

Discovering Hidden Interests from Twitter for Multidimensional Analysis

Dongjin Yu, Jingchao Sun, Yiyu Wu, Zhiyong Ni, Youhuizi Li

School of Computer Science and Technology
Hangzhou Dianzi University
Hangzhou, China

E-mail: yudj@hdu.edu.cn, wyygoup@gmail.com

Abstract—With the popularity of social networks, Twitter has become one of the dominant providers of massive quantities of information. Exploring the distributions and correlations from Twitter data helps accurate personalized recommendations. Online Analytical Processing, or OLAP, provides an intuitive form that is suitable for exploring Twitter data. Unfortunately, the traditional OLAP approaches can only deal with structured data, not unstructured textual data like tweets. The key to applying OLAP to Twitter data is to mine and build a dimension hierarchy of tweeter interests. However, the current methods can extract tweeter interests from Twitter data on a single level, but fail to obtain a hierarchy of tweeter interests with different granularities. To address this problem, we propose a LDA-based model, called MS-LDA, which combines tweeters' social relationships and tweets to extract and build the tweeters' interest dimension hierarchy. Such a dimension hierarchy can be further employed to apply OLAP techniques to Twitter data. In addition, we employ Word2vec to obtain the linguistic similarity of words in tweets, to improve its effectiveness. The extensive experiments demonstrate that our method can effectively extract the dimension hierarchy of tweeters' interests for multidimensional analysis.

Keywords: Multidimensional Analysis, Twitters, Tweets, Interests, OLAP, LDA, Tweeters.

I. Introduction

Twitter is an online social networking service that enables tweeters to send and read short messages called "tweets". Technically, Twitter data can be divided into two parts, i.e., structured data such as "id" and "location", and unstructured data which include text messages, short links and so on. In addition, tweeters have some social behaviours, such as Follow (paying attention to others), Mention ("@" others) and Retweet (forwarding others' tweets).

Online Analytical Processing, or OLAP, enables users to

analyse multidimensional data interactively from multiple perspectives using operations such as roll-up, drill-down, slicing and dicing [1]. With the explosive growth of Twitter, it has become increasingly necessary to introduce the technology of OLAP to analyse complex Twitter data in an interactive manner. Unfortunately, although OLAP technology provides an intuitive inquiry form that is consistent with human custom, it can only handle structured data, and fails to deal with scenarios related to unstructured text data such as tweets. Therefore, the key to applying OLAP to Twitter data is the question of how to identify dimensions from Twitter data. Among these, the interest dimension attracts the most attention from researchers.

Although some progress has been made in the area of online analysis of Twitter data, there are still some problems to be solved. Many state-of-the-art approaches can only identify a single dimension from tweets instead of the whole dimension hierarchy. Although some of the approaches are able to identify dimension hierarchies, this usually depends on the efforts of domain experts. The main issue this paper tries to resolve can be described as follows: *how to extract the relevant information from the unstructured Twitter data to construct dimension hierarchies of tweeters' interests to achieve effective online multidimensional analysis.*

In this paper, we propose a multi-layered semantic LDA (MS-LDA) model to achieve multidimensional analysis of Twitter data. Firstly, we acquire tweets and the social relationship among tweeters through the REST APIs provided by Twitter. Secondly, we utilize data pre-processing techniques, such as ignoring stop words and part of speech analysis, to remove the irrelevant words in tweets. Finally, we extract the dimension hierarchy of tweeters' interests based on the probability distribution of the various interests and sub-interests revealed in the tweets. To summarize, the multi-layered semantic LDA (MS-LDA) proposed in this paper can not only extract tweeters' interests, but also dig out more fine-grained sub-interests. The interests and their sub-interests thus constitute the dimensional hierarchy of interests and the final data cube together with the

other structured dimensions.

The main contributions of this paper are as follows. 1) We present a new model called MS-LDA for mining the dimension hierarchy of a tweeter’s interests from unstructured massive Twitter data, which considers not only tweets but also the social relations among tweeters. 2) We take advantage of Word2vec, a two-layer neural network model, to obtain the linguistic similarity of words to improve the effectiveness of identifying the interest dimensions. 3) We perform the multidimensional analysis of Twitter data which demonstrates the effectiveness of MS-LDA.

The rest of the paper is organized as follows. After Section 2 introduces related work, Section 3 describes the approach in detail, focusing on the interest mining and the construction of interest dimension hierarchies. In Section 4, we show the overall effect on real Twitter data, and compare our approach with others. Finally, Section 5 concludes the paper and outlines future work.

II. Related Work

During the past decade, with the increased amount of text data, such as web pages, tweets and web blogs, the question of how to apply the traditional OLAP technique to the unstructured text for business intelligence has attracted much attention from many researchers.

Numerous studies have been conducted to explore the social networks such as Twitter. Siswanto et al. [2] use the supervised learning-based classification to determine the user’s interests based on the bio and a collection of tweets. Lim et al. [3] present a framework for classifying the relative interests of Twitter users using information from Wikipedia. Pu et al. [4] present Wiki-LDA to mine user’s interests in Twitter. Xu et al. [5] propose a framework to discover users’ interests by introducing a modified author-topic model called the twitter-user model. Zhao et al. [6] introduce a Twitter-LDA model which is specifically used to extract topics from short tweet data.

Meanwhile, many researchers concentrate on extending the traditional multidimensional data model to support unstructured text such as Twitter data. For example, Maha et al. [7], propose a generic multidimensional model dedicated to the OLAP of tweets whereas N. Rehman et al. [8] develop a system for warehousing streams from Twitter. These authors also attempt to extend the established OLAP technology to enable multidimensional analysis of social media data by integrating text and opinion-mining methods into the data warehousing system [9].

In contrast to previous work, the Multi-layered Semantic LDA (MS-LDA) model proposed in this paper can not only extract tweeters’ interests but also dig out more fine-grained sub-interests. The interests and the corresponding sub-interests thus constitute the dimensional hierarchy and the final data cube for multi-dimensional analysis.

III. Mining Interests

This section discusses how to construct MS-LDA to mine tweeters’ interests from Twitter data, and how to construct interest dimension hierarchies based on the mined interests.

A. Overview of MS-LDA

LDA is a document topic generation model, which denotes that each word in the document is obtained through a particular process, that is, choosing a topic with a certain probability, and then choosing a word from this topic with a certain probability [10]. Unfortunately, due to the similarities existing between some topics, LDA cannot distinguish between some fine-grained ones. In addition, LDA does not consider the semantic features of words. In a similar way to LDA, the words and documents in MS-LDA are associated with a potential topic. However, using MS-LDA it is possible to identify multi-level topics hidden in large-scale data such as Twitter, which LDA fails to do. Moreover, MS-LDA allows the semantic similarity between words and interests to affect the process of model generation, in order to improve the recognition results of sub-interests. Here, we employ the term “interest” instead of the term “topic” to better fit the current application domain.

Tweets are different from the general texts. In addition to text information, the tweets also contain complex social relationships, such as Following, Mentioning and Retweeting [11]. To capture the above social relationships, we build a social list for each tweeter, including the original tweeters of forwarded tweets, the mentioned tweeters and the followed tweeters. We then add this social list as an influence factor into the MS-LDA model.

In summary, MS-LDA integrates tweets and the social relationships among tweeters, and is therefore suitable for multi-level interest mining. The Bayesian network graph of MS-LDA is shown in Figure 1. When MS-LDA generates a tweet, it first chooses an interest from the multinomial distribution θ_m which is generated by Dirichlet distribution with parameter α . Next, it chooses a sub-interest z' according to the degree of social impact and semantic impact. Finally, it chooses a word w from the multinomial distribution $\varphi_{k'}$ which is generated by Dirichlet distribution with parameter β' and Y' . For simplicity, Table 1 summarizes the notations used throughout this paper.

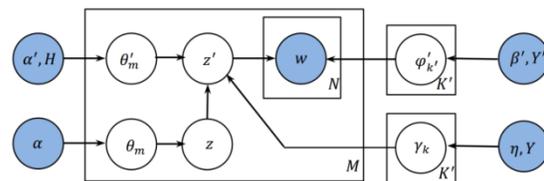


Figure 1: The Bayesian network of MS-LDA

Table 1: Notations used throughout the paper

Notation	Definition
K	Number of interests
M	Number of tweeters
V	Number of words
K'	Number of sub-interests
θ_m	Interest distribution over the tweeter m based on tweets
θ'_m	Sub-interest distribution over the tweeter m based on tweets
$\varphi'_{k'}$	Word distribution over sub-interest k'
γ_k	Sub-interest distribution over interest k
η	Hyper parameter for $\gamma_k, 1 \leq k \leq K$
α	Hyper parameter for $\theta_m, 1 \leq m \leq M$
α'	Hyper parameter for $\theta'_m, 1 \leq m \leq M$
β'	Hyper parameters for $\varphi'_{k'}, 1 \leq k' \leq K'$
w	Word in tweets
$z_{m,v} (z_{-m,v})$	Interest of word v in tweets of tweeter m (indicator before sampling)
$z'_{m,v} (z'_{-m,v})$	Sub-interest of word v in tweets of tweeter m (indicator before sampling)
$n_{m,k'} (n'_{m,\cdot})$	Co-occurrence of tweeter m and sub-interest k' (or all sub-interests)
$n_{k,v} (n_{k,\cdot})$	Co-occurrence of interest k and word v (or all words)
$n_{k,k'} (n'_{k,\cdot})$	Co-occurrence of interest k and sub-interest k' (or all sub-interests)
$n_{k',v} (n_{k',\cdot})$	Co-occurrence of sub-interest k' and word v (or all words)

In other words, for a given tweet, its joint probability distribution of all the words and their interests and sub-interests can be calculated as follows:

$$P(z, z', w | \alpha, \alpha', \beta', \eta, H, Y, Y') = P(z|\alpha)P(z'|z, \alpha', \eta, H, Y)P(w|z', \beta', Y') \quad (1)$$

B. Derivation of MS-LDA

The MS-LDA model can be derived using the Gibbs sampling method. The main steps are as follows:

Firstly, the interest probability distribution $P(z|\alpha)$, the sub-interest probability distribution $P(z'|z, \alpha', \eta, H, Y)$ and the word probability distribution $P(w|z', \beta', Y')$ can be calculated according to the Euler formula, as follows:

$$P(z|\alpha) = \left(\frac{\Gamma(K\alpha)}{\prod_k \Gamma(\alpha)} \right)^T \prod_{m=1}^M \frac{\prod_k \Gamma(n_{m,k} + \alpha)}{\Gamma(n_{m,\cdot} + K\alpha)} \quad (2)$$

$$P(z'|z, \alpha', \eta, H, Y) = \left(\frac{\Gamma(K'\alpha')}{\prod_{k'} \Gamma(\alpha')} \right)^T \prod_{m=1}^M \frac{\prod_{k'} \Gamma(H_{m,k'}(n_{m,k'} + \alpha'))}{\Gamma(H_{m,\cdot}(n'_{m,\cdot} + K'\alpha'))} \times \left(\frac{\Gamma(K'\eta)}{\prod_{k'} \Gamma(\eta)} \right)^T \prod_{k=1}^K \frac{\prod_{k'} \Gamma(Y_{k,k'}(n_{k,k'} + \eta))}{\Gamma(Y_{k,\cdot}(n_{k,\cdot} + K\eta))} \quad (3)$$

$$P(w|z', \beta', Y') = \left(\frac{\Gamma(V\beta')}{\prod_v \Gamma(\beta')} \right)^T \prod_{k'=1}^{K'} \frac{\prod_v \Gamma(Y'_{k',v}(n_{k',v} + \beta'))}{\Gamma(Y'_{k',\cdot}(n_{k',\cdot} + V\beta'))} \quad (4)$$

The interpretation of the $H_{m,k'}$, $Y'_{k',v}$ and $Y_{k,k'}$ in the above equations is as follows.

The social behaviour among tweeters reflects their interest preferences to a certain extent. $H_{m,k'}$, or the social impact, denotes the degree to which the social relationship affects the tweeter m being interested in the sub-interest k' . Let $P_{m,k'}$ be the probability of tweeter m being interested in the sub-interest k' , which can be obtained by Eq. (3), and let $S_m = \{u_1, u_2, u_3, \dots, u_{N_m}\}$ be the social list of tweeter m in which u_j represents the j -th tweeter in the social list S_m and N_m represents the total number of tweeters in the social list S_m . The *social impact* can thus be calculated according to Eq. (5), where $P_{-u_j,k'}$ represents the probability that the tweeter u_j is interested in the k' sub-interest in the previous iteration.

$$H_{m,k'} = \frac{\sum_{j=1}^{N_m} P_{-u_j,k'}}{N_m} \quad (5)$$

We consider that the higher the semantic similarity between words and the sub-interest, the greater the probability that the word belongs to the sub-interest. Let $Y'_{k',v}$ be the degree of *word/sub-interest semantic impact* of the word v belonging to the sub-interest k' . For the computation of word/sub-interest semantic impact, we pick up the most frequent n' words which belong to the sub-interest k' to compose a collection $Q'_{k'} = \{q'_{k',1}, q'_{k',2}, \dots, q'_{k',n'}\}$. Let $R'_{k'} = \{r'_{k',1}, r'_{k',2}, \dots, r'_{k',n'}\}$ be the number-of-the-words collection in which each item gives the number of occurrences of the corresponding word, and $Sim(v, q'_{k',j})$ be the semantic similarity between the word v and $q'_{k',j}$. The *word/sub-interest semantic impact* can thus be obtained from Eq. (6).

$$Y'_{k',v} = \frac{\sum_{j=1}^{n'} (r'_{k',j} \cdot Sim(v, q'_{k',j}))}{\sum_{j=1}^{n'} r'_{k',j}} \quad (6)$$

For example, when calculating the degree of semantic impact of the word ‘‘Yao_Ming’’ (short to v) belonging to the sub-interest P , we first pick up the most frequent top three words which belong to the sub-interest P , i.e., $Q'_P = \{\text{Basketball}, \text{Kobe_Bryant}, \text{NBA}\}$. Suppose $R'_P = \{30, 20, 10\}$ is the number of occurrences of the corresponding words. Thus, the degree of *word/sub-interest semantic impact* of the word ‘‘Yao_Ming’’ belonging to the sub-interest P can be calculated as:

$$Y'_{P,v} = \frac{(30 * 0.35 + 20 * 0.63 + 10 * 0.42)}{(30 + 20 + 10)} = 0.455.$$

To calculate $Sim(v, q'_{k',j})$, we utilize Word2vec, an efficient tool proposed by Google to characterize a word as a vector [12]. According to Word2vec, the similarity in the vector space can be used to represent the similarity of text semantic. For example, we can obtain the vectors of “Basketball”, “Football”, “Yao_Ming” and “Kobe_Bryant”. The similarity of “Yao_Ming” and “Basketball” calculated by Word2vec is higher than that of “Yao_Ming” and “Football”, which is consistent with the real case.

For the words w_1 and w_2 , we firstly obtain their vectors V_1 and V_2 respectively by Word2vec. Then, their semantic similarity can be calculated as follows:

$$Sim(w_1, w_2) = \frac{\sum_{i=1}^x (V_{1,i} \times V_{2,i})}{\sqrt{\sum_{i=1}^x V_{1,i}^2} \times \sqrt{\sum_{i=1}^x V_{2,i}^2}} \quad (7)$$

In a similar way to the *word/sub-interest semantic impact*, the higher the semantic similarity between sub-interest and interest, the greater the probability that the sub-interest belongs to the interest. Let $Y_{k,k'}$ be the degree of *sub-interest/interest semantic impact* of the sub-interest k' belonging to the interest k . We select the most frequent n words which belong to the interest k to compose a collection $Q_k = \{q_{k,1}, q_{k,2}, \dots, q_{k,n}\}$. Suppose $R_k = \{r_{k,1}, r_{k,2}, \dots, r_{k,n}\}$ is a collection such that each item is the number of occurrences of the corresponding word. The *sub-interest/interest semantic impact* can thus be calculated as:

$$Y_{k,k'} = \frac{\sum_{i=1}^n r_{k,i} \frac{\sum_{j=1}^n r'_{k',j} Sim(q_{k,i}, q'_{k',j})}{\sum_{j=1}^n r'_{k',j}}}{\sum_{i=1}^n r_{k,i}} \quad (8)$$

We then sample all interests and sub-interests until we achieve the stable sampling results according to the following two posterior distributions:

$$P(z'_{m,v} | w, \alpha', \beta', z'_-, H, Y') = \frac{P(z'_{m,v} | \alpha', \beta', H, Y')}{P(z'_{m,v} | \alpha', \beta', H, Y')} \quad (9)$$

$$\propto \frac{Y'_{k',v} (n_{k',v} + \beta')}{Y'_{k',\cdot} (n_{k',\cdot} + V\beta')}$$

$$P(z_{m,v} | z', z_-, \alpha, Y, \eta) = \frac{P(z_{m,v} | \alpha, Y, \eta)}{P(z_{m,v} | \alpha, Y, \eta)} \quad (10)$$

$$\propto \frac{Y_{k,k'} (n_{k,k'} + \eta)}{Y_{k,\cdot} (n_{k,\cdot} + K\eta)} \times \frac{n_{m,k} + \alpha}{n_{m,\cdot} + K\alpha}$$

Finally, we obtain:

$$\theta_m = \frac{n_{m,k} + \alpha}{n_{m,\cdot} + K\alpha} \quad (11)$$

$$\theta'_m = \frac{H_{m,k'} (n_{m,k'} + \alpha')}{H_{m,\cdot} (n'_{m,\cdot} + K'\alpha')} \quad (12)$$

$$\phi'_{k'} = \frac{Y'_{k',v} (n_{k',v} + \beta')}{Y'_{k',\cdot} (n_{k',\cdot} + V\beta')} \quad (13)$$

$$\gamma_k = \frac{Y_{k,k'} (n_{k,k'} + \eta)}{Y_{k,\cdot} (n_{k,\cdot} + K\eta)} \quad (14)$$

C. Constructing Interest Dimension Hierarchies

Following the above steps, we can obtain the tweeter-interest probability distribution θ_m , tweeter-sub-interest probability distribution θ'_m , sub-interest-word probability distribution $\phi'_{k'}$, and interest-sub-interest probability distribution γ_k . Now, we exploit these matrices to construct the interest dimension hierarchy to support OLAP.

First, we determine the relationship between sub-interests and interests according to the distribution of the interest-sub-interest probability distribution γ_k . Secondly, a word can belong to multiple sub-interests according to the distribution of sub-interest-word $\phi'_{k'}$. We choose the top 20 maximum probability words that belong to the sub-interest to be the word set of this sub-interest. Finally, we name the sub-interest according to its corresponding word set, and the interest according to the sub-interest set that belongs to it.

IV. Experiment

A. Experimental Setup

To evaluate the effectiveness of our approach, we conducted extensive experiments on real Twitter data collected by Twitter Rest API. Firstly, we selected 15 Twitter users from the Twitter homepage as the seeds and acquired all their 7,373 followers. Next, we retrieved the tweeters’ profiles (nicknames, locations, etc.), tweets (text, favourite count, retweet count, etc.), and their social relationships (follower list, fan list, etc.). Finally, we removed the tweets with fewer than six words and the duplicate tweets. In this way, we obtained 10,160,317 tweets from 6,907 tweeters.

Table 2 shows the statistics of the dataset. The number of tweets that contains “@” amounts to 62% of the total and the average number of twitter’s followers is 1565, indicating the rich social relationships.

Table 2: Dataset Statistics

	#Tweeters	#Tweets	#Tweets with @	#Topics	#Words	#Distinct words	#Follower
Total	6,907	10,160,317	6,301,995	571,068	175,040,499	541,430	10,810,638
Average of tweeters	—	1,471	912	83	25,342	78	1,565
Average of tweets	—	—	62%	—	17.2	—	—

The vector model of Word2vec was trained using the Google News data set, which consists of 100 billion words and covers three million words and phrases, each of which

has 300 dimensions. The experiments ran on a server with six cores of E5-2620 2.00GHz, 64GB memory, and the Windows 7 operating system.

B. Evaluation of Sub-interest Identification

We randomly selected 110 tweeters from the total of 6,907. Five postgraduate students manually labelled the sub-interests of the 110 tweeters independently. If they failed to reach the same sub-interests for a given tweeter, they simply labelled the tweeter with the majority vote. The performance of MS-LDA was evaluated using three widely-adopted metrics, namely precision, recall and F-measure.

$$\text{precision}_i = 100\% \times \frac{|tweeters \text{ correctly identified to } i\text{-th interest}|}{|tweeters \text{ identified to } i\text{-th interest}|} \quad (15)$$

$$\text{recall}_i = 100\% \times \frac{|tweeters \text{ correctly identified to } i\text{-th interest}|}{|tweeters \text{ actually belonging to } i\text{-th interest}|} \quad (16)$$

$$\text{f-measure}_i = 100\% \times \frac{2 \times \text{precision}_i \times \text{recall}_i}{\text{precision}_i + \text{recall}_i} \quad (17)$$

$$\text{f-measure} = \frac{1}{n} \sum_{i=1}^n \text{f-measure}_i \quad (18)$$

A tweeter often has more than one interest. We consider that a tweeter has a particular interest if his probability of having this interest is greater than a predefined threshold value. Here, we set this threshold as 0.1, or T=0.1. To evaluate its effectiveness, we compared our model with LDA and Twitter-LDA [6]. We manually labelled the sub-interests based on the results obtained by LDA and Twitter-LDA since neither LDA nor Twitter-LDA provides multi-level interests. Because a poor result would lead to a poor labelling of sub-interests and vice versa, this comparison can partially demonstrate the performance of our approach. Table 3 shows the average precision (P), recall (R) and F-measure (F) of MS-LDA compared with those of LDA and

Table 3: Comparisons of MS-LDA, Twitter-LDA and LDA

	LDA			Twitter-LDA			MS-LDA		
	P%	R%	F%	P%	R%	F%	P%	R%	F%
Average	68.69	73.82	69.89	75.09	81.14	76.18	82.24	87.11	82.96

Twitter-LDA. As it indicates, our model achieves the best values.

Unlike Twitter-LDA, MS-LDA considers the social relationships among tweeters and the word semantics of tweets. In order to investigate the effect of these two factors, we also conducted experiments without these factors. As Figure 2 shows, the F-measures of standard MS-LDA are better than those of MS-LDA which does not consider social relationships or word semantics under all circumstances. On the other hand, the F-measure of MS-LDA without considering tweeters social relationships achieves 79.54%, which is lower than that of the standard MS-LDA but higher than Twitter-LDA. Meanwhile, the total F-measure of MS-LDA without considering tweets’ word semantics achieves 72.48%, which is far lower than that of the standard MS-LDA and Twitter-LDA, but higher than the traditional LDA.

Table 4 shows the statistical significance based on *p* value of MS-LDA compared with those of LDA and Twitter-LDA. In the case of the significance value being set to 0.1, i.e., $\alpha = 0.1$, our method outperforms the other two with regard to precision, recall and F-measure.

Table 4: Statistical Significance Between MS-LDA and LDA

	P	R	F
<i>P</i> Value: MS-LDA vs LDA	0.0954	0.0934	0.0901
<i>P</i> Value: MS-LDA vs TWITTER-LDA	0.0951	0.0934	0.0951

C. Overall Effect

Based on the above data model, we can conduct various operations such as roll-up, drill-down and dicing on Twitter data. For example, Figure 3 shows the result when drilling down from “Sports”, which reveals the distribution of three different sub-interests of “Sports” in different areas. The data can be aggregated by rolling the area level up to the country level based on the “Location” dimension.

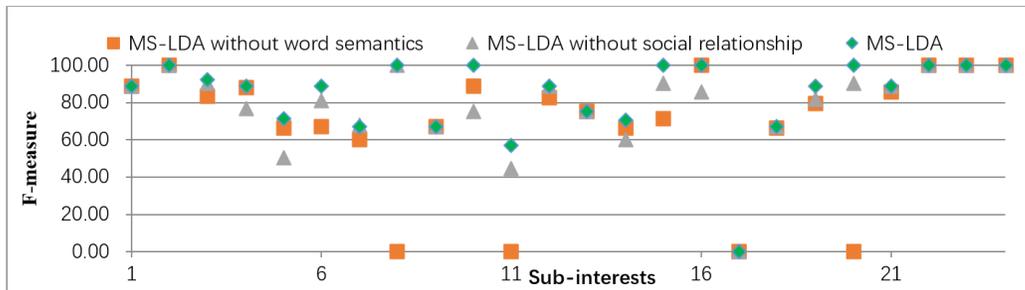


Figure 2: The effectiveness of considering social relationship and word semantics

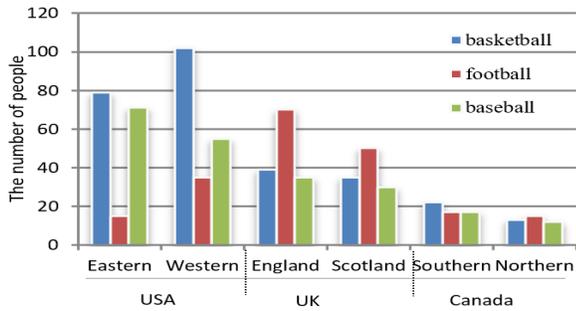


Figure 3: Distribution of sports interests in different locations

Generally speaking, the interests that tweets reveal are always changing, indicating a drift with time. Because all the tweeters' tweets are grouped according to different granularities with respect to time periods, the trained MS-LDA model can be employed to discover the drift of interest distribution of the tweeters in different periods. As Figure 4 demonstrates, in January and February tweeters pay more attention to basketball, but in April and May their interests are transferred to football. On the other hand, the numbers of tweeters who are concerned about movies remains almost stable from January to February in 2016.

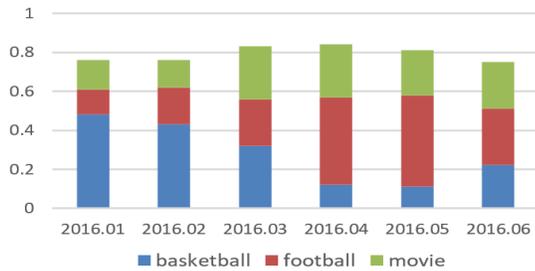


Figure 4: Changes in tweeter interests

V. Conclusion and Future Work

In this paper, we propose an improved topic model, i.e., MS-LDA, which is used to extract the dimension hierarchies of tweeters' interests, normally hidden in the large amount of unstructured Twitter data. We conducted extensive experiments on a large real data set collected by Twitter API to evaluate the effectiveness of MS-LDA. The results show that MS-LDA has a better interest recognition effect than other models.

The Word2vec model employed in this paper is trained using news provided by Google. However, the presentation of news in general is somewhat rigorous, while tweets are more colloquial. In the future, we will consider the use of Twitter data to train the Word2vec model to improve the effectiveness of MS-LDA. In addition, we plan to parallel MS-LDA to improve the running speed.

ACKNOWLEDGMENT

The work is supported by National Natural Science Foundation of China (No. 61100043), the Key Science and Technology Project of Zhejiang Province (No. 2017C01010, No. 2016F50014 and No. 2015C01040). The authors would also like to thank anonymous reviewers who made valuable suggestions to improve the quality of the paper.

References

- [1] S. Mansmann, N. U. Rehman, A. Weiler, and M. H. Scholl, "Discovering olap dimensions in semi-structured data," *Information Systems*, vol. 44, pp. 120–133, 2014.
- [2] E. Siswanto, M. L. Khodra, and L. J. E. Dewi, "Prediction of interest for dynamic profile of twitter user," in *Advanced Informatics: Concept, Theory and Application (ICAICTA), 2014 International Conference of*. IEEE, 2014, pp. 266–271.
- [3] K. H. Lim and A. Datta, "Interest classification of twitter users using wikipedia," in *Proceedings of the 9th International Symposium on Open Collaboration*. ACM, 2013, p. 22.
- [4] X. Pu, M. A. Chatti, U. Schroeder *et al.*, "Wiki-lda: A mixed-method approach for effective interest mining on twitter data," in *Proceedings of CSEDU 2016*, no. EPFL-CONF-217479, 2016.
- [5] Z. Xu, L. Ru, L. Xiang, and Q. Yang, "Discovering user interest on twitter with a modified author-topic model," in *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01*. IEEE Computer Society, 2011, pp. 422–429.
- [6] 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 *European Conference on Information Retrieval*. Springer, 2011, pp. 338–349.
- [7] M. B. Kraiem, J. Feki, K. Khrouf, F. Ravat, and O. Teste, "Olap of the tweets: from modeling toward exploitation," in *Proceedings of the 2014 IEEE Eighth International Conference on Research Challenges in Information Science (RCIS)*. IEEE, 2014, pp. 1–10.
- [8] N. U. Rehman, S. Mansmann, A. Weiler, and M. H. Scholl, "Building a data warehouse for twitter stream exploration," in *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*. IEEE Computer Society, 2012, pp. 1341–1348.
- [9] N. U. Rehman, A. Weiler, and M. H. Scholl, "Olaping social media: the case of twitter," in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ACM, 2013, pp. 1139–1146.
- [10] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.
- [11] M. K. Olorunnimbe and H. L. Viktor, "Tweets as a vote: Exploring political sentiments on twitter for opinion mining," in *Proceedings of the International Symposium on Methodologies for Intelligent Systems*. Springer, 2015, pp. 180–185.
- [12] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *Computer Science*, 2013.