Personalizing Search via Automated Analysis of Interests and Activities

Jaime Teevan
MIT, CSAIL

Susan T. Dumains
Microsoft Researcher
One Microsoft Way

Eric Horvitz
Microsoft Researcher
One Microsoft Way

Presented by Sotiris Gkountelitsas
Motivation Example
Initial Results

Google Search for "cancer"

About 224,000,000 results (0.19 seconds)

- Cancer - Wikipedia, the free encyclopedia
  Cancer (IPA: /ˈkænsər/) (medical term: malignant neoplasm) is a class of diseases in which a cell, or a group of cells, display uncontrolled growth...
  Classification - Signs and symptoms - Causes - Pathophysiology
  en.wikipedia.org/wiki/Cancer - Cached - Similar

- American Cancer Society: Information and Resources for Cancer
  Dedicated to helping persons who face cancer. Supports research, patient services, early detection, treatment and education.
  www.cancer.org - Cached - Similar

- Cancer Center: Types, Symptoms, Causes, Tests, and Treatments
  Start here for information on common types of cancer, including breast, lung, colon, skin, prostate, and ovarian cancer. Get the facts on cancer symptoms, ...
  www.webmd.com/cancer/default.htm - Cached - Similar

- Comprehensive Cancer Information - National Cancer Institute
  Accurate, up-to-date, comprehensive cancer information from the US government's principal agency for cancer research.
  www.cancer.gov - Cached - Similar

- What is Cancer? What Causes Cancer?
  Learn about the causes of cancer plus diagnosis and treatment options.
  www.medicalnewstoday.com/info/cancer-oncology/ - Cached

- Cancer care and support charity - Macmillan Cancer Support
Relevance Feedback

- Linux is a cancer that attaches itself in an intellectual ...
  - www.theregister.co.uk/2001/06...

- Cancer: Cancer News and Photos - Courant.com
  - But 9-year-old Mason Kempf of Overland Park, Kansas, who is battling inoperable brain cancer, has just one special wish, and it's one that only Mother ...
  - www.courant.com/topic/health/.../cancer-HEDA0000010.topic - Cached

- CA: A Cancer Journal for Clinicians
  - 28 Oct 2010 ... Peer-reviewed journal of the American Cancer Society providing information on cancer diagnosis, treatment, and prevention.
  - caonline.amccancersoc.org - Similar

- BBC - Health: Cancer
  - BBC Cancer Information from the BBC Health experts on the most common types of cancer and the care and treatment currently available.
  - www.bbc.co.uk/health/.../cancer/index.shtml - Cached - Add to iGoogle

- Cancer (astrology) - Wikipedia, the free encyclopedia
  - Cancer (§) is the fourth astrological sign in the Zodiac, originating from the constellation of Cancer. In western astrology, this sign is no longer ...
  - en.wikipedia.org/wiki/Cancer_(astrology) - Cached - Similar

- Association of Online Cancer Resources
  - Welcome to our cancer online community, find people who are dealing with the same health challenges as you. Join an Online Support Group. Give & Get help.
  - www.acor.org/ - Cached - Similar

Searches related to cancer:
- types of cancer
- cancer horoscope
- cancer astrology
- lung cancer
- cancer sign
- cancer definition
Revised Results
Defining the Problem

• Search engines satisfy information intents but do not discern people.

• Detailed specification of information goals.
  ▫ People are lazy.
  ▫ People are not good at specifying detailed information goals.
Solution

- Use of implicit information about the user to create profile and rerank results **locally**.
  - Previously issued queries.
  - Previously visited Web pages.
  - Documents or emails the user has read, created or sent.
BM25

- The method ranks documents by summing over terms of interest the product of the term weight ($w_i$) and the frequency with which that term appears in the document ($tf_{ij}$).

- Result = $w_i \cdot tf_{ij}$

- No Relevance information available:

$$w_i = \log \frac{N}{n_i}$$

- Relevance information available:

$$w_i = \log \frac{(r_i + 0.5)(N - n_i \cdot R + r_i + 0.5)}{(n_i - r_i + 0.5)(R - r_i + 0.5)}$$
Traditional vs Personal Profile FB

- $N' = N + R$
- $n_i' = n_i + r_i$

$$w_i = \log \frac{(r_i + 0.5)(N - n_i + 0.5)}{(n_i + 0.5)(R - r_i + 0.5)}$$
Corpus Representation \((N,n_i)\)

- We need information about two parameters:
  - \(n_i\) number of documents in the web that contain term \(i\).
    - Issue one word queries
  - \(N\) number of documents in the Web
    - Issue query with the word the.

- The Corpus can be query focused or not.

- Practical Issues:
  - Cannot always issue a query for every term (inefficient).
    - Approximate corpus using statistics from the title and snippet of every document (efficient).
User representation \((R, r_i)\)

- For the representation of the user a rich index of personal content was used that captured the user’s interests and computational activities.
- Index included:
  - Web pages.
  - Email messages.
  - Calendar items.
  - Documents stored on the client machine.
- The user representation can be query focused or not.
- Time sensitivity
  - Documents indexed in the last month vs the full index of documents
- User Interests:
  - Query terms issued in the past.
Document and Query Representation

- Document representation is important in determining both what terms \( i \) are included and how often they occur \( (tf_i) \).

- Terms can be obtained from:
  - Full text
  - Snippets
  - Words at different distances from the query terms
Evaluation Framework

• 15 computer literate participants.
• Evaluation for the top 50 search results.
• 3 possible evaluations:
  ▫ Highly relevant
  ▫ Relevant
  ▫ Not relevant
• Queries:
  ▫ Personal formulated queries.
  ▫ Queries of general interest (Busch, cancer, Web search).
• Discounted Cumulative Gain (DCG):

\[
DCG(i) = \begin{cases} 
G(1), & \text{if } i = 1 \\
DCG(i-1) + G(i)/\log(i), & \text{otherwise.}
\end{cases}
\]
Results (1/2)

The graph illustrates the average normalized DCG across different conditions:
- **Corpus Representation**:
  - Full Text
  - Web
  - Snippet

- **Query Focus**:
  - No
  - Yes

- **User Representation**:
  - No Model
  - Query
  - Web
  - Recent
  - Full Index
  - Snippet
  - Full Text
  - Near Query
  - All Words

The best combination highlighted on the graph is indicated with a red arrow.
Results (2/2)

The bar chart shows the average normalized Discounted Cumulative Gain (Norm DCG) for different methods: Random (Rand), No, RF, PS, URL, Web, and Mix. The chart includes error bars indicating the variability of the measurements. The PS Algorithm and Combined Result are represented by different colors and markers.
Conclusion

• Automatically constructed user profile to be used as Relevance Feedback is feasible.

• Performs better than explicit Relevance Feedback.

• Combined with Web Ranking improves even more the performance.
Further Exploration

• Better tuning of the profile parameters:
  ▫ Time.
  ▫ Automate best parameter combination selection.
  ▫ Additional classes of text and non text based content.
  ▫ Location.
• Thank you for your attention!!!

• Questions