Finding Diverse Needles in a Haystack of Comments - Social Media Exploration for News

Hang Zhang[*]
Max Planck Institut für Informatik
Saarbrücken,Germany
hzhang@mpi-inf.mpg.de

Vinay Setty
Max Planck Institut für Informatik
Saarbrücken, Germany
vsetty@mpi-inf.mpg.de

ABSTRACT

Use of social media platforms to express opinion and discuss various topics has been increasingly popular. Consequently, huge volume of social media data is generated by users across all these platforms, e.g. users comment on a variety of content items such as news articles, videos, images on social media. These comments are often noisy and sparse, therefore, identifying sub-topics within them to explore social media is a challenge. In this paper, we develop an effective way to distill sub-topics from all the comments related to a textual query and apply two different diversification techniques to select comments. We conduct experiments to validate our idea using seven years of Reddit comments and news events from Wikipedia Current Events Portal as queries.

CCS Concepts

•**Information systems** → *Social networks;* Social networking sites;

Keywords

Social media comments; Reddit; News

1. INTRODUCTION

With the proliferation of online social media applications, users are increasingly expressing their opinion online generating massive amounts of data. For example, in social media applications like Youtube, Instagram, Reddit users express their opinion by commenting on the videos, pictures, links to news articles and textual questions etc. posted by other users. Such comments often tend to be sparse, noisy and too long to grasp their summary. Moreover, in the comments users are known to raise a wide-range of sub-topics related to the thematic topic they are discussing about. Consequently, exploring such a large-scale comments to distill diverse sub-topics and summarize the comments is analogous to finding diverse set of needles in a haystack of comments.

Comments from subreddit forums in Reddit are of conversational style in particular. They form tree-structured discussions under threads: a comment at lower level is a reply to the one at higher level while comments at same level are more or less independent to each other. Because of the casual nature of the social media, users can be inspired by the existing comments or simply brainstormed to comment on the topical aspects that are indirectly relevant to the threads, some of the topical aspects also lead to topic drifts in the discussion. For example, a thread about *The 2014 FIFA World Cup is held in Brazil* there are discussions talking about airlines companies since people care about the expenses of flying there; For a thread *US president Barack Obama announced resumption of normal relations between Cuba and the US*, a large number of comments discussing about the tobacco and Cuban cigars. Note that in forums like Reddit these threads are usually moderated and it is safe to assume that they are relevant to the topic being discussed but since they discuss a wide variety of sub-topics related to the topic, exploring them can be overwhelming to the users.

These discussions implicitly form many topic aspects which contain hundreds or thousands of comments. Applying standard retrieval techniques such as TF-IDF [14] on comments gives significant redundancy in the results which is undesirable. Moreover, existing diversification proposed in [1, 4, 9, 8, 23] are rendered ineffective because the comments tend to be short, noisy and unedited unlike the typical web documents such as news articles.

Therefore, there is a need to distill the implicit topics from the noisy and sparse comments and select the diversified representative comments so that the social media exploration is simplified. This results in several challenges: (1) the implicit topics need to be distilled and corresponding comments need to be clustered together, (2) different topical groups need to be ranked and finally, (3) diversified representative topics need to be selected and ranked.

To address the above challenges, in this paper we propose a solution that filters, processes, ranks and diversifies the Reddit comments in the following stages: (1) given a news event as a query using a probabilistic text retrieval technique we retrieve comments which gives pseudo relevance score of the comments, (2) we then use Dirichlet Multinomial Mixture (DMM) model with collapsed Gibbs sampling which is robust to noisy and sparse comments for clustering them, (3) we then rank the clusters based on the pseudo relevance scores, finally (4) we apply two diversification techniques for maintaining the proportionality in comments and also retain the conversation structure of the comments. We conclude the paper with experiments on using Reddit comments from over seven years and using events from Wiki news as query workload.

[*] Also affiliated with Department of Computational Linguistics, Saarland University

WebSci '16, May 22-25, 2016, Hannover, Germany
© 2016 ACM. ISBN 978-1-4503-4208-7/16/05. . . $15.00
DOI: http://dx.doi.org/10.1145/2908131.2908168

2. RELATED WORK

There are a number of diversification techniques for edited documents to reduce the redundancy in the retrieval result. Carbonell and Goldstein [4] proposes Maximum Marginal Relevance which returns the next document as the one having maximum marginal relevance (MMR) given the set of already returned documents; Zhai et al. [23] generalizes the idea behind Maximum Marginal Relevance and devises an approach based on language models; Intent-Aware Selection (AI-SELECT) [1] and eXplicit Query Aspect Diversification (XQuAD) are redundancy-based explicit diversification methods of covering all query aspects by including each one of them and penalizing redundancy; There are also proportionality-based explicit diversification methods (PM-1/2) [9, 8] aiming at retrieving a result list that represents query aspects according to their popularity by promoting proportionality. These technique prove effective for long edited documents, they don't fit the retrieval task with short, sparse and unedited comments.

When it comes to the huge volume of unedited social media data, ways to approach the problem vary due to the inherent characteristics of the social media data sources. There are a number of studies on summarizing the data by either extracting the latent topical structure or generating a summary which can contribute to the redundancy reduction and diversity coverage. Zhao et al. [17] focuses on the threads from web forums by investigating the task of web forum thread summarization and generating a brief statement of each thread that involves multiple dynamic topics. Xinfan Meng et al. [15] generates an entity-centric, topic-oriented opinion summarization from Twitter which aims at generating opinion summarization in accordance with topics. Zhao et al. [18] focus on hierarchical multi-label classification of social text streams, their work can track topics with conceptual drifts over time. However, comments with various topic aspects as well as the topic drifts can grow very long in the discussion and it is hard to grasp the summary.

There are a few news exploration and diversification techniques proposed in the literature but they do not consider social media comments in to account [19, 20].

3. PROBLEM DEFINITION

So far most related research approach the problem of reducing redundancy through uncovering the latent topical structure and generating a summary for the user. However, comments linked to the threads are of conversational style so that they touch various topic aspects, some of which can also lead to topic drifts. When using a news event as query q to retrieve comments from subreddit forums in Reddit, the top m threads T, $T = \{t_1, t_2, ...t_m\}$ and comments C, $C = \{C_1, C_2, ...C_m\}$ are retrieved as the pseudo search result R, $R = T \cup C$; C_i, $(C_i \in C)$ is a set of comments $c_{i,j}$, $(c_{i,j} \in C_i)$ linked to thread t_i, $t_i \in T$, $Ci = \{c_{i,1}, c_{i,2}, ...c_{i,x_i}\}$. The search result R has significant redundancy which is undesirable. Therefore, there is a need to to distill the K implicit topics from the pseudo search result R and select the diversified representative comments as the search result R' to simplify the social media exploration.

4. SOLUTION

Our solution overview is as Figure 1: Given a news summary as a query q, we first use elastic search [10] with the probabilistic retrieval framework Okapi BM 25 [14] to retrieve and rank the threads according to the pseudo relevance score then choose the top m threads T as well as the linked comments C as pseudo search result R. Next we pre-process the comments C tactfully in order to extract interpretable topics from R. Once the K topics are extracted, we use diversification method to diversify the pseudo-relevant search result R to obtain the search result R'. One diversification method uses **Sainte-Laguë (SL) Method** to reduce the redundancy and diversify search results topically and sentimentally, the other one is **Comment-Tree Decomposition (CTD) Method**, which is particularly for tree-structured comments to reduce the redundancy while keeping the coherence and conversation style of the comment trees.

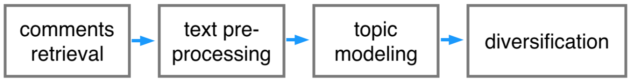


Figure 1: solution overview

4.1 Text Pre-processing

We have three steps in text pre-processing

1. We remove urls as the comment topic is often different from the topic of the referenced url. In addition, we also remove any non-alphanumeric characters.

2. Extract named entities using *Senna* [7] and we boost the importance of named entities by assigning higher weights (based on number of occurances) than other terms in the text representation. Zhao et al. [24] find out there are more entity-oriented topics in social media whereas there are more event-oriented topics in traditional media such as *New York Times*. So extraction and duplication of named entities (doubling the number of named entities in the text representation) can lead to more interpretable topics.

3. Part-of-speech tagging using *Senna* [7]. According to the part-of-speech tags, remove the stop words and keep the verbs as well as words that can potentially form noun phrases. According to centering theory [11], it's a better way to model topic drifts.

4.2 Topic Modeling

The challenge for modeling topics in comments comes from the short and noisy nature of the data. Regarding the shortness of the data, LDA [3] is not fit for the task because the arbitrary number of topics easily leads to topics that are hard to be interpreted and the output is also a multinomial distribution over topics per document (documents are individual comments in Reddit in the rest of the paper). Derived LDA Author-Topic Model [21] assumes that there is one particular topic probability distribution for every author; However, authors are everywhere on Reddit participating in the various discussions. We choose the Dirichlet Multinomial Mixture (DMM) model [16] to label each comment with one topic tag. DMM is a probabilistic generative model for documents and embodies two assumptions about the generative process: first, the documents are generated by a mixture model; second, there is one-to-one correspondence between mixture components and clusters. When generating document d, DMM first selects a mixture component (topic cluster) k according to the mixture weights (weights of clusters) $P(z = k)$. Then document d is generated by the selected mixture component (cluster) from distribution $P(d|z = k)$. Thus, we can characterize the likelihood of document d with the sum of the total probability over all mixture components:

$$P(d) = \sum_{k=1}^{K} P(d|z = k)P(z = k)$$

where, K is the number of mixture components (topic clusters). DMM [16] assumes that each mixture component (topic cluster) is a multinomial distribution over words and each mixture component (topic cluster) has a Dirichlet distribution prior:

$$P(w|z=k) = P(w|z=k,\Phi) = \phi_{k,w}$$

where $\sum_{w}^{V} \phi_{w,k} = 1$ and $P(\Phi|\vec{\beta}) = Dir(\vec{\theta}|\vec{\beta})$. DMM [16] also assumes that the weight of each mixture component (cluster) is sampled from a multinomial distribution which has a Dirichlet prior:

$$P(z=k) = P(z=k|\Theta) = \theta_k$$

where $\sum_{k}^{K} \theta_k = 1$ and $P(\Theta|\vec{\alpha}) = Dir(\vec{\theta}|\vec{\alpha})$ We use collapsed Gibbs Sampling [22] for DMM. The number of topics can be inferred automatically and it is fast to converge. Documents are randomly assigned to K clusters initially and the following information is recorded: $\vec{z}$ is the cluster labels of each document, m_z is number of documents in each cluster z, and n_z^w is the number of occurrences of word w in each cluster z, N_d is the number of words in document d, N_d^w is the number of occurrence of word w in the document d. The documents are traversed for a number of iterations. In each iteration, each document is reassigned to a cluster according to the conditional distribution of $P(Z_d = z|\vec{z}_{\neg d}, \vec{d})$, $\neg d$ means d is not contained:

$$P(Z_d = z|\vec{z}_{\neg d}, \vec{d}) \propto \frac{m_{z,\neg d} + \alpha}{D - 1 + K\alpha} \frac{\prod_{w \in d} \prod_{j=1}^{N_d^w} (n_{z,\neg d}^w + \beta + j - 1)}{\prod_{i=1}^{N_d} (n_{z,\neg d} + V\beta + i - 1)}$$

Where, hyperparameter α controls the popularity of the clusters. Hyperparameter β emphasizes on the similar words between a document and clusters. The most frequent words in each cluster can be extracted as topic words.

4.3 Diversification Methods

4.3.1 Sainte-Laguë Method

In situations where the comments do not follow conversational style, they can be treated as independent entities. Zhao et al. [24] shows that one characteristic of social media content compared with traditional news media is the amount and coverage of user opinions expressed. Therefore, in addition to the topic aspect of the comment, we also create a sentiment dimension for each comment. That way, a comment $c_{i,j}$ $(c_{i,j} \in C_i)$ is labeled with a topic tag $topic_{c_{i,j}}$ and a sentiment tag $senti_{c_{i,j}}$ by VADER [12]. The reason why we add the sentiment dimension when comments are treated as independent entities is that a single comment has only one sentiment; However, when the conversation style is considered, discussions of tree-structured comments are treated as entities so that generally discussions present all the three sentiments namely positive, negative and neutral. We choose VADER [12] for the sentiment analysis task because it uses a combination of qualitative and quantitative methods to produce empirically validate, valence-based, gold-standard sentiment lexicon which is especially attuned to social media texts. Each comment $c_{i,j}$ has a score $s_{c_{i,j}}$ given by users and a pseudo relevance score $s_{pr,i}$ that C_i inherits from t_i. For each of the threads $t_i(t_i \in T, |T| = m)$, we cluster the linked comments $C_i(C_i \in C)$ according to both of the tags $(topic_{c_{i,j}}, senti_{c_{i,j}})$; Then we first rank the all the topic-sentiment clusters according to the pseudo relevance score $s_{pr,i}$ so for each rank r_j $(r_j \in R, |R| = |T| = m)$, there are a number of N_{cl,r_j} topic-sentiment clusters cl_{i,r_j} $(cl_{i,r_j} \in Lr_j, |Lr_j| = N_{cl,r_j})$; Second we rank the comments in each of the topic-sentiment clusters cl_{i,r_j} according to the score $s_{c_{i,j}}$ given by users.

Denominator	/1	/3	/5	Seats(*)
topic A positive	50*	16.67*	10	2
topic A neutral	40*	13.33*	8	2
topic A negative	30*	10	6	1

Table 1: Sainte-Laguë method example

Here we propose the **Sainte-Laguë (SL) method** to diversify the search result by retrieving the representative comments proportionally from Lr_j. **SL method** [13] is a highest quotient method for allocating seats in party-list proportional representation used in many voting systems. After all the comments have been tallied, successive quotients are computed using equation 1 for each cluster. where, V is the total number of comments in each cluster; S is the number of seats that cluster has been allocated so far, initially 0 for all clusters.

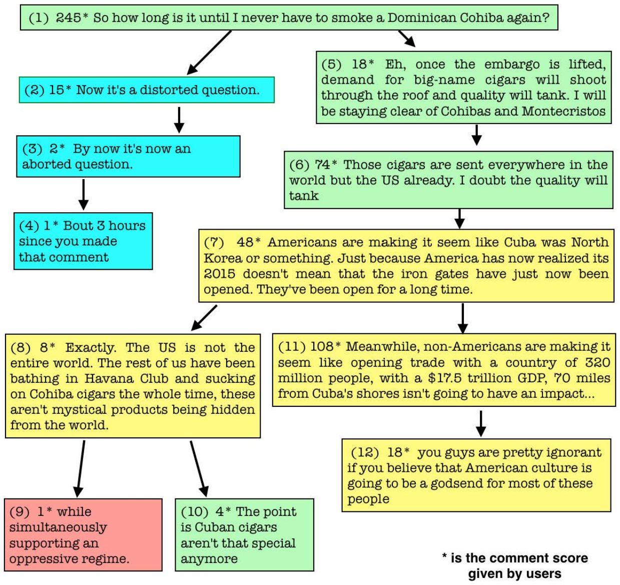
$$quotient = \frac{V}{2S+1} \qquad (1)$$

Whichever cluster has the highest quotient gets the next seat allocated, and their quotient is recalculated given their new seat total. The process is repeated until all seats have been allocated. The number of the seats is a hyper-parameter, it can be set according to users' interests. We illustrate the **Sainte-Laguë (SL) method** with an example using three clusters as shown in Table 1. In this example, five comments are expected to be retrieved (number of seats is five). The denominators in the first row are calculated as $2S + 1$ where $S = 0, 1, 2...$ respectively and the quotients in the column 2 to column 4 are the quotients calculated using formula 1. The quotients marked with '*' represent the allocated seats. For this example, two comments are from cluster *topic A positive* two comments from cluster *topic A neutral* and one comment from cluster *topic A negative* are retrieved.

We apply **SL method** to clusters L_{r_j} of all ranks. Comments c_g with higher user score $s_{c_{i,j}}$ are selected first. N_{r_j} is the number of retrieved comments (seats) at rank r_j $(r_j \in R)$ and $N_{r_j} = min(\gamma \cdot N_{cl,r_j}, |C_{i,r_j}|)$ where, γ is a positive constant limiting the number of selected comments for the threads with arbitrarily large number of comments. $|C_{i,r_j}|$ is number of comments C_i at rank j. Then the representative comments are retrieved from each topic-sentiment cluster of all ranks proportionally.

4.3.2 Comment-Tree Decomposition Method

Comments that are linked to a thread generally form a large number of tree-structured discussions. As comment-tree u in Figure 2, the comment (c) at the lower level is a reply to the one at the higher level. e.g. $c(2)$ is a reply to $c(1)$. Comments sharing the same parent comment and situating at the same level are relatively independent replies to their parent comment, e.g. $c(2)$ and $c(5)$ are the two independent replies of same level to $c(1)$. The hierarchical levels reflect the coherence and the conversational style of the discussion. Here we introduce **Comment-Tree Decomposition (CTD) Method** to decompose large-scale comment-trees u into small-scale sub-trees u' by enumerating paths from the head to its leaves at level l. We use the example in Figure 2 to illustrate the tree decomposition process. The comment tree u in the Figure 2 is a part of a discussion which has 8 level with 27 comments and it is linked to the thread *US, Cuba restore full diplomatic relations after 54 years*. The arrows point from higher levels to lower levels. We enumerate the paths from head at level 0 to all of its leaves at level 5.



(1) 245* So how long is it until I never have to smoke a Dominican Cohiba again?

(2) 15* Now it's a distorted question.

(3) 2* By now it's now an aborted question.

(4) 1* Bout 3 hours since you made that comment

(5) 18* Eh, once the embargo is lifted, demand for big-name cigars will shoot through the roof and quality will tank. I will be staying clear of Cohibas and Montecristos

(6) 74* Those cigars are sent everywhere in the world but the US already. I doubt the quality will tank

(7) 48* Americans are making it seem like Cuba was North Korea or something. Just because America has now realized its 2015 doesn't mean that the iron gates have just now been opened. They've been open for a long time.

(8) 8* Exactly. The US is not the entire world. The rest of us have been bathing in Havana Club and sucking on Cohiba cigars the whole time, these aren't mystical products being hidden from the world.

(9) 1* while simultaneously supporting an oppressive regime.

(10) 4* The point is Cuban cigars aren't that special anymore

(11) 108* Meanwhile, non-Americans are making it seem like opening trade with a country of 320 million people, with a $17.5 trillion GDP, 70 miles from Cuba's shores isn't going to have an impact...

(12) 18* you guys are pretty ignorant if you believe that American culture is going to be a godsend for most of these people

* is the comment score given by users

Figure 2: comment tree

Therefore, the comment-tree u in Figure 2 is decomposed into four sub-tree (u'_i) as following: subtree u'_1 is $c(1) \rightarrow c(2) \rightarrow c(3) \rightarrow c(4)$; subtree u'_2 is $c(1) \rightarrow c(5) \rightarrow c(6) \rightarrow c(7) \rightarrow c(8) \rightarrow c(9)$; subtree u'_3 is $c(1) \rightarrow c(5) \rightarrow c(6) \rightarrow c(7) \rightarrow c(8) \rightarrow c(10)$; subtree u'_4 is $c(1) \rightarrow c(5) \rightarrow c(6) \rightarrow c(7) \rightarrow c(11) \rightarrow c(12)$. The next step is to choose one of subtrees u'_i to represent the original tree. All the subtrees of smaller scale potentially have less redundancy compared with the original tree. we give a tree score $score(u'_i)$ to each of the subtrees and select the subtree with the highest tree score. We propose several ways of computing the tree score $score(u'_i)$ for the subtree u'_i:

1. **Comment score**: each comment $c_{i,j}$, $c_{i,j} \in u'_i$ has a score $s_{c_{i,j}}$ given by users as the number marked with * in Figure 2. tree score is the sum of the user score of the comments in the subtree: $score(u'_i) = \sum_{j=1}^{|u'_i|} s_{c_{i,j}}$.

2. **Linguistic features**: score the subtree using the diversity of the linguistic features (f) of comments ($c_{i,j}$) in the subtree: $score(u'_i) = |\{f_{c_{i,j}} | c_{i,j} \in u'_i\}|$. The different linguistic features we proposed using the techinque explained in Section 4.1 are *NP words* (words that can potentially form noun phrases), *named entities* and *bigrams*.

3. **Number of topics**: This feature refers to the number of unique topics discussed in a subtree. $score(u'_i) = |\{topic_{c_{i,j}} | c_{i,j} \in u'_i\}|$ where $topic_{c_{i,j}}$ is the topic tag of comment $c_{i,j}$.

5. EXPERIMENTAL EVALUATION

We conducted our experiment using the Reddit data (for detailed analysis see [5]) under news, world news and politics subreddits from year 2008 to 2015 which consists of $845,004$ threads and their $26,669,242$ linked comments. Then we chose 50 news event summaries from **Wikipedia Current Events Portal**[1] from year 2011 to 2014 as queries. We limit our evaluations to these 50 queries as they were most widely discussed in Reddit.

retrieval result	CG	retrieve percent
diversified result with SL	71.80 ± 44.31	16.60%
pseudo search result	50.37 ± 27.29	100%

Table 2: Sainte Laguë method experiment result

We use **elastic search** [10] with **Okapi BM 25** [14] retrieve and rank the threads and choose the top 10 threads as well as their linked comments as pseudo search result (R) for each query. There are 4436.43 comments on average for each query. Then we carried out the following tasks for the pseudo search result (R) of each query:

1. **Text pre-processing**: normalize the comments for each query following the steps in section 4.1.

2. **Topic modeling**: extract the topics from each pseudo relevant result R using the technique as described in section 4.2. We set $\alpha = 0.1$ and $\beta = 0.2$ in order to extract more interpretable topics after examining manually the all topic words; We also notice comments of different languages are clustered according to the language, which shows the robustness of our topic modeling technique. Then each comment is labeled with a topic tag.

3. **Sentiment analysis**: we use VADER as discussed in section 4.4 to give a score for each comment for the Sainte-Laguë method. We label the comment using compound score and label a positive tag when the score is between 1 and 0.1, a negative tag when it is between -0.1 and -1 or a neutral tag when it falls in between 0.1 and -0.1.

We use **Cumulative Gain (CG)** to measure the diversity and CG also penalizes redundancy: $CG[k] = \sum_{j=1}^{k} G[j]$ and $G[k] = \sum_{i=1}^{m} J(d_k, i)(1-\alpha)^{r_{i,k-1}}$ where, $r_{i,k-1}$ is the number of comments (d_j) ranked up to position $k-1$ that contain nugget n_i and $r_{i,k-1} = \sum_{j=1}^{k-1} J(d_j, i)$; $J(d, i) = 1$ if comment (d) contains nugget n_i, otherwise $J(d, i) = 0$; k is set to be 10 in our experiment because we choose top 10 threads and their linked comments; The possibility of assessor error α is set to be 0.5. Charles L.A claims [6] that CG at rank k can be used directly as diversity evaluation measure and Al-maskari [2] provides evidence that CG correlates better with user satisfaction than Normalized Discounted Cumulative Gain (nDCG).

Experiment for the Sainte-Laguë (**SL**) Method: We set $\gamma = 2.5$ to limit the number of selected comments for the threads with arbitrarily large number of comments. We use topic-sentimental tag as the nugget, which is the combination of both topic and sentiment tags to compute CG for the pseudo search result and diversified search result with SL method for each query, the average CG over 50 queries is presented in the table 2. We used the CG of the pseudo search result as the baseline.

We used the single topic tag as the nugget for Comment Tree Decomposition (**CTD**) Method to compute CG for the pseudo search result for each query. We also did parallel experiment for the different ways to compute the *tree score* for the decomposed subtrees and compare their diversified results. We set the decomposition level l at 5 and retrieve the all the selected subtrees linked to each thread. CG is computed for different ways to score the subtrees. The average CG over the diversified results and pseudo search results of the 50 queries is reported as the table 3 and we use the CG of the pseudo search result as a baseline to compare with.

[1] https://en.wikipedia.org/wiki/Portal:Current_events

	CG	retrieval percent
CTD comment score	27.11 ± 11.19	70.51%
CTD NP words	27.31 ± 10.81	70.67%
CTD named entities	27.54 ± 11.39	59.17%
CTD bigrams	26.60 ± 10.63	70.77%
CTD number of topics	28.45 ± 11.72	73.26%
pseudo search result	26.38 ± 9.8	100%

Table 3: experiment with CTD method

5.1 Discussion

The experiment result for SL method shows that diversified search results have tremendous diversity improvement with only 16.60% of comments from the pseudo search result on average. The diversity improvement is foreseeable because comments are retrieved directly from the topic sentimental clusters with proportionality. SL method proves to be effective with the expense of the coherence in the discussion. The experiment using CTD method also shows diversity improvement with approximately 70% comments from the returned results on average. Among all the different ways of scoring sub-trees, using *number of topic* to score the sub-tree achieved the highest diversity. With *named entities* as a linguistic feature to score the sub-tree, only 59% the comments from the pseudo search result to present improved diversity, it indicates that people are more interested in talking about person, organization, location etc, which is also consistent with the Zhao et al. [24] that there are more entity-oriented topics in social media. *NP words* can reflect the topic drifts, result shows it achieved slight diversity improvement. CTD with *comment score* and *bigrams* also preserve the original diversity with around 70% comments from pseudo search results. In summary, CTD also demonstrates the effectiveness to reduce the redundancy and improve the diversity of pseudo search results using less comments; Small-scale comment trees also maintain the coherence and conversation style of the discussion.

6. CONCLUSION

In this paper, we proposed novel techniques to distill the diverse interpretable topics from pseudo search result using topic model with effective text processing. We studied the characteristics of Reddit comments and introduced two diversification methods namely Sainte-Laguë (SL) Method and Comment Tree Decomposition (CTD) Method to reduce the redundancy and diversify the returned results. SL method treat comments as entities while CTD preserve the conversational style of the discussions. We also report the experiment results of the two methods. Both of methods prove to be effective diversification technique.

7. REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *ICWSM*, pages 5–14, 2009.

[2] A. Al-Maskari, M. Sanderson, and P. Clough. The relationship between ir effectiveness measures and user satisfaction. In *SIGIR*, pages 773–774. ACM, 2007.

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

[4] J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, pages 335–336. ACM, 1998.

[5] D. Choi, J. Han, T. Chung, Y.-Y. Ahn, B.-G. Chun, and T. T. Kwon. Characterizing conversation patterns in reddit: From the perspectives of content properties and user participation behaviors. In *COSN*, pages 233–243. ACM, 2015.

[6] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *SIGIR*, pages 659–666. ACM, 2008.

[7] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537, 2011.

[8] V. Dang and B. W. Croft. Term level search result diversification. In *SIGIR*, pages 603–612. ACM, 2013.

[9] V. Dang and W. B. Croft. Diversity by proportionality: an election-based approach to search result diversification. In *SIGIR*, pages 65–74. ACM, 2012.

[10] C. Gormley and Z. Tong. *Elasticsearch: The Definitive Guide*. " O'Reilly Media, Inc.", 2015.

[11] B. J. Grosz, S. Weinstein, and A. K. Joshi. Centering: A framework for modeling the local coherence of discourse. *Computational linguistics*, 21(2):203–225, 1995.

[12] C. Hutto and E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *AAAI Conference on Weblogs and Social Media*, 2014.

[13] A. Lijphart and B. Grofman. *Electoral laws and their political consequences*. Agathon Press, 1986.

[14] C. D. Manning, P. Raghavan, H. Schütze, et al. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

[15] X. Meng, F. Wei, X. Liu, M. Zhou, S. Li, and H. Wang. Entity-centric topic-oriented opinion summarization in twitter. In *SIGKDD*, pages 379–387. ACM, 2012.

[16] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using em. *Machine learning*, 39(2-3):103–134, 2000.

[17] Z. Ren, J. Ma, S. Wang, and Y. Liu. Summarizing web forum threads based on a latent topic propagation process. In *CIKM*, 2011.

[18] Z. Ren, M.-H. Peetz, S. Liang, W. van Dolen, and M. de Rijke. Hierarchical multi-label classification of social text streams. In *SIGIR*, pages 213–222. ACM, 2014.

[19] V. Setty, S. Bedathur, K. Berberich, and G. Weikum. Inzeit: Efficiently identifying insightful time points. *Proc. VLDB Endow.*, 3(1-2):1605–1608, 2010.

[20] J. Singh, W. Nejdl, and A. Anand. History by diversity: Helping historians search news archives. In *CHIIR*, pages 183–192. ACM, 2016.

[21] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths. Probabilistic author-topic models for information discovery. In *SIGKDD*, pages 306–315. ACM, 2004.

[22] J. Yin and J. Wang. A dirichlet multinomial mixture model-based approach for short text clustering. In *SIGKDD*, pages 233–242. ACM, 2014.

[23] C. X. Zhai, W. W. Cohen, and J. Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *SIGIR*, pages 10–17. ACM, 2003.

[24] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. In *Advances in Information Retrieval*, pages 338–349. Springer, 2011.