Are Topically Diverse Documents Also Interesting?

Hosein Azarbonyad, Ferron Saan, Mostafa Dehghani, Maarten Marx, and Jaap Kamps
University of Amsterdam

Motivation

Text **interestingness** is a measure of assessing the quality of documents from users' perspective which shows their willingness to read. In some research, text interestingness is measured based on its **topical diversity**. In this paper, we investigate the relation between interestingness and topical diversity.

Main Research Question:

- Are topically diverse documents also interesting?

Main Findings:

- In general there is a relatively low correlation between interestingness and topical diversity.
- There are two extreme categories of documents:
  - Highly interesting, but hardly diverse (focused interesting documents).
  - Highly diverse but not interesting documents.

  Removing these two extreme types of documents, there is a positive correlation between interestingness and diversity.

Text Diversity and Text Interestingness

- **Text topical diversity:**

$$div(D) = \sum_{i=1}^{T} \sum_{j=1}^{T} p_i^D p_j^D \delta(i,j),$$

where,

$$\delta(i,j) = \frac{ArcCos(CosineSim(i,j))}{\pi}$$

- **Text Interestingness**: a study on parliamentary proceedings

$$I(D) = \sum_{i=1}^{N} w_i * f_i$$

where f_i's are features:
  - Based on intensity of debates:
    - Number of switches between speakers
  - Based on quantity and quality of key players in the debates
    - The percentage of present members
    - Whether the prime minister is present
    - Whether the deputy prime minister is present
    - Number of speakers who are party leaders
  - Based on the length of debates:
    - Word count of debates
    - Closing time of debates

  and w_i's are corresponding weights of features which are taken from trained model reported in [1] .

- Correlation of Debates' Topical Diversity and Interestingness:
  - Pearson's product-moment correlation coefficient

[1] Hogenboom, A., Jongmans, M., Frasincar, F., Structuring political documents for importance ranking. NLDB2012, pp. 345-350

Data Collection

Dutch parliamentary proceedings

- To train an LDA model
  - from 1999 to 2011
  - 20,547 debates
- To measure the correlation of diversity and interestingness
  - from 2006 to 2010
  - 6,575 debates

Canadian parliamentary proceedings

- To train an LDA model
  - from 1994 to 2014
  - 9,053 debates
- To measure the correlation of diversity and interestingness
  - from 2004 to 2014
  - 7,823 debates

Experiments

Exp. 1: Measuring Topical Diversity of Debates

Table: Top three diverse debates in Dutch and Canadian parliaments

Canadian proceedings			Dutch proceedings		
Topic	**#Speeches**	**Diversity**	**Topic**	**#Speeches**	**Diversity**
competitiveness	140	0.224	kingdom relations	20	0.222
industry,science,technology	105	0.218	housing, integration	40	0.219
closed containment	72	0.217	transportation	24	0.216

- Diverse debates have a high number of speeches in Canadian. proceedings, but a low number of speeches in the Dutch proceedings.

Exp. 2: Measuring Interestingness of Debates

Table: Top three interesting debates in Dutch and Canadian parliaments

Canadian proceedings			Dutch proceedings		
Topic	**#Speeches**	**Interestingness**	**Topic**	**#Speeches**	**Interestingness**
government,budget	331	0.52	pension	823	0.86
government orders	325	0.51	economic crisis	681	0.74
crime	314	0.50	war in Iraq	454	0.74

- Unlike diverse debates, interesting ones are mostly focused on a few topics.
- Number of speeches in interesting debates is high (since number of speaker switches is an important feature).

Exp. 3: The Correlation Between Interestingness and Diversity

Table: The correlation of debates' interestingness and diversity on Dutch and Canadian proceedings (▲ indicates the significance using t-test, two-tailed, $p - value < 0.05$)

Interestingness	Canadian	Dutch
Interestingness(all features)	0.13▲	0.11▲
Interestingness(speaker switches)	0.11▲	0.03
Interestingness(prime minister)	0.08▲	0.14▲
Interestingness(deputy prime minister)	0.06▲	0.1▲
Interestingness(closing time)	-0.12▲	-0.01

- There is a relatively low correlation between diversity and interestingness in both Dutch and Canadian datasets.
- There is a negative correlation between closing time of debates and their diversity.

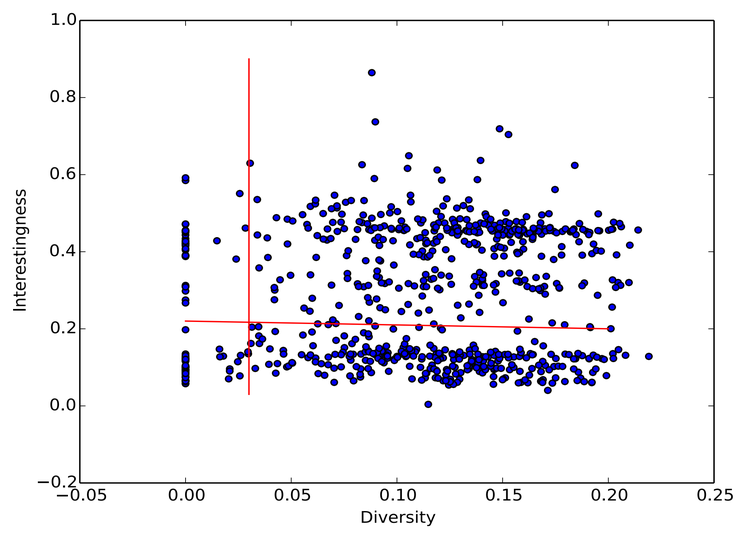

Figure: Scatter plot of interestingness (y-axis) against diversity (x-axis) on debates from 2006 to 2010 on Dutch parliamentary proceedings. Each point in the plot corresponds to a debate.

- Most of diverse documents have low value of interestingness. (left part of the plot)
- There are a few debates with high value of interestingness and very low value of diversity. (top right part of the plot)
- Removing the aforementioned parts (indicated by red lines in the figure) the correlation of diversity and interestingness (using all features) increases to 0.35.

Conclusion

- **Diversity** and **interestingness** metrics are not necessarily reflecting the same characteristics of documents.
- There is a relatively low correlation between text interestingness and diversity.
- Removing extreme cases (interesting but not diverse documents and diverse but not interesting documents) interesting documents are also topically diverse.

ExPoSe:Exploratory Political Search, Funded by NWO (project# 314-99-108)

NWO Netherlands Organisation for Scientific Research

University of Amsterdam

{dehghani,h.azarbonyad}@uva.nl