BIG DATA ANALYTICS

Topic-Sensitive PageRank

ESSEC BUSINESS SCHOOL

Topic-Sensitive PageRank

- Improvement to PageRank
- Weights certain pages more heavily because of their topic
- Alter the way random surfers behave:
  - prefer to land on a page that is known to cover the chosen topic
- Can also be applied to negate the effects of "link spam"

Motivation

- Distinct interests may be expressed using the same term:
  - E.g., "jaguar": the animal, the automobile, a version of the MAC operating system ...
- If we can deduce that the user is interested in automobiles $\Longrightarrow$ we can do a better job
- A private PageRank vector for each user?

General Idea

- **The topic-sensitive PageRank**:
    - one vector for each topic
    - bias the PageRank to favor pages of that topic
    - classifies users according to their interest in each of the topics
- We lose some accuracy...
- But we store only a short vector for each user

Using Topic-Sensitive PageRank (1)

- How to integrate topic-sensitive PageRank into a search engine?

  1. Choose the topics and create specialized PageRank vectors
  2. Pick a teleport set for each of these topics
  3. Use it to compute the topic-sensitive PageRank vector
  4. Determine the set of topics that are most relevant for a particular search query
  5. Use the PageRank vectors for that topic

Using Topic-Sensitive PageRank (2)

Selecting the topic set:

- use the top-level topics of the Open Directory (human classification)

To determine the set of topics that are most relevant for a particular search query:

- Allow the user to select a topic from a menu
- Infer the topic(s):
  - the words from the recent Web pages or recent queries
  - the information about the user:
    - bookmarks
    - stated interests on Facebook
    - ...

Classifying documents by topic: main idea

- **Topics are characterized by words that appear surprisingly often in documents on that topic**
  - e.g. neither "fullback" nor "measles" appear very often in documents on the Web
    - "Fullback": pages about sports
    - "Measles": pages about medicine.

Classifying web pages by topic (1)

- A large, random sample of the Web: the background frequency of each word
- A large sample of pages known to be about sports:
  - Identify the words that appear significantly more frequently in the sports sample than in the background
  - To avoid misspelling: put a floor on the number of times a word appears

Classifying web pages by topic (2)

- $S_1, S_2, ..., S_k$: the sets of words that have been determined to be characteristic of each of the topics
- P: be the set of words that appear in a given page P
- Compute the Jaccard similarity between P and each of the S_i's.
- Classify the page: topic with the highest Jaccard similarity
  - All Jaccard similarities may be very low
  - $\implies$ pick reasonably large sets S_i to cover all aspects of the topic

Back to Topic Sensitive PageRank

- Classify the pages the user has most recently retrieved
- Blend the topic-sensitive PageRank vectors
- Same procedure on the bookmarked pages or combine both

References

- J. Leskovec, A. Rajaraman and J. D. Ullman *Mining of Massive Datasets* (2014), Chapter 5
- O. Klopp, M. Panov, S. Sigalla and A. Tsybakov. Assigning Topics to Documents by Successive Projections. (2021) `https://arxiv.org/pdf/2107.03684.pdf`