TU Delft

Delft University of Technology

Same, Same, but Different

Algorithmic Diversification of Viewpoints in News

Tintarev, Nava; Sullivan, Emily; Guldin, Dror; Qiu, Sihang; Odjik, Daan

DOI
10.1145/3213586.3226203

Publication date
2018

Document Version
Accepted author manuscript

Published in
UMAP '18 Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization

Citation (APA)
Tintarev, N., Sullivan, E., Guldin, D., Qiu, S., & Odjik, D. (2018). Same, Same, but Different: Algorithmic Diversification of Viewpoints in News. In UMAP '18 Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization (pp. 7-13). New York, NY: Association for Computing Machinery (ACM). https://doi.org/10.1145/3213586.3226203

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

Same, Same, but Different

Algorithmic Diversification of Viewpoints in News

Nava Tintarev
TU Delft
n.tintarev@tudelft.nl

Emily Sullivan
TU Delft
E.E.Sullivan-Mumm@tudelft.nl

Dror Guldin
University of Amsterdam
dror.guldin@gmail.com

Sihang Qiu
TU Delft
S.Qiu-1@tudelft.nl

Daan Odjik
Blendle Research
daan@blendle.com

ABSTRACT

Recommender systems for news articles on social media select and filter content through automatic personalization. As a result, users are often unaware of opposing points of view, leading to informational blindspots and potentially polarized opinions. They may be aware of a topic, but only be exposed to one viewpoint on this topic. However, recommender systems have just as much potential to help users find a plurality of viewpoints. In this spirit, this paper introduces an approach to automatically identifying content that represents a wider range of opinions on a given topic. Our offline results show positive results for our distance measure with regard to diversification on topic and channel. However, our user study results confirm that user acceptance of this diversification also needs to be addressed in tandem to enable a complete solution.

CCS CONCEPTS

• **Information systems** → **Learning to rank**; **Information retrieval diversity**; *Personalization*; *Similarity measures*; • **Human-centered computing** → Empirical studies in HCI;

KEYWORDS

News recommendation, diversity based ranking, framing

ACM Reference Format:
Nava Tintarev, Emily Sullivan, Dror Guldin, Sihang Qiu, and Daan Odjik. 2018. Same, Same, but Different: Algorithmic Diversification of Viewpoints in News. In *UMAP'18 Adjunct: 26th Conference on User Modeling, Adaptation and Personalization Adjunct, July 8–11, 2018, Singapore, Singapore*. ACM, New York, NY, USA, Article 4, 7 pages. https://doi.org/10.1145/3213586.3226203

1 INTRODUCTION

Recommender systems play an important role in helping to mediate many of our everyday decisions. They help us filter and rank content automatically when the volume is too large to handle for human curation. They do this by learning from our past interactions and inferring our interests.

In recent years, a common criticism of recommender systems has been that they may serve to create *filter bubbles* (see [20]) for users by censoring their choices over time and effectively polarising their preferences; see [3, 16, 18]. This is however disputed, for example, Flaxman et al. found evidence that recent technological changes both increase and decrease various aspects of polarisation [7]. This suggests that there may be design choices for recommender systems that could decrease polarisation. Existing recommendation algorithms focus strongly on relevance, and even those which consider diversity do not consider that some content may be more challenging for a given user.

Addressing these issues likely requires both algorithmic and interface solutions to helping users consume content that is both relevant and diverse. In this paper, we broaden the discussion of what traditionally is seen as diversity in the recommender systems domain, to be more suited to the news recommendation context. We do this by making the following key scientific contributions:

(1) We develop a new distance measure for diversity *within a topic*, enabling diversity while maintaining topic relevance.
(2) Using this measure we introduce an adaptation to an existing diversity based ranking technique, Maximal Marginal Relevance (MMR). This enables us to compose lists of diverse recommendations, with increasing information content further down the list.
(3) We evaluate these results in both an offline evaluation, and a user-study using real-world news articles from a commercial news aggregator.

2 RELATED WORK

Recommender systems are playing an increasingly key role in opinion-forming domains, such as news discovery, web search, and social networking. In what follows, we introduce two common responses to filter bubbles in the literature: algorithm and user centered approaches. Ultimately, this paper proposes a primarily algorithm-based approach, while also considering the role of user acceptance of diverse content in order to combat the problematic aspects of filter bubbles.

2.1 Algorithm-based approaches

A common approach to filter bubbles is to develop *diversity-aware recommendation algorithms*. By focusing on recommendation diversity, novelty, and *relevance*, it may be possible to ensure the user receives a broader set of recommmendations; see [1, 2, 4, 23, 30]. For example, Ziegler et al. proposed a topic diversification approach based on taxonomy-based dissimilarity [30]. As may be anticipated, using simple dissimilarity also impacted accuracy negatively. An alternate set of approaches which re-rank a list of top items was found to improve diversity without a great loss in accuracy [2]. Smyth and Bridge found that diversity based on the hamming distance based on whether or not the items had been rated (over a number of users)

UMAP'18 Adjunct, July 8–11, 2018, Singapore, Singapore
2018. ACM ISBN 978-1-4503-5784-5/18/07…$15.00
https://doi.org/10.1145/3213586.3226203

helped retrieve a target item most often and most efficiently [23]. Abbassi et al. found for their more user-centric clustering approach that users preferred to be exposed to items in a diversified set of clusters, but with a less diversified set of items inside each cluster [1]. This work was elaborated in a study where participants made recommendations for a fictitious friend, and could apply diversity in terms of author, theme, and genre for a sequence of books [25].

2.2 User-centered approaches

An alternative approach is to help users understand the justification for a recommended item or the recommendation space better, e.g., through creating *diversity-aware interfaces* [24]. While improving recommendation diversity can go some way to coping with the filter bubble, it is far from a complete solution. It does not help to educate users about the effects of the phenomenon or increase their awareness of the filter bubble itself, for example. In this regard the work of [16] is pertinent, showing how visualization was found to increased users' awareness of the filter bubble, understandability of the filtering mechanism and to users having a sense of control over their data stream [16]. The same paper also allowed other types of filtering such as categories or topics to which the user was exposed. Other work has looked at ways of addressing limited information access and filter bubbles [16, 26]. In one study, users were able to control which people in their immediate and extended network contributed to their information feed on Twitter. The interface increased users' sense of transparency and control [12, 26]; see also the work of [19, 27] for related ideas.

3 USE CASE

The motivating scenario is a user who is focused on a certain point of view, unaware of related viewpoints which are perceived as important by other groups. Figure 1 outlines the workflow for the system developed in this paper.

The system uses articles retrieved from a commercial news aggregator for quality journalism that includes major US news outlets, such as the New York Times, Wall Street Journal, and the Washington Post. First, the system identifies disputed topics from articles within a given time frame (e.g., the last month); topics that have different coverage in different news sources, and might therefore be considered to reflect different agendas. From these, the user selects a topic which is interesting and relevant to them, e.g., *Tax Reform*.

Next, the system retrieves a set of candidate news articles on the topic from the news aggregator. Then it re-ranks these articles according to both diversity and relevance using a distance function applied to a Maximal Marginal Relevance re-ranking algorithm (described in Section 4.2). The result is a recommended list of articles that are representative of a diversity of view-points, or "framings" of the same topic. To support further control, and to enable evaluation, the user can fine-tune the parameters of the distance function, and consequently the resulting recommendations.

4 DIVERSITY-BASED RANKING

To obtain a diverse ranking of articles, we re-rank articles taking into account their similarity to each other. To measure this similarity between articles, we compute linguistic features and meta-data from both a linguistic resource (LIWC) (Section 4.1.1), and the news aggregation service (Section 4.1.2), that have a wide distribution across news sources. We combine these features into a linear weighted sum. This supports the definition of a novel distance function (Section 4.1.3), which in turn allows us to define the way in which items in a list are different from each other. We use a specific re-ranking method called Maximal Marginal Relevance that uses this measure. This allows us to change the order of articles in a way which considers both diversity and relevance.

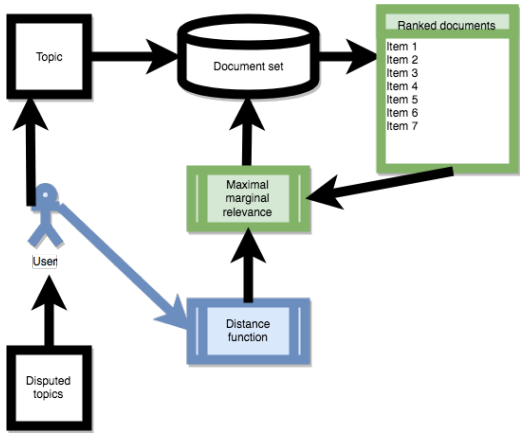


Figure 1: System workflow

4.1 Distance function

4.1.1 Linguistic Inquiry and Word Count (LIWC2015). LIWC is a tool for computerized text analysis, and includes rich dictionaries and summary variables for analyzing the style in which text is written [21]. The summary variables are taken from previously published findings and converted to percentiles based on standardized scores from large comparison samples. Due to prior commercial agreements, the precise algorithms are not available. Note that the four summary variables have been re-scaled so that they reflect a 100-point scale. The used variables and definitions from the operators' manual are:

- **Analytical thinking**. A high number reflects formal, logical, and hierarchical thinking; lower numbers reflect more informal, personal, and narrative thinking [22].
- **Clout**. A high number suggests that the author is speaking from the perspective of high expertise, and is confident; low Clout numbers suggest a more tentative, humble, even anxious style [11].
- **Authenticity.** Higher numbers are associated with a more honest, personal, and disclosing text; lower numbers suggest a more guarded, distanced form of discourse [17].
- **Emotional tone**. A higher number is associated with a more positive, upbeat style; a low number reveals greater anxiety, sadness, or hostility. A number of around 50 suggests either a lack of emotionality or of ambivalence [6].

Using the LIWC summary variables, we analyzed 19,136 US news articles published and offered on a news aggregation service from November 2017. Figure 2 shows the LIWC score distribution for all

the articles. Figure 3 shows the distribution of 386 articles on the topic Tax Reform.

Figure 2: LIWC scores of all US news articles

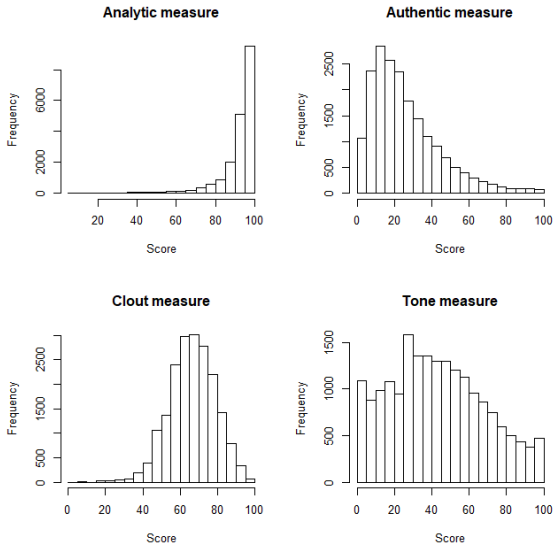
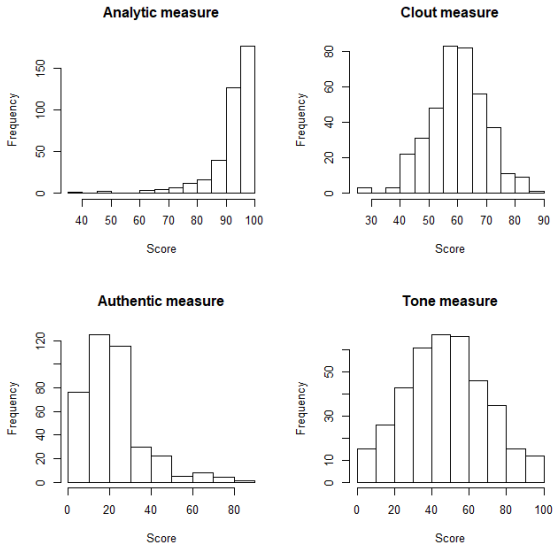


Figure 3: LIWC scores of US articles on Tax Reform



We can see in these figures that *emotional tone* and *clout* had the widest distribution, and can be considered more suitable to represent different points of view. In contrast, articles tend to score high in analytic and low in authenticity, especially when focused in on a controversial topic like tax reform. This led us to only include tone and clout in our diversity metrics.

We expect that the high analytic and low authenticity scores are a result of our focus on news articles. It could very well be that the analytic measure would be more discriminating for other types of articles, such as movie reviews.

4.1.2 Commercial meta-data. This research was conducted in collaboration with a news aggregation provider specializing in personalized and diverse recommendations. They automatically extract meta-data and stylometrics about the news articles, including:

- Automatically detected topics using Latent Dirichlet allocation (LDA)
- Gravity of the topic (light, neutral, heavy). A story about a cat may be light, whereas a story about a disaster may be heavy.
- Feel. This is a feature that is related to sentiment (positive, neutral, negative).
- Topic complexity (easy, medium, complex). A science article may be more complex than a human interest story.
- Linguistic complexity. This reflects how hard an article might be to read, longer words or sentences are indicators of higher linguistic complexity.
- Article source. This indicated the original news source, e.g., Wall Street Journal, Economist.
- Article channel. This indicated the broad categorization of an article, e.g., sports, politics, entertainment etc.

Figure 4 shows the distribution for these features. Of these Gravity, Topic complexity, Feel, LDA, and Channel were found to be most suitable for diversification. To represent the features Gravity (light, neutral, heavy), Complexity (easy, medium, complex) and Feel (positive, neutral, negative), we used three discrete values (0, 0.5, 1). To allow us to combine these features they were normalized (c.f. Table 1). Min-max normalization was applied to the distances.

Figure 4: Distributions of stylometrics of all US news articles

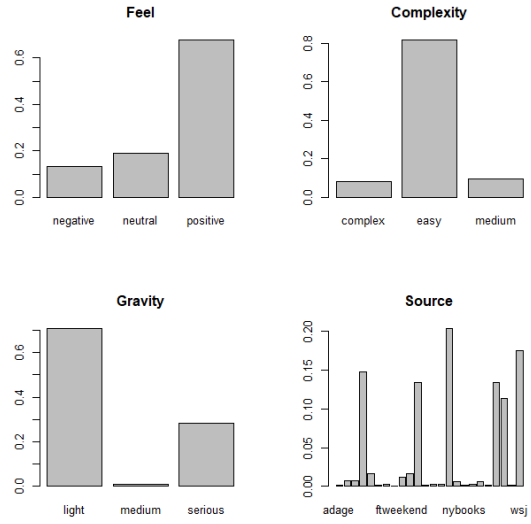


Table 1: Different features used in distance function

Feature	Weight	Distance
Tone	0.25	"Tone distance"
Clout	0.20	Euclidean distance
Gravity	0.17	Euclidean distance
Complexity	0.11	Euclidean distance
Feel	0.11	Euclidean distance
LDA	0.09	Kullback-Leibler
Channel	0.07	Cosine Similarity

4.1.3 Combining features. The distance function is a linear weighted average of LIWC, and meta-data features from the news aggregator. Table 1 summarizes the resulting weighting. These weightings are an informed guess, and are later evaluated offline using an exhaustive gridsearch (c.f., Section 5). The gridsearch allowed us to evaluate the weights for the different features, in a systematic fashion, and compare these 'optimal' results with our hand-crafted weighting. We will see later that the hand-weighted features in fact resulted in slightly *higher* diversity; outperforming the offline gridsearch. The motivation for the manual weighting is as follows.

Tone was given the highest weight because it had the widest distribution among the other measures. Furthermore, tone seems to be the most conceptually discerning. We found that higher scores on tone gave positive reasons in favor of something; whereas lower scores on tone gave negative reasons. Whether an article has a positive or negative normative valiance strongly suggests a diversity in viewpoint, at least in terms of the way that the article reaches its conclusion. Further, we defined *Tone distance* as the Euclidean distance after transforming the "Tone" axis (1-100) to a two-dimensional space, with an additional axis that counts for the distance from "total neutrality" (a score of 50). This transformation was made to incorporate the idea that "very positive" and "very negative" articles are similar in terms of being "very emotional".

Clout was given the second highest weight. Clout measures how assertive or confident the author of an article comes across in her writing. An article that has a more tentative approach to reaching a conclusion shows a strong difference than an article that is overly confident. Part of the reason that it is important for readers to have access to a variety of viewpoints is to be more reflective about important complex issues. Thus, having access to diversity in levels of author confidence is important.

Gravity was given the third highest weight. Since gravity tracks how light or heavy a story is, it is a good indicator of the sort of stakeholders a given article discusses and how the topic relates to these diverse stakeholders. For example, in our data-set there is a satirical article about how the new tax reform bill affects dog owners, this will give the reader a different take on the tax reform issue compared to an article about its affects on low-income families.

Feel and *Complexity* were given equal weight. While each measure is still important for diversity, we saw these as complementary to the other measures. Feel is complementary to tone, and complexity is complementary to LDA and gravity.

Lastly, *LDA* and *Channel* were given the lowest weights since we wanted to make sure that articles did not diverge so much as to suggest a different topic all together.

4.2 Maximal marginal relevance (MMR)

Having defined a novel distance function, we are now able to apply methods for ranking, resulting in a list of recommendations. Current ranking methods usually order items by declining relevance. In order to get a balanced and diverse selection of items for users, we chose Maximal Marginal Relevance (MMR) [5] to provide a linear combination of relevance and diversity. The linear combination is called "marginal relevance", which means the item has high marginal relevance if it is both relevant to the first-picked document and has high diversity to previously selected items. MMR can be expressed as follows.

$$MMR \triangleq \max_{D_i \in R \setminus S} [\lambda \, (Rel(D_i)) - (1 - \lambda) \max_{D_j \in S} (1 - Div(D_i, D_j))] \quad (1)$$

Where R is the ranked list of relevant documents; S is the list of selected documents in R; $R \setminus S$ is therefore the list of as yet unselected documents in R; function Rel demonstrates the quantified relevance metric of each document in R; and function Div is the diversity metric between different documents. Given the above definitions, MMR can generate a new ranked list using relevance function, diversity function and an adjustment parameter λ in $[0, 1]$. When $\lambda = 1$, MMR computes a completely relevance-ranked list. And a maximal diversity-ranked list will be generated if $\lambda = 0$. In our system, the default value of λ is 0.75 and users can adjust the value of λ on web-based user interface to obtain a personalized selection of documents.

5 OFFLINE EVALUATION

To assess the diversification approach we conducted an offline evaluation using an exhaustive grid search. This allowed us to evaluate different parameters, notably the weights for the different features, in a systematic fashion and compare these results with our hand-crafted weighting. These were evaluated using an intra-list diversity measure using Channels and Topics features, which allowed us to evaluate the list on distinct properties from the ones used to create the diversified lists.

5.1 Datasets

The grid search was conducted using a subset of articles regarding a tax reform bill from all US-based outlets offered on a news aggregation service from month of November, 2017. This subset contains 386 articles from 17 diverse news outlets, such as the Wall Street Journal, Washington Post, New York Times and Bloomberg Business Week.

5.2 Evaluation Metric

The evaluation metric for each combination in the gridsearch was Intra-List Diversity. This measure was chosen as it represents *how* the diversification manifests. An alternative measure of novelty, such as the one proposed by [14], is less useful in the news domain where most items are expected to be new to a user.

Intra-list diversity (ILD) is used to evaluate the diversity of ranked lists, which is defined as follows:

$$Diversity = \frac{\sum_{i=1}^{n} \sum_{j=i+1}^{n} Distance(D_i, D_j)}{n \times (n-1)/2} \quad (2)$$

Table 2: Comparison of gridsearch and hand-crafted feature weights.

	Gravity	Complex.	Feel	Clout	Tone	λ
Optimal	0.2	0.2	0.2	0.8	0.6	0.6
Crafted	0.17	0.11	0.11	0.2	0.25	0.3

Where function *Distance* is given by distance between two items (D_j and D_j). So, ILD is defined to be the average diversity distance between all pairs of items in the ranking list.

The decision to give an equal contribution to both Channels and Topics was meant to treat "diversification" as represented by both the human labelled and the automatically detected "framing".

Distance was defined by the meta-data defined by the news aggregator, with equal contributions from the Channels and Topics distance measures:

$$Distance(D_i, D_j) = 0.5 \times Distance_{Channels} + 0.5 \times Distance_{LDA} \quad (3)$$

where $Distance_{Channels}$ is given by cosine distance, based on cosine similarity on channels:

$$Distance_{Channels} = 1 - C_s(Channels_{D_i}, Channels_{D_j}) \quad (4)$$

where C_s means cosine similarity. $Distance_{LDA}$ is given by the Kullback-Leibler divergence on the LDA topics.

5.3 Methodology

The grid search covered 5 parameters from LIWC and the news aggregation provider meta-data (Gravity, Complexity, Feel, Clout, Tone), and it included the parameter Lambda (λ) from the MMR algorithm, which gives us a balance between relevance and diversity. The search was conducted in intervals of 0.2 for each parameter. 4096 combinations were examined.

5.4 Results: Parameter sensitivity

Table 2 compares the optimal combination found via the grid search with a hand-crafted solution based on the 5 parameters previously identified. Recall that LDA and Channel were omitted from the computation to enable us to study the effect of the distance measures on our diversity function. Surprisingly, we found that the gridsearch (ILD = 0.31) resulted in a lower diversity on our diversity measure than the hand-crafted setting suggested in Section 4.1.3 (ILD = 0.41). These results are comparable, and suggest that the hand-crafted feature weights are suitable for experimentation for the user study described in Section 6.

6 USER STUDY

In Section 4.1.3 we adjusted our distance function to the discriminatory ability of the different feature parameters. We also evaluated the sensitivity of the parameters for influencing diversity in Section 5. The ultimate test however remains with regards to user perceptions; Are people drawn to read these more diverse articles?

6.1 Procedure

6.1.1 Materials. Candidate disputed topics were identified from news published in American news sources in November 2017. 17 political varied news sources were used, such as The Economist, The New York Times, and the Wall Street Journal.

A topic was considered "disputed" if it was mentioned in several news sources, with a significantly different LIWC score between at least one pair of outlets (e.g., the average tone of articles regarding "Donald Trump" had a difference of 83.05 between two sources).

6.1.2 Design. The experiment was conducted as a within-subjects design to compare one baseline article (relevance ranking only) with another more diverse article (re-ranked for diversity using our optimal recipe). Each participant was presented with one article and asked to imagine they just read that article and are interested in reading more about the topic.

The order of presentation of the baseline and diverse article was randomized. The order of topics was randomized across a manually selected subset of disputed topics: Tax Reform (455 articles); Roy Moore (87); Russia (357); Trump (1166).

Participants were then presented with two more articles on the same topic. One article is taken from the baseline list, and the other is from our diversified list: both selected at rank 2. They were shown the title, source of the article, and a brief summary supplied by the news aggregation provider. They were then asked to choose, in a forced choice, which article they would be interested to read next.

Participants were asked a series of additional questions: How closely do they follow U.S. news? How *interesting* did they find the first article presented? How *difficult* they found making the forced choice? Each of these questions were ranked on a numeric scale from 1 to 5 (1= never or not at all interesting or difficult, 5 = often or very interesting or difficult).

6.1.3 Hypotheses. H1: Participants will pick the baseline article more than the diverse article. The tendency that people are drawn to news that is attitude-confirming is well documented [8, 13, 15, 29]. Thus, while the diverse article may contribute more interesting content, we expect participants to be conservative and at least initially prefer the news they would normally receive.

H2: Participants that do not find the first article interesting will be more likely to choose the diversified article. In order to test whether the original article contained attitude-confirming information for the participant, we asked what level of interest the participant had in the original article [10]. If the interest level in the original article is low, this suggests that there is a benefit in supplying a more diversified article, and the participant is more likely to choose it.

H3: Participants that follow U.S. news more closely will chose the baseline article more often. The impact of filter bubbles on those who follow news closely is mixed. On the one hand, people seek out different viewpoints to bolster their own view and may spend more time reading diverse viewpoints in order to be critical [9]. On the other hand, the well documented cases of confirmation bias suggest higher preference for ones own view. For users who are experienced in the topic, we expect a stronger confirmation bias – choosing the baseline over the diversified article.

H4: Participants will find the decision difficult, or be unsure. The diversity metric we are interested in studying is diversity in viewpoint. This means that the diverse articles are still relevant to the topic and similar to the baseline. The forced choice is between different approaches/viewpoints to the topic. The fact that both articles are relevant to the topic will make the choice difficult.

6.2 Results

Given the smaller sample size we focus on descriptive statistics.

6.2.1 Participants. 15 participants took part in the evaluation. The age of participants ranges from 23 to 41 with education level ranging from no college to having a Ph.D. or other advanced degree. 14 of the participants are male, 1 female. Participants were balanced across the topics: Trump (4), Tax Reform (3), Russia (4), and Roy Moore (4). The level of US news experience for the participants ranges. (33%) of participants have a high level of US news experience, (40%) have a moderate level, (26%) have a low level.

H1: Participants will pick the baseline article more than the diverse article. In line with H1, we found that more participants (66%) chose the baseline, while fewer (33%) chose the diversified article. This suggests initial resistance to stepping outside one's filter bubble to read diversified viewpoints on a given topic.

H2: Participants that do not find the first article interesting will be more likely to choose the diversified article. We found that of those who had a low interest level of the original article, answering with a 1 or 2, 40% chose the baseline; while 60% chose the diversified article. Of those who had a high interest level of the original article, answering with a 4 or 5, 100% chose the baseline. Of those who had an moderate interest level of the original article, answering with a 3, 71% chose the baseline; 29% chose the diversified article.

H3: Participants that follow U.S. news more closely will chose the baseline article more often. We found that of those who follow US news closely, answering with a 4 or 5, only 20% chose the diversified article. Of those who do not follow US news, answering with a 1 or 2, 50% chose the diversified article. Of those who moderately follow US news, answering with a 3, 33% chose the diversified article.

H4: Participants will find the decision difficult, or be unsure. We found that the average difficulty of the forced choice was 2.8 (/5) suggesting a neutral level of difficulty with a standard deviation of 1.01. Of those who chose the baseline article the average difficulty for choosing was a 2.8 with a standard deviation of 1.14. Of those who chose the diversified article the average difficult was still 2.8, but with a standard deviation of .84.

7 DISCUSSION

This section discusses the limitations of the approach, and implications of the findings in both the offline evaluation, and user study.

7.1 Offline evaluation

Our offline evaluation suggests that our hand-crafted distance function, using linguistic and stylometric terms, influences diversity in terms of topic and channel. However, the weaker result for the grid-search compared to a hand-crafted solution suggests that a more granular evaluation, with smaller steps than 0.2, could improve diversification of automated methods further.

The dataset studied in this paper was limited in size, and a more refined diversity function is likely to result from a larger scale evaluation. Further evaluations are in progress to compute diversity for larger datasets, and for a wider range of topics. Having identified the workflow and algorithm for diversification (MMR), conducting this work is now a matter of (execution) time rather than the development of new ideas.

Human curation has a limited capacity especially in the news domain, where there is a high rate of release. Therefore an algorithmic component is strictly required to enable personalized news provision. However, offline studies are limited in the sense that they give no sense of whether the diversified news will be read or not. The user study summarized below highlighted some limitations of automatic diversification.

7.2 User Study

Due to the smaller sample size and gender imbalance of this study, more investigation needs to be done in order to make strong conclusions regarding the trends surrounding user choice for diversification of viewpoints. That said, there are some interesting trends.

Our initial findings (H1) suggest that filter bubbles are in part due to the interests of individuals, not merely lack of awareness or access to diversified viewpoints. In particular, as H3 predicted, there was a higher rate of users with topical expertise choosing the baseline. This may suggest that breaking filter bubbles is more difficult compared to introducing diversified articles *before* one enters into a filter bubble. This is a worrying finding when considering results that suggest that falsehoods diffuse faster than truth [28].

Contrary to our prediction with H4, participants did not find the forced choice difficult. This suggests that the diversification might have been on an acceptable level – articles were not similar enough to create a hard choice for the participants, but still different enough that the choice was not trivial. It may also suggest that the participants did not wish to give the choice too much thought.

In future work we plan to, in addition to the forced choice, ask participants which of the forced choice articles they think is more diverse to the original article. This would give more of an insight as to whether our diversity metrics correspond to *perceived diversity*.

8 CONCLUSION

In this paper, we develop a new distance measure for diversity *within a topic*, using linguistic and stylometric distance measures. Using this measure we introduce an adaptation to an existing diversity based re-ranking technique, Maximal Marginal Relevance (MMR), to compose lists of diverse recommendations. Finally, we evaluate these results in both an offline evaluation, and a user-study. Our offline results show positive results for our distance measure with regard to diversification on topic and channel. However, our user study results confirm that user acceptance of this diversification also needs to be addressed in tandem to enable a complete solution. This work demonstrates a first feasibility of algorithmic definitions of diversity relating to different points of view. It also reiterates the importance of solutions that consider user perceptions when tackling the challenges of diversification and filter bubbles.

REFERENCES

[1] Zeinab Abbassi, Vahab S. Mirrokni, and Mayur Thakur. 2012. *Diversity Maximization Under Matroid Constraints.* Technical Report. Department of Computer Science, Columbia University.

[2] Gediminas Adomavicius and YoungOk Kwon. 2011. Improving Aggregate Recommendation Diversity Using Ranking-Based Techniques. *IEEE Transactions on Knowledge and Data Engineering* 24 (2011), 896–911.

[3] Eytan Bakshy, Solomon Messing, and Lada A. Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348 (2015), 1130–1132.

[4] Derek Bridge and John Paul Kelly. 2006. Ways of Computing Diverse Collaborative Recommendations. In *Adaptive Hypermedia and Adaptive Web-based Systems*. 41–50.

[5] Jaime G. Carbonell and Jade Goldstein. 1998. The Use of MMR and Diversity-Based Reranking for Reodering Documents and Producing Summaries. In *Proceedings of the 21st meeting of International ACM SIGIR Conference*. 335–336.

[6] Michael A Cohn, Matthias R Mehl, and James W Pennebaker. 2004. Linguistic markers of psychological change surrounding September 11, 2001. *Psychological science* 15, 10 (2004), 687–693.

[7] Seth Flaxman, Sharad Goel, and Justin M Rao. 2016. Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly* 80, S1 (2016), 298–320.

[8] R Kelly Garrett. 2009. Echo chambers online?: Politically motivated selective exposure among Internet news users. *Journal of Computer-Mediated Communication* 14, 2 (2009), 265–285.

[9] R Kelly Garrett, Dustin Carnahan, and Emily K Lynch. 2013. A turn toward avoidance? Selective exposure to online political information, 2004–2008. *Political Behavior* 35, 1 (2013), 113–134.

[10] Shanto Iyengar and Kyu S Hahn. 2009. Red media, blue media: Evidence of ideological selectivity in media use. *Journal of Communication* 59, 1 (2009), 19–39.

[11] Ewa Kacewicz, James W Pennebaker, Matthew Davis, Moongee Jeon, and Arthur C Graesser. 2014. Pronoun use reflects standings in social hierarchies. *Journal of Language and Social Psychology* 33, 2 (2014), 125–143.

[12] Byungkyu Kang, Nava Tintarev, Tobias Hollerer, and John O'Donovan. 2016. What am I not seeing? An Interactive Approach to Social Content Discovery in Microblogs. In *SocInfo*. 279–294.

[13] Silvia Knobloch-Westerwick, Benjamin K Johnson, and Axel Westerwick. 2014. Confirmation bias in online searches: Impacts of selective exposure before an election on political attitude strength and shifts. *Journal of Computer-Mediated Communication* 20, 2 (2014), 171–187.

[14] Neal Lathia, Stephen Hailes, Licia Capra, and Xavier Amatriain. 2010. Temporal diversity in recommender systems. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 210–217.

[15] Miriam J Metzger, Ethan H Hartsell, and Andrew J Flanagin. 2015. Cognitive dissonance or credibility? A comparison of two theoretical explanations for selective exposure to partisan news. *Communication Research* (2015), 0093650215613136.

[16] Sayooran Nagulendra and Julita Vassileva. 2014. Understanding and controlling the filter bubble through interactive visualization: a user study. In *Conference on Hypertext and Social Media*. ACM, 107–115.

[17] Matthew L Newman, James W Pennebaker, Diane S Berry, and Jane M Richards. 2003. Lying words: Predicting deception from linguistic styles. *Personality and social psychology bulletin* 29, 5 (2003), 665–675.

[18] Dimitar Nikolov, Diego FM Oliveira, Alessandro Flammini, and Filippo Menczer. 2015. Measuring online social bubbles. *PeerJ Computer Science* 1 (2015), e38.

[19] John O'Donovan, Barry Smyth, Brynjar Gretarsson, Svetlin Bostandjiev, and Tobias Höllerer. 2008. PeerChooser: Visual Interactive Recommendation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. ACM, New York, NY, USA, 1085–1088.

[20] Eli Pariser. 2011. *The filter bubble: What the Internet is hiding from you.* Penguin Books.

[21] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The development and psychometric properties of LIWC2015.* Technical Report.

[22] James W Pennebaker, Cindy K Chung, Joey Frazee, Gary M Lavergne, and David I Beaver. 2014. When small words foretell academic success: The case of college admissions essays. *PloS one* 9, 12 (2014), e115844.

[23] Barry Smyth and Paul McClave. 2001. Similarity vs. Diversity. In *4th International Conference on Case-Based Reasoning*.

[24] Nava Tintarev. 2017. Presenting Diversity Aware Recommendations: Making Challenging News Acceptable. (2017).

[25] Nava Tintarev, Matt Dennis, and Judith Masthoff. 2013. Adapting recommendation diversity to openness to experience: A study of human behaviour. In *International Conference on User Modeling, Adaptation, and Personalization*. Springer, 190–202.

[26] Nava Tintarev, Byungkyu Kang, T. Höllerer, and John O'Donovan. 2015. Inspection Mechanisms for Community-based Content Discovery in Microblogs. In *Recsys Workshop on Interfaces and Human Decision Making for Recommender Systems (IntRS)*.

[27] Nava Tintarev, Shahin Rostami, and Barry Smyth. 2018. Knowing the Unknown: Visualising Consumption Blind-Spots in Recommender System. In *ACM SIGAPP Symposium On Applied Computing*.

[28] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.

[29] Axel Westerwick, Steven B Kleinman, and Silvia Knobloch-Westerwick. 2013. Turn a blind eye if you care: Impacts of attitude consistency, importance, and credibility on seeking of political information and implications for attitudes. *Journal of Communication* 63, 3 (2013), 432–453.

[30] Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, and Georg Lausen. 2005. Improving Recommendation Lists Through Topic Diversification. In *WWW'05*. 22–32.

ACKNOWLEDGMENTS

This project is made possible through the NWO ICT with Industry initiative, making it possible for academics to work with industry on real world problems jointly for a week in the great environment of the Lorentz Center. This paper would not have been possible without the help of Reza Aditya Permadi and Andreas Christian Pangaribuan who were involved with system development at the workshop. We also warmly thank Blendle Research for their invaluable time and data contributions to this project.