Are Topically Diverse Documents Also Interesting?

Hosein Azarbonyad[1], Ferron Saan[2], Mostafa Dehghani[1], Maarten Marx[1], and Jaap Kamps[1]

University of Amsterdam, Amsterdam, The Netherlands

[1] {h.azarbonyad,dehghani,maartenmarx,kamps}@uva.nl
[2] ferron.saan@gmail.com

Abstract. Text interestingness is a measure of assessing the quality of documents from users' perspective which shows their willingness to read a document. Different approaches are proposed for measuring the interestingness of texts. Most of these approaches suppose that interesting texts are also topically diverse and estimate interestingness using topical diversity. In this paper, we investigate the relation between interestingness and topical diversity. We do this on the Dutch and Canadian parliamentary proceedings. We apply an existing measure of interestingness, which is based on structural properties of the proceedings (eg, how much interaction there is between speakers in a debate). We then compute the correlation between this measure of interestingness and topical diversity.

Our main findings are that in general there is a relatively low correlation between interestingness and topical diversity; that there are two extreme categories of documents: highly interesting, but hardly diverse (focused interesting documents) and highly diverse but not interesting documents.When we remove these two extreme types of documents there is a positive correlation between interestingness and diversity.

Keywords: Text Interestingness, Text Topical Diversity, Parliamentary Proceedings

1 Introduction

The availability of user-generated text-based reviews stimulated research in automatically computing the interestingness of texts [3, 4]. In [3] it is shown that text interestingness is highly correlated with topical diversity on e-books and e-commerce products description datasets. In this paper, we further investigate the relation between interestingness and topical diversity of texts. Our main research question is: *Are topically diverse documents also interesting?*

To answer this question, we independently measure interestingness and topical diversity of texts and compute their correlation. We carry out our research on the parliamentary proceedings of The Netherlands and Canada and measure the interestingness of the debates in these proceedings using the method proposed in [5] and their topical diversity using the method proposed in [1]. Parliamentary proceeding have structural measures of interestingness which are independent from the textual content. This makes them well suited to answer our research question. Our experiments show that interestingness and diversity reflect different characteristics of documents and in general there is a relatively low correlation between the two properties.

The rest of this paper is organized as follows. In Section 2, we describe the methods used for measuring text's diversity and interestingness. The results and analysis are presented in Section 3. Finally, Section 4 concludes the paper with a breif discussion on the possible future research directions.

2 Methods

In this section we describe how we measure interestingness and topical diversity of debates.

Measuring Debates' Topical Diversity Different approaches are proposed for measuring the topical diversity of texts [1, 3]. Most of these approaches first extract topics of documents using LDA [2] and then estimate the diversity of documents using the extracted topics. We use the method proposed in [1] for estimating the diversity of documents. This approach estimates the diversity of texts using Rao's coefficient [6]: for a document D,

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

where T is the set of topics; p_i^D and p_j^D are the probability of assigning topics i and j to document D, and $\delta(i,j)$ is the distance (dissimilarity) of topics i and j. This method first learns an LDA topic model and then uses that model to assign a probability distrobution over topics to documents. Different distance functions have been employed in [1]. However the used functions are not proper distance metrics. So, we use the normalized angular distance which is a distance metric and holds the properties of a metric for measuring the distance of topics [7]:

$$\delta(i,j) = \frac{ArcCos(CosineSim(i,j))}{\pi} \tag{2}$$

where $CosineSim$ is the cosine similarity of topics i and j. $ArcCos(CosineSim(i,j))$ is the arc cosine of cosine similarity of topics i and j. To calculate the similarity of topics we identify a topic i with the vector consisting of all p_i^D for all documents D in our collection. The similarity of two topics is then the cosine similarity of their vectors.

Measuring Debate's Interestingness Interestingness of a text could be defined in different ways [3, 5]. Derzinski and Rohanimanesh [3] showed that texts' interestingness is highly correlated with its topical diversity. To measure the correlation of interestingness and diversity of documents we first need to estimate the interestingness of documents. To do so, we use the method proposed in [5] and estimate an interestingness value for each document. They define the interestingness of a document as ”the probability that the public finds a document of great importance”. They focused on measuring the interestingness of debates in parliamentary proceedings. Since the interestingness of texts in parliamentary proceedings is measurable using this method, we employ the approach proposed in [5] to measure the interestingness of debates in parliamentary proceedings.

This method uses features extracted from debates for learning a supervised method to assign interestingness values to debates. The used features are categorized into three groups: features based on intensity of debates, features based on quantity and quality of key players in the debates, and features based on the length of debates. From the first category we use the number of switches between speakers in the debates. From the second category we use the most important features: the percentage of members present in the debate, whether the prime minister is present in the debate or not, whether the deputy prime minister is present in the debate or not, and the number of speakers who are floor (party) leaders as well. From the last category we use two most important features: word count of debates and closing time of debates. The importance of features are determined using weights of features in the model trained and reported in [5]. We use weighted linear combination of mentioned features to estimate the interestingness of a debate D:

$$I(D) = \sum_{i=1}^{7} w_i * f_i; \tag{3}$$

where f_i is a feature and w_i is the weight of f_i in the trained model reported in [5] for assigning interestingness values to debates and the sum is taken over the mentioned seven features.

Correlation of Debates' Topical Diversity and Interestingness We express the correlation between our two variables of interest by Pearson's product-moment correlation coefficient.

3 Analysis

In this section we first describe the datasets and different setings and pre-processings we did, and then we analyze the text interestingness and topical diversity and their correlations on these datasets.

3.1 Datasets and Experimental Setup

We use two datasets to analyze the correlation of texts' diversity and interestingness: Dutch and Canadian parliamentary proceeding. These datasets are publicly available at `http://search.politicalmashup.nl`. From the Dutch parliamentary proceedings we use the debates from 1999 to 2011 to train an LDA model. This dataset contains 20,547 debates from parliament. For measuring the correlation of diversity and interestingness, we select a period of parliament from 2006 to 2010 and calculate the correlation on the debates of this period. This period contains 6,575 debates. We also remove the procedural debates which do not contain speeches of parliament members. From Canadian proceedings we choose the debates from 1994 to 2014 to train an LDA model. This subset of dataset contains 9,053 debates. We calculate the correlation of diversity and interestingness on a subset of this dataset from 2004 to 2014 which contains 7,823 debates.

Table 1. 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

Table 2. 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

We set the number of topics of LDA to 50. The LDA models are trained on the lemmatized nouns in the documents only. Words with less than five occurrences and 100 words with highest frequencies and 100 words with highest document frequencies in the corpus are considered as stop words and removed from documents. We also do the same feature normalization done in [5] before calculating text interestingness.

3.2 Results

Measuring topical diversity of debates Table 1 shows the information of top three most diverse debates in the Dutch and Canadian parliaments. The most diverse debate in the Canadian parliament is a debate on study of competitiveness. In this debate, members discussed different issues related to farming, agriculture, and petroleum which made this debate very diverse. The most diverse debate in the Dutch proceedings is a debate in which parliament members asked questions from minister of Interior and Kingdom Relations. Table 1 also shows that diverse debates have a high number of speeches in Canadian proceedings, but a low number of speeches in the Dutch proceedings.

Measuring interestingness of debates Table 2 shows the top three most interesting debates in the Dutch and Canadian proceedings. Unlike diverse debates, interesting ones are mostly focused on a few topics. Also, since number of speaker switches is the most important feature in the interestingness prediction model, the number of speeches in interesting debates is high.

The correlation between interestingness and diversity Table 3 shows the correlation of debates' diversity and interestingness. There is a relatively low correlation between diversity and interestingness in both Dutch and Canadian datasets. In fact, these two metrics are reflecting different characteristics of documents. The results also show that there is a negative correlation between closing time of debates and their diversity. In fact, the debates that take more time are very focused on a few topics. Figure 1 shows the scatter plot of interestingness against diversity on Dutch proceedings. From this

Table 3. The correlation of debates' interestingness (all features) and diversity on Dutch and Canadian proceedings ($^{\blacktriangle}$ indicates the significance using t-test, two-tailed, $p - value < 0.05$)

Interestingness	Canadian	Dutch
Interestingness(all features)	0.13$^{\blacktriangle}$	0.11$^{\blacktriangle}$
Interestingness(speaker switches)	0.11$^{\blacktriangle}$	0.03
Interestingness(prime minister)	0.08$^{\blacktriangle}$	0.14$^{\blacktriangle}$
Interestingness(deputy prime minister)	0.06$^{\blacktriangle}$	0.1$^{\blacktriangle}$
Interestingness(closing time)	-0.12$^{\blacktriangle}$	-0.01

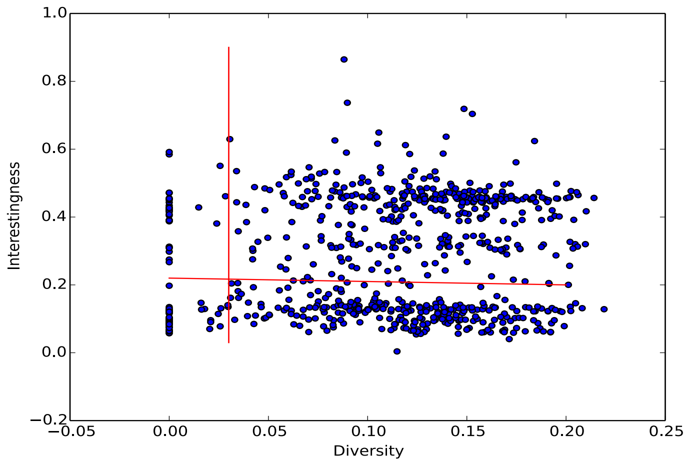

Fig. 1. Scatter plot of interestingness (y-axis) against diversity (x-axis) on debates from 2006 to 2010 on Dutch parliamentary proceedings. Each point in the plot corresponds to a debate.

figure it can be seen that most of diverse documents have low value of interestingness (the right bottom part of the plot). These are the debates which cover lots of topics but are not interesting from the users' perspective. Also there are a few debates with high value of interestingness and very low value of diversity (left part of the plot). Besides these two types of debates, we can see from Figure 1 that there is a slight positive correlation between interestingness and diversity (top right part of the plot). If we remove the debates from the first and second category (indicated by red lines in the figure) and just consider the top right points in the Figure 1, the correlation of diversity and interestingness (using all features) increases to 0.35. This results indicates that other than extreme cases (interesting but not diverse documents and diverse but not interesting documents) interesting documents are also topically diverse.

4 Conclusion

We have investigated the correlation between text interestingness and topical diversity. For the analysis, we focused on Dutch and Canadian parliamentary proceedings. The results show that the correlation of interestingness and diversity over whole documents is very low. Also, based on our results there are three major types of documents based on the correlation of diversity and interstingness: interesting focused documents; uninteresting diverse documents, and both interesting and diverse documents. The documents of the first two categories are extreme ones which there is no clear correlation between their interestingness and diversity values. It would be interesting to investigate more on the documents of these two categories and analyse their properties to see what is the main reason behind the low correlation of interestingness and diversity on them.

Our results indicated that over the whole dataset there is a relatively low correlation between text interestingness and diversity. However, in previous studies it has been concluded that text interestingness and diversity are highly correlated [3]. There are some possible explanations: We used a method for measuring the interestingness of documents which is independent of the content of documents. However, text diversity is dependent to the content of the documents. Also, [5] used a manually selected debates to train the interestingness prediction model. The chosen debates are the debates which contain the information needed to estimate the interestingness. However we conducted our evaluations on whole debates. Therefore, the used interestingness measure may not be a proper measure to assess the interestingness of all kind of debates. Another reason for getting the low correlation value on debates is that based on our analysis, some of the topics of the LDA model trained on debates are not pure and contain words which should basically belong to different topics. Also, there are some general topics which contain procedural words and are not very informative. So, the impure and general topics make the diversity value estimated for debates very noisy.

Acknowledgements This research was supported by the Netherlands Organization for Scientific Research (ExPoSe project, NWO CI # 314.99.108; DiLiPaD project, NWO Digging into Data # 600.006.014) and by the European Community's Seventh Framework Program (FP7/2007-2013) under grant agreement ENVRI, number 283465.

5 References

[1] K. Bache, D. Newman, and P. Smyth. Text-based measures of document diversity. In *KDD 2013*, pages 23–31, 2013.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[3] M. Derzinski and K. Rohanimanesh. An information theoretic approach to quantifying text interestingness. In *NIPS MLNLP workshop*, 2014.

[4] D. Ganguly, J. Leveling, and G. J. Jones. Automatic prediction of text aesthetics and interestingness. In *COLING 2014*, 2014.

[5] A. Hogenboom, M. Jongmans, and F. Frasincar. Structuring political documents for importance ranking. In *Natural Language Processing and Information Systems*, volume 7337, pages 345–350. 2012.

[6] C. Rao. Diversity and dissimilarity coefficients: A unified approach. *Theoretical Population Biology*, 21(1):24 – 43, 1982.

[7] S. Van Dongen and A. J. Enright. Metric distances derived from cosine similarity and pearson and spearman correlations. *arXiv preprint arXiv:1208.3145*, 2012.