Exploration of News Corpora

By studying actor evolution & graph visualization over time

by
Ravee Malla
2008CS50224

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

Information Need: *How did Bill Clinton win the 1992 elections?*

- How to frame the query?
- Suffers from lack of temporal & contextual coherence
- Impossible to get a sense of issue relevance
- Where do I start?
Motivation

Information Need: *How did Bill Clinton win the 1992 elections?*

- Set all relevant actors as filters \{George Bush, Ross Perot, Bill Clinton\}
- Set the time period in which articles are to be studied \{1992\}
- Immediately reveals how their relationships evolved temporally
Problem Statement

- Structure news articles as a directed graph for ease of analysis
- Derive meaning/context/connections from articles
- Create an easy to use news browsing tool
- Planned as an extension of work done in [1]
A Framework for Exploration of News Corpora by Actor Evolution and Interaction [1]

- **Actors**: Dominant entities appearing in a news article
- Place a node for each article labelled by it’s actors
- Model interactions as set of 5 *transformations*

- Score each transformation and pick those above a threshold
Our Dataset & Approach

- Consider the articles from NYTimes archives
- Filter out articles related to 1992 US Elections
- Extract actors for each article using Open Calais API [3]
  - returns the dominant entities, topics, social tags, relationships
- Construct the graph of these articles as discussed in [1]
  - Analyze the graph for interesting properties
  - Improve the edge coherence of graph
Results: Structure of the graph

- Linear chains in the graph
- Represent fast developing breaking news
- Articles of ‘central’ importance in the graph
- Summary articles

Bill Clinton & George Bush interacting with other actors

Article summarizing about all that happened the past week
Results: Graph degree distribution

- Do degree distributions tell us something about significance of particular events and how to summarize graphs?
Results: Actor Interactions

- Unpopular actors fade out quickly
- Popular actors interact amongst each other
- The actor ‘Bill Clinton’ and ‘President’ are the 2 most interacting entities
Future Directions

- Current code will be rewritten so as to try other heuristics of generating graphs and summarization techniques
- Implement thematic tagging of nodes and using the tags to create independent themes from the graph
- Modeling news as an open/closed system.
  - Open system: Eg. Global Warming
  - Closed system: Eg. Cricket News
- Query based summarization
- Learn better summarization by user feedback
Tool Demonstration & Interface

- We can click on a node to get the corresponding article
- If we pick out a path of articles, they are connected by a coherent story
- Other visualization interfaces need to experimented
Thank You!

Questions & Comments.

References
[3] api.opencalais.com