A tool for Exploring a News Corpus

Through **entity/topic interaction**, graph analysis, document summarization & news visualization

By
Rahul Goyal : 2008CS50222
Ravee Malla : 2008CS50224

under the guidance of
Dr. Maya Ramanath
Dr. Amitabha Bagchi

Motivation

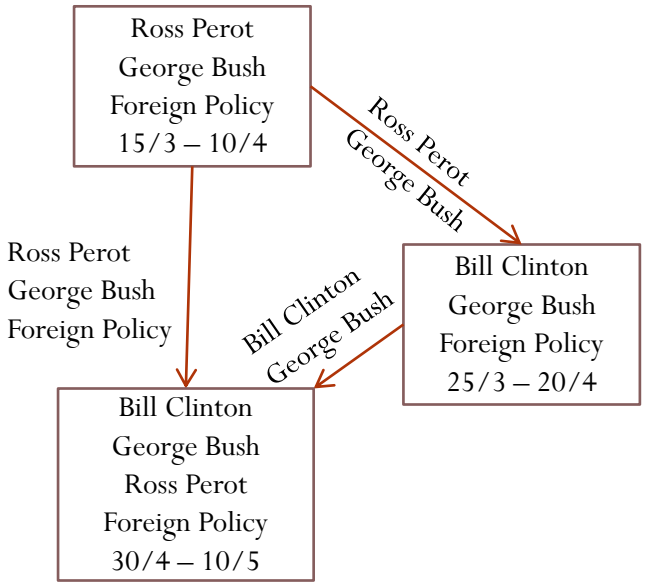
- Why are we building a tool for news browsing?
  - News browsing is one of the primary uses of the internet
  - News on the internet is presented as a set of text documents, classified into categories based on
    - Sections of a news (Sports/Nation/International)
    - People/Events (that are talked about, *manually* tagged)
    - News sources (NY Times, LA Times, TOI, etc)
  - Searching for relevant news in this maze is difficult
    - Relevant in time? In actors? In context? Time-scales?
  - Mimic the structure of a conventional news paper

Past Work

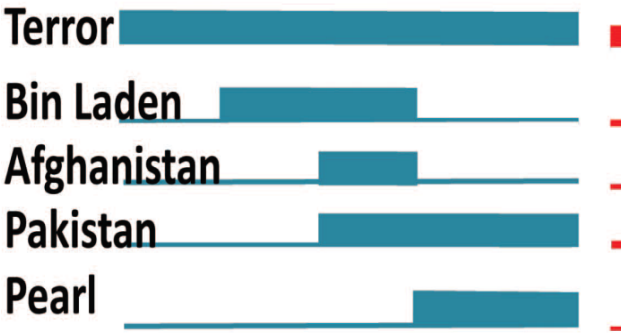
- A Framework for Exploration of News Corpora by Actor Evolution and Interaction [1]
  - Visualizing news articles as a graph, with each node for an article with the dominant actors connecting nodes throughout
  - Focus is **still on the article**, hard to visualize what is happening to an actor(s)
  - # of article links **grow very quickly** with larger sets
- Connecting the dots between news articles [2]
  - Take 0, 1 or 2 articles & form coherent chains
  - Computationally expensive, offline process
  - No way to guide/control the article selection process, so we can't be sure of getting the complete picture

Past Work

Graph Representation



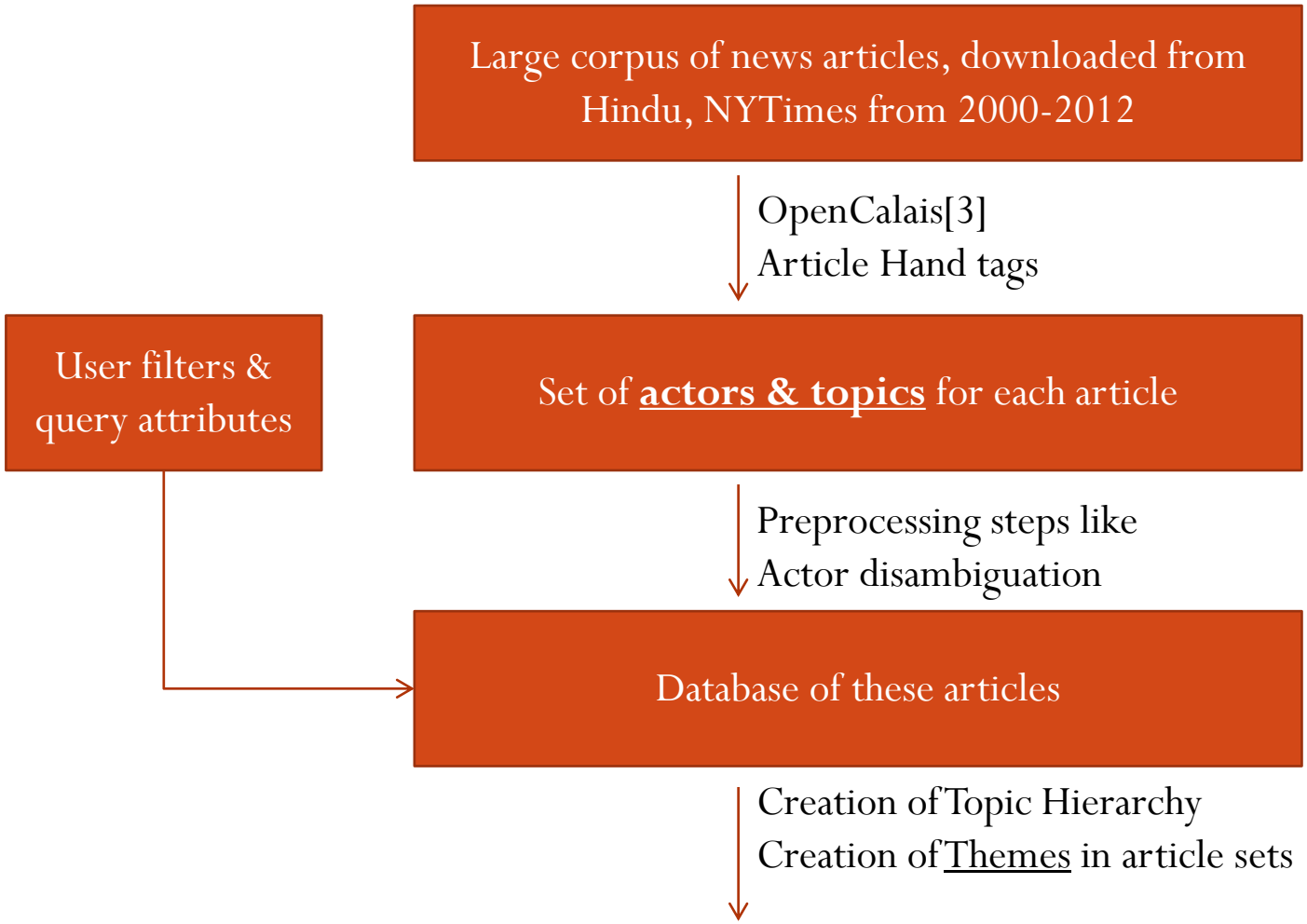
Timeline coherence chain



A dual relation exists between these 2 representations.

Can we combine them both?

Broad Framework

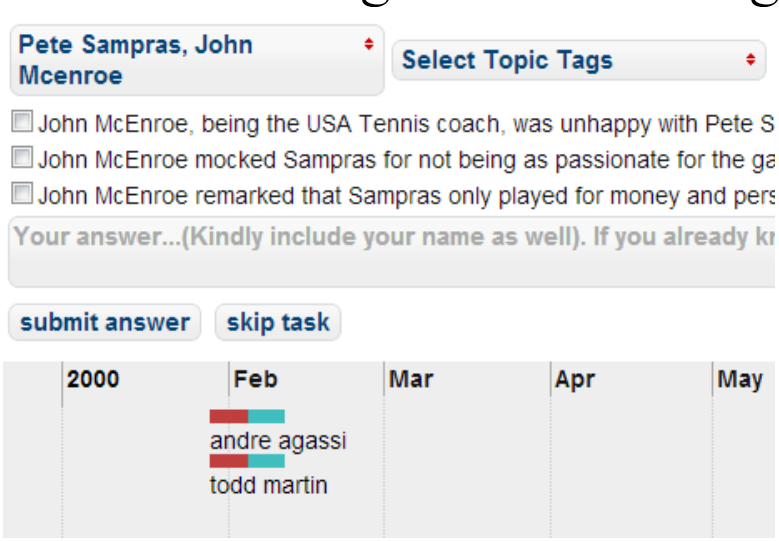


The problem of Clean Tagging

- It is inherently hard
  - Entity Disambiguation (R Singh, Raman Singh)
  - Broad tags vs Narrow tags (*World* vs *Wimbledon Day 2*)
- Tried out LDA, Author-Topic Models, PLSI, TFIDF but all have their disadvantages
- However, there are ways to counter this effectively
  - Search on an ontology
  - Edit distance cleaning of tags like *Mr. Manmohan Singh*
- However, more & more content pushed online, especially **online news**, is being tagged manually

Interface Visualization 1

- Key concept: **Actors, Topics & Themes on a timeline**
- For every actor appearing in one or more articles
  - Draw a track on timeline for it, showing all the events (in the form of articles) and related topics
- Actor/Topic interactions are captured by filtering on them, and looking at the resulting articles



Interface Visualization 2

- But it is confusing to go through another actor to study the visualization of the filtered parameters. Also it leads to redundancy
- AK, SG, PB & NG seem to be appearing in the same topics Infact, they appear due to the **same article set.** So can we aggregate these actors & topics?
- Moreover, can dissimilar interactions, in the same time period, be shown separately?

**A particular task. Topics are color coded.*

Creating Themes out of Article sets

- Define a theme to be a partition of articles in a time window on the basis of their actor/topic labels
- Algorithm CreateTheme(S):
  - P = most_popular_actor_in_S
  - S_P = {articles containing P}
  - Themes = CreateThemes(S – S_P)
  - return (Themes ∪ {S_P})
- Works best when every actor has a preferential interaction with a fixed set of actors
  - Experimentally, among the actors that co-occur with P, *only a small fraction* also occur in the articles that don't contain P
- Idea of Actor – Topic Hierarchies (sub-actors & sub-topics)

Future Work

- Get user feedback by way of task solving
  - Kindly invite you to try out the tool and give us feedback!
- Integrate more features
- Make a complete end-to-end system which downloads news in real-time, and pushes on to the interface
  - Possibly integrated with twitter/news/sentiment analysis results from other projects

Interface Demonstration

Thank You!

Questions & Comments.

References

[1] R. Choudhary, S. Mehta, A. Bagchi, and R. Balakrishnan. Towards characterization of actor evolution and interactions in news corpora. In *Advances in Information Retrieval*.

[2] Shahaf, D., Guestrin, C.: Connecting the dots between news articles. Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining

[3] api.opencalais.com