Information Diversity

Marko Grobelnik

Jozef Stefan Institute

Ljubljana, Slovenia

Web Epistemics Workshop – Bielefeld, Feb 16th 2012
Dictionary definition of “diversity”

**di·verse**  
 adj  
1. differing from one another: UNLIKE <people with diverse interests>  
2. composed of distinct or unlike elements or qualities <a diverse population>  
   — di·verse·ly adverb

- **adjective**
  1. of a different kind, form, character, etc.; unlike: a wide range of diverse opinions.
  2. of various kinds or forms; multiform.
Diversities coming out of “content”

• **Topic**
  – What topic is text about? (categorization, segmentation)

• **Social**
  – Who is writing? (publisher, author)
  – Who is being written about? (people, organizations)
  – Publisher’s influence

• **Geographical**
  – Where the content was produced?
  – Which geography is being addressed?

• **Opinion**
  – Sentiment (positive/negative/neutral)
  – Opinions (beyond polarized sides)
  – Reporting bias (differences)

• **Knowledge**
  – Fact coverage (difference in sources)
  – Relationship extraction (how entities are related)

• **Cross-lingual / Multi-lingual**
  – What language is being used?
  – Cross-lingual story linking

• **Context**
  – Temporal context (story linking, discussion threads, trends in other dimensions)
  – Background knowledge (knowledge bases, ontologies, …)
  – Contrasting with other sources (e.g. main-stream vs. twitter)
Diversities coming out of “usage”

• Demographic context
  – Who is the user? (age, gender, job, income)
• Topic
  – What are user interests? (predefined, calculated)
  – What is user searching for? (query logs)
• Geography
  – Where a user is coming from? (home, accessing)
• Access method
  – How a user is accessing data? (web browser, mobile, forum, phone, email)
• Social context
  – With whom a user is connected to? (social network, communication)
• Time
  – When a user is accessing information (absolute, day of week, hour of day)
• Historical
  – What a user was doing in the past?
  – How user activities change trough time? (trends)
DIVERSITY IN WEB SEARCH

Bostjan Pajntar, Jozef Stefan Institute
Marko Grobelnik, Jozef Stefan Institute
What is the most common tasks where we manipulate text in everyday life?
– “Internet search”!

...but – how smart is search technology today?
– ...not too smart!
– It is sophisticated, but not smart...
Example: searching for “Jaguar”

- Query “jaguar” has many meanings...
- ...but the first page of search engines doesn’t provide us with many answers
- ...there are 84M more results
Context sensitive search with

http://searchpoint.ijs.si

Query

Conceptual map

Search Point

Dynamic contextual ranking based on the search point
Blaz Fortuna, Jozef Stefan Institute
Carolina Galleguillos, UC San Diego
Nello Cristianini, University of Bristol

NEWS BIAS DIVERSITY
News are rolling...

...can we model a point of view?
Experimental setup

- **Time period:** March 31st 2005 – April 14th 2006

- **Size of collections:**

<table>
<thead>
<tr>
<th>Source</th>
<th>No. of news</th>
</tr>
</thead>
<tbody>
<tr>
<td>Al Jazeera</td>
<td>2142</td>
</tr>
<tr>
<td>CNN</td>
<td>6840</td>
</tr>
<tr>
<td>Detroit News</td>
<td>2929</td>
</tr>
<tr>
<td>International Herald Tribune</td>
<td>9641</td>
</tr>
</tbody>
</table>

- **Number of discovered matches:**

<table>
<thead>
<tr>
<th></th>
<th>AJ</th>
<th>CNN</th>
<th>DN</th>
<th>IHT</th>
</tr>
</thead>
<tbody>
<tr>
<td>AJ</td>
<td>–</td>
<td>816</td>
<td>447</td>
<td>834</td>
</tr>
<tr>
<td>CNN</td>
<td>816</td>
<td>–</td>
<td>1103</td>
<td>2437</td>
</tr>
<tr>
<td>DN</td>
<td>447</td>
<td>1103</td>
<td>–</td>
<td>895</td>
</tr>
<tr>
<td>IHT</td>
<td>834</td>
<td>2437</td>
<td>895</td>
<td>–</td>
</tr>
</tbody>
</table>
Bias example

**UK soldiers cleared in Iraqi death** – Seven British soldiers were acquitted on Thursday of charges of beating an innocent Iraqi teenager to death with rifle butts. A judge at a specially convened military court in eastern England ordered the adjudicating panel to return ‘not guilty’ verdicts against the seven because he did not believe there was sufficient evidence against them, the Ministry of Defence said. . . .

**British murderers in Iraq acquitted** – The judge at a court-martial on Thursday dismissed murder charges against seven soldiers, from the 3rd Battalion, the Parachute Regiment, who’re accused of murdering Iraqi teenager; claiming there’s insufficient evidence to secure a conviction, The Associated Press reported Thursday. . . .
Prediction of news source

- **The task**: given a pair of news articles describing the same event, can we predict the news source for each?
- In this experiment we focused on CNN and Al Jazeera

- SVM linear classifier was used for prediction
  - Evaluation was done using 10-fold cross-validation
  - Significance of results was tested against random matches
- We used SVM feature selection [Brank et. al.] to extract most important classification keywords.
Detecting News Reporting Bias

• We compared **CNN** and **Aljazeera** reports about the same events from the war in Iraq
  – ...300 aligned articles describing the same story from both sources

• The same topics are expressed in both sources with the following keywords:
  – **CNN** with:
    • Insurgents, Troops, Baghdad, Iran, **Militant**, Police, **Suicide**, **Terrorist**, United, National, Hussein, **Alleged**, Israeli, Syria, Terrorism...
  – **Aljazeera** with:
    • Attacks, Claims, **Rebels**, Withdrawing, Report, **Fighters**, President, **Resistance**, Occupation, Injured, Army, Demanded, Hit, Muslim, ...
### Keywords differences on topic level

<table>
<thead>
<tr>
<th><strong>Topic</strong></th>
<th><strong>CNN</strong></th>
<th><strong>AJ</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Iran, nuclear, palestinian, Israel, gaza, EU, enrichment, IAEA</td>
<td>EU, Iran, Rice, militant, Aceh, diplomats, monitoring, encouraging resume, Rafsanjani, research, atomic, Russian, sanctions, reference</td>
<td></td>
</tr>
<tr>
<td>Iraq, Baghdad, Hussein, shiite, trials, insurgents, troops</td>
<td>insurgents, Hussein, attorney, Kember, family, British</td>
<td></td>
</tr>
<tr>
<td>settlers, Hamas, barriers, Israeli, clashes, Hezbollah, farms, suffer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Palestinian, Gaza, Israel, Sharon, Hamas, Abbas, militant</td>
<td>militant, Israel, pullout, missiles, launch, Putin, Beirut, jews</td>
<td></td>
</tr>
<tr>
<td>Tehran, resume, research, atomic, Rafsanjani, Ahmadinejad, reference</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Iraqi, Palestinian, Baghdad, Iran, Gaza, nuclear, shiite, Hamas</td>
<td>militant, insurgents, terrorists, forces, cross, source, Hussein</td>
<td></td>
</tr>
<tr>
<td>shia, Israeli, fighters, sunnis, squad, farms, occupation, gunmen</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lebanon, Syria, Hariri, assassination, beirut, opposition</td>
<td>Rafik, cooperation, son, rice, Hezbollah, Syria, Hussam, form</td>
<td></td>
</tr>
<tr>
<td>Lebanese, Rafiq, christian, opposition, Aoun, Baath, assassination</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
News source similarity maps

Based on source vocabulary differences

BEP

Based on covered events

Topic
TOPICAL TEMPORAL DIVERSITY

Blaz Fortuna, Jozef Stefan Institute
Marko Grobelnik, Jozef Stefan Institute
Topic Trends Tracking of the documents including “Clinton”

Result set

Query

Topic Trends Visualization

Topics description

US Elections

US Budget

NATO-Russia

Mid-East conflict

Table:

<table>
<thead>
<tr>
<th>Topic Name</th>
<th>Articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>dole, campaign, republican</td>
<td>96</td>
</tr>
<tr>
<td>budget, billion, mexico</td>
<td>152</td>
</tr>
<tr>
<td>trade, tax, house</td>
<td>427</td>
</tr>
<tr>
<td>nato, yeltsin, russia</td>
<td>86</td>
</tr>
<tr>
<td>israel, palestinians, peace</td>
<td>215</td>
</tr>
</tbody>
</table>
WW2 query “Pearl Harbor” from NYTimes archive
WW2 query “Belgrade” from NYTimes archive
WW2 query “Normandy” from NYTimes archive
Challenges

• People in the world have very diverse views
  – ...can model this?

• Some current activities:
  – **Cross-lingual technology** enabling bridging inter-cultural views
  – Aligning **diverse views to the history** through news archives (avoiding interpretations)
  – Modeling divisions in Wikipedia edit stream along many dimensions