An excursion into

Visualization

Metro Maps

Dafna Shahav WWW 2012

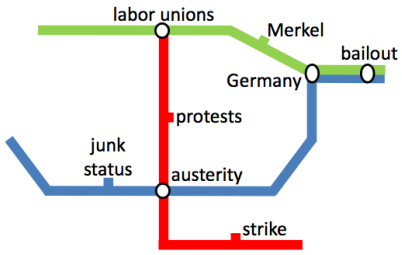


Figure 1: Greek debt crisis: a simplified metro map

Goal

- Corpus of news articles (test case)
- Query on corpus
- Automatically organize results into "metro map"
  - Line
    - "coherent narrative thread"
    - one aspect of story
  - Relationships between lines
    - Intersect
    - Overlap
    - Branches

Formal definition

Metro map is a pair (G, P)

G is directed graph

P set of paths in G

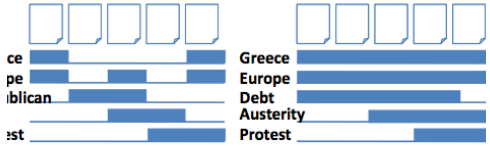
Such that

Each edge E of G must be on at least one path in P

Define properties that make good maps

Coherence

- Concept of importance of word for two consecutive documents in a line
- Choose small set words for scoring
- Define optimization problem



re 2: Word patterns in Chain A (left) and B (rig
correspond to the appearance of a word in the
depicted above them.

5

Coverage

- "cover important aspects of story but encourage diversity"
- Coverage feature (think word) in documents
  - Eg tf.idf
- Coverage feature in map:

  1- (product over all docs (1 – coverage feature in document))
- Coverage of a map for corpus:

  weighted sum over features of coverage of feature in map
  - Simple example weight: word frequency in dataset

6

Connectivity

- Number of lines of in set of paths that intersect.

7

Putting it together

- Maximize coverage under the constraints:
  - Coherence ≥ chosen threshold
- Maximize connectivity among those maps that maximize coverage

8

Example (condensed)

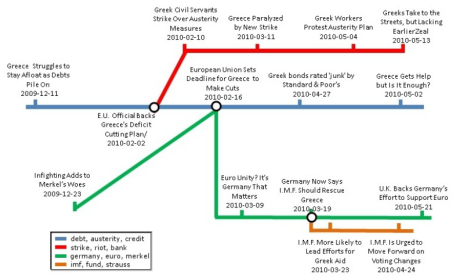


Figure 5: An example of our results (condensed to fit space). This map was computed for the query 'Gree* debt'. The main storylines discuss the austerity plans, the riots, and the role of Germany and the IMF in the crisis.

9

Study with experts judging

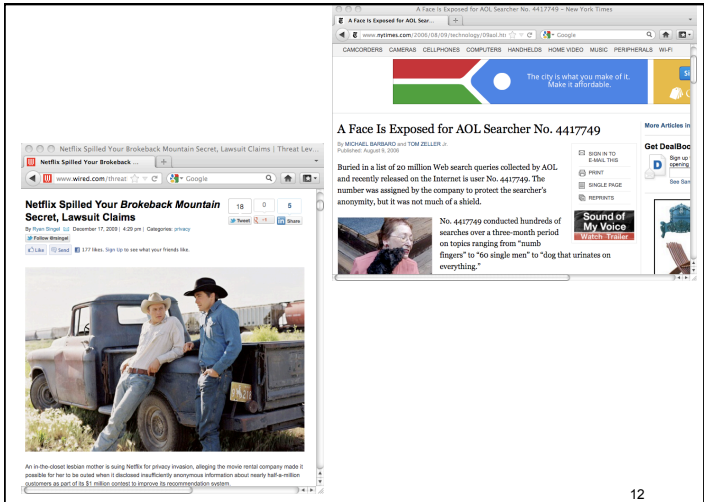
- Experts choose top 10 events for topic
- Measure fraction of important events retrieved
- Vary number of lines used (add to optimization)
- 3 topics: Chile, Haiti,Greece

Lines	3	4	5	6
Chile	80%	100%	100%	100%
Haiti	50%	70%	80%	80%
Greece	30%	60%	60%	70%

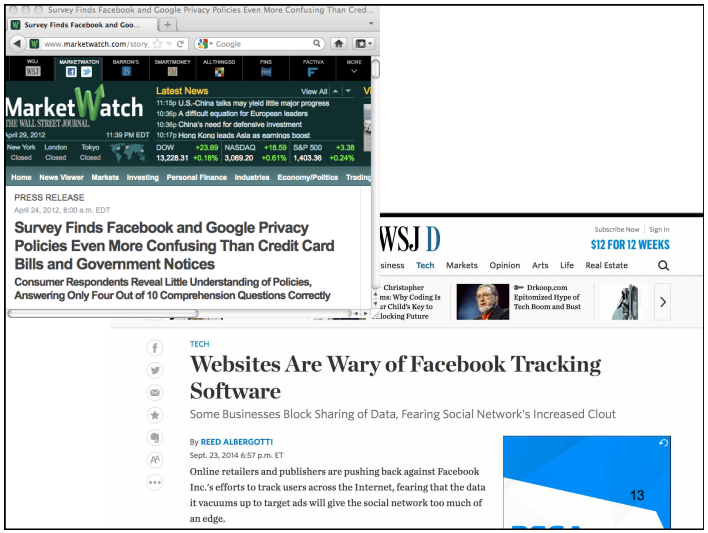
10

Privacy

11



12



Exposing users: techniques

Look at

[You Might Also Like: Privacy Risks of Collaborative Filtering](), Calandrino, J.A, Kilzer, A., Narayanan, A., Felten, E.W., and Shmatikov, V., *IEEE Sym. on Security and Privacy (SP)*, 2011, pp. 231 - 246.

- Various item-to-item collaborative filtering methods
- Practical algorithms

14

Set up

- attacker and target user
- attacker to infer unobservable transaction by target user
  - e.g. item purchased or rating given item
- attacker uses "auxilary information" about some transactions of target user
- attacker only observes
  - does not enter ratings/ make transactions
  - no fake users

15

Sources of auxiliary information

- provided by target system
  - e.g. public ratings by user
- "third-party sites"
  - partner with target site
  - e.g. embed playlist on blog
- other sites
  - user places related content
  - e.g. Facebook user profile

16

"Generic Inference Attacks"

- Auxiliary information
  - target system provides lists of related items
  - target system provides item-to-item covariance matrix used by collaborative filtering
- Auxiliary information & Active attack
  - target system uses k-nearest neighbor recommender

17

Using related items

- system gives list of related items for each item based on user selection
- auxiliary items: attacker knows certain items associated with target user
- attacker
  - monitors related-items lists of auxiliary items
  - scores changes in lists:
    - new items appear or items move up on lists
  - if score for an item above threshold, infer item added to target user's record

18

Using covariance matrix

- item-item covariance matrix M available
  - Hunch.com questions to users
- user record containing items interacted with
- auxiliary information: attacker knows subset A of items associated with target user u
  - new item in record for u => covariances beween new item and (some) items in A goes up
  - subset unique to target user?

19

Using covariance matrix, cont.

- attacker
  - monitors changes in covariance submatrix
    - columns for A
    - rows A U {candidate new items}
  - scores changes in submatrix
  - if score for an item above threshold, infer item added to target user's record
- Lots of details concerning update delays in paper

20

Active attack: for kNN recommender systems

- Example target system
  - similarity measure on users
  - find k most similar users to user u
  - rank items purchased by one or more of k most similar users
    - ranking by number times purchased
  - recommend items to u in rank order

21

kNN recommender systems, cont.

- auxiliary information: subset of m items target user U has purchased
  - claim m of about O(log (# users)) suffices
- attacker
  - creates k sibyl users
  - puts m auxiliary items on sibyls' histories
    - "high probability" kNN of each sibyl is other k-1 sibyls and U
  - infer that any items recommended by system to any of sibyls and not one of m aux items is item U has purchased

22

Evaluation

- use
  - yield: number inferences per user per observation period
  - accuracy: percentage of inference that are correct
- need "ground truth"
- Several studies in paper
  - Hutch.com, LibraryThing, Last.fm

23

used on Amazon

- no ground truth
- API provides "Customers who bought x also bought y" and sales rank of items
- chose customers: top reviewers but not among top 1000 reviewers
- auxiliary info: entire set items previously reviewed by chosen customers
  - avg ~120 per customer
  - misses items purchased w/out reviewing

24

Inference for Amazon

- collected data for 6 mo
- only considered customers who reviewed in 6mo. before or during data collection
- each item, each user: retrieved top 10 most related items
- infer: customer purchased t if t appears or rises in related-items list associated with at least K auxiliary items for the customer
  - K parameter
- evaluate with case studies
  - find item later reviewed

25

Privacy issues in search, recommendations, and other information services

In Practice:

- What is privacy?
- Kinds of problems?
- What problems are of concern?
- How address?

26