International Journal of Digital Information and Wireless Communications (IJDIWC) 5(3): 185-193
The Society of Digital Information and Wireless Communications, 2015 (ISSN: 2225-658X)

A Framework for Comments Analysis and Visualization

D. Ramamonjisoa, R. Murakami, B. Chakraborty
Faculty of Software and Information Science
Iwate Prefectural University, IPU
Takizawa, Japan
david@iwate-pu.ac.jp,g231n032@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. Individual users are flooded with thousands of comments in some discussion webpages, so the reading task became costly in time and processing reducing their daily life productivity. Commenting activity can be addictive. The online social activity can bring an unpleasant experience for the comments readers. In this paper, we propose a framework to support readers to analyze and visualize opinion and topics on the comments. We describe experiments on product reviews comments and news comments.

KEYWORDS

Comments analysis; comments visualization; architecture; sentiment analysis; text 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 presents a framework for supporting comments readers on their task by providing users' comments analysis, and visualization of opinions and topics.

In this paper, we first introduce the general framework and then detail the methods used. 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 April 28th 2015 concerning about the riot, looting in Baltimore and the mysterious death of Freddie Gray (http://yhoo.it/1Pa6ckY). Within a few hours after the publication of this article online, 38000 comments are already posted to the Yahoo news site. For someone who wants to grasp the 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 2200 thumb-ups and has the username “Ms Lazy Thump”, 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.

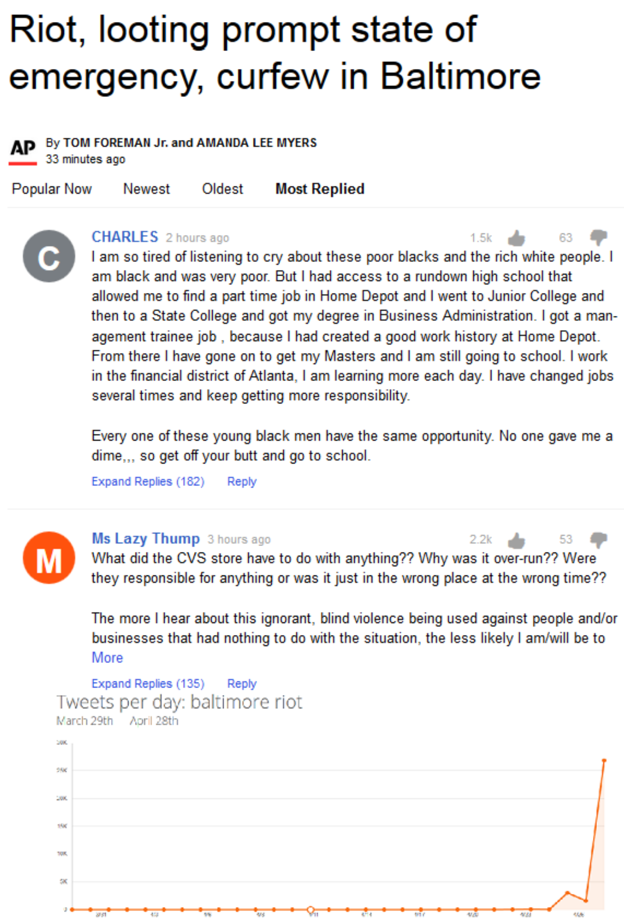

Figure 1. Comments example from Yahoo News

A survey of research related to comment analysis shows that the studied comments are mainly on reviews of products or movies [5]. They are generally related to the marketing research. For the sentiment analysis, research papers have focused on finance related comments and health related comments [7], [8]. Those comments are valuable for the companies because they can move the market up or down. Those comments may also contain important information for decision-making.

In any case, a comment is quantified according to its quality. A good quality comment is the one with good writing style, without extreme sentiments, and respecting the posting rule policies. On some news websites, moderator tasks are carried out by a filtering system including the moderator rules and machine learning models [9].

In general, readers want to find in comments something new not mentioned in the article. A complementary information, a joke, a sarcasm or gossip is an example. As stated in the figure 1, M’s comment is very liked and most replied because it has new information not mentioned in the main article.

3 PROBLEM AND COMMENT SCORING FORMALIZATION

Is it possible to overview and understand the discussion within comments without reading them all and eliminating noise by using a system with a sufficient knowledge, machine learning, and information retrieval programs.

Given an URL of an article and comments, a system should output a key comments, topics and sentiment score.

3.1 Simple Structure

Each comment is considered as a set of sentences including emoji and url address. A preprocessing program is used to filter the comment in order to eliminate any reply within it. Only sentences written by the user is retained. In this case, there is no redundancy of the comment except when the user does it voluntarily.

The data model is described as follows (and more details can be found in [7]):
Users’ comments or blog posts are designated as comment collections. The model of the

comments (as each comment is a specific short document) collection is described below:
$C = \{c_i\}$ where
$c_i = (commentID, \mathrm{time}, \mathrm{title}_i, content_i)$
The model is transformed into a feature-based vector representation. The feature is a function $f(c_i)$ such as term frequency, inverse document frequency, or tfidf (the product of the previous two features). Part-of-speech features, statistical properties can also be included. The set of features used for model is defined as follows:

$$F = \{f_i(c_j)\} \quad (1)$$

These features can be used for the topic modeling, key comments extraction or the machine learning algorithms to train a sentiment analysis and predict new comment sentiment.
The basic data matrix for the topic extraction is based on tfidf feature. A bag-of-word model was then constructed by attaching a weight to each extracted word. The content of the document is then a set of tuple keywords and weights as used in many information retrieval (IR) tasks:

$$content_i = \{(k_{ij}, w_{ij})\} \quad j \in [1..n], \quad (2)$$

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:

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

For the refinement, a natural language processing (NLP) task conducted to extract important keywords such as nouns or adjectives from the *content*i of each *c*i.

3.2 Readers Graph Structure

The comment board can have a discussion threading or record the reply-to structure. To deal with this structure, we use a model based on the ReQuT (Reader, Quotation, and Topic) model in the paper [11]. The comment author is set as a reader.
Given an article and the set of comments $C = \{c_I\}$ associated with it, a directed weighted graph can be constructed such as $G_R = (V_R, E_R, W_R)$. A comment c_i has a reader $r_a \epsilon V_R$. When a reader r_b mentions a reader r_a in one of his/her comments, an edge $e_R(r_b, r_a) \epsilon E_R$ is created. The edge weight noted as $W_R(r_b, r_a)$ is the ratio between the number times r_b mention r_a against all times r_b mention other readers (including r_a). The reader authority is obtained by using the PageRank algorithm in the following equation, where R is the set of comment readers (commenters) and d = 0.85 a damping factor.

$$PR(r_a) = \frac{1-d}{|R|} + d * \sum_{r_b} W_R(r_b, r_a) * PR(r_b)$$

$$RM(w_k) = \sum\nolimits_{c_j \leftarrow r_a} tf(w_k, c_j) * PR(r_a) \quad (3)$$

$RM(w_k)$ (equation 3) is the reader measure of a word w_k and $tf(w_k, c_j)$ is the term frequency of w_k in c_j authored by r_a ($c_j \leftarrow r_a$).

The next section describes the framework.

4 COMMENTS READING SUPPORT FRAMEWORK

We design and develop a prototype system to deal with the comments reading task. The general architecture of the system is depicted in the figure 2.
The item article and comments are the input of the system, then the system process those data to provide key comments, topic words and topic phrases, the overall sentiment of those comments. The system has several modules including the topic extractor, comment ranking program, and sentiment analyzer.

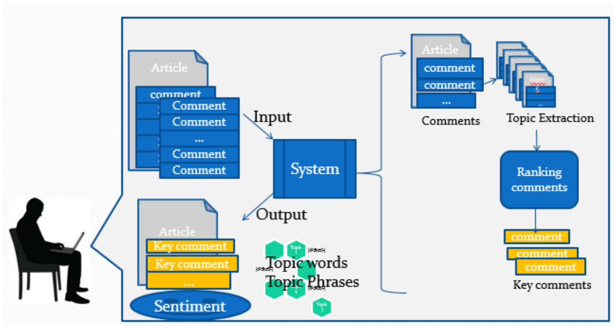

Figure 2. Comments reading support framework

5 TOPICS EXTRACTION

The first analysis on the comments concerns the extraction of topics. Topics are the thematic summary of the comment collection. In other words, it answers the question what themes are those comments discussing.
The topic modeling is used to extract T topics out of the comments collection. That is, we have a set of comment "documents" $C = \{c_1, c_2, \dots c_n\}$ and a number of topics $T = \{t_1, t_2, \dots, t_m\}$. 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].

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.
We used a toolkit implemented in Python language named *scikit-learn* to solve the optimization problem based on projected gradient methods [13].
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. [14] 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.

6 KEY COMMENTS EXTRACTION

6.1 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.

6.2 Key Comments selection with the Reader authority

The reader measure defined in the section 3 is used to select the key comments as the top 10 highest score.
The comment score according to the reader authority is as follows:

$$score(c_i) = \sum_{w_k \epsilon c_i} RM(w_k) \qquad (4)$$

7 SENTIMENT ANALYSIS

The method used for the sentiment analysis is based on document level analysis. Each comment is assumed as a document and the document is processed by using features for its subjectivity and polarity (positive, neutreal or negative). A supervised learning is realized to classify each comment into its polarity. We use Naïve Bayes classification model and train the classifier on the product reviews comments. We build the training set by querying an online sentiment analyzer and check those results by ourselves.
In the figure 3, an example of the training and test for the classifier on the iPad mini 3 reviews comments data. As the number of comments increases, the accuracy of the classifier is not improved.

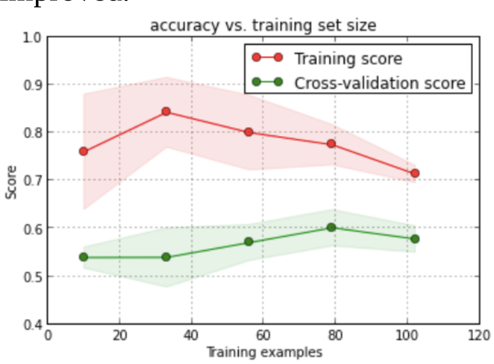

Figure 3. iPadmini 3 comments sentiments prediction accuracy

8 EXPERIMENTS

Our experiments were based on Yahoo news comments, Guardian News comments and CNET product reviews comments datasets. The Yahoo most read and commented news dataset was obtained from the authors of the paper [10]. The recent news on the social demonstrations in the US which turned into violence is chosen to our case study. The press release of iPadmini on Guardian News, iPadmini3, iPhone 6 Plus and iWatch apple products on CNET are also studied.

8.1 Most commented news on Yahoo News dataset

In our previous study, we tested the topic modeling and key comments extraction with the Yahoo News during 2012 [15]. We continue to analyze the Yahoo News for the current events. In this paper, we show the analysis of the riots in Baltimore (US) in April 2015. 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. Yahoo News article has a unique content identification which can be used to query a Yahoo API to obtain all comments.
When the HTML data are processed, user can 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. We obtain the result within a few seconds.
We present topics computed from the riot news article. There are more than 20,000 commenters and 36,800 comments. The news title is as "*Riots in Baltimore over man's death in police custody.*"
The news reports the riot and the mystery death of the black person dealing drug. Table 2 describes the topic list results from phrases LDA.

Table 2. phrases LDA results

#	Topics phrases LDA			
	police officers	150	innocent people	55
	black community	148	peaceful protest	51
	black people	139	Mr Gray 50	
	Freddie Gray	111	break the law	50
	white people	98	Possession of a controlled	
	riot and loot	84	dangerous substance with	

African Americans	79	intent to distribute	47
law enforcement	76	burning and looting	47
unlawful possession of a		Martin Luther King	47
controlled dangerous		hard work	46
substance	74	police force	45
looting and burning	71	police brutality	41
black white	66	committing crimes	41
National Guard	61	United States	37
destroy property	61	business owners	37
black man	58	Middle Working	
don t care	55	Class Americans	36

The phrases LDA results show the most frequent phrases in the comments collection. The parameters in the setting are tuned to the highest threshold and minimum support for the frequent phrases mining.

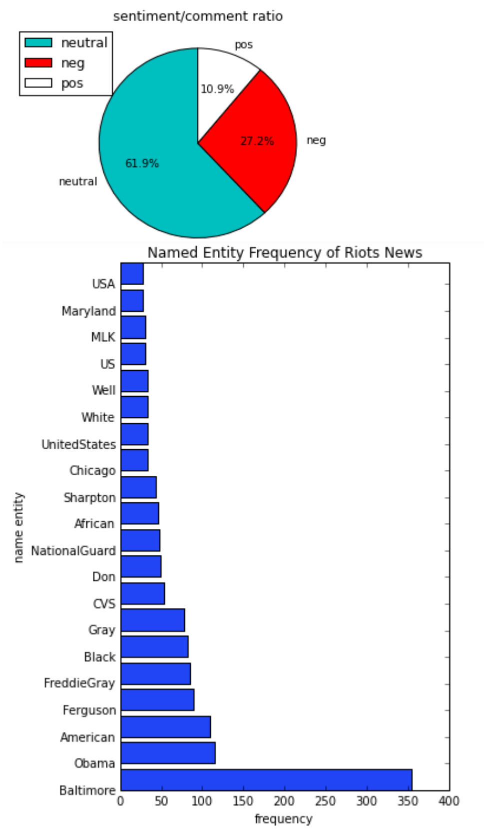

Figure 4. (a) sentiment proportion on Riots , (b) Most frequent name entities on Riots in Baltimore

8.2 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. [17]. 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 (http://bit.ly/1zTTXSU). The comments are reviewing the iPad mini and produced over 2 days from October 24th 2012. There were 161 comments in total.

We used the python modules NLTK and scikit-learn [13] 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.

For this comments data, we extracting topics with multi-words (phrases) LDA modeling. The result is presented as phrases cloud as in Figure 5.

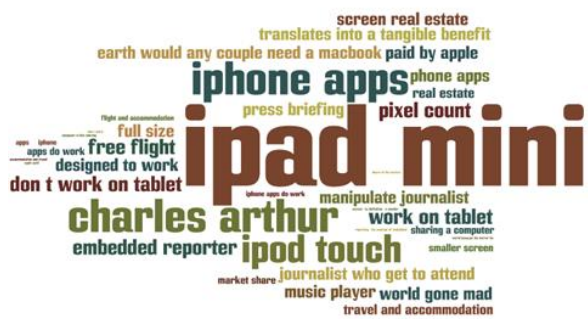

Figure 5. Topics based on phrases (multi-words)

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."

The highest score on the reader measure (see section 6.2) which represents the key comment extraction is the following comment.

Author: super8, Content:

"Low resolution, low PPI, low spec processor - all this can be yours for only £100 MORE than a much higher spec Nexus 7 or Amazon Kindle Fire HD.

You're falling into the trap of just looking at specs. Most users don't care about specs unless it translates into a tangible benefit (or problem).

The iPad has a 35% bigger screen than a Nexus 7, but it's lighter. On the other hand it is somewhat wider. Whether this is better or worse is nothing to do with specs - it depends on your personal preferences. Personally I would prefer the extra 35%, but perhaps others wouldn't. For me, the 7.9" spec is better than the 7" spec, but that's me.

The Nexus 7 has a 230ppi vs 170ppi for the iPad Mini. This is a partly a result of having a smaller screen. By definition, a smaller screen with a given resolution will have a higher PPI. Is this difference visible to most users given that many other aspects such as saturation, colour fidelity, black point also impact the quality of the screen? Had Apple made the iPad Mini 15% smaller they could have pushed the PPI up at the expense of screen size.

The iPad Mini is 1024 x 768 and this is something Apple won't change (unless they go retina with 2048 by 1536) because it means that all iPad apps will work without needing to be changed. This is a big deal for developers. So within that context Apple make a choice. Higher DPI or bigger screen. They clearly think a 7.9" screen is a better tradeoff.

As for the processor, does the Nexus 7 processor translate into a noticeable benefit? Is the Nexus 7 smoother at scrolling? How does that processor affect battery life? Users judge a processor not by some number but by how it affects the actual performance of the object in their hand."

An extract of the most popular comments by the readers is as the following.

"Low resolution, low PPI, low spec processor - all this can be yours for only £100 MORE than a much higher spec Nexus 7 or Amazon Kindle Fire HD. Bargain! Don't worry, it's very good." **104 Likes**

This specific comment has the following features.

- Sentiment (polarity=0.237 [-1,1], subjectivity=0.526 [0,1]) => positive sentiment and subjective
- Online Sentiment Analyzer: negative
- Sarcastic comment : very difficult to detect and judge
- Readers like such a kind of comment on product review because it provides more information on the product by comparing it to the apple's competitor product which has better specification and cheaper.

8.3 Experimental Result from CNET reviews

In this section, we present the experiment on the apple products reviews comments on CNET.com. The products we studied are iPadmini 3 (http://cnet.co/1ougkZC), iPhone 6 plus (http://cnet.co/1Bq33mP), and iWatch (http://cnet.co/1qBUBh6).

figure 5a

figure 5b

Figure 5. (a) Five topics on iPad mini 3 , (b) Five topic phrases result on the iPhone 6 Plus

A simple run of a sentiment analyzer to the iphone 6 plus comments shows that the comments sentiment is slightly positive (see figure 6b).

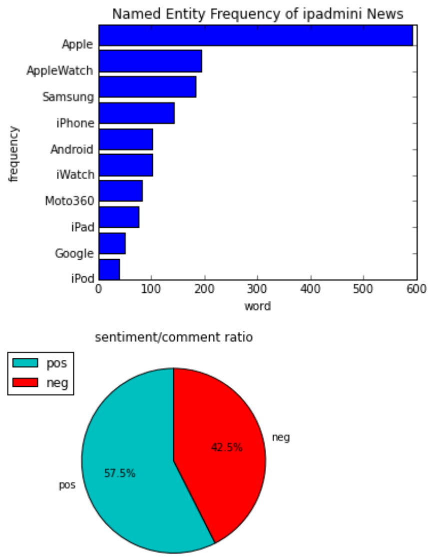

Figure 6. (a) Most frequent name entities on iWatch, (b) sentiment proportion on the iPhone 6 Plus

9 CONCLUSIONS

We conducted experiments for analyzing and visualizing users' comments with a comments analysis framework by focusing on topic extraction based on topic phrases mining, key comments extraction based on two methods (comments and topic similarity measure and reader authority measure), and finally sentiment analysis and visualization.

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.

In all aspect of comment analysis, readers like surprise or new information from comments complementing the main article. The brevity, writing style and insightful are also important to the comment quality.

Social news article comments and product reviews comments are different. The themes discussed in social news comments are for everyone, however the product reviews comments themes are more experts oriented discussions using technical terms. The number of posted comments on the news article is huge in comparing to the product reviews ones.

The extension of this research is to use the framework for content recommendation application such as in [3], [7], [16]. 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.

There are also future researches on the visualization techniques allowing user to navigate the topics and facilitating the discussions described in interactive visual analysis tool ForAVis [18], interative topic modeling for conversations ConVisIT [19], conversation modeling [20], data driven approach to storytelling [21].

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. In: 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. In: Proc. of WWW2013, Rio de Janeiro, Brazil. pp.61--62 (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, No 17 (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, Issue 4, No 68 (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, Issue 2, No 23, (2012)

[7] R. Feldman. Techniques and applications for sentiment analysis. In: Communications of the ACM, vol.56, issue 4, pp.82--89 (2013).

[8] Bo Pang and Lillian Lee. Opinion Mining and Sentiment Analysis. In Foundations and Trends in Information retrieval. Vol. 2, No1-2, pp.1--135, 2008.

[9] Maggie Xiong. Sentiment Analysis for Comment Moderation.In the Sentiment Symposium SS14 (sentimentsymposium.com), (2014).

[10] 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).

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

[12] Elham Khabiri and James Caverlee and Chiao-Fang Hsu. Summarizing User-Contributed Comments. Association for the Advancement of Artificial Intelligence (AAAI), pp.534--537, (2011).

[13] 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, É. Duchesnay. Scikit-learn:Machine Learning in Python. Journal of Machine Learning Research 12:2825--2830 (2011).

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

[15] David Ramamonjisoa, Riki Murakami, Basabi Chakraborty. Comments Analysis and Visualization Based on Topic Modeling and Topic Phrase Mining. In Proc. EBW 2015, ESG Paris, France (2015).

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

[17] Z. Ma, A. Sun, Q. Yuan, G. Cong. Topic-driven reader comments summarization. In: Proc. Of CIKM'12 pp.265--264 (2012).

[18] Franz Wanner, Thomas Ramm, Daniel A. Keim. ForAVis: explorative user forum analysis. In: Proceeding WIMS '11 the International Conference on Web Intelligence, Mining and Semantics, Article No. 14 (2011)

[19] E. Hoque and G. Carenini. ConVisIT: Interactive Topic Modeling for Exploring Asynchronous Online Conversations. In: IUI '15 Proceedings of the 20th International Conference on Intelligent User Interfaces. pp 169-180 (2015).

[20] C. Wang, M. Ye, and B. A. Huberman. From user comments to on-line conversations. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. pp 244-252 (2012).

[21] Jessica Hullman, Nicholas Diakopoulos, Elaheh Momeni, Eytan Adar. Content, Context, and Critique: Commenting on a Data Visualization Blog. In: Proceeding of CSCW '15 the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing. pp. 1170-1175 (2015)