Wiki-LDA: A Mixed-Method Approach for Effective Interest Mining on Twitter Data

Xiao Pu[1,2], Mohamed Amine Chatti[3], Hendrik Thüs[3] and Ulrik Schroeder[3]
[1]*École Polytechnique Fédérale de Lausanne , Lausanne, CH*
[2]*Idiap Research Institute, Martigny, CH*
[3]*Informatik 9 (Learning Technologies), RWTH Aachen University, Aachen, DE*
xiao.pu@idiap.ch, {chatti, thues, schroeder}@cs.rwth-aachen.de

Keywords: learning analytics, educational data mining, personalization, adaptation, learner modelling, interest mining, topic modelling, Twitter.

Abstract: Learning analytics (LA) and Educational data mining (EDM) have emerged as promising technology-enhanced learning (TEL) research areas in recent years. Both areas deal with the development of methods that harness educational data sets to support the learning process. A key area of application for LA and EDM is learner modelling. Learner modelling enables to achieve adaptive and personalized learning environments, which are able to take into account the heterogeneous needs of learners and provide them with tailored learning experience suited for their unique needs. As learning is increasingly happening in open and distributed environments beyond the classroom and access to information in these environments is mostly interest-driven, learner interests need to constitute an important learner feature to be modeled. In this paper, we focus on the interest dimension of a learner model and present Wiki-LDA as a novel method to effectively mine user's interests in Twitter. We apply a mixed-method approach that combines Latent Dirichlet Allocation (LDA), text mining APIs, and wikipedia categories. Wiki-LDA has proven effective at the task of interest mining and classification on Twitter data, outperforming standard LDA.

1 INTRODUCTION

Recently, there is an increasing interest in learning analytics (LA) and Educationa Data Mining (EDM). LA focuses on the development of methods for analyzing and detecting patterns within data collected from educational settings, and leverages those methods to support the learning experience. A systematic overview on LA and its key concepts is provided by (Chatti et al., 2012) and (Chatti et al., 2014) through a reference model for LA based on four dimensions, namely data, environments, context (what?), stakeholders (who?), objectives (why?), and methods (how?).

EDM is concerned with developing methods to explored the unique types of data that come from educational settings and, using these methods to better understand students and the setting in which they learn (Romero et al., 2010). From a technical perspective, EDM is the application of data mining techniques to educational data (Baker, 2010).

There are many applications or tasks in educational environments that have been addressed in LA and EDM research. A key area of application is learner (student) modelling, as a result of a focus on adaptive intelligent web-based educational systems, including intelligent tutoring system (ITS) and adaptive hypermedia system (AHS) (Baker, 2010; Chatti et al., 2012; Romero et al., 2010). A learner model represents information about learner's characteristics or states, such as knowledge, motivation, meta-cognitation, and attitudes (Baker, 2010). A learner model is also a representation of information about an individual learner that is essential for adaptation and personalization tasks (Chatti, 2010). The six most popular and useful features in learner modelling include the learner's knowledge, interests, goals, background, individual traits, and context (Brusilovsky and Millan, 2007). Different data mining techniques have been used to build a learner model (Romero et al., 2010). The majority of the proposed approaches, however, have focused on the modelling of the learner's knowledge. This can be explained by the fact that knowledge has constituted the most important part of the learner model in ITS and AHS. In contrast, these systems have paid little attention to learner's interests.

We believe that future learner modelling applica-

tions will increasingly focus on the interest dimension of a learner model, as a result of a shift in focus in the last few years from centralized learning system (e.g. ITS, AHS, LMS) to open and networked learning environments, such as personal learning environments (PLEs) and massive open online courses (MOOCs). These environments deal with large volume of data from a wide variety of sources beyond the ITS/LMS. The data comes from formal as well as informal learning channels (Chatti et al., 2012). As access to information in these environments is mostly interest-driven, learner interests need to constitute an important learner feature to be modelled in order to help learners overcome the information overload problem as well as to support adaptation, personalization, and recommendation tasks.

Detecting learner's interest is also crucial for lifelong learner modelling. (Kay and Kummerfeld, 2011) define a lifelong learner model as a store for the collection of learning data about an individual learner. The authors note that to be useful, a lifelong learner model should be able to hold many forms of leaning data from diverse sources. This data can come in different formats, distributed across space, time, and media. The capacity to mine learner's interests across different learning contexts would provide more effective personalized learning experiences for lifelong learners.

Recognizing the importance of the interest dimension in the learner modelling task, we propose in this paper an innovative approach to effectively mine learner's interests in social networks (e.g. Twitter). We apply a mixed-method approach that combines Latent Dirichlet Allocation (LDA), texting mining APIs, and wikipedia categories.

2 RELATED WORK

Recently, the ability to discover topics or interests of Internet users from information provided on their personal profiles on social media has become increasingly important and necessary. In particular, relevant to our own work, there has been few recent and for the most part different approaches to discover users' topics of interest on Twitter. Content analysis on Twitter introduces unique challenges to the efficacy of topic models on short, messy text. Tweets are constrained to a 140 characters in length and are written in informal language with misspelling, acronyms and non-standard abbreviations, unlike the standard written English on which many supervised models in machine learning and natural language processing (NLP) are trained and evaluated (Mehrotra et al., 2013; Ramage et al., 2010). Hence, effectively modeling content on Twitter requires techniques that can adapt to this uncommon data. In the following, we give an overview of related work in this field of research.

(Michelson and Macskassy, 2010) present a simple non-machine learning approach to discover Twitter users' topics of interest by examining the entities they mention in their tweets. Their approach leverages a knowledge base to disambiguate and categorize the entities in the Tweets, then develop a "topic profile" which characterizes users' topics of interest, by discerning which categories appear frequently and cover the entities. In their work, the goal is to support clustering and searching of Twitter users based on their topics of interest. The authors, however, note that the noisy and ambiguous nature of Twitter makes finding the entities within the tweets quite challenging.

(Puniyani et al., 2010) perform an exploratory analysis of the content of Twitter, using Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to uncover latent semantic themes. They show that these latent topics are predictive of the network structure. The latent topics predict which other microbloggers a user is likely to follow, and to whom microbloggers will address messages.

(Mehrotra et al., 2013) state that the application of standard LDA to Twitter content produces mostly incoherent topics. The authors propose that a solution to this problem is tweet pooling; i.e. merging related tweets together and presenting them as a single document to the LDA model. They investigate different tweet pooling schemes to improve topics learned from Twitter content without modifying the basic machinery of LDA. Finally they make a comparison and conclude that the novel scheme of Hashtag-based pooling leads to drastically improved topic modelling over Unpooled and other schemes.

(Zhao et al., 2011) note that standard LDA does not work well with the messy form of Twitter content. The authors present a Twitter-LDA model slightly different from the standard LDA to discover topics from a representative sample of the entire Twitter. They propose to use one topic per tweet, and argue that this is better than the basic LDA scheme and the author-topic model. The authors then use the proposed model to empirically compare the content of Twitter and a traditional news medium - the New York Times. They note that Twitter can be a good source of topics that have low coverage in traditional news media. And although Twitter users show relatively low interests in world news, they actively help spread news of important world events.

(Ramage et al., 2010) propose Labeled LDA (L-LDA) as variation of LDA based on a partially supervised learning model. Unlike LDA which returns

topics that are latent (i.e., simply numbered distributions over words), L-LDA associates a document with easily-interpretable topics. The authors apply L-LDA to map the content of the Twitter feed into dimensions. These dimensions correspond roughly to substance, style, status, and social characteristics of posts.

(Quercia et al., 2012) focus on the task of document classification in Twitter (i.e., given a Twitter profile and a set of possible topics, determine which topics best fit the profile's tweets). The authors use Labeled LDA (L-LDA) and compare it to the competitive baseline of Support Vector Machines (SVM). They determine the possible topics in the training documents by using text classification APIs. As a result, they conclude that L-LDA generally performs as well as SVM, and it clearly outperforms SVM when training data is limited, making it an ideal classification technique for infrequent topics and for (short) profiles of moderately active users. L-LDA can accurately classify a profile with topics for which it has seen only small amounts of training data and greatly outperforms SVMs at determining how similar a pair of profiles is, implying that L-LDA's techniques of inference are preferable to the linear classification of SVM when dealing with rich, mixed-topic documents such as Twitter profiles.

This related research suggests a number of interesting methods that could be used for content analysis on Twitter. However, these methods are only capable of generating single-word interests. For instance, it is not possible to generate the keyphrase *educational data mining* as a possible interest. Instead, only single keywords - in our example *educational*, *data*, and *mining* - could be generated. This is in general a key limitation of standard LDA and its variations in the literature to date. In this paper, we propose Wiki-LDA as a novel method for significatly improving LDA topic modelling on Twitter. Wiki-LDA leverages LDA, text mining APIs, and Wikipedia categories in order to produce meaningful single-word as well as keyphrase interests and accurately classify them into related topics.

3 CONCEPTUAL APPROACH

Our overall approach breaks into nine high level steps, as depicted in Figure 1:

1. Collect and store Tweets from Twitter as training set . This was done by crawling Tweets from popular user accounts which are listed under the major topic classifications on Twitter. This training data set was then pre-processed and indexed via the Lucene[1] text information retrieval library;
2. Transform the text data to a vector representation and finally to a Mahout[2]-readable matrix format. Implement the Latent Dirichlet Allocation (LDA) algorithm for training. From this process a "bag-of-words" collection of data is obtained and then stored into MySQL database for inference;
3. Implement a dynamic crawling methods for test users, in which their crawled Tweets are used for prediction. Pre-processing of the crawled Tweets was done, also with the Lucene library;
4. Implement LDA prediction for each test user based on the "bag-of-words" result from the LDA training process to predict possible topic distribution;
5. Use text analysis APIs to generate keywords and keyphrases from the test user profiles;
6. Send results from the respective APIs to Wikipedia in order to obtain all the related categories of each specific keyword or keyphrase;
7. Collect the analyzed categories from Wikipedia and use them as input for LDA in order to determine the possible topic distribution for each keyphrase generated from the APIs;

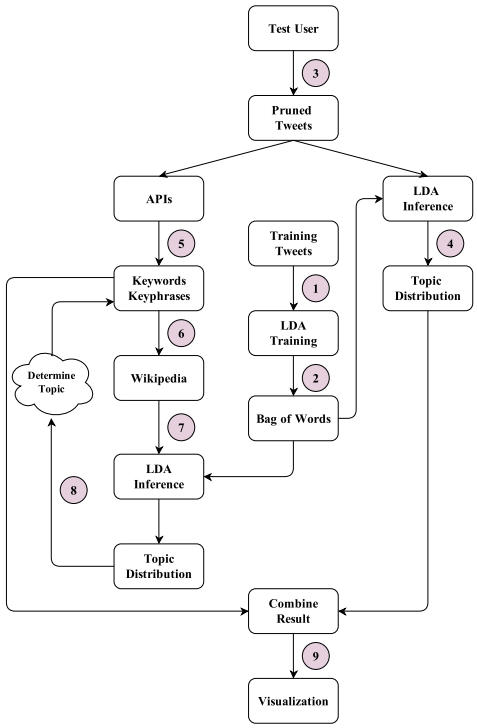

Figure 1: Wiki-LDA: Conceptual Approach

[1] http://lucene.apache.org/core

[2] http://mahout.apache.org/

8. Combine all results for each specific topic;
9. Visualize the final results to the user.

4 IMPLEMENTATION

The implementation of our approach can be roughly divided into three major parts:

- Training Set Creation - Crawling of user data from popular social networks, which in our specific case was Twitter. This part was implemented via the Twitter API[3] in Java. The API enabled us to collect user Tweets to form both the training and test data.
- Training Process - Training of the LDA machine learning algorithm using the crawled data.
- Prediction Process - Constructing a model to predict single-word as well as keyphrase interests of new users.

4.1 Training Set Creation

As a first step in the training phrase, we selected the 9 most popular abstract topics as published on the Twitter Website. These include: Art & Literature, Business, Food & Drink, Government, Health, Science & Technology, Music, Sport, Travel. We then crawled tweets from about 4-5 users in each topic to form the training set. Due to the limitation of the crawling API from Twitter, in which the maximum amount of Tweets that can be crawled for a single user per request is restricted to the 20 recent ones, we manually crawled about 150-200 additional tweets for each user over a time span of one month. We chose users for each topic based on the recommended popular users in each topic provided by Twitter. For instance, we chose @*FinancialTimes*, *BBCBusiness*, etc. For the topic topic "Business". Hence, we had a corpus of about $800 - 1000$ tweets in each topic that can be used for training.

4.2 Training Process

Our goal was to automatically identify the topics that Twitter users are interested in based on their tweets. We mainly used the Latent Dirichlet Allocation (LDA) (Blei et al., 2003) for this purpose. LDA is an unsupervised machine learning technique to identify latent topic information from large document collections. It uses a "bag of words" assumption, which treats each document as a vector of word counts. Based on this assumption, each document is represented as a probability distribution over some topics, while each topic is represented as a probability distribution over a number of words.

We ran LDA with 1000 iterations of Gibbs sampling with predefined $K = 9$. The topic-word distribution was updated each time in the Gibbs sampling process until the distribution converges. Table 1 shows an excerpt of the topic-word distribution that we obtained as a result of the training phase.

Art & Literature	**Business**	**Food & Drink**	**Government**	**Music**
review	percent	restaurant	president	album
books	bank	food	obama	music
theater	market	recipe	insurance	rock
cartoon	business	dinner	immigration	songs
book	trade	recipes	economy	rocking
novel	prices	dish	care	sound
art	bills	cooking	leaders	hear
library	boss	cheese	government	jazz
museum	opinion	soup	coverage	piano
writer	financial	chefs	enrollment	band
Science & Technology		**Sport**	**Travel**	**Health**
google		game	travel	healthy
app		NBA	traveler	insurance
search		team	destinations	calories
apple		sport	city	risk
online		league	visiting	care
android		basketball	hotel	fats
startup		soccer	tips	weight
computer		trade	beach	sleep
internet		lakers	passengers	cancer
update		crazy	weather	help

Table 1: Top 10 words analyzed out in our LDA implementation

4.3 Prediction Process

After training the LDA model using the training set above, the next step was to predict possible topic distributions for test users depending on the resulting topic-word distribution. As pointed out in section 3, a key limitation of standard LDA is that it does not allow to generate keyphrase interests. This would lead to an interest list which is less coherent and interpretable. To address this issue, we developed a novel method, called *Wiki-LDA* for significantly improving LDA topic modelling on Twitter. Wiki-LDA extends the standard LDA by leveraging text analysis APIs and Wikipedia categories. In the following, we discuss the Wiki-LDA approach in more details.

Since Twitter is too sparse for traditional topic modelling, we followed an author-wise pooling approach to gather together all the tweets from a given user into a single document (Mehrotra et al., 2013). Our model, thus learns the latent topics that characterize users, rather than tweets.

We then applied online text analysis tools to the collected tweets of the test user. We used Alchemy API[4]

[3] https://dev.twitter.com

[4] http://www.alchemyapi.com/api/

and OpenCalais API[5] to extract possible keywords and keyphrases from the tweet data. Table 2 presents the results analyzed from Twitter user "@google" by the mentioned text analysis tools.

	AlchemyAPI	OpenCalais
Keywords	android	System software
	google	Software
	disney	PlayOnLinux
	googleio	Web 2.0
	techcrunch	Cloud clients
	seattle	Cross-platform software
	googleplay	Embedded Linux
	asia	Smartphone
	googlesearch	Google
	obama	Gmail
	percy harvin	Android

Table 2: List of keywords extracted from APIs for @google

The next step was to classify the extracted keywords and keyphrases into related topics using LDA. This would be a straightforward exercise if the keyphrases contain some words which exist in the LDA training results, but this process would present some problems if the words are totally new to LDA. Our aim was to increase the probability that a generated keyword or keyphrase is accurately classified by LDA. To achieve this, we used Wikipedia API, which provides all possible categories corresponding to a particular keyword or keyphrase. After crawling all possible categories based on a given keyword or keyphrase query, we collect these categories and use them as input for LDA.

Figure 2 illustrates a sample process for the classification of the extracted keyphrase "percy harvin" (i.e. a keyphrase generated by Alchemy API for Twitter user "@google") by combination of Wikipedia and LDA. The complete classification procedure works as follows:

1. If the system finds extracted keywords/keyphrases (in our example "percy harvin") from text analysis APIs which cannot be analyzed by original LDA, it automatically input these keywords/keyphrases to the Wikipedia API;
2. The Wikipedia API returns all categories associated with "percy harvin" to the system. Here the Wikipedia categories associated with "percy harvin" include: "American football wide receivers", "sports clubs established in 1961", "Sports in Minneapolis Minnesota", etc;
3. The system receives these categories and splits them into single words; in our example the collection of all words for "percy harvin" are : ["American", "football", "wide", "receivers", "sports", "clubs", "established", etc];
4. The system uses this collection of words as input to LDA. After calculation, LDA gives one topic distribution for each word. Here for the word collection derived from the categories of "percy harvin", the distribution is: Sports 0.67, Goverment 0.03, Music 0.13, etc;
5. Finally, we choose the topic with the highest probability from the distribution provided by LDA (in our example "Sports") as the possible topic of the original keyphraseinput input "percy harvin".

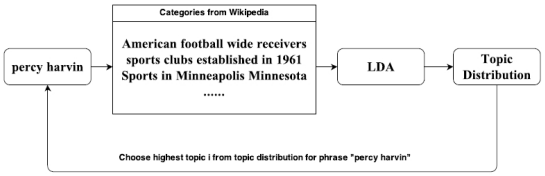

Figure 2: Sample classification process for keyphrase "percy harvin"

Table 3 is a sample result of the related keywords and keyphrases for topic "Science & Technology" from user "@google". The standard LDA algorithm could only analyze single keywords (e.g. "coming", "google"). The term extraction step by using the Alchemy API and OpenCalais API resulted in more keywords and keyphrases that couldn't be directly analyzed by standard LDA (e.g. "PlayOnLinux", "System software"). Harnessing Wikipedia categories as explained above, has led to an accurate classification of these keywords and keyphrases to the topic "Science & Technology". The analysis and classification results are visualized through a graphical user interface, as depicted in Figure 3.

	Most Related Topic	classified words in this topic
Standard-LDA	Sci. & Tech.	google coming
Wiki-LDA	Sci. & Tech.	System software (openCalais) Smartphone (openCalais) PlayOnLinux (openCalais) google (AlchemyAPI) googleio (AlchemyAPI) techcrunch (AlchemyAPI) coming (LDA) google (LDA)

Table 3: Comparison of classification results with standard LDA and Wiki-LDA

The complete process of the Wiki-LDA approach for interest mining and classification in Twitter is shown in Figure 4. The system uses the Twitter API

[5]http://www.opencalais.com/documentation/opencalais-documentation

to collect the Tweets a user. After a pre-processing step which uses the Lucene library for tokenization, removal of stop words, and stemming of the input data, the system uses the result data set simultaneously as input for the LDA prediction algorithm and the text extraction APIs. The standard LDA prediction part produces the topic distribution for the user based on the input data. The APIs extract keywords and keyphrases which are then used by the system as input for the Wikipedia API to gather all possible categories for each extracted keyword and keyphrase. The bag of category words are then given to LDA again in order to determine the possible topic for each extracted keyword and keyphrase as discussed in the example above. The analysis results from standard LDA and Wiki-LDA are then merged into a single interest list representing the final topic distribution for the Twitter user.

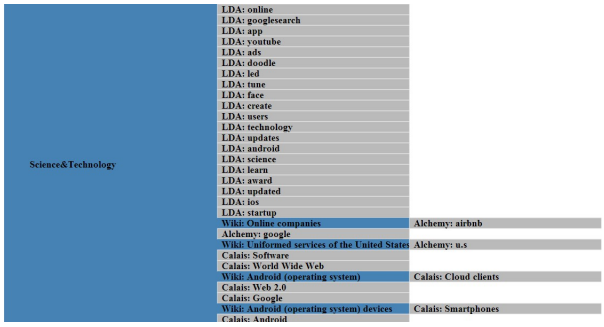

Figure 3: Visualization of interests related to topic Sci&Tech for test user @google

Figure 5 depicts a comparison between the analysis and classification results generated with standard LDA and Wiki-LDA. It shows that Wiki-LDA enables to extract and accurately classify more interest keywords and keyphrases as compared to the naive application of LDA.

5 EXPERIMENTAL EVALUATION

In this section we describe the details of the experiment conducted to gauge the quality of the analysis and classification results achieved by Wiki-LDA. The experiment evaluation was performed through quantitative measures as well as personal interviews with Twitter users.

5.1 Classification Evaluation

We selected four Twitter users for evaluation of the Wiki-LDA approach, as shown in Table 4.

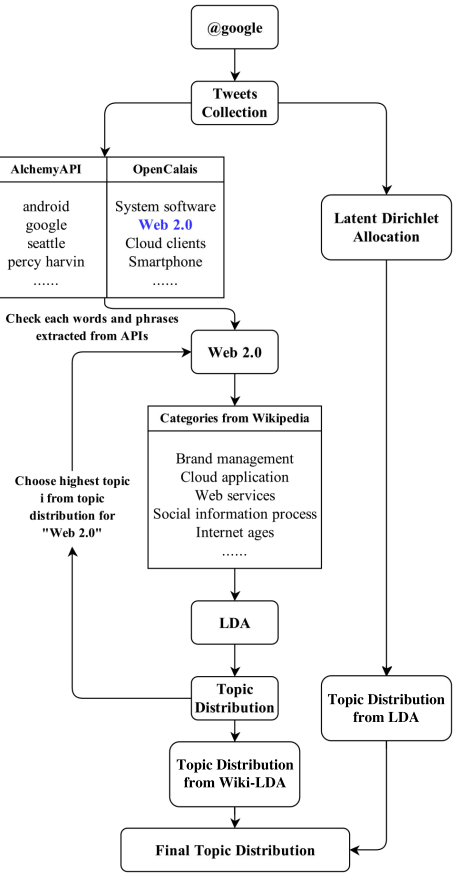

Figure 4: Overall prediction process for test user @google

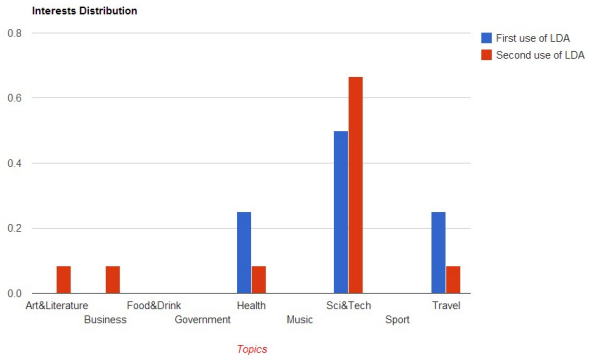

Figure 5: Comparison of analysis and classification results with standard LDA and Wiki-LDA for test user @google

Twitter User	# Collected Tweets
@Oldaily	128
@BarckObama	141
@DailyHealthTips	320
@NBA	112

Table 4: Tweets from test users

We ran the Wiki-LDA algorithm over the Tweets of the four users and extracted the possible topic distribution for each test user, with related keywords and keyphrases. To evaluate the performance of Wiki-LDA, we manually computed the *precision* and *recal* of topics for each test user, where recall is the percentage of the extracted interests that are indeed correct, and precision is the percentage of the correct extracted interests out of all extracted interests. We then combined precision and recall in the composite metric of F-measure (or F1 score): $F1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$. Table 5 summarizes a comparison between the F1 Score achieved by standard LDA and Wiki-LDA. The evaluation shows that Wiki-LDA has led to significantly improved interest mining results on the Twitter data used for our experiment.

Twitter User	F1 Score	
	Standard-LDA	Wiki-LDA
@oldaily	.696	.918
@DailyHealthTips	.936	.979
@BarackObama	.931	.985
@NBA	.746	.850

Table 5: Classification Evaluation

In order to show the improvement of results more clearly, Table 6 shows the specific topics with highest probability for each test user, and top-5 relevant words which are extracted by both original LDA and Wiki-LDA. From the results, we can see that, in the prediction part of the system, the Wiki-LDA model can correctly analyze and classify not only single keywords, but also keyphrases, thus making the interest mining task more accurate and meaningful.

Twitter User	Most related Topics	classified words in this topic	
		Standard-LDA	Wiki-LDA
@oldaily	Sci.&Tech.	google learn online create research	Online education Educational software George Siemens E-learning learn
@BarackObama	Government	coverage insurance senate act covered	Patient Protection Affordable care act primary campaign health president
@DailyHealthTips	Health	foods worst diet hair healthy	hair care baldness human skin color human skin weight
@NBA	Sport	game score NBA season lakers	National basketball association Cleveland Cavaliers lakers game

Table 6: Extraction and classification results with standard LDA and Wiki-LDA

5.2 Personal Interviews

Personal interviews were conducted with four Twitter users. Table 7 shows an excerpt of the interests of each user extracted by both standard LDA and Wiki-LDA. The list of interests generated by Wiki-LDA included not only keywords but also keyphrases, in addition to keywords which did not appear in the training set. These interests were presented to the users who were asked to gauge the consistency of the results. In general, the user feedback was that the interests generated by Wiki-LDA are more accurate, meaningful, and coherent than those generated by standard LDA. This result was further confirmed by the computation of F1 score based on the users' responses, as summarized in Table 8. Overall, the evaluation results indicate that the Wiki-LDA model is a better choice than standard LDA for interest mining on Twitter data.

Test Users	Hobby	Extracted keywords	
		Standard-LDA	Wiki-LDA
@sadiksha	Computer Science	comming google photo	google PlayOnLinux System software .....
@Xia41258659	Cuisine, Travelling, Music	restaurant around food	laura jansen Matcha dinner .....
@LZYuan_1981	Reading, Travelling, Music	case books writer	Culture library song .....
@vividxiao	Music, Travelling	listening rock deal	Pairs Ladygaga piano .....

Table 7: Extracted interests for test users

Twitter User (Volunteer)	F1 Score for Interests analysis	
	Standard-LDA	Wiki-LDA
@Sadiksha	.143	.571
@Xia41258659	.574	.857
@LZYuan_1981	.588	.824
@vividxiao	.256	.749

Table 8: Results from personal interviews

6 CONCLUSION AND FUTURE WORK

Learning Analytics (LA) and Educational Data Mining (EDM) are concerned with developing methods for exploring data coming from educational environments to resolve educational research issues. Learner modelling is a crucial task in these emerging research areas. In this paper, we focused on the interest dimension in the learner modelling task, which is

crucial in today's learning environments characterized by openness and autonomy. We presented the conceptual, implementation, and evaluation details of Wiki-LDA, as a mixed-method interest mining approach that combines Latent Dirichlet Allocation (LDA), text extraction APIs, and wikipedia categories in order to effectively mine user's interests in Twitter. Through the combination of machine learning, information retrieval, and knowledge bases, we were able to mitigate the obvious limitation of the small size of the training data set and to extract not only keywords but also keyphrases as possible interests.

Overall, the evaluation results showed that Wiki-LDA clearly outperforms standard LDA in terms of the meaningfulness and coherence of the extracted interests as well as the accuracy of the classification of the interests in related topics. Hence, this work provides a novel method for significantly improving interest mining on Twitter data.

While our early results are encouraging for generating the interest profile of a Twitter user, there are still a number of areas we would like to improve. The first, and most important are is defining a large training corpus, which is crucial for a machine learning task. We have crawled tweets from 3-4 user accounts from Twitter for each abstract topic as training set. A logical next step to improve is hence to gather many more Tweets from more users, and improve the range of possible abstract topics in order to classify more latent words.

Moreover, the Wiki-LDA algorithm has still room for improvement. One technical limitation of LDA is the need to fix the possible number of topics K before learning. To improve on this one can consider the possibility of letting K to be infinity in LDA and determine the number of topics through a separate learning process.

Another important area to improve is our evaluation. We plan to perform a larger scale experiment in a real learning environment which will allow us to thoroughly evaluate our interest mining approach.

ACKNOWLEDGEMENTS

The first author acknowledges the support of the Swiss National Science Foundation through the MODERN Sinergia Project (www.idiap.ch/project/modern).

REFERENCES

Baker, R. (2010). Data mining for education. *International Encyclopedia of Education*, 7:112–118.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Brusilovsky, P. and Millan, E. (2007). User models for adaptive hypermedia and adaptive educational systems. In Brusilovsky, P., Kobsa, A., and Nejdl, W., editors, *The Adaptive Web, LNCS 4321*, chapter 1, pages 3–53. Springer-Verlag Berlin Heidelberg.

Chatti, M. A. (2010). The laan theory. In *Personalization in Technology Enhanced Learning: A Social Software Perspective*, pages 19–42. Aachen, Germany: Shaker Verlag.

Chatti, M. A., Dyckhoff, A. L., Schroeder, U., and Thüs, H. (2012). A reference model for learning analytics. *International Journal of Technology Enhanced Learning*, 4(5/6):318–331.

Chatti, M. A., Lukarov, V., Thüs, H., Muslim, A., Yousef, A. M. F., Wahid, U., Greven, C., Chakrabarti, A., and Schroeder, U. (2014). Learning analytics: Challenges and future research directions. *e-learning and education journal (eleed)*, 10.

Kay, J. and Kummerfeld, B. (2011). Lifelong learner modeling. In Durlach, P. J. and Lesgold, A. M., editors, *Adaptive Technologies for Training and Education*, pages 140–164. Cambridge University Press.

Mehrotra, R., Sanner, S., Buntine, W., and Xie, L. (2013). Improving lda topic models for microblogs via tweet pooling and automatic labeling. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 889–892. ACM.

Michelson, M. and Macskassy, S. A. (2010). Discovering users' topics of interest on twitter: a first look. In *Proceedings of the fourth workshop on Analytics for noisy unstructured text data*, pages 73–80. ACM.

Puniyani, K., Eisenstein, J., Cohen, S., and Xing, E. P. (2010). Social links from latent topics in microblogs. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, pages 19–20. Association for Computational Linguistics.

Quercia, D., Askham, H., and Crowcroft, J. (2012). Tweetlda: supervised topic classification and link prediction in twitter. In *Proceedings of the 3rd Annual ACM Web Science Conference*, pages 247–250. ACM.

Ramage, D., Dumais, S. T., and Liebling, D. J. (2010). Characterizing microblogs with topic models. *ICWSM*, 10:1–1.

Romero, C., Ventura, S., Pechenizkiy, M., and Baker, R. S. (2010). *Handbook of educational data mining*. CRC Press.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., and Li, X. (2011). Comparing twitter and traditional media using topic models. In *Advances in Information Retrieval*, pages 338–349. Springer.