

Personalizing Web Search using Long Term Browsing History

Nicolaas Matthijs (University of Cambridge, UK)

Filip Radlinski (Microsoft, Vancouver)

What is personalized web search ?

Web Images Videos Maps News Shopping Gmail more ▼ nicolaasmathijs@googlemail.com | Web History | Settings ▼ | Sign out

Google IR Search SafeSearch off ▼
About 460,000,000 results (0.47 seconds) Advanced search

Everything [International Rectifier - The Power Management Leader](#)

Manufacturer of power semiconductors (MOSFET, IGBT, Diodes and Thyristors).
★ Show stock quote for IRP
HEXFET Power MOSFETs - Careers - Contact Us - Application Notes
www.irf.com/indexnew.html - Cached - Similar

The web
Pages from the UK

Any time
Past 2 months
▼ More search tools

▼ More search tools

[Infrared - Wikipedia, the free encyclopedia](#)

Infrared radiation (IR radiation) is electromagnetic radiation with a wavelength between 0.7 and 300 micrometres, which equates to a frequency range between ...
en.wikipedia.org/wiki/Infrared - Cached - Similar

[Information retrieval - Wikipedia, the free encyclopedia](#)

Information retrieval (IR) is the science of searching for documents, for information within documents, and for metadata about documents, as well as that of ...
en.wikipedia.org/wiki/Information_retrieval - Cached - Similar
★ Show more results from en.wikipedia.org

[Glasgow Information Retrieval Group](#)

Information Retrieval Group at the Department Computing Science, University of Glasgow, Scotland, UK.
ir.dcs.gla.ac.uk/ - Cached - Similar

[HM Revenue & Customs: Home Page](#)

Includes information about the UK's customs and tax department and includes the prototype on-line VAT (Value Added Tax) returns system.
Self Assessment - VAT - Online services - PAYE for employers
www.hmrc.gov.uk/ - Cached - Similar

[Met Office: Europe: Infrared satellite imagery](#)

These images come from satellites which remain above a fixed point on the Earth (i.e. they are "geostationary"). The infrared image shows the invisible ...
www.metoffice.gov.uk/satpics/latest_IR.html - Cached - Similar

[BANK OF IRELAND Share Price | BIR.IR - Yahoo! Finance UK](#)

BANK OF IRELAND. (Ish: BIR.IR / ISIN IE003060259) ... More Headlines for BIR.IR. Add BIR.IR Headlines to My Yahoo! ...
uk.finance.yahoo.com/q?s=BIR.IR - 13 hours ago - Cached - Similar

Web Images Videos Maps News Shopping Gmail more ▼ nicolaasmathijs@googlemail.com | Web History | Settings ▼ | Sign out

Google IR Search SafeSearch off ▼
About 460,000,000 results (0.47 seconds) Advanced search

Everything [International Rectifier - The Power Management Leader](#)

Manufacturer of power semiconductors (MOSFET, IGBT, Diodes and Thyristors).
★ Show stock quote for IRP
HEXFET Power MOSFETs - Careers - Contact Us - Application Notes
www.irf.com/indexnew.html - Cached - Similar

The web
Pages from the UK

Any time
Past 2 months
▼ More search tools

▼ More search tools

[Infrared - Wikipedia, the free encyclopedia](#)

Infrared radiation (IR radiation) is electromagnetic radiation with a wavelength between 0.7 and 300 micrometres, which equates to a frequency range between ...
en.wikipedia.org/wiki/Infrared - Cached - Similar

[Information retrieval - Wikipedia, the free encyclopedia](#)

Information retrieval (IR) is the science of searching for documents, for information within documents, and for metadata about documents, as well as that of ...
en.wikipedia.org/wiki/Information_retrieval - Cached - Similar
★ Show more results from en.wikipedia.org

[Glasgow Information Retrieval Group](#)

Information Retrieval Group at the Department Computing Science, University of Glasgow, Scotland, UK.
ir.dcs.gla.ac.uk/ - Cached - Similar

[HM Revenue & Customs: Home Page](#)

Includes information about the UK's customs and tax department and includes the prototype on-line VAT (Value Added Tax) returns system.
Self Assessment - VAT - Online services - PAYE for employers
www.hmrc.gov.uk/ - Cached - Similar

[Met Office: Europe: Infrared satellite imagery](#)

These images come from satellites which remain above a fixed point on the Earth (i.e. they are "geostationary"). The infrared image shows the invisible ...
www.metoffice.gov.uk/satpics/latest_IR.html - Cached - Similar

[BANK OF IRELAND Share Price | BIR.IR - Yahoo! Finance UK](#)

BANK OF IRELAND. (Ish: BIR.IR / ISIN IE003060259) ... More Headlines for BIR.IR. Add BIR.IR Headlines to My Yahoo! ...
uk.finance.yahoo.com/q?s=BIR.IR - 13 hours ago - Cached - Similar

Web Images Videos Maps News Shopping Gmail more ▼ nicolaasmathijs@googlemail.com | Web History | Settings ▼ | Sign out

Google IR Search SafeSearch off ▼
About 460,000,000 results (0.47 seconds) Advanced search

Everything [International Rectifier - The Power Management Leader](#)

Manufacturer of power semiconductors (MOSFET, IGBT, Diodes and Thyristors).
★ Show stock quote for IRP
HEXFET Power MOSFETs - Careers - Contact Us - Application Notes
www.irf.com/indexnew.html - Cached - Similar

The web
Pages from the UK

Any time
Past 2 months
▼ More search tools

▼ More search tools

[Infrared - Wikipedia, the free encyclopedia](#)

Infrared radiation (IR radiation) is electromagnetic radiation with a wavelength between 0.7 and 300 micrometres, which equates to a frequency range between ...
en.wikipedia.org/wiki/Infrared - Cached - Similar

[Information retrieval - Wikipedia, the free encyclopedia](#)

Information retrieval (IR) is the science of searching for documents, for information within documents, and for metadata about documents, as well as that of ...
en.wikipedia.org/wiki/Information_retrieval - Cached - Similar
★ Show more results from en.wikipedia.org

[Glasgow Information Retrieval Group](#)

Information Retrieval Group at the Department Computing Science, University of Glasgow, Scotland, UK.
ir.dcs.gla.ac.uk/ - Cached - Similar

[HM Revenue & Customs: Home Page](#)

Includes information about the UK's customs and tax department and includes the prototype on-line VAT (Value Added Tax) returns system.
Self Assessment - VAT - Online services - PAYE for employers
www.hmrc.gov.uk/ - Cached - Similar

[Met Office: Europe: Infrared satellite imagery](#)

These images come from satellites which remain above a fixed point on the Earth (i.e. they are "geostationary"). The infrared image shows the invisible ...
www.metoffice.gov.uk/satpics/latest_IR.html - Cached - Similar

[BANK OF IRELAND Share Price | BIR.IR - Yahoo! Finance UK](#)

BANK OF IRELAND. (Ish: BIR.IR / ISIN IE003060259) ... More Headlines for BIR.IR. Add BIR.IR Headlines to My Yahoo! ...
uk.finance.yahoo.com/q?s=BIR.IR - 13 hours ago - Cached - Similar

What is personalized web search ?

Present each user a different ranking tailored to their personal interests and information need

Web Images Videos Maps News Shopping Gmail more ▼ nicolaasmattijs@googlemail.com | Web History | Settings ▼ | Sign out

Google IR Search SafeSearch off ▼

About 45,100,000 results (0.39 seconds) Advanced search

Everything
Images
More

The web
Pages from the UK

Any time
Past 2 days

All results
Timeline
Fewer shopping sites
More shopping sites

More search tools

Infrared - Wikipedia, the free encyclopedia ☆ - 13:41
Infrared radiation (IR radiation) is electromagnetic radiation with a wavelength between 0.7 and 300 micrometres, which equates to a frequency range between ...
Overview - Origins of the term - Different regions in the infrared
en.wikipedia.org/wiki/Infrared - Cached - Similar

Infrared photography - Wikipedia, the free encyclopedia ☆
Top: tree photographed in the near infrared range. Bottom: same tree in the visible part of the spectrum. Infrared image of the Mississippi River crossed by ...
en.wikipedia.org/wiki/Infrared_photography - 3 hours ago - Cached - Similar

Show more results from en.wikipedia.org

Infrared Waves ☆
27 Mar 2007 ... Infrared light lies between the visible and microwave portions of the electromagnetic spectrum. Infrared light has a range of wavelengths, ...
science.hq.nasa.gov/kids/imagers/emis/infrared.html - Cached - Similar

infrared ☆
Infra-Red is interesting, because it is easily generated and doesn't suffer electromagnetic interference, so it is nicely used to communication and control, ...
www.ustr.net/Infrared/Infrared1.shtml - Cached - Similar

Met Office: British Isles: Infrared satellite imagery ☆
These images come from satellites which remain above a fixed point on the Earth (i.e. they are "geostationary"). The infrared image shows the invisible ...
www.metoffice.gov.uk/satpics/latest_uk_ir.html - Cached - Similar

Sponsored links

InfraRed Lamps
Discount Infra Red bulbs here buy now save time, save money
www.lamps-on-line.com

Fluke Thermal Imager
Try before you buy! from Acutest
The New Fluke Ti25 Thermal Imager
www.acutest.net

See your ad here >

Web Images Videos Maps News Shopping Gmail more ▼ nicolaasmattijs@googlemail.com | Web History | Settings ▼ | Sign out

Google IR Search SafeSearch off ▼

About 7,130,000 results (0.16 seconds) Advanced search

Everything
Videos
Books
Blogs
More

The web
Pages from the UK

Information retrieval - Wikipedia, the free encyclopedia ☆
Information retrieval (IR) is the science of searching for documents, for information within documents, and for metadata about documents, as well as that of ...
History - Overview - Performance measures - Model types
en.wikipedia.org/wiki/Information_retrieval - Cached - Similar

Information Retrieval - University of Glasgow ☆
Computing Science ...
An online book by CJ van Rijsbergen, University of Glasgow.
www.dcs.ole.ac.uk/kaikh/Preface.html - Cached - Similar

Sponsored links

Information Retrieval
Search internal websites, file shares, databases, and more.
www.google.co.uk/enterprise

Related to web information retrieval:

Web Data Extraction
Fast, powerful, reliable and easy web data extraction. Free Trial!
www.AutomationAnywhere.com/extract

Information Retrieval - The Journal of Information Retrieval is an international forum for theory, algorithms, and experiments that concern search and ...
www.springer.com/computer/...%26+information+retrieval/.../10791 - Cached

Journal of Information Retrieval - SpringerLink ☆
www.springerlink.com/link.asp?id=103814 - Similar

Information retrieval: algorithms and heuristics - Google Books Result
David A. Grossman, Ophir Frieder - 2004 - Computers - 332 pages
The authors answer these and other key information retrieval design and implementation questions. This book is not yet another high level text.
books.google.co.uk/books?isbn=1402030037...

= Personalized web search

Web Images Videos Maps News Shopping Gmail more ▼ nicolaasmattijs@googlemail.com | Web History | Settings ▼ | Sign out

Google IR Search SafeSearch off ▼

About 905,000 results (0.23 seconds) Advanced search

Everything
News
More

The web
Pages from the UK

Any time
Past 3 weeks

re search tools

International Rectifier - The Power Management Leader ☆
Manufacturer of power semiconductors (MOSFET, IGBT, Diodes and Thyristors). Show stock quote for IRF
www.irf.com/indexnsw.html - Cached - Similar

HEXFET Power MOSFETs
Careers
Contact Us
Application Notes

Parametric Search
Sales Representative
DC-DC
About IR

More results from irf.com >

International Rectifier - HEXFET® Power MOSFET Product Information ☆
International Sites: | Chinese 简体中文 | Korean 한국어 | Japanese 日本語 | About International Rectifier | Contact Us | Privacy ...
www.irf.com/product-info/hexfet/ - Cached - Similar

International Rectifier - Wikipedia, the free encyclopedia ☆
International Rectifier Corp. (NYSE: IRF) is a manufacturer of power semiconductors (Integrated Circuits, MOSFET, and IGBTs), located in El Segundo, ...
en.wikipedia.org/wiki/International_Rectifier - Cached - Similar

International Rectifier - High Reliability ☆
International Sites: | Chinese 简体中文 | Korean 한국어 | Japanese 日本語 | About International Rectifier | Contact Us | Privacy ...
hrel.irf.com/ - Cached - Similar

International Rectifier datasheet pdf catalog ☆
12, 104MT, THREE PHASE AC SWITCH, International Rectifier ... 35, 10BQ030PF6, SCHOTTKY RECTIFIER, International Rectifier ...
www.datasheetcatalog.com/internationalrectifier/1/ - Cached - Similar

Sponsored links

Official Digi-Key Site
Search Our Huge Selection of Quality Electronic Components Here!
uk.digikey.com

See your ad here >

Related Work

- **Many approaches**
- **Clickthrough-based approaches**
 - PClick (Dou et al., 2007)
 - Promote URLs previously clicked by the same user for the same query
- **Profile-based approaches**
 - Teevan et al., 2005
 - Rich model of user interests
 - Built from search-related information, previously visited web sites, documents on hard-drive, e-mails, etc.
 - Re-rank top returned search results

Goal

- Improve on existing personalized web search techniques
 - ✓ Combine a profile-based approach with a clickthrough-based approach
 - ✓ Selection of new features
 - ✓ Build an improved user representation from long term browsing history
- Improve on the evaluation methodology
 - ✓ Find out whether search personalization makes a difference in real life
 - ✓ Improve search result ranking without changing search environment
- Develop tool used by real people



Search Personalization Process

1) User Interest Extraction

- User Profile representing user's interests
 - ✓ List of weighted terms
- List of all visited URLs and number of visits
- List of all search queries and results clicked

2) Result re-ranking

- Change order of results to better reflect user's interests
- Get first 50 results for query on Google
- Re-rank based on user profile by giving a score to each snippet

User Profile Extraction

Step 1: Term List Generation

- Don't treat web pages as normal flat documents but as structured documents
 - Use different sources of input data
 - Title unigrams
 - Metadata description unigrams
 - Metadata keywords
 - Full text unigrams
 - Extracted terms (Vu et al., 2008)
 - Extracted noun phrases (Clark et al., 2007)
 - Specify how important each data source is (weight vector)
 - Combination of data sources
- => List of terms to be associated with the user

User Profile Extraction

Step 2: Term List Filtering

- No filtering
- WordNet based POS filtering
- Google N-Gram corpus based filtering

=> Filtered list of terms

User Profile Extraction

Step 3: Term Weighting

- TF

$$\vec{F}_{t_i} = \begin{bmatrix} f_{title_{t_i}} \\ f_{mdesc_{t_i}} \\ f_{text_{t_i}} \\ f_{mkeyw_{t_i}} \\ f_{terms_{t_i}} \\ f_{nphrases_{t_i}} \end{bmatrix}$$

$$w_{TF}(t_i) = \vec{F}_{t_i} \cdot \vec{\alpha}$$

- TF-IDF

$$w_{TFIDF}(t_i) = \frac{1}{\log(DF_{t_i})} \times w_{TF}(t_i)$$

- pBM25 (Teevan et al., 2005)

$$w_{pBM25}(t_i) = \log \frac{(r_{t_i} + 0.5)(N - n_{t_i} + 0.5)}{(n_{t_i} + 0.5)(R - r_{t_i} + 0.5)}$$

= User Profile: list of terms and term weights

Search Personalization Process

1) User Interest Extraction

- User Profile representing user's interests
 - List of weighted terms
- List of all visited URLs and number of visits
- List of all search queries and results clicked

2) Result re-ranking

- Change order of results to better reflect user's interests
- Get first 50 results for query on Google
- Re-rank based on user profile by giving a score to each snippet

Result re-ranking

Step 1: Snippet scoring

- Matching

$$score_M(s_i) = \sum_{z=1}^{N_{s_i}} f_{t_z} \times w(t_z)$$

- Unique Matching

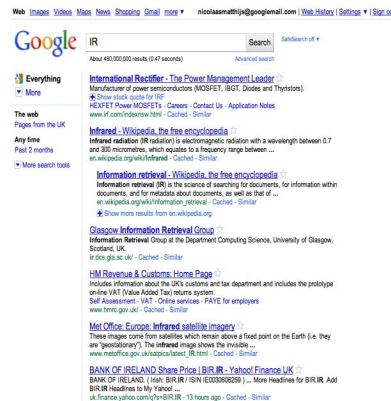
$$score_{UM}(s_i) = \sum_{z=1}^{N_{s_i}} w(t_z)$$

- Language Model

$$score_{LM}(s_i) = \sum_{z=0}^{N_{s_i}} \log \left(\frac{w(t_z) + 1}{w_{total}} \right)$$

Step 2: Keep Google rank into account

Step 3: Give extra weight to previously visited pages



Web Images Videos Maps News Shopping Gmail more • nicolasmathis@gmail.com | Web History Settings • Sign out

Google IR Search

About 460,000,000 results (0.47 seconds)

Advanced search

Everything

More

Show links open for IR

The web

Pages from the UK

Any time

Plus 2 months

More search tools

International Rectifier - The Power Management Leader
Manufacturer of power semiconductors (MOSFET, IGBT, Diodes and Thyristors).
HEXFET Power MOSFETs - Contact Us - Application Notes
www.internationalrectifier.com - Cached - Similar

Infrared - Wikipedia, the free encyclopedia
Infrared radiation (IR radiation) is electromagnetic radiation with a wavelength between 0.7 and 300 micrometers, which results in a frequency range between ...
en.wikipedia.org/wiki/infrared - Cached - Similar

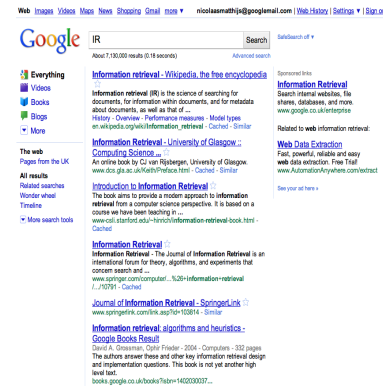
Information retrieval - Wikipedia, the free encyclopedia
Information retrieval (IR) is the science of searching for documents, for information within documents, and for metadata about documents, as well as that of ...
en.wikipedia.org/wiki/information_retrieval - Cached - Similar

Glasgow Information Retrieval Group
Information Retrieval Group at the Department Computing Science, University of Glasgow, Scotland, UK
info.cs.gla.ac.uk - Cached - Similar

HM Revenue & Customs: Home Page
Includes information about the UK's customs and tax department and includes the prototype on-line VAT Online Advice Tool returns system.
Self Assessment - VAT - Online services - PAYE for employers
www.hmrc.gov.uk - Cached - Similar

Max Office Europe: Infrared satellite imagery
These images come from satellites which remain above a fixed point on the Earth (i.e. they are "geostationary"). The infrared image shows the invisible ...
www.maxoffice.co.uk/satelliteimages_IR.html - Cached - Similar

BANK OF IRELAND Share Price | BRIR - Yahoo! Finance UK
BANK OF IRELAND (Insh:BRIR / ISIN:IE003060269) ... More Headlines for BRIR: Add BRIR:Headline to My Tabbed ...
uk.finance.yahoo.com/cv=BRIR - 15 hours ago - Cached - Similar



Web Images Videos Maps News Shopping Gmail more • nicolasmathis@gmail.com | Web History Settings • Sign out

Google IR Search

About 119,000 results (0.18 seconds)

Advanced search

Everything

More

Show links open for IR

The web

Pages from the UK

Any time

Plus 2 months

More search tools

Information retrieval - Wikipedia, the free encyclopedia
Information retrieval (IR) is the science of searching for documents, for information within documents, and for metadata about documents, as well as that of ...
en.wikipedia.org/wiki/information_retrieval - Cached - Similar

Information Retrieval - University of Glasgow
An online book by G. van Rijbergen, University of Glasgow.
www.dcs.gla.ac.uk/Kath/IR/ir.html - Cached - Similar

Introduction to Information Retrieval
The book aims to provide a modern approach to information retrieval from a computer science perspective. It is based on a course we have been teaching at ...
www.csl.stanford.edu/~brnch/information-retrieval-book.html - Cached - Similar

Information Retrieval
Information Retrieval: The Journal of Information Retrieval is an international forum for theory, algorithms, and experiments that concern search and ...
www.springer.com/computer_4230/information-retrieval_7_1/2021 - Cached - Similar

Journal of Information Retrieval - SpringerLink
www.springerlink.com/link.asp?10-102814 - Similar

Information retrieval, algorithms and heuristics
Google Books Result
David A. Grossman, Optic Finder - 2004 - Computers - 332 pages
The authors answer these and other key information retrieval design and implementation questions. This book is not yet another high level text.
books.google.co.uk/books?id=140200037...

Evaluation

- Difficult problem
- Most previous work
 - Small number of users evaluating relevance of small number of queries (Teevan et al., 2005)
 - Simulate personalized search setting using TREC query and document collection
 - After-the-fact log based analysis (Dou et al., 2007)
- Wanted to find out whether it yields a real difference in real-life usage
- Ideally: real-life usage data from lots of users over long time
- Unfeasible: high number of parameters

=> 2 step evaluation process

Evaluation: Capturing Data

- Need users and data to work with
- Full browsing history
- Not publicly available

⇒ Firefox add-on



AlterEgo Search 0.2

AlterEgo Search personalization based on your browsing history

- 41 users / 3 months
- 530,334 page visits / 39,838 Google searches

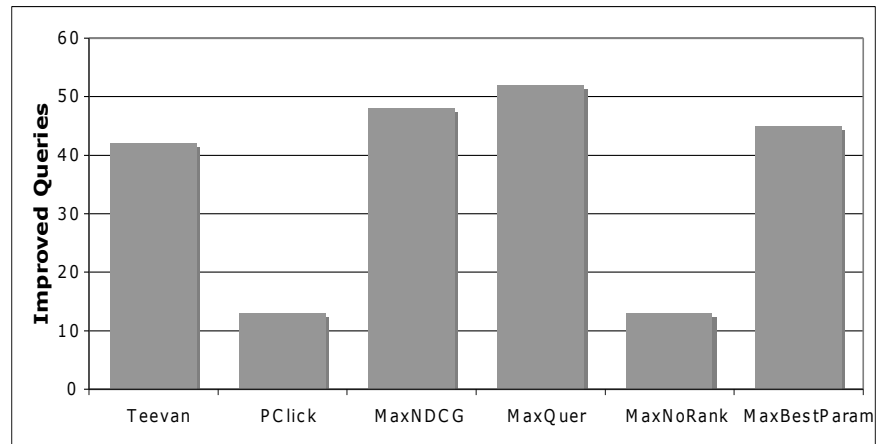
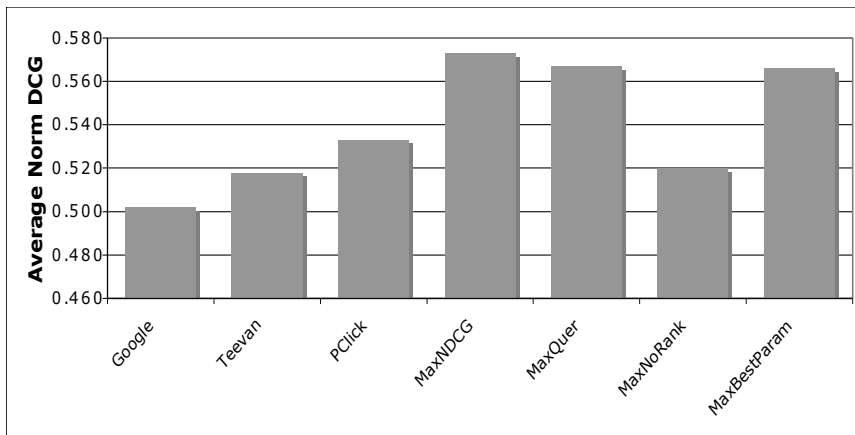
Step 1: Offline Relevance Judgments

- Identify most promising parameter configurations
- Offline evaluation session
- 6 users assess the relevance of the top 50 results for 12 queries
- Assess all possible combinations of all parameters
- Calculate NDCG score for each ranking (Jarvelin et al., 2000)

$$NDCG@10 = \frac{1}{Z} \sum_{i=1}^{10} \frac{2^{rel_i} - 1}{\log_2(1 + i)}$$

Step 1: Results

- 15,878 profile + re-ranking combinations investigated
- Compared to 3 baseline systems (Google, PClick and Teevan)
- 4,455 better than Google | 3,335 better than Teevan | 1,580 better than Pclick
- Identified 4 most promising personalization approaches



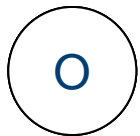
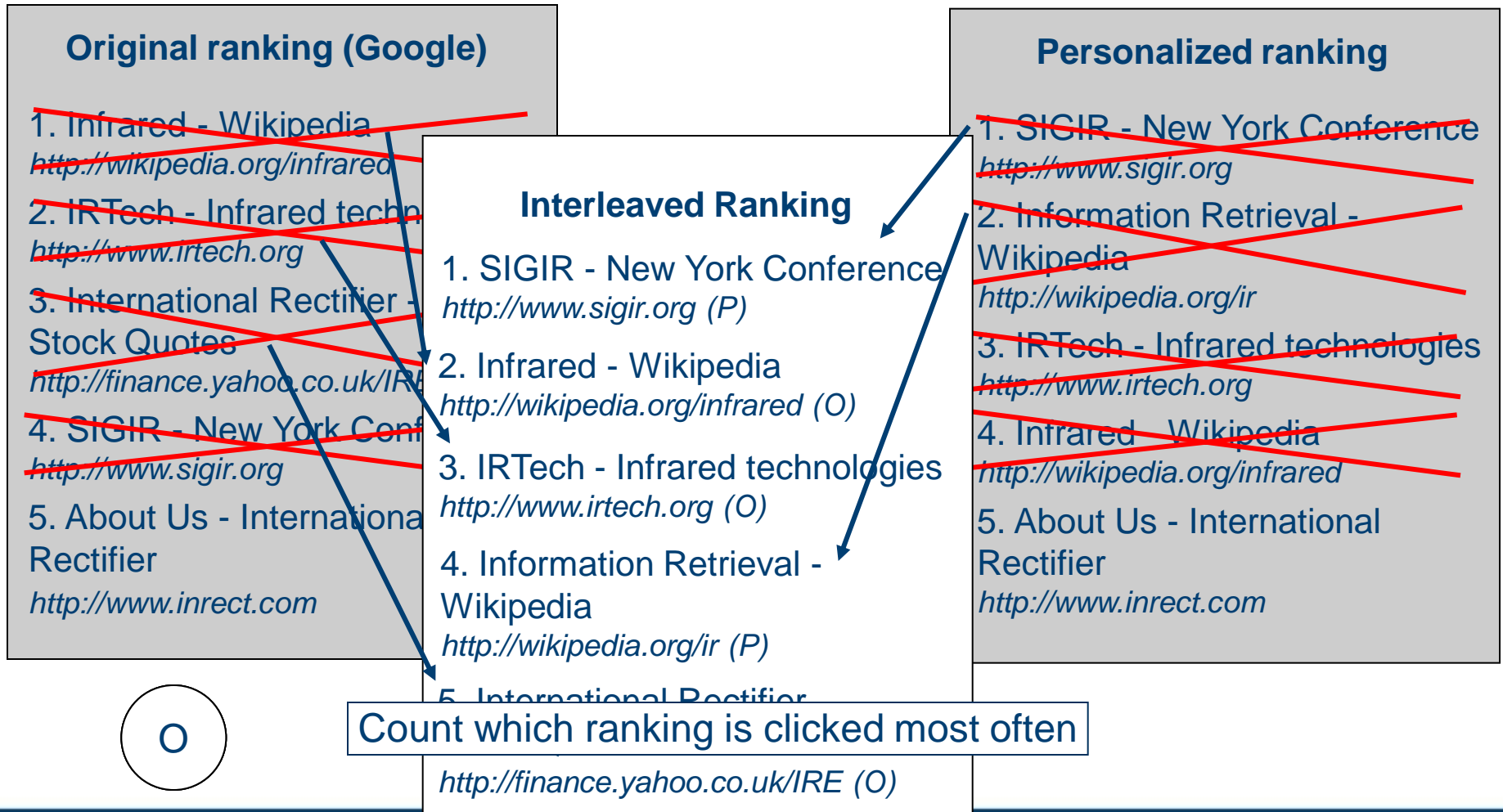
Step 1: Results

- Treating web pages as a flat document does not work.
- Advanced NLP techniques and keyword focused approaches work best.
- One re-ranking method outperforms all of the other ones:
 - ✓ LM
 - ✓ extra weight to visited URLs
 - ✓ keeping the Google rank into account

Step 2: Online Interleaved Evaluation

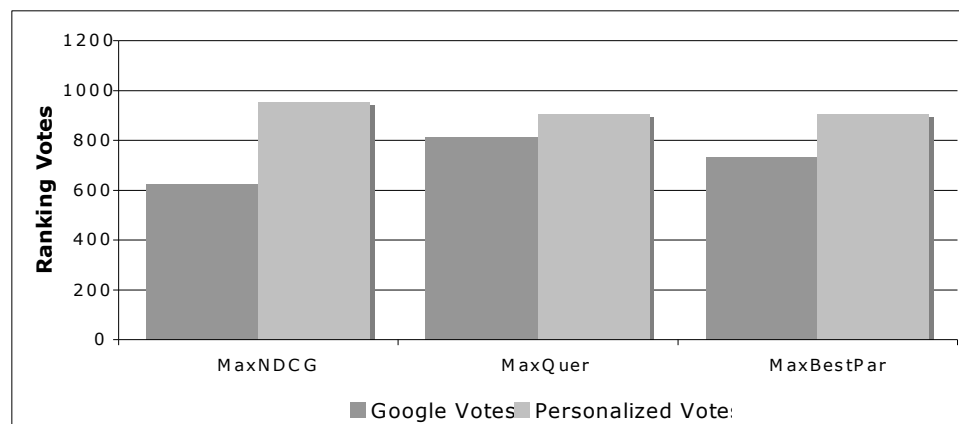
- Assess the selected personalization techniques
- Extend Firefox add-on to do personalization in user's browser as they go
- Interleaved evaluation using Team-Draft Interleaving algorithm (Radlinski et al., 2008)
- Shown to accurately reflect differences in ranking relevance (Radlinski et al., 2010)

Step 2: Online Interleaved Evaluation



Results

- 41 users / 2 weeks / 7,997 queries
- MaxNDCG significantly ($p < 0.001$) outperforms Google
- MaxBestPar significantly ($p < 0.01$) outperforms Google
- MaxQuer significantly ($p < 0.05$) outperforms Google



- Run on all queries: 70% of queries untouched, 20% improved, 10% worse
Average improvement of 4 ranks. Average deterioration of 1 rank.
- One strategy is consistently the best:
TF-IDF, RTitle, RMKeyw, RCCParse, NoFilt - LM, Look At Rank, Visited

Future Work

- Expand set of parameters
 - ✓ Learning optimal weight vector
 - ✓ Using other fields
- Temporal information
 - ✓ How much browsing history should be used?
 - ✓ Decay weighting of older items
 - ✓ Page visit duration
- Other behavioral information
- Use extracted profile for other purposes

Conclusion

- Outperform Google and previous best personalization strategies
- Build an improved user profile for personalization
 - ✓ Not treat web pages as flat documents
 - ✓ Use more advanced NLP techniques
- Improve upon the evaluation methodology
 - ✓ First large online comparative evaluation of personalization techniques
 - ✓ Investigate whether personalization makes difference in real life usage
 - ✓ Done in academic setting, no large datasets available
- Tool that can be downloaded and used by everyone
 - ✓ Code is open sourced, very clean and readable

Questions