General Topic Annotation in Social Networks: A Latent Dirichlet Allocation Approach

Amir H. Razavi[1], Diana Inkpen[1],
Dmitry Brusilovsky[2], and Lana Bogouslavski[2]

[1] School of Electrical Engineering and Computer Science
University of Ottawa
{araza082,diana}@eecs.uottawa.ca
[2] Business Intelligence Solutions
dmitry@bisolutions.us

Abstract. In this article, we present a novel document annotation method that can be applied on corpora containing short documents such as social media texts. The method applies Latent Dirichlet Allocation (LDA) on a corpus to initially infer some topical word clusters. Each document is assigned one or more topic clusters automatically. Further document annotation is done through a projection of the topics extracted and assigned by LDA into a set of generic categories. The translation from the topical clusters to the small set of generic categories is done manually. Then the categories are used to automatically annotate the general topics of the documents. It is remarkable that the number of the topical clusters that need to be manually mapped to the general topics is far smaller than the number of postings of a corpus that normally need to be annotated to build training and testing sets manually. We show that the accuracy of the annotation done through this method is about 80% which is comparable with inter-human agreement in similar tasks. Additionally, using the LDA method, the corpus entries are represented by low-dimensional vectors which lead to good classification results. The lower-dimensional representation can be fed into many machine learning algorithms that cannot be applied on the conventional high-dimensional text representation methods.

Keywords: Latent Dirichlet Allocation (LDA), Automatic document annotation, Text representation, Topic extraction.

1 Introduction

Today, modern social networks and services have become an increasingly important part of how users spend their time in the online world. Social networking sites are now increasingly becoming social networking services, and they bring more and more information to the users through their available communication tools. In the meanwhile, in order to present the best set of features of a social network or service and also to have a proper control on such a vast interface, automatic social network analysis has an important role. In the same time, the concepts of social media are being actively adopted by the enterprises; many of them are implementing their own

O. Zaïane and S. Zilles (Eds.): Canadian AI 2013, LNAI 7884, pp. 293–300, 2013.
© Springer-Verlag Berlin Heidelberg 2013

enterprise social media platforms. The market of enterprise social media collaboration software is fast growing.

In this paper, we focused our efforts on the unstructured part of the social networks which is the textual postings and related comments. We present a semi-supervised method to extract general topics from a social network corpus and then annotate the postings using the general topics. The general topic annotation can be used for further conceptual analysis of the textual content of the social network. In the proposed method, we annotate a subset of the social network threads (posts and their comments) automatically and then we evaluate the annotation quality (by comparing it to the labels assigned manually) to show that is reliable enough to be used in the social network text analysis task. Our method applies Latent Dirichlet Allocation (LDA) on a corpus to initially infer some topical word clusters which will be then used for document annotation and representation at the next stage. Each extracted topical cluster is interpreted by a human judge and will be projected into a generic categorical topic which will be then the label of the document (a thread in our case).

2 Background

In 2003, Blei, Ng and Jordan presented the Latent Dirichlet Allocation (LDA) model and a Variational Expectation-Maximization algorithm for training their model. Those topic models are a kind of hierarchical Bayesian models of the applied corpus [1]. The model can unveil the main themes of the applied corpus, which can potentially use to organize, search, and explore the documents of the corpus. In LDA topic modeling, a "topic" is a distribution over a fixed vocabulary of the corpus and each document can be represented by several topics with different weights. The number of topics and the proportion of vocabulary that create each topic are considered as two hidden variables of the model. The conditional distribution of these variables given an observed set of documents is regarded as the main challenge of the model.

Collapsed Variational Bayes (CVB) inference [2] also analytically marginalizes the topic proportions and is regarded as an alternative deterministic inference for LDA. The proposed inference algorithm can improve the accuracy and efficiency of the standard Bayesian inference for LDA.

Griffiths & Steyvers [3] applied a derivation of the Gibbs sampling algorithm for learning LDA models. They showed that the extracted topics capture a meaningful structure of the data. The captured structure is consistent with the class labels assigned by the authors of the articles. The paper presents further applications of this analysis, such as identifying "hot topics" by examining temporal dynamics and tagging some abstracts to help exploring the semantic content.

Since then, the Gibbs sampling algorithm was shown as more efficient than other LDA training methods, e.g., variational EM and Expectation-Propagation [4]. This efficiency is attributed to a famous attribute of LDA namely, "the conjugacy between the Dirichlet distribution and the multinomial likelihood". This means that the conjugate prior is useful since the posterior distribution is the same as the prior, and it makes inference feasible and causes that when we are doing sampling, the posterior sampling become easier. Because of this, the Gibbs sampling algorithms was applied for inference in a variety of models which extend LDA [5] [6] [7].

Hoffman et al [8] introduced a new derivation named "Online LDA" which is a stochastic gradient optimization algorithm for topic modeling. The algorithm iteratively subsamples a small number of documents from the entire corpus and then updates the topics by the new inferences. Since in this method we do not need to store topic proportions for the entire corpus, it is much more memory conservative than the standard approach. Furthermore, the authors show that since the algorithm updates topics more frequently, it converges faster than the other methods. However, when it runs over a large corpus, it does not scale up to appropriate large numbers of topics. Adaptive scheduling algorithm [9] can also be regarded as applicable extensions of this model.

3 Data Set

In our research, we were looking for sources of social network data that include textual postings and related comments, in which the main posting could be connected to the corresponding comments in order to form a thread; the connection is done via parent/posting identifier (id) information items. It was challenging and rather time consuming to find such datasets. We selected a set of post/comment textual data that we extracted from the well-known "Friendfeed" social network media.

Initially, through a multi-level filtering task, a large amount of data (~ 23 GB in compressed format) was collected from "Friendfeed.com"; from which we extracted the information items useful to our research, including main postings and their related comments which are linked based on Post_id. Then we integrated the main postings (12,450,658) to their corresponding comments (3,749,890) in order to create same topic threads. At the next stage, we filtered out all the threads with no comments (with Null comments).

The source data was in more than 11 different languages; therefore we run a language identification tool in order to select a subset including only postings and related comments in English. There were many postings/comments mixed in English and another language; this represented another challenge at this stage. Hence, we decided to remove the threads that were partially commented in other languages and kept only threads that were entirely in English at the final stage. We also filtered out all threads smaller than 120 characters or with less than three comments.

The built data-set included more than 24000 usable threads as input for our topic detection task. A randomly selected subset of 500 threads was chosen to be manually annotated in order to be used for training/ testing a variety of classifiers as a proof for the applicability of our general topic annotation method of the threads. The class labels (general topics) were selected and generalized manually *based on the topics extracted automatically by the LDA method* [1]. The final set of general topics contained the following 10 categories:

`consumers, education, entertainment, life_stories, lifestyle, politics, relationships, religion, science, social_life, technology.`

[1] Will be explained more in the "LDA Topical Modeling" section.

Additionally, a random subset of ~4000 threads including the initial 500 threads plus 3500 unlabeled threads (that we call *background resource*) has also been selected for estimating the LDA models that are needed for the topic detection. (The method will be explained later in Section 4.2.)

4 Methodology

4.1 Preprocessing

In the preprocessing stage, initially all the different headers, internet addresses, email addresses and tags were filtered out. Then all the delimiters such as spaces, tabs or newline characters, have been removed from postings, whereas the expressive characters like: " - . ' ' ! ? " were kept. Punctuations (such as quotes, " ") could be useful for determining the scope of speaker's messages. This step considerably reduces the size of feature space and prevents the system from dealing with a large number of unrealistic tokens as features for our classifiers and LDA estimation/inferences.

Two types of stop-words removal were performed: static stop words removal and corpus based dynamically stop words removal. For the first one, we tokenized the posts/comments individually to be passed to the static stop-word removal step that is based on an extensive list of stop-words which has been already collected specifically for the applied dataset (i.e., social network).

In the second one, additional stop words were determined based on their frequency, distribution and the tokenization strategy over the corpus (i.e., unigrams, bigrams, 3 or 4 grams). We removed tokens with very high frequency relative to the corpus size where those appear in every topical class (i.e., those are almost useless for the topic identification task). The output of this stage passed to the stemming process through the Snowball[2] stemming algorithm.

4.2 LDA Topical Modeling

For our goal of general topic extraction from social network threads, we developed a method based on the original version of LDA [1]. LDA is a generative probabilistic model of a corpus. The basic idea is that the documents are represented as a weighted relevancy vector over latent topics, where a topic is characterized by a distribution over words. We applied and modified the code originally written by Gregor Heinrich [10] based on the theoretical description of Gibbs Sampling. A remarkable attribute of the chosen method is that lets a word to participate in more than one topical subset based on its different senses/usages in its context.

The subset that we used for running the LDA algorithm consisted of 4000 threads (500 labeled and 3500 background source) which already passed the preparation and filtration processes (the pre-processing). In this way, each thread is represented by a number of topics in which each topic contains a small number of words inside (i.e., each topic consists in a cluster of words); and each word can be assigned to more than one topic across the entire input data (e.g., polysemous words can be in more than one topic). Therefore, the number of topics and the number of words inside each topic are

[2] `http://snowball.tartarus.org/`

two parameters of the method that can be adjusted according to the input data. In this research, the values of the parameters have been empirically set to 50 topic clusters, and maximum 15 words in each cluster. Then, the LDA method assigns some groups of words (the 50 groups of 15 words inside each group) as topics, with different weights, for each text (in our case each thread).The topical cluster of words are interpreted and assigned to a real topical phrase/word, manually.

For example, the following topical cluster: {"Google", "email", "search", "work", "site", "services", "image", "click", "page", "create", "contact", "connect", "buzz", "Gmail", "mail"} which is a real example extracted by the LDA model estimation process from the explained corpus, initially has been interpreted (manually) as "Internet" topic and at the next level of the *topic generalization* was placed under the "`technology`" and "`social_life`" categories.

Similarly to the above example, all the 50 topical clusters extracted by the LDA method were manually mapped to the previously listed 10 generic and human-comprehensible topics. We observed that the 10 class labels (general topics) are distributed unevenly over the dataset of 500 threads, in which we had 21 threads for "consumers", 10 threads for "education", 92 threads for "entertainment", 28 threads for "incidents", 90 threads for "lifestyle", 27 threads for "politics", 58 threads for "relationships", 31 threads for "science", 49 threads for "social_activities", and 94 threads for "technology". Thus, the baseline of any classification experiment over this dataset may be considered as 18.8%, for a trivial classifier that puts everything in the most frequent class, "technology". However, after balancing the above distribution through over/under sampling techniques, the classification baseline lowered to 10%. The last step was performed via the Synthetic Minority Oversampling Technique (SMOTE) [11] over the class labels with frequencies lower than average, and random under-sampling method over those which have frequencies higher than average. We sustained those extra steps in order to obtain an evenly distributed dataset and do not deal with an unbalanced data classification task and its side effects.

Since the LDA modeling does not assign a single general topic (e.g., "entertainment") to each tread, the assignment of the general topics (i.e., one of the 10 class labels) is a further task that will be done through a separate classification process.

4.3 Topic Classification

As mentioned before, the training/testing dataset for the supervised classification task consisted of 500 manually annotated threads annotated with the 10 general categories enumerated in section 3. For this dataset, we initially applied a variety of Bag of Word (BOW) representations (i.e., binary, frequency and TF-IDF[3] based methods) in order to create the best discriminative representation over the entire 500 threads dataset. After removing stop-words and stemming as explained in section 4.1., we obtained 6573 words as the feature set for the general topic classification task.

[3] The TF-IDF (term frequency versus inverse document frequency) method was selected which is a classic method that gives higher weights to terms that are frequent in a document but rare in the whole corpus.

As the second and axillary representation of the same data, we used the topical cluster relevancy vector of the each thread[4] (calculated using the LDA technique) to obtain a low-dimensional representation of the threads. We evaluated that representation of the data and reserved for the complementary comparison between the two representations. Then we integrated the two representations mentioned above into one representation, which consisted of 6623 features (words and 50 topics) to test the classification (automatic annotation) performance over the integrated representation. As part of the supervised learning core of the system, we trained a variety of classifiers, in order to evaluate the general topic annotation performance of the method.

5 Results and Discussion

We run our comparing classification experiments on the 500 filtered Friendfeed threads. We conducted the classification evaluations using stratified 10-fold cross-validations (this means that the classifier is trained on nine parts of the data and tested on the remaining part, then this is repeated 10 times for different splits, and the results are averaged over the 10 folds). We performed several experiments on a range of classifiers and parameters for each representation to check the stability of a classifier's performance. We changed the "Seed", random parameter of the 10-fold cross-validation in order to avoid the accidental "over-fitting". In order to resolve any conjecture of over-fitting, the final evaluation of the method has been performed on a set of four pre-set classifiers included: Complementary Naïve Bayes (NB), Multinomial Naïve Bayes, Support Vector Machine (SVM) (SMO in Weka) and Decision Trees (DT) (J48 in Weka). They were chosen because Naïve Bayes is known to work well with text, because SVM is a very good performer in general, and because DT's output in readable for humans.

Table 1. Comparison of the classification evaluation measures for different representation methods

Evaluation measure → / Representation/ Classifier used ↓	TP Rate Wtd. Avg.[8]	FP Rate Wtd. Avg.	Precision Wtd. Avg.	Recall Wtd. Avg.	F-Measure Wtd. Avg.	Accuracy %
BOW(TF-IDF)/ CompNB	0.772	0.025	0.744	0.772	0.743	77.22
LDA Topics/ Adaboost (j48)	0.693	0.034	0.679	0.693	0.684	69.33
BOW(TF-IDF) +LDA/ SVM(SMO)	0.8	0.022	0.786	0.8	0.79	80.00

[4] Each vector contains only 50 features corresponding to the 50 LDA clusters.

The evaluation measures calculated by the most stable classifier over the three representations are shown in table 1. This performance is acceptable, considering that manual general topic annotation is an uncertain task (even for the human beings). The uncertainty has roots in the following three aspects: 1) the topics in our list of 10 categories are sometimes too general; 2) the nature of the social network scattered postings (informal text using abbreviations that are not clear for everybody, etc.); 3) the subjectivity of the manual annotations; the reasons for some discrepancies between human annotations (with the same problem definition) could be tracked in their different personality, mood, background and some other subjective conditions. Human judgment is subjective and is not necessarily the same, among different people upon the same case. According to the related literature, when documents are annotated by more than one human annotator the expected agreement between judges is normally around 60-85% on different datasets [12], [13], [14]. Therefore, it is helpful to have a standard annotation system that always annotates based on some constant definitions, patterns and rules, as our automatic system does.

Our "general topic detection" method can be applied for trend detection purposes in any collaborative writing web sites in which people add or modify contents, in the style of posts/comments. It could also be handy for some web-logs or some specialist forums. It could also be adapted for some kinds of message categorization or even spam detection for any type of text messaging services on the internet or even on cellular phones.

6 Conclusion and Future Work

We designed and implemented an efficient "general topic detection" method over the "Friendfeed" social network textual dataset. The system applies LDA topical modeling estimation/inference for the topic detection purpose. The method also gets benefit from some classification algorithms for the purpose of general topic detection. The system is useful as standard general topic annotation applications, mostly in messaging services and collaborative writing web sites. Moreover, the performance of the system is similar to a range of comparable tasks.

There are many advantages of our method, including:

1) The LDA method automatically assigns topics to the posts/comments (via a small group of words clustered together). Then we manually interpret and generalize the clusters into small number (e.g., 10) of high-level classes (showed in section 4.2.). The remarkable advantage of this method is that the number of topical groups that need to be manually mapped to the general topics are far smaller than the number of postings of a social network corpus (or any corpus in general) that would need to be annotated to build training and testing sets manually.

2) In the LDA representation each document (thread) is represented by the LDA weighted membership distribution of the topical word clusters; hence any other high dimensional vector representation of any collection of documents can be also replaced by its LDA weighted membership distribution in order to reduce the dimensionality and consequently dealing with the curse of dimensionality. The lower dimensional representation can be used for any supervised/unsupervised machine learning algorithm which cannot be applied on high-dimensional data.

3) We observed that the quality of the topical clusters of the LDA algorithm improves simply by adding the 3500 background source data (threads extracted from the same corpus) to the original 500 threads selected for the supervised learning. This means that consequently the performance of our automatic general topic detection method is improved using *unlabeled* background source data.

One limitation of the current design is that it is *case insensitive*; it could be developed based on *case sensitive* texts in order to extract more specific topical keywords/phrases of the contents.

In future work, we are planning to replace the manual interpretation of the LDA topical word clusters with an automatic topic assignment. This idea could be realized by getting benefits from resources such as "Wordnet Domains".

References

1. Blei, D., Ng, A., Jordan, M.: Latent Dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
2. Teh, Y.-W., Newman, D., Welling, M.: A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. In: Procs. NIPS (2006)
3. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Sciences 101, 5228–5235 (2004)
4. Minka, T., Lafferty, J.: Expectation propagation for the generative aspect model. In: Proceedings of UAI (2002)
5. Wang, X., McCallum, A.: Topics over time: A non-markov continuous-time model of topical trends. In: Proceedings of KDD (2006)
6. Blei, D.M., McAulie, J.: Supervised topic models. In: Procs. of NIPS (2007)
7. Li, W., McCallum, A.: Pachinko allocation: Dag-structured mixture models of topic correlations. In: ICML (2006)
8. Hoffman, M., Blei, D., Bach, F.: Online learning for latent Dirichlet allocation. In: Proceedings of NIPS (2010)
9. Wahabzada, M., Kersting, K.: Larger residuals, less work: Active document scheduling for latent dirichlet allocation. In: Gunopulos, D., Hofmann, T., Malerba, D., Vazirgiannis, M. (eds.) ECML PKDD 2011, Part III. LNCS, vol. 6913, pp. 475–490. Springer, Heidelberg (2011)
10. Heinrich, G.: Parameter estimation for text analysis, Technical Report (For further information please refer to JGibbLDA at: `http://jgibblda.sourceforge.net/`)
11. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Synthetic Minority Oversampling Technique. Journal of Artificial Intelligence Research 16, 321–357 (2002)
12. Wang, A., Hoang, C.D.V., Kan, M.-Y.: Perspectives on crowdsourcing annotations for natural language processing. Language Resources and Evaluation, 1–23 (2012)
13. Ferschke, O., Daxenberger, J., Gurevych, I.: A Survey of NLP Methods and Resources for Analyzing the Collaborative Writing Process in Wikipedia (2012)
14. Fleischmann, K.R., Templeton, C., Boyd-Graber, J., Cheng, A.-S., Oard, D.W., Ishita, E., Koepfler, J.A., Wallace, W.A.: Explaining Sentiment Polarity: Automatic Detection of Human Values in Texts (2012) (to appear)