A Method of Automatic Annotation for Medical Record Text Based on Latent Dirichlet Allocation

Xinyu Jin [1, a], Qiliang Jin [2,b] and Yuze Li[3,c]

Zheda Road 38th, Zhejiang University Hangzhou, China

[a]jinworker@126.com, [b]jinqiliang2007@163.com, [c]uncleleeee @gmail.com

Keywords: medical record text; semantic analysis; Latent Dirichlet Allocation; BM25;

Abstract. With the rapid development of medical information, medical data, especially medical record text, are difficult to intelligent analyses, because these data have loose grammar structure. Latent semantic analysis technology in the field of text mining in recent years made extensive research and application, and Latent Dirichlet Allocation(LDA), put forward by Blei, is a method to solve those difficulties. This paper proposed an improved LDA based on BM25 mixture weights method to analyze Chinese medical record text and had a good performance.

Introduction

In China, there is a lot of unstructured text of the medical record text, and medical terms in these records are always unstandardized and same medical concept may have different description. Therefore, extracting the useful information in the text is difficult, the same to intelligent analysis. Through the manual annotation of semantic net, Lord and his partners put similarity measurement method on the annotation of gene function and established contact between cells and gene products with the semantic hierarchy[1][2]. These researches are on the existing manual annotation of the semantic net. However, Latent semantic analysis technology can find potential semantic association by itself with existing semantic association. Chute and his partners realize the classification and retrieval of medical records through Latent Semantic Indexing(LSI)[3]. Bo Li suggested an automatic annotation method of medical text based on latent semantic model and semantic tree[4]. In this paper, based on the weighted BM25-LDA, introduces a method of extracting the implicit information in the text and automatic labelling the record text.

Latent Dirichlet Allocation

Latent semantic analysis, a branch of semantic analysis, starts from the LSI[5]. By definition of matrix decomposition vector space model, LSI obtains latent semantic information. On the basis of this, Hofmann put forward Probabilistic Latent Semantic Indexing(pLSI)[6]. Blei proposed LDA, makes up for the pLSI that it doesn't define in the document generation probability and other defects[7]. The graphical model representation of LDA is shown in Fig.1.

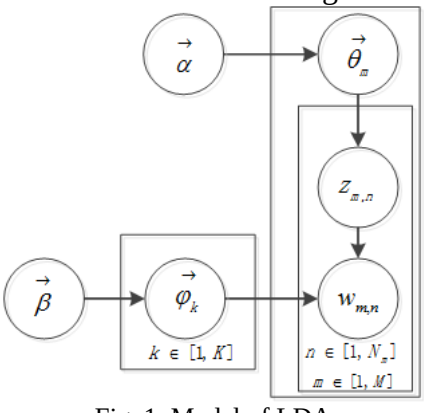

Fig. 1. Model of LDA

In LDA, a document is generated through the following process:

1. Choose length $N \sim Poisson(\xi)$

2. Choose $\theta \sim Dirichlet(\alpha)$
3. For each word w_n in the document
   a. Choose a topic $z_n \sim Multinomial(\theta)$
   b. Choose a word w_n ruled by $p(w_n|z_n;\beta)$, a multinomial distribution under the condition of z_n

Gibbs Sampling of the LDA Model

Currently, the main training algorithm of LDA is Gibbs sampling, put forward by Griffths[8]. $w_{m,n}$ is the n-th word of m-th document, and the latent semantic of this word is $z_{m,n}$. For sampling, we mark the n-th word of m-th document as coordinate i, and $-i$ means all the coordinates but i. We use conditional probability $p(z_i|z_{-i},w)$ to simulate $p(z|w)$, and the equation is as Eq. 1[10]:

$$
\begin{aligned}
p(\mathbf{z}_i = k \mid \vec{\mathbf{z}}_{-i}, \vec{\mathbf{w}}) &= \frac{p(\vec{\mathbf{w}} \mid \vec{\mathbf{z}})}{p(\vec{\mathbf{w}}_{-i} \mid \vec{\mathbf{z}}_{-i}) p(\mathbf{w}_i)} \cdot \frac{p(\vec{\mathbf{z}})}{p(\vec{\mathbf{z}}_{-i})} \\
&\propto \frac{\Delta(\vec{\mathbf{n}}_z + \vec{\boldsymbol{\beta}})}{\Delta(\vec{\mathbf{n}}_{z,-i} + \vec{\boldsymbol{\beta}})} \cdot \frac{\Delta(\vec{\mathbf{n}}_m + \vec{\boldsymbol{\alpha}})}{\Delta(\vec{\mathbf{n}}_{m,-i} + \vec{\boldsymbol{\alpha}})} \\
&\propto \frac{n_{k,-i}^{(t)} + \beta_t}{\sum_{t=1}^{V} n_{k,-i}^{(t)} + \beta_t} \cdot \frac{n_{m,-i}^{(t)} + \alpha_k}{[\sum_{k=1}^{K} n_m^{(t)} + \alpha_k] - 1}
\end{aligned}
\tag{1}
$$

The multinomial distribution of the t-th word of the k-th latent semantic is as follows[10]:

$$
\varphi_{k,t} = \frac{n_k^{(t)} + \beta_t}{\sum_{k=1}^{K} n_m^{(k)} + \beta_t} \tag{2}
$$

The multinomial distribution of the k-th latent semantic of m-th document is as follows[10]:

$$
\theta_{m,k} = \frac{n_m^{(k)} + \alpha_k}{\sum_{k=1}^{K} n_m^{(k)} + \alpha_k} \tag{3}
$$

Application

Preprocessing. Medical record text is a series of inspection, diagnosis and treatment of a patient by his doctor. There are two parts in the process. The first is Chinese word segmentation. There are several algorithms in this field: based on a dictionary matching, based on the statistical model, and based on knowledge understanding[8]. In this paper, we choose hybrid segmentation: use special dictionary to segment, and use word segmentation model which trained by mandarin corpus to segment. The second is to build the word vector of medical text.

Automatic Annotation. The semantic model after train by LDA based on BM25 weighted contains two type of information:

1. Multinomial distribution of vocabulary of each semantic $\{\varphi_1, \cdots, \varphi_k\}$
2. Multinomial distribution of each semantic of each document $\{\theta_1, \cdots, \theta_k\}$

Automatic annotation means using the most property semantic express each element of $\{\varphi_1, \cdots, \varphi_k\}$, and two problem need to be deal with:

1. Semantic label form and the generating method of topic
2. Relevance of topic and semantic label

The process is as follows:

for $\varphi_k \in \{\{\varphi_1, \cdots, \varphi_k\}$

{

Generate semantic tagging candidate set $\{\{l_1, \cdots, l_{n_k}\}$

Get score of each candidate in $\{\{l_1, \cdots, l_{n_k}\}$ with scoring function

Choose the highest correlation labels as the topic of φ_k

}

Result. We take medical record text from a hospital of Hangzhou, China. And the result is shown in Table 1.

vocabulary of semantic (top 10)	Automatic label	Manual label
dizziness, chronic, spirit, head, customs, consciousness, pathological, soft, the nervous system, well	The nervous system Consciousness. sober	a description of the nervous system
miscellaneous, ry-wet,breathing,rale,pathological,double lung,superficial,the trachea,sound,shortness of breath	breathing the trachea rale	some inspection and description of the lungs
auscultation, noise, valves, regular, heart rate, pathological, negative, heart, chest tightness, premature beat	Rhythm of the heart negative regular	some examination and description of the heart

Table 1. The comparison of manual annotation and machine annotation (translated from Chinese).

Table 1 shows that machine can express the topic type. And we get a tip that this algorithm can extract the information of medical record text in a certain extent.

Conclusion

This paper introduces medical record text and the difficulties of records analysis. Then we describe semantic analysis and LDA. And we use LDA based on BM25 weighted to test the machine labeling and give the result. The result tells that LDA may help us classify the meaning of text paragraphs. And we can get a computer-aided system to help doctors improving the efficiency of diagnosis and treatment.

Acknowledgements

Research is supported by the opening foundation of the State Key Laboratory for Diagnosis and Treatment of Infectious Diseases, The First Affiliated Hospital of Medical College， Zhejiang University, grant NO. 2014KF06，supported by Software technology innovation comprehensive 2013 pilot project in zhejiang province and the National science and technology major projects, a new generation of broadband wireless mobile communication network NO.2013ZX03005013.

References

[1] Lord P W, Stevens R D, Brass A, et al. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation.[J]. Bioinformatics, 2003, 19(10):1275-1283(9).

[2] Ashburner M, Ball C A, Blake J A, et al. Gene Ontology: tool for the unification of biology[J]. Nature Genetics, 2000, 25(1):25-29.

[3] Chute C G, Yang Y, Evans D A. Latent Semantic Indexing of medical diagnoses using UMLS semantic structures.[C]// Proceedings / the ... Annual Symposium on Computer Application [sic] in Medical Care. Symposium on Computer Applications in Medical Care1991:185-9.

[4] Bo Li. Analysis Model of Medical Text and Image based on LDA and LSA and its Application. Jilin University, 2012.(in Chinese)

[5] Dumais S, Furnas G, Landauer T, et al. Latent semantic indexing[C]. Proceedings of the Text Retrieval Conference. 1995.

[6] Hofmann T. Probabilistic latent semantic indexing[C]. Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1999: 50-57.

[7] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. the Journal of machine Learning research, 2003, 3: 993-1022.

[8] Griffiths T. Gibbs sampling in the generative model of Latent Dirichlet Allocation[J]. Standford University, 2002.

[9] Wenfeng Cao. Key technologies of Chinese word segmentation. Nanjing University of Science and Technology, 2009.(in Chinese)

[10]Yuze Li. Research and Apply on Patient Record Text Mining Based on Latent Semantic Analysis. Zhejiang University, 2015.(in Chinese)