2023FY, A1A2, 4840-1031

Monday 14:55-16:40

Web Engineering (ウェブ工学)

https://tinyurl.com/webeng23

Masashi Toyoda

E-mail: mtoyoda@acm.org

The Web as a Platform

- Sharing information
  - Docs, news, images, movies, files...
- Mass media
  - TV, movie, newspaper, radio, magazine...
- Services
  - E-commerce, trading, gaming...

The Web as Data Sources

- Texts
- Images, movies
- Graphs
  - Hyperlink, social network, P2P network...
- Geospatial data
  - Map queries, check-in...
- Time series
  - User traffic, sensor monitoring...

Goal

- Learn recent trends in Web related research including
  - Search engine, Information Retrieval
  - Natural Language Processing
  - Machine Learning, Data Mining
  - Social Media/Network
  - Big Data
  - Cyber-Physical Systems
  - Internet of Things (IoT)

Format

- Seminar Style (Online)
- Two or Three students present research papers (published on top conferences/journals) at each session
  - Presentation: 20 min.
  - Q&A: 10 min.

Organization of sessions

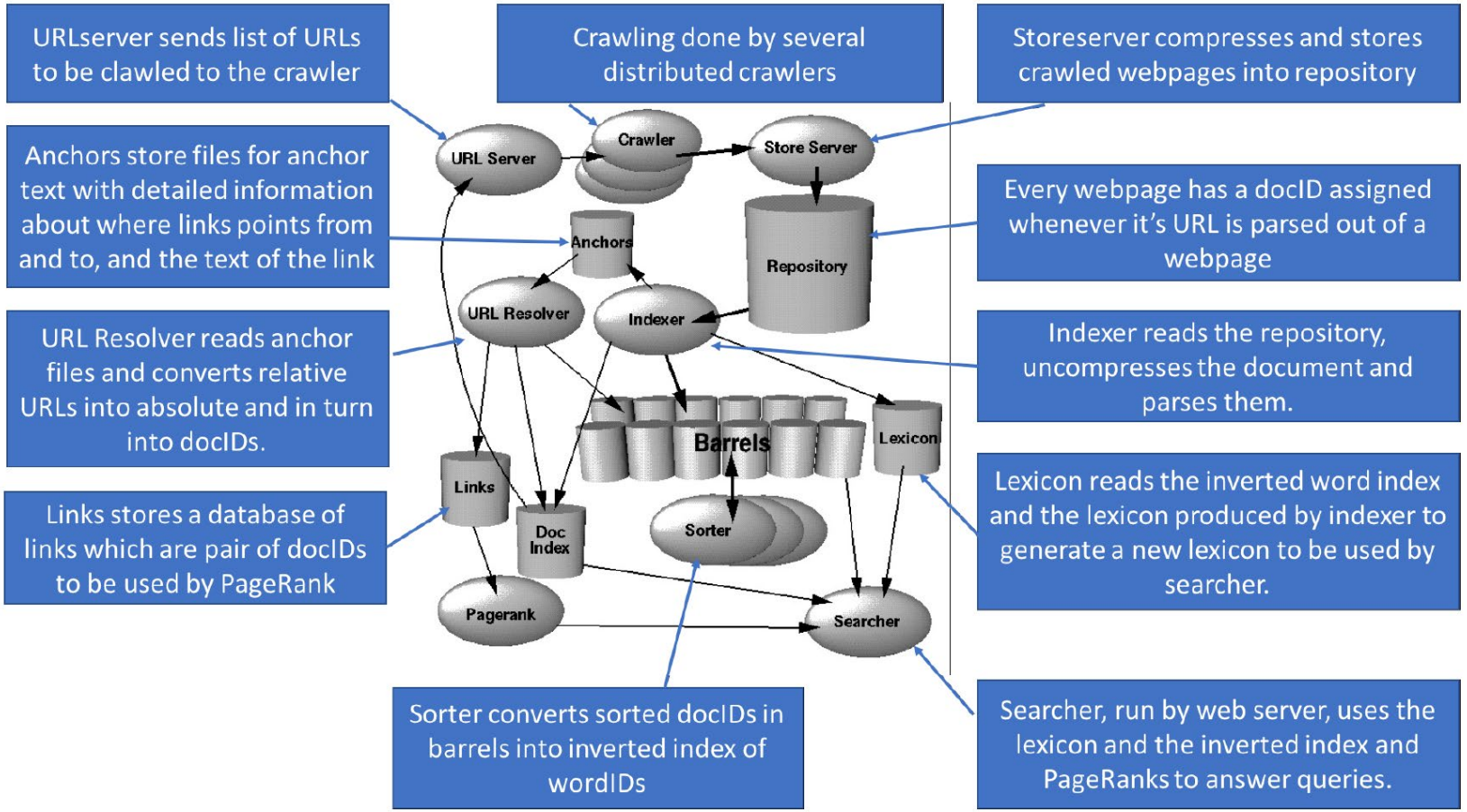
- First half (Session 1 – 4)
  - Test of time award papers from WWW, KDD, etc.
- Second half (Session 5 – 11)
  - Recent research papers
    - Web: WWW, WSDM
    - Database: SIGMOD, VLDB, ICDE
    - IR: SIGIR
    - Data mining: KDD, ICDM
    - AI: AAAI, ICWSM
    - NLP: ACL, EMNLP
    - Etc.

Test of Time Award Papers

- If you want to present in the first half sessions, you need to choose a test of time award paper in WWW, KDD, SIGIR, etc.
  - WWW Seoul Test of Time Awards https://www.iw3c2.org/ToT/
  - KDD Test of Time Awards http://www.kdd.org/awards/kdd-test-of-time-award
  - SIGIR Test of Time Awards http://sigir.org/awards/test-of-time-awards/

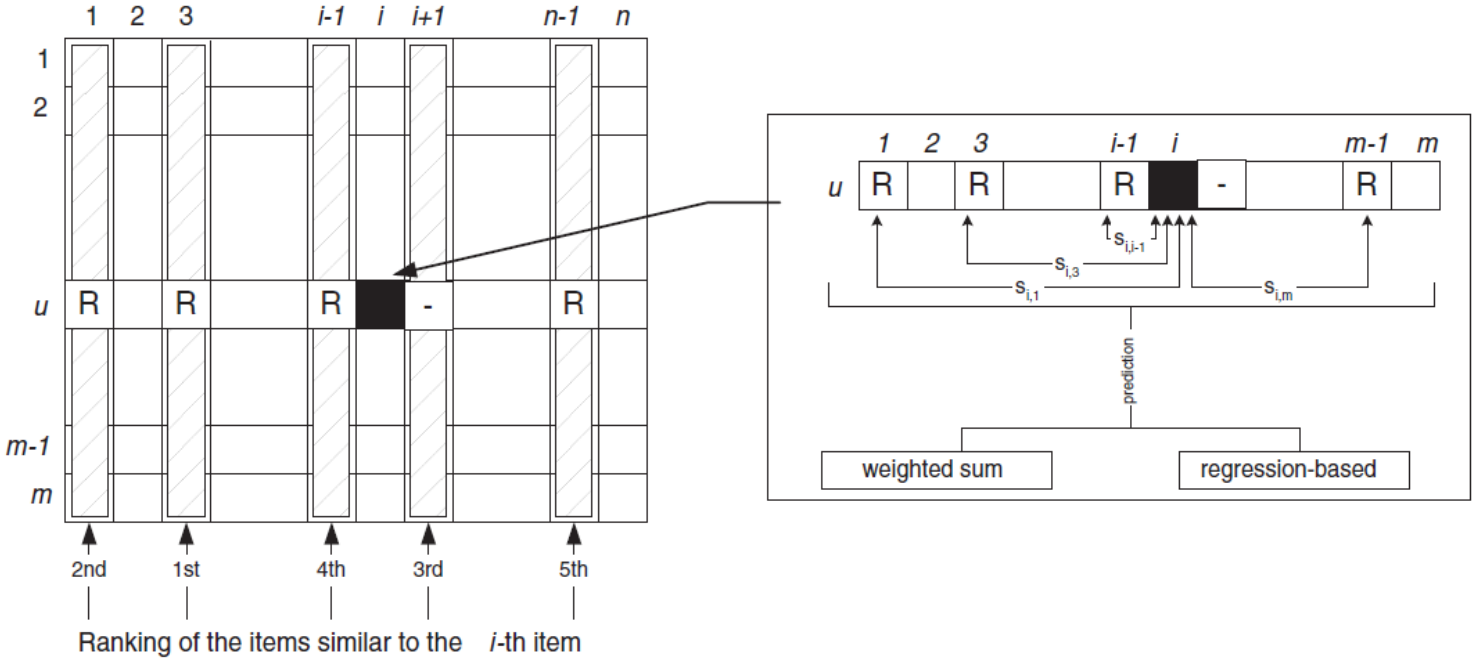
WWW2015: Test of Time Awards
The Anatomy of a Large-Scale Hypertextual Web Search Engine [S. Brin, L. Page, WWW1998]

- Google's architecture in the early days



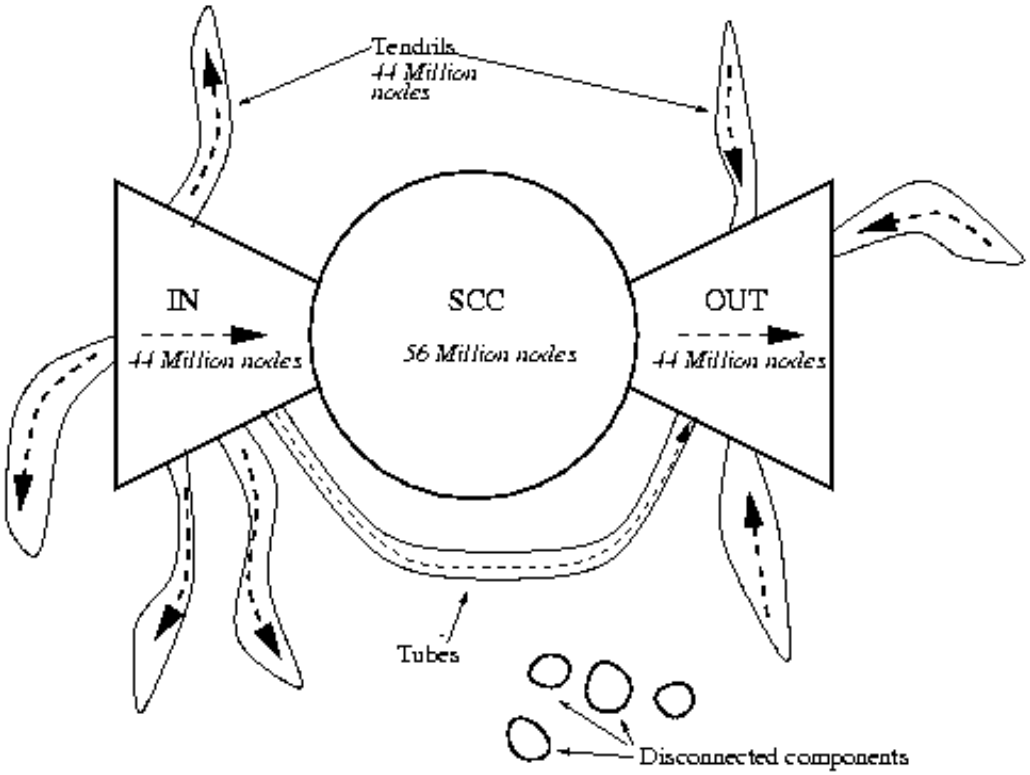
WWW2016: Test of Time Awards
Item-based collaborative filtering recommendation algorithms [B. Sarwar+, WWW1998]

- Many collaborative filtering systems now use this approach



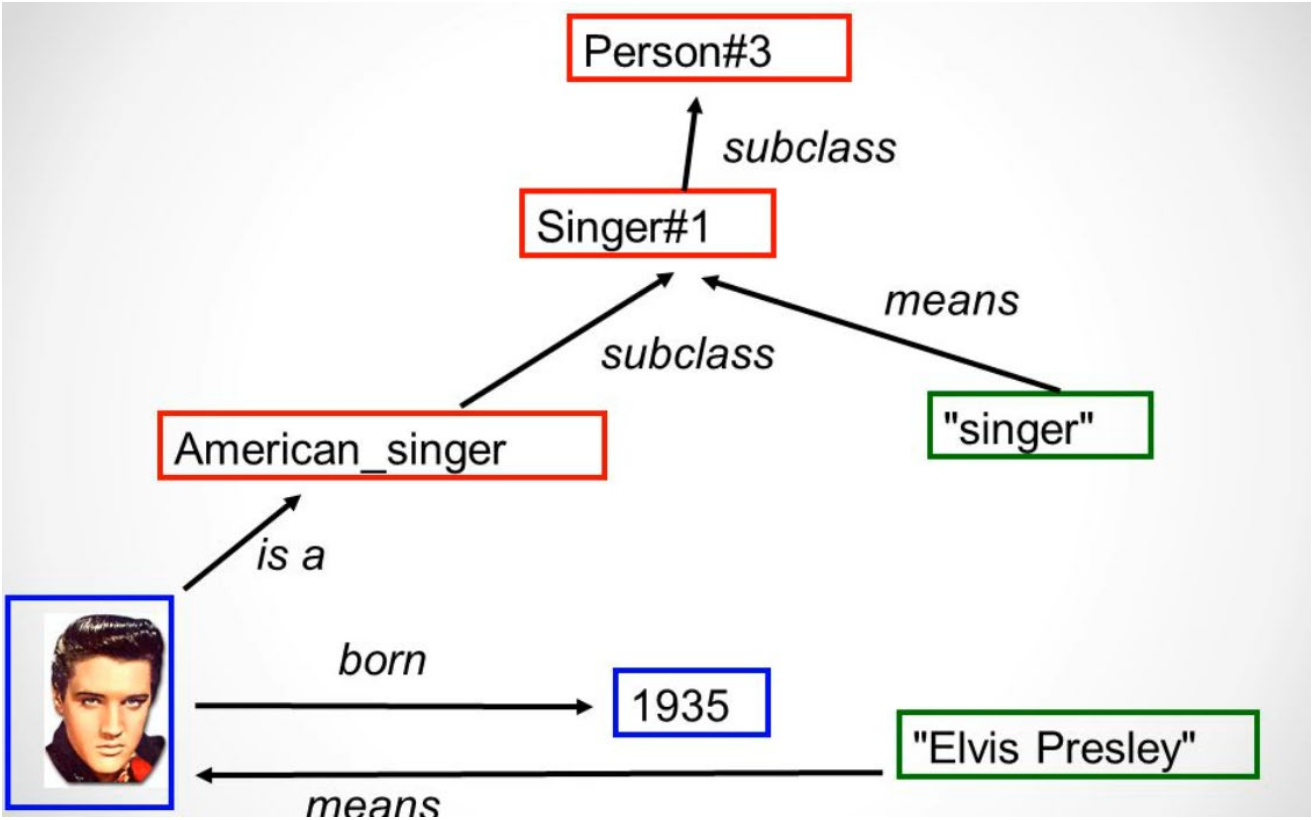
WWW2017: Test of Time Awards
Graph Structure in the Web [A. Broder+, WWW2000]

- The first large-scale empirical study of the dynamics of the Web



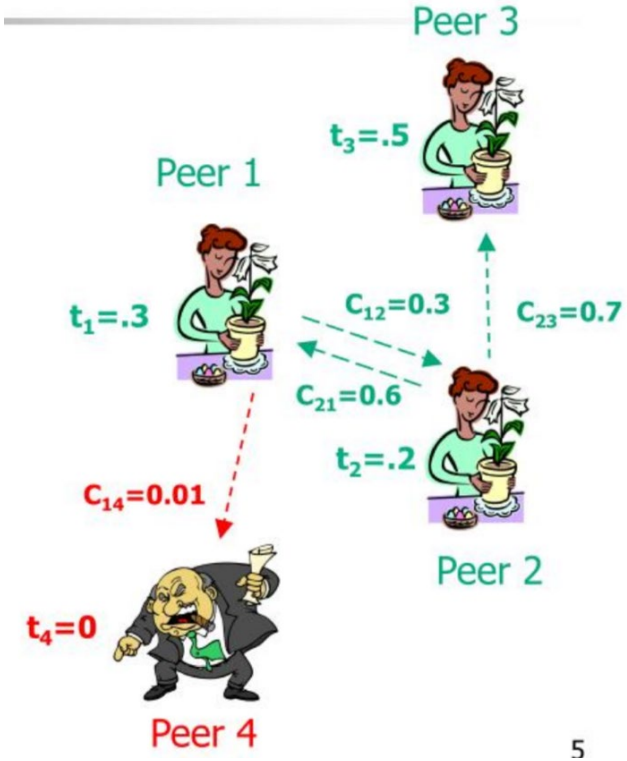
WWW2018: Test of Time Awards
YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia [F. Suchanek+, WWW2007]

- One of the first projects to extract semantic knowledge graph at large scale from Wikipedia



WWW2019: Test of Time Awards
The EigenTrust Algorithm for Reputation Management in P2P Networks [Kamvar+, WWW2003]

- In P2P file sharing networks, reduce inauthentic files distributed by malicious peers
  - Goal: Identify sources of inauthentic files by gibing each peer a trust value based on its previous behavior
  - Algorithm: Propagate local trust values to friends, friends of friends, and so on
    - Propose a distributed algorithm to calculate this eigen trust value



http://tinyurl.com/webeng23

WWW2020: Test of Time Awards
Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews [Dave+, WWW2003]

- Aggregate product reviews in various sites
  - Extract product attributes and aggregate pos/neg about them

the _productname... (+0.00) the nr70... (+16.94) the memory... (+13.30) the palm... (+13.40) the screen... (+2.97) the jog... (+4.19) the same... (+2.62) the cli... (+46.01) the device... (+2.36) the new... (+4.88) the best... (+5.18) the first... (+2.73) the cradle... (-1.47) the clie's... (-0.05) the most... (+4.61) the handheld... (+4.95) the top... (+6.14) the pda... (+4.15) the only... (+5.85) the m505... (+2.79) the other... (+1.88) the software... (+4.05) the unit... (+2.86) the hotsync... (-2.11) the camera... (+1.56) the graffiti... (+3.61) the palmos... (+4.78) the stylus... (+0.28) the standard... (+1.73) the nr70v... (+0.40) the speaker... (+3.18) the next... (-0.50) the ability... (+0.55) the size... (+3.68) the t615c... (+1.02) the bottom... (+0.28) the ms... (+1.44) the handera... (+2.21) the peg... (+2.46) the original... (+1.14) the user... (+1.08) the right... (+1.70) the default... (+2.00) the keyboard... (-0.76) the market... (+1.93) the NUMBER... (+0.00) the us... (+2.12) the way... (+1.00) the left... (+0.61) the sync... (+0.58) the color... (+1.15) the picture... (+0.40) the speed... (+1.31) the fact... (+1.32) the visor... (+1.33) the mobile... (+2.34) the pictures... (-0.25) the ir... (+0.47)

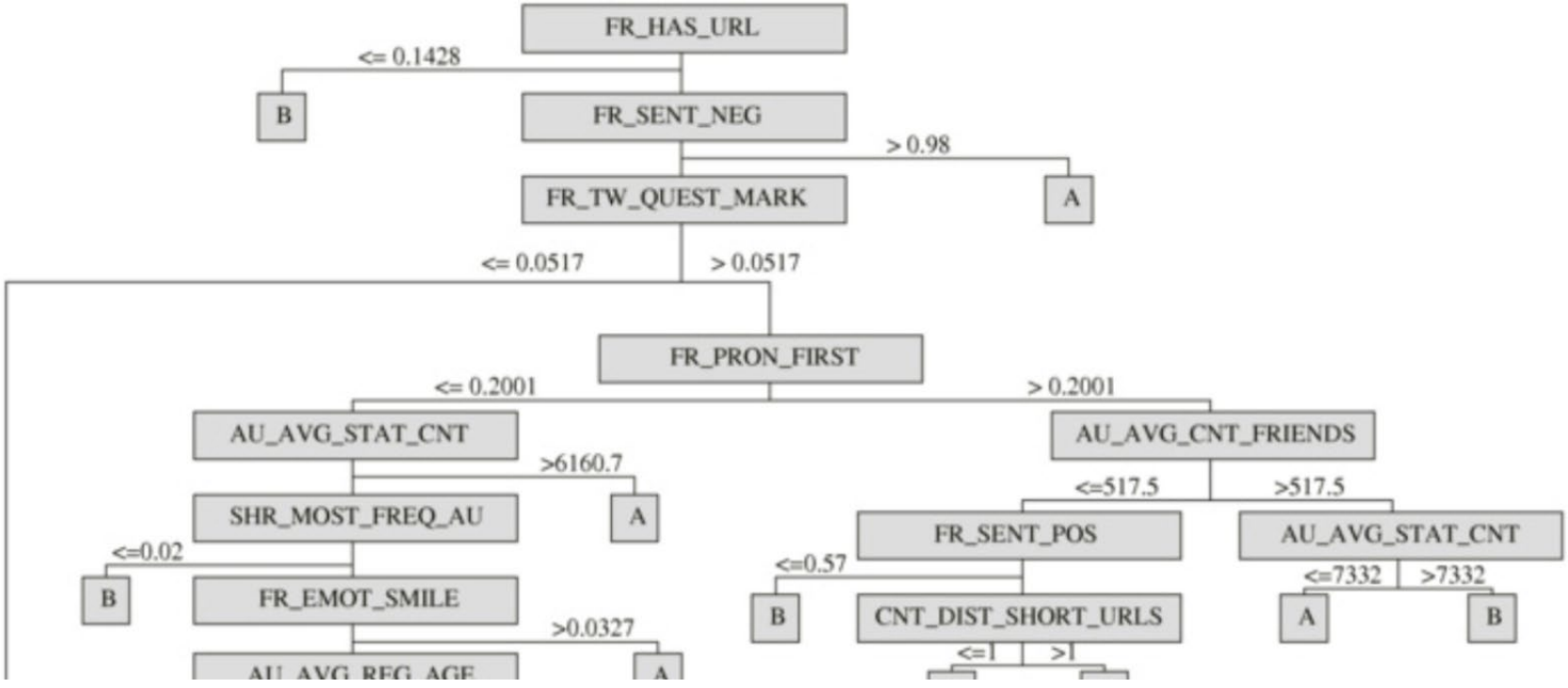
the nr70...

+2.39: The PhotoStand application is one of the only apps preinstalled on the CLIE to take full advantage of the stunning 320x480 display of the NR70
+1.54: The Memory Stick support in the NR70 series is well implemented and functional, although there is still a lack of non-memory Memory Sticks for consumer consumption
+1.44: Unlike the more recent T series CLIE's, the NR70 does not require an add-on adapter for MP3 playback, which is certainly a welcome change
+1.32: While not as fast as processors found in other handheld platform designs, this CPU allows the NR70 to perform responsively while yielding longer battery life than handhelds with faster processors
+1.28: As with every Sony PDA before it, the NR70 series is equipped with Sony's own Memory Stick expansion

http://tinyurl.com/webeng23

WWW2021: Test of Time Awards
Information Credibility on Twitter [Castillo+, WWW2011]

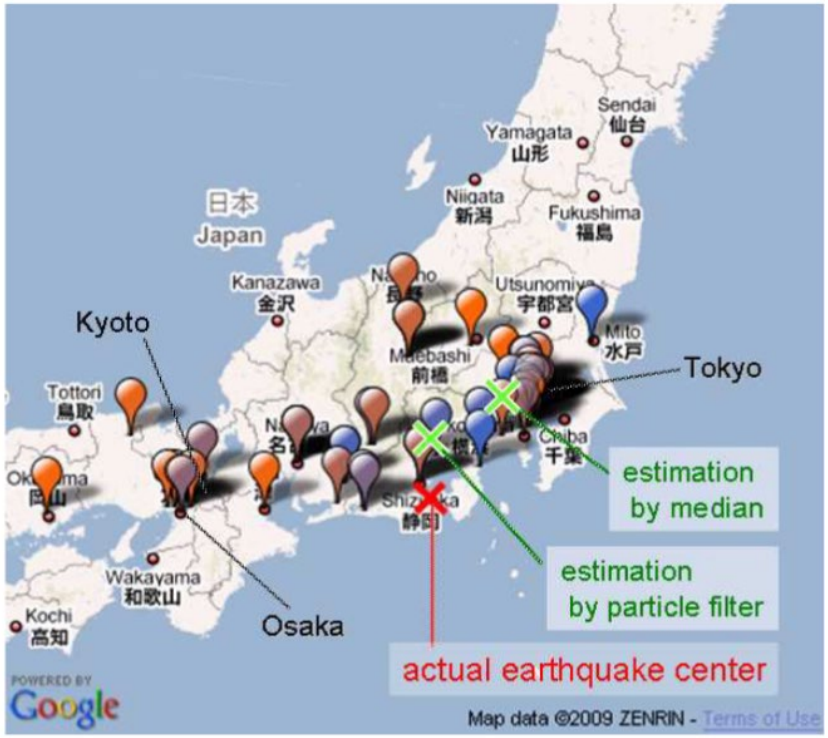
- Classify tweets related to newsworthy events by their **believability**
  - Users believe those tweets are to be true or false

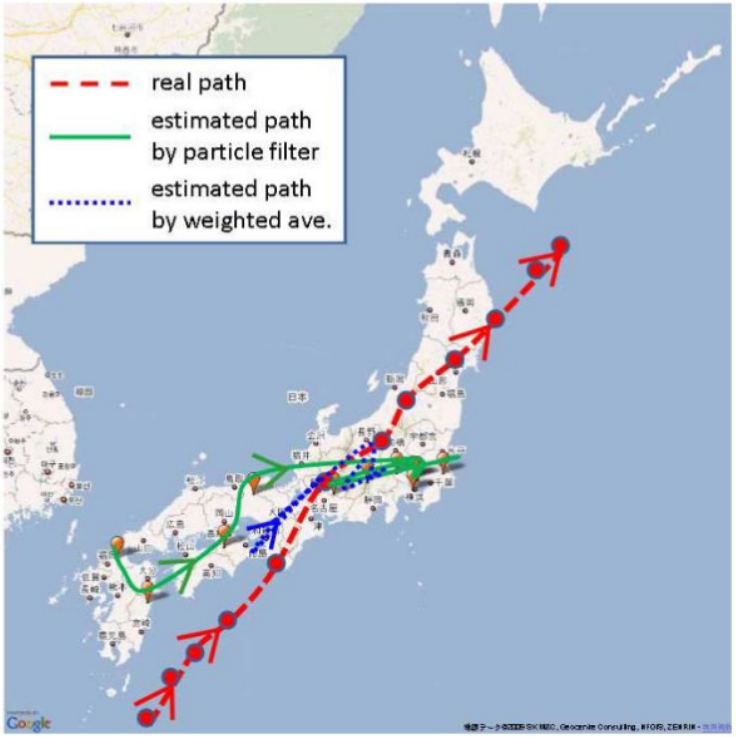


http://tinyurl.com/webeng23

WWW2022: Test of Time Awards
Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors [Sakaki+, WWW2010]

- Detecting real world events using Twitter as Sensor





http://tinyurl.com/webeng23

WWW2023: Test of Time Awards
A Contextual-Bandit Approach to Personalized News Article Recommendation
[Li+, WWW2010]

- A learning algorithm sequentially selects articles to serve users based on contextual information
  - Users and articles, while simultaneously adapting its article-selection strategy based on user-click feedback

Figure 1: A snapshot of the "Featured" tab in the Today Module on Yahoo! Front Page. By default, the article at F1 position is highlighted at the story position.

Recent research papers

- If you want to present in the last half sessions, you need to choose:
  - Choose full research papers (over 8 pages) in top conferences/journals published in recent 4 years (2020 – 2023)
  - **DO NOT** choose workshop, poster, and preprint (e.g. arXiv.org) papers
  - Topics of the papers should include experiments with a large-scale data collected from the Web or real world
  - Cross-disciplinary papers are recommended

Presentation

- Talk: 20 min., QA: 10 min.
- Language: Japanese or English
- Slides should include:
  - Importance of the paper in the research area
  - Motivation, purpose, and goal
  - Novelty (compared with related work)
  - Method
  - Experimental results
  - Strong points and weak points

Slack

- We use slack.com to announcement and chat in sessions
- I will check your questions and comments in sessions and mark them if they are “good” (Mark will become scores for your credits)

- After receiving your e-mail, I will invite you to **webeng2023.slack.com**
- Your name in slack should be “Your_Name_StudentID” (Please do no use nicknames)

Credits

- Point system:
  - **100** points (優)
  - **90** points (良)
  - **80** points (可)
  - Less than 70 points (不可)
- Scores:
  - Presentation: **50** pts. (required)
  - Non-trivial question/comment: **10** pts./session (Do not count multiple times in a session)
  - Attendance: **0** pts.

Schedule

http://tinyurl.com/webeng23

Date		Session
10/16	Mon	Guidance
10/23	Mon	Guidance
10/30	Mon	Session 1: Test of time award papers
11/6	Mon	Session 2: Test of time award papers
11/13	Mon	Session 3: Test of time award papers
11/20	Mon	Session 4: Test of time award papers
11/27	Mon	Session 5: Recent papers
12/4	Mon	Session 6: Recent Papers
12/11	Mon	Session 7: Recent papers
12/18	Mon	Session 8: Recent papers
12/25	Mon	Session 9: Recent papers
1/15	Mon	Session 10: Recent papers
1/22	Mon	Session 11: Recent papers
1/29	Mon	Session 12: Recent papers

Assignment

- Today, we decide presenters for the 1st and 2nd sessions
- All students should e-mail following info. **by 27 Oct**.
  - Student ID Number, Name, e-mail for slack
  - 3 candidate dates for presentation, and
  - 3 papers you want to present with complete bib info.
    - Authors, Title, Conference/Journal, pages, year
- Assignment & Schedule will be released **on 30 Oct**.
- Note:
  - Students who select earlier day have more chance to present a preferred paper
  - Let me know ASAP, if you change the paper to present