A Hybrid Latent Dirichlet Allocation Approach for Topic Classification

Chi-I Hsu
Information Management Department
Kainan University
Luzhu Dist., Taoyuan City, Taiwan
imchsu@mail.knu.edu.tw

Chaochang Chiu
Information Management Department
Yuan Ze University
Chungli Dist., Taoyuan City, Taiwan
imchiu@saturn.yzu.edu.tw

***Abstract*—Many classification techniques can automatically summarize text into topics and accordingly identify topic terms from the online reviews. Among these techniques Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA) are some of the most often employed approaches. LDA is a probability generated model that projects a document into the topic space using Dirichlet Distribution, and each topic is a collection of words of the probability distribution. As the LDA extracted topics are often implicit, this study first applies LDA to examine the topics of online reviews for game apps in a supervised way. To improve the topic classification performance for LDA, this study proposes a hybrid LDA approach to use Genetic Algorithm (GA) in discovering optimal weights for LDA topics.**

Keywords—Latent Dirichlet Allocation; Genetic Algorithm; Topic Classification

I. INTRODUCTION

The topic detection and opinion polarity classification for online comments are major issues to be tackled for opinion mining. The text itself, in particular for Chinese contents, is unstructured and relatively too complicated to be analyzed directly. For a term frequency approach in analyzing Chinese contents, words segmentation and keywords extraction are essential to convert the texts into a term vector as a digital form to be computationally processed.

When deciding on the topic or category of an online review, we can manually determine the terms that are commonly used on a topic, and then use computer algorithms to automatically calculate the frequency and relative ratio of these common terms that appear in web reviews. (such as TF-IDF value) to determine a topic or category this online review belongs.

Today many classification techniques can automatically summarize text into topics and accordingly identify topic terms from the online reviews. Among these techniques Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA) are some of the most often employed approaches. LDA is a probability generated model that treats a document as a mix of different topics and projects a document into the topic space using Dirichlet Distribution, and each topic is a collection of words of the probability distribution.

Topics extraction as a major application of LDA, some studies using LDA to generate topics or themes for a document or image. For example, see [1], [2], and [3]. However, LDA usually sums up implicit topics without knowing the significance and meaning of themes [4]. As the LDA extracted topics in an implicit way, this study therefore proposes a supervised LDA classification mechanism, in which GA is used to optimize the vector of weights for the LDA topic set which is a document-topic matrix. Through the fitness function and evolution characteristics of GA, the topics LDA generated may gradually approximate to their actual class.

This study develops web crawlers to automatically collect online reviews by game app users on Google Play store. These review contents are processed by text mining techniques. This study first applies LDA to examine the topics of online reviews for game apps in a supervised way. The proposed hybrid LDA approach are further applied and discussed.

II. LITERATURE SURVEY OF LDA

LDA is a probability generated model proposed by Blei et al. [5], via the Dirichlet Distribution this model randomly projects a document into the topic space, and each topic contains multiple words. LDA is the extension of Probability Latent Sematic Analysis (PLSA). Proposed by Hofmann [6], PLSA adopt Aspect Model as the main framework, using probability density function to calculate the proportion of topics in a document, and the probability of words contained in a topic, through this method to find the latent semantic relationships between documents, topics, and words.

The LDA has been used in a number of areas to address different types of issues. LDA application in feature selection, such as Wang and Wong [7] use LDA to assist online personalized recommendations. LDA application in traceability, such as Xing and Croft [8] apply LDA to customized data retrieval; Lukins et al. [9] use LDA to produce the links between topics, queries and reports. LDA application in topic extraction, such as: Krestel et al [1] used LDA to generate recommendation tags; Xudong and Hui [2] tried to

978-1-5090-5795-5/17/$31.00 ©2017 IEEE

divide human actions into different categories using unsupervised Latent Dirichlet Markov clustering; and Baldi et al. [3] attempted to mark the latent topics of software arising from the LDA and claimed the topics to some extent corresponding to the software development cycle changes.

However, the topics of the LDA extracted may not be effectively understood or given a valid explanation or description, that is to say, the topics which LDA automatically summed up are often implicit, or can not articulate the significance and meaning of themes [4]. In the end, LDA topic have practical significance or management implications for managers or developers? Regarding this problem, the survey of Hindle et al. [10] found that only half the chance (50%) can successfully label or describe topics LDA automatically generated.

As the LDA extracted topics are implicit, this study therefore proposes a supervised LDA classification mechanism, unlike Panichella et al. [11] use GA to select appropriate LDA parameters (such as the prior probabilities α for document to generate topics, the prior probability β for topic to generate words, the number of topics K). The study aims to propose a supervised LDA classification mechanism that uses GA to optimize the weight vector of the LDA topic set which is a document-topic matrix to improve the accuracy of LDA topic classification.

III. The Research Method

A. Research Process

The research process is shown in Figure 1. First, crawlers and text mining techniques are used for collecting online reviews (e.g. user comments on Google Play) and for text pre-processing such as Chinese word segmentation, feature selection et al. Second, develop a classification framework according to various discussion topics and build the experiment data sets from the collected online reviews. Third, apply LDA to perform semantic analysis and the proposed supervised LDA-GA for classification prediction. Finally, the experiment results of LDA and LDA-GA can be compared and discussed from the semantic point of view, and then the accuracy results of LDA-GA and other common data classification methods such as decision trees, C4.5, support vector machines (SVM), K-nearest neighbor (KNN) can also be compared.

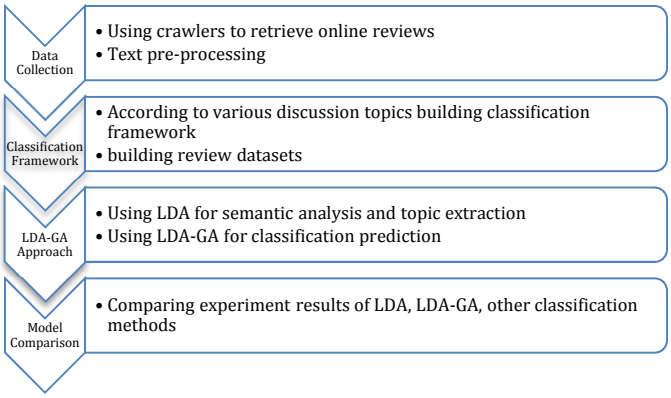

Fig. 1. Research process

B. LDA Approach

Because the topics extracted by LDA are often implicit, this study proposes a supervised LDA-GA approach for classification prediction of online reviews. LDA original matrix representation is shown in Figure 2, where the data set D is the documents to words matrix generated by text pre-processing of the collected online reviews. LDA algorithm can project data set D to potential topic space to obtain the documents-topics matrix θ, and the topics-words matrix Φ.

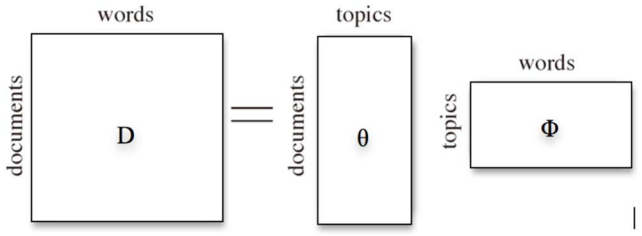

Fig. 2. LDA matrices schematic

C. LDA-GA Approach

The LDA-GA approach proposed in this study, mainly uses GA to find the optimal weights for the topic set, that is, using the GA evolution algorithm to change the θ matrix of LDA. Figure 3 is a schematic LDA-GA, in which the GA optimization is used to determine the GA weight ĝ. The data Set D is a n×m matrix; and ĝ is the transpose array of GA weight of k × 1 dimension; the documents-topics matrix θ is n×k dimension as shown in Table 1, where P_{ij} is the probability of document i related to topic j ; Finally, the topics-words matrix φ is m×k dimension.

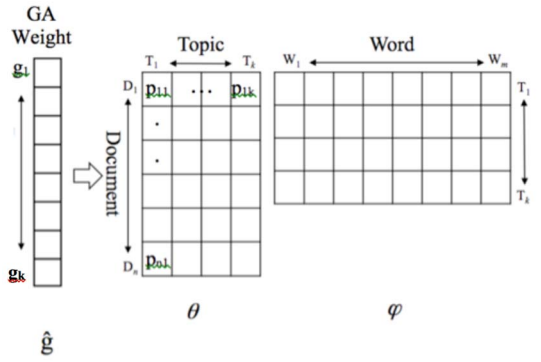

Fig. 3. LDA-GA matrices schematic

TABLE I. The Document to Topic Matrix

	$Topic_1$	$Topic_2$	$Topic_3$	…	$Topic_r$	…	$Topic_K$
D_1	$P_{1,1}$	$P_{1,2}$	$P_{1,3}$	…	$P_{1,r}$	…	$P_{1,K}$
D_2	$P_{2,1}$	$P_{2,2}$	$P_{2,3}$	…	$P_{2,r}$	…	$P_{2,K}$
…	…	…	…	…	…	…	…
D_n	$P_{n,1}$	$P_{n,2}$	$P_{n,3}$	…	$P_{n,r}$	…	$P_{n,K}$

The fitness function of GA is shown in equation (1), wherein the dataset D have n documents (i = 1 ~ n); for the ith document, Score (i) is the score function for correct or incorrect classification, the topic with the greatest probability Arc Max (P_{ij}) is the predicted classification O_i'. Score (i) = 1 if the predicted classification is equal to the target classification O_i, otherwise Score (i) = 0. Through this fitness function, LDA-GA can generate new θ' matrices, and then repeat iterations to achieve the best classification results with the constraint of accuracy > 0.5.

$$\text{Fitness function} \quad \text{Maximize} \quad \sum_{i=1}^{n} Score(i)$$
$$\text{where} \quad Score(i) = 1 \quad \text{if} \quad O_i' \left[Arc\ Max_{j=1}^{k}(p_{ij}) \right] = O_i \qquad (1)$$
$$\text{else} \quad Score(i) = 0$$

IV. EXPERIMENT

According to Apps Annie ranking for Taiwan Game apps downloading in May 2015 [12], online user comments regarding the ranking are retrieved from Google Play store, with a total number of 679,070 comments covering a time frame of 2010/11/01~2015/05/02.

Our previous study identified eight topics for categorizing the game apps opinions, including (a) popularity / marketing, (b) word of mouth / recommendation, (c) audio / visual, (d) operation / interface, (e) game content, (f) transaction costs, (g) software stability, and (h) overall [13]. The results indicate game players have most concerns regarding (e) game content such as story, map, mission, skill, scoring; (g) software stability such as version, server, maintain, connection, network; (h) overall such as time killing, motivation, interest in general and etc.

By systematic sampling N comments for each topic, this study first adopts LDA4j open software for topic classification [14]. LDA4j is a Java implementation of LDA. An example of user comment for game app "Taichi Panda" (太極熊貓) which is labeled as the topic "word of mouth / recommendation" is shown in Figure 4. Some sampled-txt files as the document input of LDA4j are shown in Figure 5.

The confusion matrix of correct and predicted topics for LDA (N=1000) is shown in Table 2. With accuracy >0.5 or value > 500, we can only successfully identify three LDA automatically generated topics. They are 90% (900/1000) for Topic 1, 63.9% (639/1000) for Topic 3, and 63.7% (637/1000) for Topic 2.

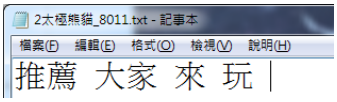

Fig. 4. An example of user comment in a txt file

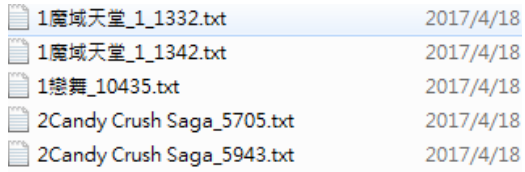

Fig. 5. Some sampled-txt files as input for LDA4j

The confusion matrix of correct and predicted topics for LDA (N=300) is shown in Table 3. With accuracy >0.5 or value > 150, we also can successfully identify three LDA automatically generated topics. They are 94.3% (283/300) for Topic 1, 54.7% (164/300) for Topic 2, and 53.7% (161/300) for Topic 3.

To improve the accuracy of topic classification, LDA-GA is developed to generate new θ' matrices. The confusion matrix for LDA-GA (N=300) is shown in Table 4. GA helps successfully identify four LDA automatically generated topics with accuracy >0.5 or value > 150. They are 94.3% (283/300) for Topic 1, 68% (204/300) for Topic 3, 55.7% (167/300), for Topic 2, and 54.7% (164/300) for Topic 6. With the aid of GA to find the optimal topic weights, this hybrid approach helps improve the topic classification performance for LDA.

TABLE II. LDA CONFUSION MATRIX(N=1000)

Correct	Predicted							
	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8
Topic 1	900	20	5	4	52	0	4	15
Topic 2	89	637	13	25	9	17	203	7
Topic 3	1	15	639	10	10	174	12	139
Topic 4	3	144	20	347	126	6	139	215
Topic 5	20	319	157	43	307	32	83	39
Topic 6	11	5	12	305	139	363	140	25
Topic 7	6	36	32	259	82	266	92	227
Topic 8	7	54	354	52	273	25	162	73

TABLE III. LDA CONFUSION MATRIX(N=300)

Correct	Predicted							
	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8
Topic 1	283	5	2	4	2	0	0	5
Topic 2	6	164	7	20	7	1	2	93
Topic 3	0	9	161	14	7	2	96	7
Topic 4	0	57	3	79	30	104	20	7
Topic 5	8	107	25	64	38	18	16	24
Topic 6	3	3	3	61	83	124	16	7
Topic 7	2	7	5	28	44	45	142	27
Topic 8	2	10	134	26	53	2	7	66

TABLE IV. LDA-GA CONFUSION MATRIX(N=300)

Correct	Predicted							
	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8
Topic 1	283	4	2	1	0	0	10	0
Topic 2	5	167	14	3	0	4	107	0
Topic 3	1	6	204	0	3	5	41	3
Topic 4	2	72	13	76	7	48	79	7
Topic 5	8	107	41	9	5	43	86	5
Topic 6	13	3	15	4	33	164	66	33
Topic 7	4	15	38	7	4	41	134	4
Topic 8	3	8	156	0	21	5	106	21

V. CONCLUSION AND FUTURE DEVELOPMENT

Because the general unsupervised learning uses TF-IDF word statistics to know the representation of a word in the document, the TF-IDF value of each word is used for feature selection, and the Vector Space Model (VSM) is used to represent a document for subsequent document classification. However, the TF-IDF value only represents the relationship between a single word and the document, but it can not find the relationship between words, topics and documents.

LDA is able to project the documents into k topic spaces and express the relationship between words, topics and documents with probability. In this study, a hybrid LDA classification mechanism is proposed, in which the topic weights are optimized by GA to improve the accuracy of LDA topic classification. From the semantic point of view, this approach can illustrate the LDA results including the meaning of topics and the words representing a topic.

In the experiment, this study measures the proportion of positives that are correctly identified to evaluate the topic classification accuracy. Limitation in this research is the lack of adequate considering on the precision rate which is the fraction of classified documents that are relevant. In the future, the experiment design can be further improved to compare LDA, LDA-GA, and other data classification methods commonly found in data mining techniques such as decision trees, SVM, and KNN for model evaluation and discussion.

ACKNOWLEDGMENT

This study is supported by the Ministry of Science and Technology, Taiwan, R.O.C., under contract no. MOST 105-2410-H-424-004-. Special thanks to Shi-Jia Ye, Ming-Hao Sung, and Te-Sheng Chu for their assistance in the text processing and system implementation.

REFERENCES

[1] Fankhauser Krestel, R.A. Nejdl, P.A. Wolfgang, "Latent dirichlet allocation for tag recommendation," Third ACM Conference on Recommender Systems, New York, New York, USA. 2009, pp. 61-68.

[2] Z. Xudong, and L. Hui, "Unsupervised human action categorization using latent dirichlet markov clustering," 4th International Conference on Intelligent Networking and Collaborative Systems (INCoS), 2012.

[3] P.F. Baldi, C.V. Lopes, E.J. Linstead, and S.K. Bajracharya, "A theory of aspects as latent topics," 23rd ACM SIGPLAN Conference on Object-oriented Programming Systems Languages and Applications, OOPSLA '08, New York, NY, USA, ACM, 2008, pp. 543-562.

[4] J.C. Campbell, A. Hindle, and E. Stroulia, "Ch6 Latent dirichlet allocation: extracting topics from software engineering data," The Art and Science of Analyzing Software Data, T. Zimmermann, T. Menzies, C. Bird, Eds. Morgan Kaufmann, 2015.

[5] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent dirichlet allocation," Journal of Machine Learning Research, Vol. 3, No. 4-5, 2003, pp. 993-1022.

[6] T. Hofmann, "Probabilistic latent semantic indexing," Fifteenth Annual Conference on Uncertainty in Artificial Intelligence, 1999, pp. 289-296.

[7] H. Wang, and K. Wong, "Recommendation-assisted personal web," IEEE Ninth World Congress on Services (SERVICES), 2013, pp. 136-140.

[8] W. Xing, and W.B. Croft, "LDA-based document models for ad-hoc retrieval," 29th Annual International ACM SIGIR Conference, 2006, pp. 178-185.

[9] S.K. Lukins, N.A. Kraft, and L.H. Etzkorn, "Source code retrieval for bug localization using latent dirichlet allocation," 2008 15th Working Conference on Reverse Engineering, WCRE '08, Washington, DC, USA, IEEE Computer Society, 2008, pp.155–164.

[10] A. Hindle, C. Bird, T. Zimmermann, and N. Nagappan, "Relating requirements to implementation via topic analysis: Do topics extracted from re- quirements make sense to managers and developers?" ICSM, IEEE Computer Society, 2012, pp. 243-252.

[11] A. Panichella, B. Dit, R. Oliveto, M. Di Penta, D. Poshyvanyk, and A.De Lucia," "How to effectively use topic models for software engineering tasks? An approach based on genetic algorithms," 2013 International Conference on Software Engineering, IEEE Press, 2013, pp.522–531.

[12] Appannie, Rankings of Taiwan Game Apps, [Online]. Available: www.appannie.com/indexes/all-stores/rank/overall/?_ref=header. [Accessed May 2, 2015].

[13] C. Hsu, and C. Chiu, "The factor analysis and heuristic N-phrase rule for users' opinion mining and visualization. 2016 International Conference on Robotics, Control, and Automation, ICRCA 2016, Tokyo, Japan, 2016.

[14] LDA4j, LDA introduction and JAVA implementation, [Online]. Available: http://www.hankcs.com/nlp/lda-java-introduction-and-implementation.html [Accessed April 18, 2017].