Topic Modeling on Users's Comments

David Ramamonjisoa
Faculty of Software and Information Science
Iwate Prefectural University, IPU
Takizawa, Japan

***Abstract*—User-contributed comments are increasing exponentially on the Social Web, they are found widely in the social media sites (internet discussion fora or news providers). This paper describes an experiment for topic modeling on users' comments in social media. The future application of the method is discussed.**

Keywords: topic modeling; users's comments;

I. 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].

A social knowledge task gathers and records the concerns of people and problems following an event. These concerns may consist of the general trends or needs of individuals or groups participating in video comments or blog articles. This is because users' comments can produce a new consensus among users and this consensus has a bearing on users' thoughts. Thanks to text mining techniques, topic trends or users' needs can be analyzed and summarized autonomously.

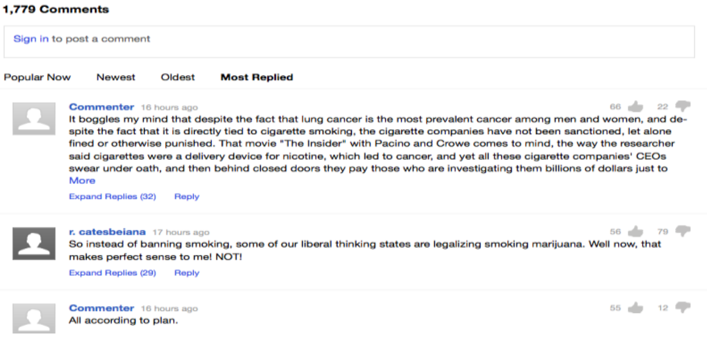
Comments have also spams and trolls that pollute the social network. They can be a cyberbullying and threat to the real society [3].

This paper describes an experiment on users's comments topic modeling.

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 by tackling the spamming and trolling problems.

II. What are comments?

According to the news on February 5th 2014 on Yahoo most popular news, the figure 1 comments example has a headline "***Cancer rates will surge 57 percent in next 20 years, report says.***" The report was from the World Health Organization (WHO).



1,779 Comments

Commenter 16 hours ago

It boggles my mind that despite the fact that lung cancer is the most prevalent cancer among men and women, and despite the fact that it is directly tied to cigarette smoking, the cigarette companies have not been sanctioned, let alone fined or otherwise punished. That movie "The Insider" with Pacino and Crowe comes to mind, the way the researcher said cigarettes were a delivery device for nicotine, which led to cancer, and yet all these cigarette companies' CEOs swear under oath, and then behind closed doors they pay those who are investigating them billions of dollars just to

r. catesbeiana 17 hours ago

So instead of banning smoking, some of our liberal thinking states are legalizing smoking marijuana. Well now, that makes perfect sense to me! NOT!

Commenter 16 hours ago

All according to plan.

Figure 1. Comments example from Yahoo News

At the time we took this snapshot, there were 1779 comments. A comment has three attributes such as the userID, the time stamp and the content. The commentator uses his/her userID to post a comment but his real identity is not revealed.

III. Approaches

The data model for the experiments is described as follows:

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:

$$D = \{c_i\}$$

$$where \quad c_i = (docID,\ time\_stamp,\ title_i,\ content_i)$$

A natural language processing (NLP) task was processed to extract important keywords such as nouns or adjectives from the $content_i$ of each c_i. A bag-of-word model was constructed by attaching a weight to the extracted words. A weight may be just the frequency of a term, tf results. The content of the document is then a set of tuple keywords and weights as follows as in information retrieval (IR) techniques:

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

$$n \text{ number\_of\_keywords\_in\_the\_content},$$

$$w_{ij} > \tau(\text{threshold})$$

A document collection is therefore a table where rows consist of the weights of each keyword in each document and 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:

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

$$m : \mathrm{number\_of\_documents\_in\_the\_collection}$$

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

IV. 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 procedure 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.

A. *Latent Dirichlet Allocation (LDA) basics*

Topic modeling based on Latent Dirichlet Allocation (LDA) is a generative model that can be used to identify the underlying topics that documents are generated from. The document weights come from a Dirichlet distribution (a distribution that produces other distributions) and those weights are responsible for allocating the words of the document for the topics of the collection. The document weights are hidden variables, also known as latent variables [4].

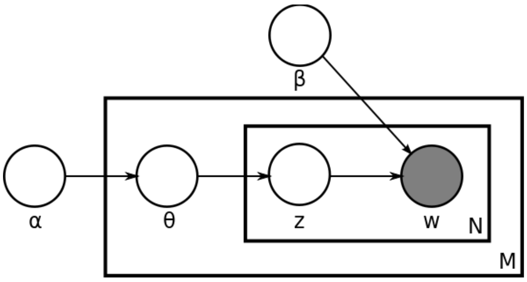


Figure 2. LDA as a graphical model

The figure 2 shows a plate notation representing the LDA model.

With plate notation, the dependencies among the many variables can be captured concisely. The boxes are "plates" representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document. M denotes the number of documents, N the number of words in a document. Thus:

α is the parameter of the Dirichlet prior on the per-document topic distributions,

β is the parameter of the Dirichlet prior on the per-topic word distribution,

θ_i is the topic distribution for document i,

Φ_k is the word distribution for topic k,

z_{ij} is the topic for the jth word in document i, and

w_{ij} is the specific word.

The central inference problem for LDA is determining the posterior distribution of the latent variables given the document:

$$p(\theta, \beta, \mathbf{z}|\mathbf{w}) = \frac{p(\theta, \mathbf{z}, \mathbf{w}, \beta)}{p(\mathbf{w})}$$

To solve this equation, Gibbs sampling or other approximation is applied.

B. *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 $T = \{t_1, t_2, \ldots, t_m\}$. Any 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 topic modeling LDA based topic clustering approach is an assignment from each comment to a cluster [5].

C. *Non-negative Matrix Factorization (NMF) for topic modeling*

Another method used for solving the topic modeling problem is the 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 I. 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 in Table I, 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 then 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

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

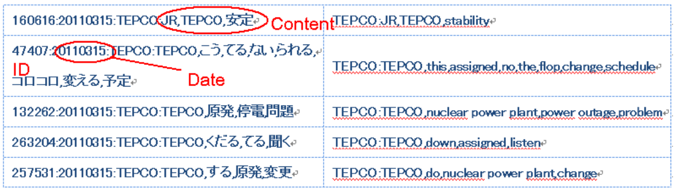
We used a toolkit implemented in Python language named scikit-learn to solve the optimization problem based on projected gradient methods [6].

V. Experiments

Our experiments were based two datasets. The first one is a video streaming comments dataset. This dataset was obtained from previous experiment in [7]. The second one is the Yahoo most read and commented news dataset from the paper [8].

A. *Dataset : Tokyo Electric Power Plant accident interview comments*

The first dataset is a collection of users' comments (see example in the figure 3) obtained from NICOVIDEO during live press conferences of the Tokyo Electric Power Company (TEPCO) between March 15th and March 20th, 2011. On average, there were 2400 comments a day and 14450 in total. Each comment is limited to 128 Japanese characters.



160616:20110315:TEPCO:JR,TEPCO,安定	TEPCO:JR,TEPCO,stability
47407:20110315:TEPCO:TEPCO,こう,てる,ない,られる, コロコロ,変える,予定	TEPCO:TEPCO,this,assigned,no,the,flop,change,schedule
132262:20110315:TEPCO:TEPCO,原発,停電問題	TEPCO:TEPCO,nuclear power plant,power outage,problem
263204:20110315:TEPCO:TEPCO,くだる,てる,聞く	TEPCO:TEPCO,down,assigned,listen
257531:20110315:TEPCO:TEPCO,する,原発,変更	TEPCO:TEPCO,do,nuclear power plant,change

Figure 3. Example of comments during the live Video Streaming

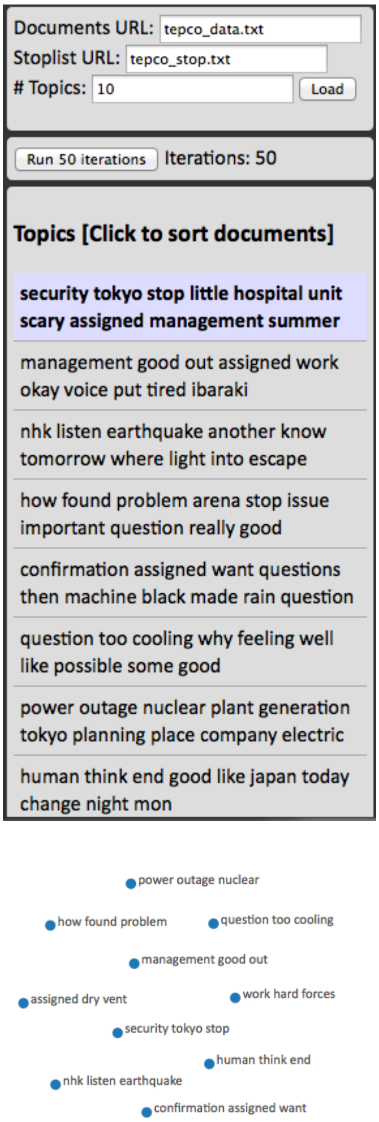
B. *Experimental Result from dataset TEPCO*

In this section, we present the results of our first experiment. The topic modeling LDA based topic clustering is implemented with the jsLDA (javascript LDA) developed by Dr. David Mimno [9]. The dataset is converted to the MALLET format and then put to the modeling tool jsLDA. There are 15000 comments in this experiment during the same period. Results of the modeling are described with the figure 4. Figure 4 shows the topics and documents interface after 50 iterations. It also depicts the topics clusters and topics correlation graphs.

C. *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 the 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 III. This preprocessing is very crucial because all Yahoo News articles HTML files don't have the same format. Some have javascripts 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 3 topics, 10 keywords for each topic and limit the maximum keywords to 10000 in comments corpus. We obtain the result within few seconds after the execution.



Documents URL: tepco_data.txt
Stoplist URL: tepco_stop.txt
# Topics: 10

Iterations: 50

Topics [Click to sort documents]

- security tokyo stop little hospital unit scary assigned management summer
- management good out assigned work okay voice put tired ibaraki
- nhk listen earthquake another know tomorrow where light into escape
- how found problem arena stop issue important question really good
- confirmation assigned want questions then machine black made rain question
- question too cooling why feeling well like possible some good
- power outage nuclear plant generation tokyo planning place company electric
- human think end good like japan today change night mon

Figure 4. Output of the topic modeling where topics are a set of terms and their correlation graphs according to a correlation threshold

VI. Conclusion and future work

We conducted experiments for modeling topics on users's comments. Next step is to extend the topic modeling by including the hierarchy and time series in the LDA or using NMF to detect keywords unrelated to the news article.

The ultimate goal of this research is to find some methods to analyze user comments and, like humans do, identify spams and/or trolls within those comments in order to ignore them or process them. Methods such as Bayesian network classifier and Support Vector Machine are already used for spam detection in e-mails [3][10][11] but they are not yet adapted to the users' comments. When topics on comments are defined, it may be possible to define an approach for off-topic detection. The existing approach is done manually by allowing users to evaluate comments and ranking them to obtain the most relevant and disregard the least popular ones.

Acknowledgment

We thank Dr Basabi Chakraborty for her valuable advices and providing us the TEPCO dataset. The student who translated the dataset in English is Tatsuya Suzuki graduated in 2013 and the other student who is working on Yahoo News dataset is Riki Murakami from the same university.

References

[1] Siersdorfer, S. et al., 2010. How Useful are Your Comments? Analyzing and Predicting YouTube Comments and Comment Ratings. Proc. of WWW2010, Raleigh, North Carolina, USA, 2010. pp. 891—900.

[2] Shmueli, E. et al., 2012. Care to Comment? Recommendations for Commenting on News Stories. Proc. of WWW2012, Lyon, France, 2012. pp.429—438.

[3] P. Galán-García, et al., 2013. Supervised Machine Learning for the Detection of Troll Profiles in Twitter Social Network: Application to a Real Case of Cyberbullying. in Proc of the 6th International Conference on Computational Intelligence in Security for Information Systems (CISIS). Salamanca, Advances in Intelligent Systems and Computing Volume 239, 2014. Spain. pp 419-428

[4] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. In Journal of Machine Learning Research, 3:993–1022, January 2003.

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

[6] Fabian Pedregosa et al.: Scikit-learn:Machine Learning in Python. Journal of Machine Learning Research 12:2825-2830, 2011

[7] Ramamonjisoa D. et al., 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.

[8] Zongyang Ma et al.: Topic-driven reader comments summarization. In Proc. Of CIKM'12, 2012.

[9] Mimno D., 2013: javascript LDA,
http://mimno.infosci.cornell.edu/jsLDA/index.html

[10] Gilad Mishne, et al., 2005. Blocking Blog Spam with Language Model Disagreement. Journal: ACM Transactions on Multimedia Computing, Communications, and Applications - TOMCCAP , pp. 1-6, 2005

[11] Igor Santos, et al., 2012. Enhanced Topic-based Vector Space Model for Semantics-aware Spam Filtering. Journal: Expert Systems With Applications - ESWA , vol. 39, no. 1, 2012.