Refining and Personalizing Searches

1

Themes

- Explicit feedback versus search history
- Personalized history versus crowd history

2

Refining and Personalizing: Targets

- collection
  - focused crawling
- query ←
- satisfying documents
  - increase set?
- ranking ←

3

Refine initially: query

- Help user get better query
- Commonly, query expansion
  - add synonyms
    - Improve recall
    - Hurt precision?
    - Sometimes done automatically – with care
  - Modify based on prior searches same query
    - Not automatic
    - All prior searches - eg. suggested search terms
      vs
    - *your* prior searches

4

Refining after search

- Use user feedback
  or pseudo-feedback
    - Approximate feedback with first results
  or implicit feedback
    - e.g. clicks

- change ranking of current results
- search again with modified query
- change ranking for future searches

5

Explicit user feedback

- User must participate
- User marks (some) relevant results
  or
- User changes order of results
  - Can be more nuanced than relevant or not
  - Can be less accurate than relevant or not
    - Example: User moves 10th item to first
      - says 10th better than first 9
      - Does not say which, if any, of first 9 relevant

6

Implicit user feedback

- Click-throughs
  - Use as relevance judgment
  - Use as reranking:
    When click result, moves it ahead of all results didn't click that come before it
  - Problems?

- Better implicit feedback signals?

7

User feedback in classic vector model

- User marks top p documents for relevance
  p = 10 to 20 "typical"
- Construct new weights for terms in query vector
  - Modifies query
  - Could use just on initial results to re-rank

8

Deriving new query for vector model

For collection C of n doc.s

- Let C_r denote set all relevant docs in collection,

Perfect knowledge Goal:

Vector $\mathbf{q}_{opt}$ =

$1/|C_r|$ * (sum of all vectors d_j in C_r) -

$1/(n- |C_r|)$ * (sum of all vectors d_k not in C_r)

centroids

9

Deriving new query for vector model: Rocchio algorithm

Give query **q** and relevance judgments for a subset of retrieved docs

- Let D_r denote set of docs judged relevant
- Let D_{nr} denote set of docs judged not relevant

Modified query:

Vector $\mathbf{q}_{new}$ = $\alpha\mathbf{q}$ +

$\beta/|D_r|$ * (sum of all vectors $\mathbf{d_j}$ in D_r) -

$\gamma/(|D_{nr}|)$ * (sum of all vectors $\mathbf{d_k}$ in D_{nr})

For tunable weights α, β, γ

10

Remarks on new query

- α: importance original query
- β: importance effect of terms in relevant docs
- γ: importance effect of terms in docs not relevant

- Usually terms of docs not relevant are least important
  - Reasonable values α=1, β=.75, γ=.15
- Reweighting terms leads to long queries
  - ***Many*** more non-zero elements in query vector $\mathbf{q}_{new}$
  - Can reweight only most important (frequent?) terms
- Most useful to improve recall
- Users don't like: work + wait for new results

11

Simple example user feedback in vector model

- **q** = (1,1,0,0)
- Relevant: **d1** = (1,0,1,1)
  d2 = (1,1,1,1)
- Not relevant: **d3**=(0,1,1,0)
- α, β, γ = 1
- $\mathbf{q}_{new}$ = (1,1,0,0) + (1, 1/2, 1, 1) - (0,1,1,0)
  = (2, 1/2, 0, 1)

Term weights change New term

Observe: Can get negative weights

12

Explicit feedback: Re-ranking

- Can disambiguate within given results
  - jaguar car versus jaguar animal
- Can modify rankings for future searches
- Algorithms usually based on machine learning
  - Learn ranking function that best matches partial ranking(s) given
- Simpler strategies:
  - use for repeat of same search
    - user reorder or select best
    - Google experiment circa 2007

13

Behavior History

- Going beyond behavior on ***same*** query.
- Personal history versus Crowd history
  - Crowd history
    - Primarily search history
      - Google's claim Bing copies
  - Personal history
    - characterize behavior
    - characterize interests: topics
    - what use?

14

Personal History: sources

- Your searches
- Your social networks
  - Your content
- Other behavior – browsing, mail?, …

15

Collaborative history

- History of people "like" you
- How get?
  - For "free": social networks
    - friends, lists, …
  - Deduce: Crowd history + personal history
    - recommendations
- How characterize?
  - Shared behaviors
  - Shared topics

16

Social Networks and Obtaining Information

17

Social networks

- Catch-all term for
  - social networking sites
    - Facebook
  - microblogging sites
    - Twitter
  - blog sites (for some purposes)
- Now interested in social networks in content sense.
  - not totality of Web
  - not Wikipedia encyclopedia pages
  - yes Wikipedia talk pages?

18

Ways we can use social networks to find information

- Search site
- Aggregate site information to get trends
- ➢ Use site information as meta-information for search
- ➢ Use site properties as meta-information for search

19

Use site information as meta-information for search

- disambiguate queries (Teeven et al 2011 suggested)
  - search Twitter with query
  - analyze content of matching tweets to identify most current, most popular meaning
- factor in ranking URLs (Dong et. al. 2010 studied)
  - harvest URLs mentioned in tweets
  - associate a URL with tweeted text surrounding it
- other uses for tweet text?
- similar analyses of social networking sites such as Facebook?

20

Use site properties as meta-information for search

- interactions: friends, followers, likes, retweets, more?
- Uses
  - ranking by popularity of content
  - ranking by influence of author
- temporal relevance
  - ranking
  - discover URLs faster (Dong et. al. 2010)

21