VIDI

Visualizing Discussion Forums

Mitja Trampuš
Marko Grobelnik
Jožef Stefan Institute, Slovenia
Exploratory analysis

Why analyze?

• To confirm/reject hypotheses: Is it true that men are more pro-abortion than women?

• To hypothesize on new interesting facts: What is special about abortion (in the context of this forum)?
Using VIDI

- Javascript webpage plugin => accessible to everyone
- Javascript injects a **toolbar** inside the webpage
- Invoked by the user or the forum admin
Toolbar at a glance
Topical Atlas
Topical Timeline

Analysis:
- Suggestions
- Atlas
- Timeline

Status: Ready

Pattern detection:
- Create pattern

Selected items:
- talk.politics.misc
- comp.lang

From: 1990-01-01
To: 2010-09-01

Number of messages per week

Keywords:
- AUG, PEOPLE, NEWS, WRITE, OBAMA, BLACK, DEAD, BECK, TAX, PAKISTAN, WILLIAM, PRICE, BUSH, MUSLIM, HA, AMERICAN, WAR, BANK, FOXTROT, PARTY
- AUG, A0, JUL, JUN, YAHOO, RFISC, LIE, OBAMA, NET, RIGHTARD, GMAIL, ANUS, RETARD, STUPID, LAMONT, ANDXORNOT, DOUG, RAY, SONIC, LYING
- JEWISH, BUSH, NEWS, JUL, YAHOO, THERMITE, SHOT, SEMITE, ANTI, COLLAPSE, PDT, KING, MEE, NANO, WRITE, STEEL, ETERNAL, AUG
The nice things about VIDI

• English, Slovene, Slovak, Hungarian
• Scales well: tested on 5 000 000 posts
• A few seconds up to a minute for an interactive chart

• Technical details + demo @ poster
Overview

• Data acquisition
• Data (pre)processing
• Visualizations and underlying algorithms

• User interface
• Architecture
  How it all fits together
Data Acquisition

• Local cache updated every 10 minutes

• INePA, LATA:
  Access to databases via SQL sockets
  – Adaptation needed

• index.hu:
  Web crawler (“fake user”)
Have you seen <b>any</b> members of the European Parliament?

- **Charset normalization**
  ¿Have you seen <b>any</b> members of the European Parliament?

- **Markup removal (HTML, bbCode)**
  ¿Have you seen any members of the European Parliament?

- **Tokenization**
  - URLs
  - “Big tokens”, frequent n-grams

[P], Have, you, seen, any, members, of, the, European_Parliament, [P]
Preprocessing

• Stopword removal
  – Heuristics: “wheeeeeeeeee
  – removed words: have, you, any, of, the
  seen, members, European Parliament

• Lemmatization
  see, member, European Parliament

• Mapping to IDs, TF-IDF vectors computation
  (0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, ...
Named Entity Extraction

- Performed in “real time”, sent to FZI’s pattern editor in real time
- **Heuristic**: a name is any capitalized word not at the beginning of the sentence, or any known name
- **Consolidation**: clustering
  - Distance metric: Levenshtein (declinations, typos)
Browsing Suggestions

- Cosine distance between TF-IDF vectors
- For scalability: cache vectors for groups of posts (i.e. whole threads)
Topical Timeline

• Clustering
  – k-means with multiple random restarts
  – cosine distance between TF-IDF vectors
  – hierarchical recursive
  – for scalability: TF-IDF vector cutoffs

• Future improvement: interpolation

• Flash with ActionScript
Topical Atlas

• MDS (MultiDimensional Scaling)
  – iterative procedure

• Dealing with high number of posts
  – Clustering
    – SVD (Singular Value Decomposition)

• Flash with ActionScript
Toolbar GUI

- Built using GWT (Google Web Toolkit)
- Technically part of the hosting page (HTML injection via DOM manipulation)
- For per-user deployment: *Bookmarklets*, i.e. javascript-bound bookmarks
Architecture
(for visualizations)

- DB (PostgreSQL)
- Data preprocessing
- Data acquisition
- Scheduler
- Web service front-end, real-time analysis (python)
- Real-time analysis (C++)
- GUI (Java)
- Visualizations (Flash/AS)
- HTTP GET / XML
- JSON/P
- Server
- Client