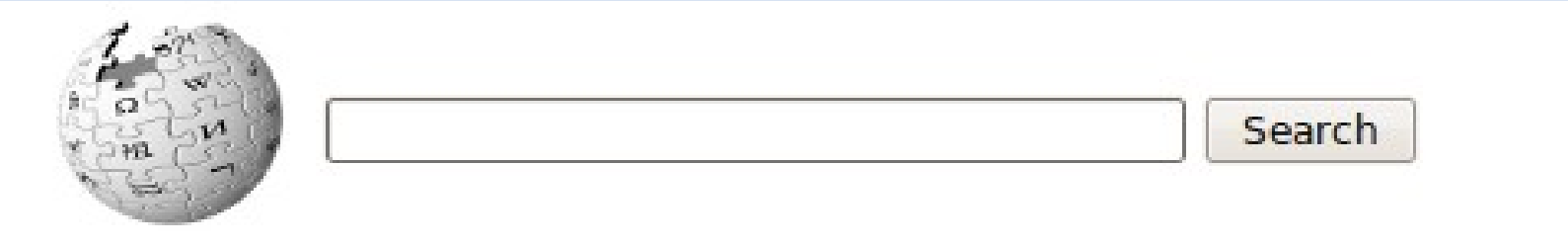
by Robert Stojnic (aka rainman)

Search statistics

- Users submit around 500MB of text in search queries every day
- Special:Search is most viewed single page
- Average of 200 searches / second (as in Special:Search, not “Go” button) - en.wp ~50%
- Important to have an independent search engine
- And yet, not many people use search - I started reconstructing it in 2007

The ideal search box

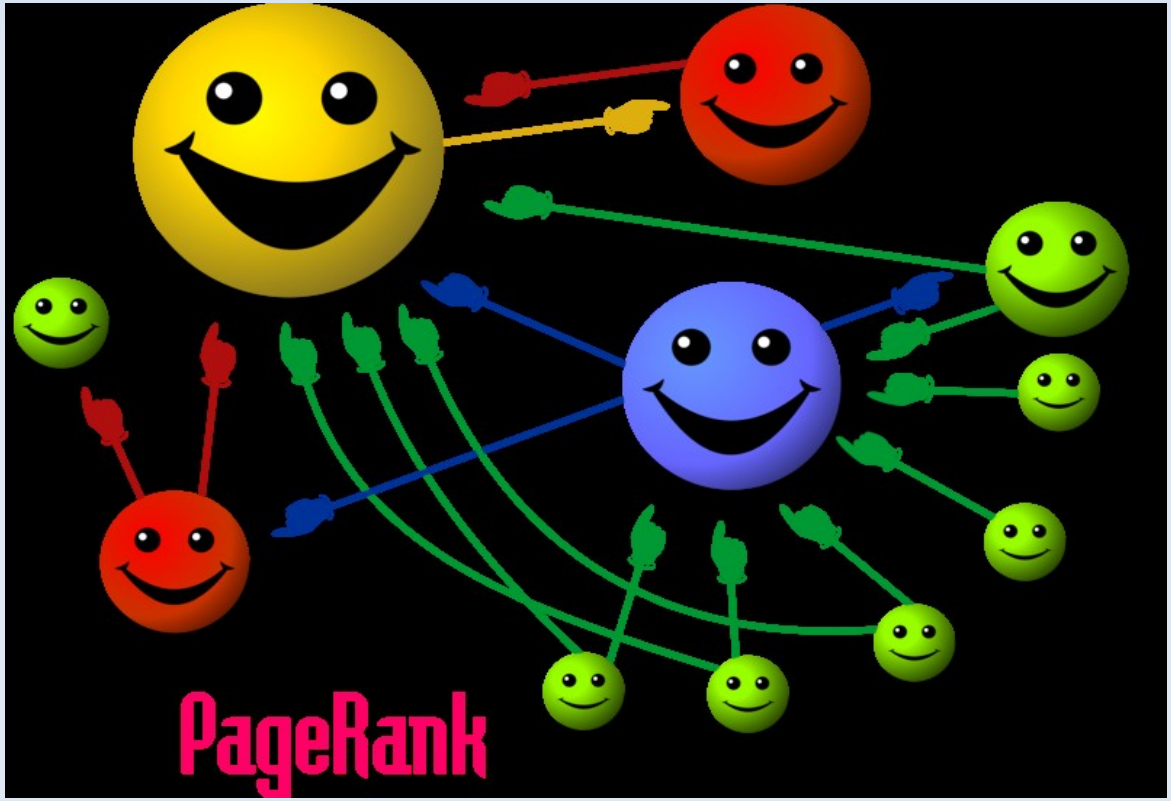
- No extra options, buttons, tabs, links, warnings, or requests (ahem, looks familiar?)



- Just type in some words associated with what you want, and most relevant article pops up (mirror human memory) - not directory or category based

The PageRank™ myth 1/2

- ”Wikipedia should just use PageRank”
- stationary distribution of a ”random surfer”



The PageRank™ myth *2/2*

- Works if links are added without constrain, but in Wikipedia, the most linked pages are the general concepts like years, countries ...
- Naively incorporating PageRank into search results puts too much weight on these articles
- However, using number of links leading to article proved to be good for sorting results when query matches the **article title**

So why is Google still better

- Much more money, time and smartness involved
- Can use whole of the internet to find the right association with the article
- More advanced information extraction - look at google snippets, when they are right they are spot-on
- But we can still do a fair job, given we have the right kind of algorithms in place

Building open-source search engine

- Lucene is the most widely used search API
- Uses classical approach: words statistically overexpressed in articles are most descriptive of article (e.g. article on Russia will contain word ”Russia” more times than a random article)
- better results than random sorting (e.g. MySQL search results) but still quite bad (pre-2007 Wikipedia)

Extracting information 1/2

- We want to be able to predict if a query is a good association for the article - how?

1) Full title matches: argentina climate ~ Climate of Argentina

2) Morphological: demographics of uk ~ Demography of United Kingdom *(stemming)*

3) Beginning of article descriptive of ”what the article is about”

4) Sections are sub-themes: serbia climate ~ Serbia#climate

Extracting information *2/2*

5) Link text and redirects are alternative names for article - useful but potentially dangerous - Wikibomb [[Evil|Scientology]]

6) Context of the page - articles whose links co-occur in paragraphs are related. e.g. Douglas Adams and The Hitchhikers Guide to the Galaxy

7) Words close together in article more relevant

=> Putting it all together = black magic

Did you mean ...

- is NOT spell checking:
  noble prize winners -> nobel prize winners
- it IS suggesting similar queries. Outline:
  - find similar words using n-grams, edit distance, double metaphones; language agnostic
  - try to modify the query to fit whole titles, frequently occurring groups of words, titles and links from beginning of article
  - sometimes too aggressive, use article snippets to infer when user likely found what he wants

For the advanced user

- prefix: - limit search to pages beginning with ... e.g. search archived talk pages
- intitle: - limit search to titles only
- incategory: - limit search to certain category (works only for categories added in main article text)
- afgan*, *stan - prefix and suffix search
- sarah~ - fuzzy search

The state of search

- Wikimedia search cluster - 12 search servers (search, text snippets, did you mean...)
- Special:Search redesigned by usability project
- Relatively stable and usable, but still much room for improvement:
  - make use of article traffic statistics
  - better relatedness metric - smartwikisearch
  - debug corner cases - chemical formulas, usage of link text, stemming, synonyms, did you mean...
  - index other tables: categories, templates, etc

Other search backends

- So far we talked only about MWSearch + lucene-search, aimed at big sites as Wikipedia
- SphinxSearch: sphinx search engine, incremental updates, relevancy based on closeness of words, dictionary-based did you mean...
- EzMwLucene: vanilla lucene with incremental updates and attachment searching
- mySQL: default and pretty sucky

Conclusions & Future work

- Wikipedia structure is more ordered than that of WWW, we use and need more smart ways to extract informations and associations
- Can we efficiently use LSI or other advanced techniques? Get research people involved!
- Making good search is not easy, and is frequently underestimated
- Increase usability of MediaWiki by making a decent default search engine