Social Information Extraction using probabilistic topic model

Dehao Li and Qi Shao

Abstract

Social micro blogging service has been one of the most important web services to publish and share different kinds of information. Information retrieval and extraction of the social information is essential work and important in the semantic web area. However, only a few studies have emphasized on social network analysis (SNA) which has risen to a field having important implications. Natural Language Processing (NLP) is very efficient and intelligent for processing amounts of textual information. Probabilistic Topic Model (PTM) (Steyvers & Griffiths, 2007) is one kind of statistical human language processing methodologies which needs sets of corpus as if indicating the desirable relations and dependencies. In this study, we will research addresses the SNA issue in relation to a new application of a classical PTM to demonstrate the feasibility of applying PTM to extract valuable information from social information.

Keywords: Social Information, Social Network Analysis, Information Extraction, Natural Language Processing, Probabilistic Topic Model.

1. Introduction

The recent decades witnessed a rapid proliferation of textual information available in digital form in a myriad of repositories. Social network is mostly referred to as a combination of Information Technologies to make communication and conversations into an interactive format. As a part of social network, social micro blogging service has been one of the most important web services to publish and share different kinds of information. The trend of utilizing social information is to leverage different features which reflect valuable information, including both textual contents and social structure. Social text contains rich information about author of the information, such as what the user is talking about, what the user is going through and so on, so social text is potentially useful in determining specific and common characteristics of an individual user.

The task of Information Extraction is to identify a predefined set of concepts in a specific domain, ignoring other irrelevant information, where a domain consists of a corpus of texts together with a clearly specified information need (Feilmayr, 2011). In contrast to the amount of studies in social network focusing on data storage and data retrieval, only a few studies have emphasized on SNA which has risen to a field having important implications. SNA is facilitated because the method allows large contexts and crowd sourced approaches, consisting of distinct textual phrases, to be mapped to similar prior experiential knowledge(Wassell, Rubin, & Frost, 2011). To achieve an extension of such an area, this research addresses the SNA issue in relation to a new application of a classical natural language processing (NLP) model, topic probabilistic model. In this study, we focus on a specific model (PTM) for investigating the information extraction of social information.

2. Objectives

This research aims to propose a simple NLP model based on social information and analyze the correlation of different sentences of social information to cluster users according to the connotative interests of their micro blogs.

3. Methodology

To achieve our objectives, four research issues are to be addressed. First, a list of Weibo users should be randomly chosen and their published information will be the raw data sample of our research. Second, raw data should be transformed into mathematic models by using probabilistic topic model. Then mathematic methodologies should be used to calculate the correlated topics. In addition, the correlated topics should be inflected to natural language and the users will be clustered into different groups according to the different topics.

Figure 1 shows the graphical model for what we will refer to as the “standard topic model” or Latent Dirichlet Allocation (LDA) (Chemudugunta & Steyvers, 2007). There are D sentences and sentence d has Nd words. α and β are fixed parameters of symmetric Dirichlet priors for the D sentence-topic multinomials represented by θ and the T topic-word multinomials represented by φ. In the generative model, for each sentence d, the Nd words are generated by drawing a topic t from the sentence-topic distribution p(z|θd) and then drawing a word w from the topic-word distribution p(w|z = t, φt). Point estimates for the θ and φ distributions can be computed conditioned on a particular sample, and predictive distributions can be obtained by averaging over multiple samples.

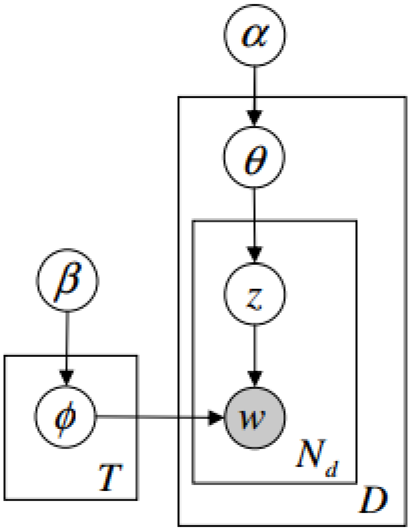

Figure 1. Standard topic model

The simple NLP which is mentioned above is based on vector space model (VSM). In the sample, VSMs in which each sentence of individual's social information is regarded as a vector are generated for each individual user. Each word in the sentence corresponds to a term. The value of a particular term which is called term frequency is the probability of its appearance in the sentence. A set of N sentences can be represented as an M*N matrix (e.g. Table 1). The VSM for each user calculates the similarity of vectors in a sentence and estimated the correspondence of the sentences. Suppose there are k topics in a sentence, we use linear algebra method to reduce the dimensionality of the matrices to make their ranks equal to k. If the two matrices of different sentences are linearly independent, the two sentences can be determined to have no correlations. Otherwise, the two sentences are correlated and the dimensionality k can be interpreted as the number of hidden topics in the sentences.

In Table 1, the similarities between sentence 2 (S_2) and sentence 3 (S_3) can be calculated as follows:

$$\mathrm{sim}(\overrightarrow{S_2}, \overrightarrow{S_3}) = \frac{0 \times 1 + 1 \times 0 + 1 \times 0 + 0 \times 0 + 0 \times 0}{|\overrightarrow{S_2}| \times |\overrightarrow{S_3}|} = 0$$

$\overrightarrow{S_2}$ and $\overrightarrow{S_3}$ are linearly independent, so it is determined that sentence 2 and sentence 3 have no similarities.

Table 1. Similarities between example sentences

Matrix	S_1	S_2	S_3	S_4	S_5	S_6
basketball	1	0	1	0	0	0
soccer	0	1	0	0	0	0
athlete	1	1	0	0	0	0
medal	1	0	0	1	1	0
man	0	0	0	1	0	1

The matrix decomposition is represented in Figure 2. The probabilities of each term are based on a corpus. The potential problem of such a model is that two sentences which have a synonym respectively may be regarded as linearly independent. To avoid the potential problem, the corpus is helped to maintain the consistency of different synonyms. If two terms are not synonyms, but are related in some other way that is important, you can designate them as related terms. The related terns should have the similar term frequency according to the corpus. By the way, polysemy should also be noticed in corpus for the reason that it has different tern frequency in different topics of different contexts. Choosing efficient corpus can help this research to achieve more accurate results.

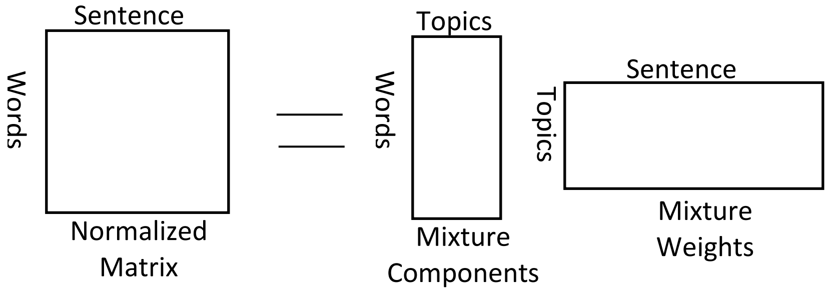

Figure 2. Matrix decomposition of sentences

4. Novelty

The largely unexplored problem of clustering user patterns is addressed in an unsupervised fashion using probabilistic topic models, and more specifically LDA (Jayagopi & Gatica-Perez, 2010). The main purpose of using PTM to analyze social information is to cluster different social individual into different groups according to their similar characteristic. PTM also addresses the problem of how to best build a vector space, which is an important gap of utilizing PTM. Instead of applying commonly used corpus depended vector space construction, this study develops a series of specific information based optimization on vector space constructing.

5. Conclusion and Significance

This study represents a contribution to the extension of PTM utilization. Our method to characterize social users by constructing vector space and then mining them using topic models is promising, allowing for the possibility of learning models to analyze group interests on large social corpora in an unsupervised way, and therefore saving a potentially huge annotation effort. The PTM model automatically discovered the topics based on co-occurrence of social information, and any meeting slices can be described as a probabilistic mixture over the discovered topics(Jayagopi & Gatica-Perez, 2010). In this regard, it contributes to the social network application by a practical implementation of PTM utilization.

PTM is widely used in information engineering and the studies adapts PTM to measure an external variable of interest, a difficult task for unsupervised learning that must be carefully validated. The information extraction problem is well addressed by teaming computer scientists with other scholars to use topic models to help explore, visualize, and draw hypotheses from their data. By working with scholars in diverse fields, PTM can be developed as a new interdisciplinary computational methodology for working with and drawing conclusions from archives of texts.

References

Chemudugunta, C., Smyth, P. and Steyvers, M. (2007). Modeling General and Specific Aspects of Documents with a Probabilistic Topic Model. *Proceedings of the 2006 Advances in Neural Information Processing Systems conference*, 19.

Feilmayr, C. (2011). Text Mining-Supported Information Extraction: An Extended Methodology for Developing Information Extraction Systems. Paper presented at *the 2011 22nd International Workshop on.Database and Expert Systems Applications* (DEXA).

Jayagopi, D. B., and Gatica-Perez, D. (2010). Mining group nonverbal conversational patterns using probabilistic topic models. *IEEE Transactions on Multimedia, 12*(8), 790-802.

Steyvers, M. and Griffiths, T. (2007). Probabilistic topic models. *Handbook of latent semantic analysis, 427*(7), 424-440.

Wassell, A., Rubin, S. H., and Frost, E. G. (2011). Integrated social information engineering. Paper presented at the *2011 IEEE International Conference on Information Reuse and Integration* (IRI).