Comments Analysis and Visualization Based on Topic Modeling and Topic Phrase Mining

D. Ramamonjisoa, R. Murakami, B. Chakraborty
Faculty of Software and Information Science
Iwate Prefectural University, IPU
Takizawa, Japan
david@iwate-pu.ac.jp,g031j149@s.iwate-pu.ac.jp,basabi@iwate-pu.ac.jp

ABSTRACT

The number of user-contributed comments is increasing exponentially. Such comments are found widely in social media sites including internet discussion forums and news agency websites. In this paper, we summarize the current approaches to text analysis and the visualization tools which deal with opinion and topics mining of those comments. We then describe experiments for topic modeling on users' comments and examine the possible extensions of methods on visualization, sentiment analysis and opinion summarization systems.

KEYWORDS

Topic modeling; user comments; opinion and sentiment mining.

1 INTRODUCTION

Many websites provide commenting facilities for users to express their opinions or sentiments with regards to content items, such as, videos, news stories, blog posts, etc. Previous studies have shown that user comments contain valuable information that can provide insight on web documents and may be utilized for various tasks [1],[2],[3],[4].

User comments are a kind of user-generated content. Their purpose is to collect user feedback, but they have also been used to form a community discussing about any piece of information on the internet (news article, video, live talk show, music, picture, and so on). The commenting tool becomes a social gathering software where commenters share their opinions, criticism, or extraneous information.

In some websites, user comments analysis is an information retrieval task which consists of comments filtering, comments ranking, and comments summarization [5].

A social knowledge task should allow users to realize the analysis autonomously or semi-autonomously, and visualize results in a succinct manner to leverage the user tasks. Topics within comments also should be extracted and summarized as in graphs or clouds. Thanks to text mining techniques, topic trends or users' needs can now be analyzed and summarized autonomously like in large text corpora such as TopicNets [6].

This paper describes an experiment on users' comments analysis. The experiment also includes visualization with topic modeling and topic phrase mining.

In this paper, we first present the methods used during the experiment and the datasets. Next, we detail the experiments and discuss the results obtained. Finally, we draw conclusions and discuss future work.

2 WHAT ARE COMMENTS

Figure 1 depicts the most commented news article on Yahoo site on October 17th 2014 concerning about the Ebola outbreak and the Obama administration policy. Within a few hours after the publication of this article online, 4108 comments are already posted to the Yahoo news site. For someone who wants to grasp the

ISBN: 978-1-941968-08-6

content and summary of those comments, it needs an enormous effort of time and reading capacity. Yahoo News site provides a comment rating tool. For this example, the most appreciated comment is marked with 549 thumb-ups and has the username "clare", but it is not enough to understand or overview all those comments of this article. Through further analysis, we want to know what kind of sentiment is given by all those comments what are the main topics, and how do these topics relate each other, and so on.

Figure 1. Comments example from Yahoo News

3 APPROACHES

The data model for our experiments is described as follows (and more details can be found in [7]): Users' comments or blog posts are designated as document collections. The model of the comments (as each comment is a specific short document) collection is described below:

$D = \{c_i\}$ where

$c_i = (docID, time_{stamp}, title_i, content_i)$

A natural language processing (NLP) task conducted to extract important keywords such as nouns or adjectives from the *content*$_i$ of each c_i. A bag-of-word model was then constructed by attaching a weight to each extracted word. A weight can be the frequency of a term, i.e., *tf*. The content of the document is then a set of tuple keywords and weights as used in many information retrieval (IR) tasks:

$$content_i = \{(k_{ij}, w_{ij})\} \quad j \in [1..n],$$

n number of keywords in the content, $w_{ij} > \tau$

A document collection is therefore a table where the rows consist of the weights of each keyword in each document and the columns list the documents. This document list is arranged as time-series data so that old posts and comments are the first element of the list and the newest comments and posts are the last. The document table is formalized as follows:

$$D^T = [content_i(row) \times c_i(column)] i \in [1..m],$$

The next section describes the method used for the topic modeling.

4 TOPIC MODELING

Topic modeling is a method for analyzing large quantities of unlabeled data. A topic is a probability distribution over a collection of words. A topic model is a formal statistical relationship between a group of observed and latent (unknown) random variables that specifies a probabilistic model to generate the topics. The central goal of a topic is to provide a "thematic summary" of a collection of documents. In other words, it answers the question what themes are those documents discussing.

4.1 LDA based Topic Clustering

The topic modeling is used to extract T topics out of the comments collection. That is, we have a set of comment "documents" $C = \{c_1, c_2, \ldots c_n\}$ and a number of topics $T = \{t_1, t_2, \ldots, t_m\}$. A document c_i can be viewed by its topic distribution. For example, $\Pr(c_1 \in t_1)$=0.50 and $\Pr(c_1 \in t_2)$=0.20 and so on. The default topic modeling based on LDA is a soft clustering. It can be modified into hard clustering by considering each comment as belonging to a single topic (cluster) t_r,

$r = argmax_r \Pr(t_r|c) = argmax_r \Pr(c \mid t_r) \Pr(t_r)$,

where r is the number of topics that has the maximum likelihood for each comment. Hence, the output of the LDA based topic clustering approach is an assignment from each comment to a cluster [8].

4.2 Non-negative Matrix Factorization (NMF) for Topic Modeling

ISBN: 978-1-941968-08-6 ©2015 SDIWC

Another method used for solving the topic modeling problem is NMF. NMF was developed based on a traditional technique called *latent semantic indexing* (LSI). The LSI is a topic modeling which includes negative weights on its output. Negative weights on keywords or topics are difficult to interpret in comparing to the results of the LDA model where weights are probability distribution and all positives. NMF takes as input the document table described in the previous section and converts it into a sparse matrix. Then, NMF solves a matrix decomposition problem given a particular rank value corresponding to the number of topics. NMF, as its name suggests, imposes non-negativity constraints on every element of the resulting matrices so that it can maintain interpretability. The output of the NMF program is a list of keywords for each topic as in LDA except that weights are not probability distribution. The formulation of the NMF method is as follows.

Table 1. Notations

Notation	Description
n	The number of keywords
m	The number of documents
k	The number of topics
$X \in \mathbb{R}^{n\times m}$	A keyword-by-document matrix
$W \in \mathbb{R}^{n\times k}$	A keyword-by-topic matrix
$H \in \mathbb{R}^{k\times m}$	A topic-by-document matrix

According to the notations listed in Table 1, given a nonnegative matrix $X \in \mathbb{R}^{n\times m}$, and an integer $k \ll \min(n,m)$, NMF finds a lower-rank approximation given by $X \approx WH$, where $W \in \mathbb{R}^{n\times k}$ and $H \in \mathbb{R}^{k\times m}$ are nonnegative factors. NMF is therefore an optimization problem as to minimize the distance between two nonnegative matrices X and WH with respect to W and H, subject to the constraints $W, H \geq 0$. The square of the Euclidian distance between X and WH is given by

$$\|X - WH\|^2$$

We used a toolkit implemented in Python language named *scikit-learn* to solve the optimization problem based on projected gradient methods [9].

4.3 Topic Phrases Mining PhraseLDA

Topics from the previous methods are difficult to interpret. The list of unigrams as a result of the topic modeling is an ambiguous representation of the topic. Phrase topics or multi-words keyword are easier to interpret. They are widely used in the library databases and most published scientific journals. A new algorithm and program were developed by El-Kishky et al. [10] to extend LDA and use n-grams (multi-words) instead of words in topics. Although this algorithm is expensive in terms of computational time, it is appropriate to use it with our comments corpus composed with short texts and a few hundreds of comments.

4.4 Key Comment Selection within Cluster of Comments

For each topic obtained by the topic modeling, a set of comments are associated. We define the key comment as the top of the comments by ranking them within their clusters. The ranking method is realized by comparing each comment vector (a bag of words) to the list of words which form the topic vector. We use cosine distance for the comparison. The most similar to the topic is the key comment.

5 EXPERIMENTS

Our experiments were based on Yahoo news comments and Guardian News datasets. The Yahoo most read and commented news dataset was obtained from the authors of the paper [10]. The news and comments are collected from April 29th 2012 to May 12th 2012.

Figure 2 shows the distribution of the number of comments per news article. The minimum number of comments per news is 14 and the maximum is 34700. The mean value of the number of comments is 1472 comments per news.

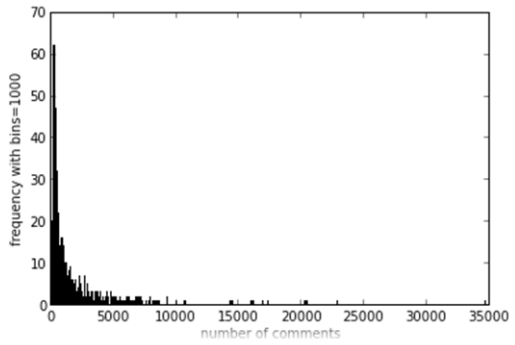

Figure 2. Distribution of Number of Comments per News article

5.1 Experimental Result from Yahoo News dataset

The Yahoo News dataset is an interesting one because it is a large corpus of data and all comments were in English and a comment can have several sentences. There are 1005 news articles and for each article, there are many comments as similar in Figure 1. The data are in HTML files so we implemented a preprocessing program based on scrapper library in Python to extract only the texts articles and texts comments according to the data model in Section 3. This preprocessing is very crucial because all Yahoo News articles HTML files don't have the same format. Some have Javascript codes and advertisement links.

When the HTML data are processed, user can enter a news filename and then compute the topics in the news article and in the comments data. In our setting, we extract only 5 topics composed of 10 keywords for each topic and limit the maximum words to 10000 in comments corpus. We obtain the result within a few seconds.

In the following example, we present topics computed from the most commented news article in the dataset. There are 18446 commenters and 34700 comments. The news article is about an incident that happens to a tanned woman and her daughter during her tanning session. The news title is as "*Tanning booth mom calls arrest for taking daughter, 5, a misunderstanding; 'It was a sunburn,' dad says | The Sideshow*." She was arrested because police believed that she brought her daughter with her in the tanning salon which is forbidden by the American law. The article has stirring up controversy among readers where it turned out to be a bad joke about the white Americans who want to have dark skins. Table 2 describes the topic list results from LDA and NMF.

Table 2. LDA and NMF results

#	Topics LDA	Topics NMF
0	like look chocolate face jolson crispy mammy obama would love	look like old chocolate al 44 burnt leather jolson turd
1	tanning child salon booth daughter would tan woman bed mother	black white people want trying person jackson michael lady lol
2	black white people skin want woman like look jackson tan	woman need ha help mental problem issue sick obviously ugly
3	look like think tan mommy mom woman fried little doe	think good doe look really mirror attractive actually don lord
4	look like leather face old woman something year lady shoe	tanning tan booth mom salon was child bed just daughter
5	jersey new kid oompa loompa get shore mom child people	skin cancer color say leather ha going future doesnt die

NMF topics are slightly better than the LDA ones. The words in Figure 3 (a) represent the most frequent words in the comments collection. The bigger the size of a word, the higher is its frequency. Figure 3(b) shows the top 50 topic phrases obtained with the topic phrase mining algorithms. The parameters in the setting are tuned to the highest threshold and minimum support for the frequent phrases mining.

Figure 3. (a) Summary unigrams example from the voyant-tool.org , (b) top 50 topic phrases result from phrase mining

A simple run of a sentiment analyzer to this news article comments shows that the comments sentiment is negative.

5.2 Experimental Result from Guardian News dataset

The Guardian newspaper online provides comments facility for each news. Readers must register to post a comment. Comments are lightly moderated and checked for spam or vandalism. The comments data presented in this paper is first used in a previous study by Llewellyn et al. [12]. The comments concern the feedback to an article entitled "*iPad mini features: what tablet users like – and what the analysts say. Data from Nielsen surveying existing tablet owners shows a skew away from price and towards features*" written by Charles Arthur. The comments are reviewing the iPad mini and produced over 2 days from October 12th 2012. There were 161 comments in total.
We used the python modules NLTK and scikit-learn [9] to process the data. K-means clustering is applied to the data by using the TFIDF features (term must appear in 2 or more comments) and LSA as a dimension reduction.

Table 3 shows the list of comments and their obtained cluster label. Figure 4 depicts the 7 clusters in two dimensions.

Table 3. Comments and Clusters from the dataset

	0	cluster_label
0	this would have killed the market 18 month ago...	7
1	yes	5
2	apple need to get over their margin . no they ...	7
3	apple already ha 60 - 70 (depending upon whic...	7
4	this would have killed the market 18 month ago...	7

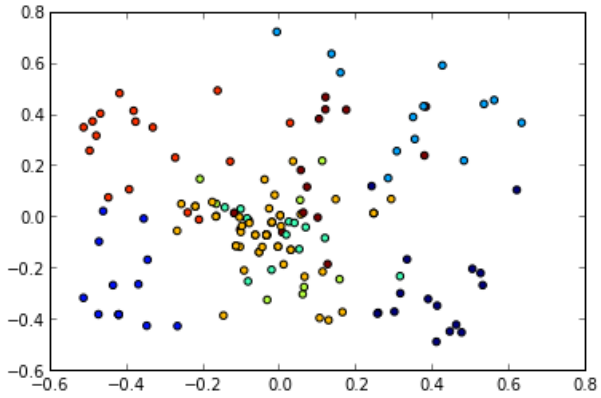

Figure 4. Clustered comments with 7 colors clusters after 2D projection

For this comments data, we conducted another experiment by extracting topics with multi-words (phrases) rather than the classical unigram approach in the LDA modeling. The result is presented as phrases cloud as in Figure 5.

Figure 5. Topics based on phrases (multi-words)

This figure shows the topics more meaningful and readable to humans in comparing to the unigram tag clouds.

An example of the key comment from the topic2 = [apple, nexus, tablet, price, product, market, people, buy] is the following:

"This isn't technology journalism at all. Its celebrity journalism. Apple has done an excellent job of promoting their products in the same way Hollywood flacks do. The general disconnect is that many (not all) Apple customers buy the products as a lifestyle choice. If a great athlete wears a particular brand of shoe, then his/her fans are likely to buy that same brand in emulation. The decision is not made based on the merits of the shoe, but the brand. I don't mind that. Apple is very successful at selling products for a premium because of the branding as hip, cool, and avant-garde. I'm one of the dull techno-drones that Apple loves to position against, so I will probably never own their product. But that doesn't mean they won't be successful. Apple's problem is that shared by all luxury brands: Maintain revenue and growth without diluting the brand. I suspect that their new tablet will not win over new customers, but capture more money from existing Apple devotees."

6 CONCLUSIONS

We conducted experiments for analyzing and visualizing users' comments with topic modeling, comments clustering, topic phrases mining and visualization tool.

Topic phrases are the most suited to the comments dataset. The extracted topic phrases are very easy to interpret and reflect very well to the summary of the comments.

Topics can also be used for content recommendation application such as in [3]. Semantic topics organization enables the user to selectively browse comments on a topic and focus only on those set of comments. This set of comments is then used for different recommendation schemes to the user's interests. A prediction system for user preferences can be developed.

The future work concerns the evaluation of the must read comment recommendation system and topics relatedness over some criteria such as time or semantic relation, commenter id and other comment references within comments such as the ReQuT model (Reader, Quotation, Topic) described in [13].

ACKNOWLEDGMENTS

The authors gracefully acknowledge Assoc. Prof. Aixin Sun at Nanyang Technological University Singapore for providing the Yahoo News comments dataset and the paper proof reading.

REFERENCES

[1] S. Siersdorfer, S. Chelaru, W. Nejdl, J. San Pedro, 2010. How Useful are Your Comments? Analyzing and Predicting YouTube Comments and Comment Ratings. Proc. of WWW2010, Raleigh, North Carolina, USA, pp. 891—900, 2010.

[2] E. Shmueli, A. Kagian, Y. Koren, R. Lempel, 2012. Care to Comment? Recommendations for Commenting on News Stories. Proc. of WWW2012, Lyon, France, pp.429—438, 2012.

[3] V. Jain and E. Galbrun., 2011. Topical Organization of User Comments and Applications to Content Recommendation. Proc. of WWW2013, Rio de Janeiro, Brazil, 2013.

[4] S. Siersdorfer, S. Chelaru, J. San Pedro, I. Sengor Altingovde, W. Nejdl, 2014. Analyzing and Mining Comments and Comment Ratings on the Social Web.In Journal ACM Transactions on the Web (TWEB), volume 8, issues 3, June 2014.

[5] M. Potthast, B. Stein, F. Loose, S. Becker, 2012. Information Retrieval in the Commentsphere. In ACM Transactions on Intelligent Systems and Technology, Vol. 3, No 4, September 2012.

[6] B. Gretarsson, J. O'Donovan, S. Bostandjiev, T. Höllerer, A. Asuncion, D, Newman, P. Smyth 2012. TopicNets: Visual Analysis of Large Text Corpora with Topic Modeling. In ACM Transactions on Intelligent Systems and Technology, Vol. 3, No. 2, February 2012

[7] D. Ramamonjisoa, D. Suzuki, and B. Chakraborty, 2013. Extracting and Visualizing People's Needs and Topic Trends from Users' Comments on Video Streaming Sites or Blog Posts. Proc. of e-Society, Lisbon, Portugal, pp.421–426, 2013.

[8] Elham Khabiri and James Caverlee and Chiao-Fang Hsu. Summarizing User-Contributed Comments. Association for the Advancement of Artificial Intelligence. 2011.

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay. Scikit-learn:Machine Learning in Python. Journal of Machine Learning Research 12:2825-2830, 2011.

[10] A. El-Kishky, Y. Song, C. Wang, C.R. Voss, J. Han, 2015. Scalable Topical Phrase Mining from Text Corpora, in the Proc. Of Very Large Databases (VLDB) Endowment, vol.8, pp.305-316, 2014-2015.

[11] Z. Ma, A. Sun, Q. Yuan, G. Cong, 2012. Topic-driven reader comments summarization. In Proc. Of CIKM'12, 2012.

[12] C. Llewellyn, C. Grover, J. Oberlander, 2014. Summarizing Newspaper Comments. In Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media, pp 599-602, March 2014.

[13] M. Hu, A. Sun, E. P. Lim. Comments-Oriented Blog Summarization by Sentence Extraction. In Proc. Of CIKM 2007. Lispoa, Portugal, pp.901-904.