ISSN: 2348 - 2117

International Journal of Engineering Technology and Computer Research (IJETCR)
Available Online at www.ijetcr.org
Volume 3; Issue 2; Page No. 59-62

LIMTM: A FRAMEWORK FOR ASSIMILATING LINK BASED IMPORTANCE INTO SEMANTICALLY COHERENT CLUSTERS OF CORRELATED WORDS

Kanagadurga Natarajan [1], M.Sivasundaravinayagamoorthy [2]

[1] Final M.E Computer Science and Engineering, Department of Computer Science and Engineering, A.V.C College of Engineering, Tamil Nadu, India.

kanagadurga17@gmail.com

[2] Assistant Professor, Department of Computer Science and Engineering, A.V.C College of Engineering, Tamil Nadu, India.

ABSTRACT

As more information becomes available, it is getting harder and harder to find what we are looking for like finding a needle in the haystack. In today's world where most of the information is electronically stored new tools which help to organize, search and understand information are need of the hour. Topic can be described as "a recurring pattern of co-occurring words". Topic models are used to discover hidden topic based patterns. Using those discovered topics collection of documents can be annotated. Using those annotations documents can be organized, understood, summarized and searched. Topic modeling has become a well known text mining method and is widely used in document navigation, clustering, classification and information retrieval. Given a set of documents, the goal of topic modeling is to discover semantically coherent clusters of correlated words known as topics, which can be further used to represent and summarize the content of documents. By using topic modeling, documents can be modeled as multinomial distributions over topics instead of those over words. Topics can serve as better features of documents than words because of its low dimension and good semantic. Interpretability Topic modeling has become a widely used tool for document management. However, there are few topic models distinguishing the importance of documents on different topics. A framework LIMTM (Link based Importance into Topic Modeling) is used to incorporate link based importance into topic modeling. Specifically, ranking methods are used to compute the topical importance of documents.

Keywords: Text Corpus, Topic Modeling, Link based Importance, Ranking, and Log Likelihood

INTRODUCTION:

Topic modeling is a widely known text mining method and is conventionally used in document navigation, clustering, classification and information retrieval because of its propitious application performance. The objective of topic modeling is to uncover semantically coherent clusters of correlated words known as topics from a collection of documents. The discovered topics can be further used to represent and summarize the content of documents. The customary topic models include PLSA (Probabilistic Latent Semantic Analysis) and LDA (Latent Dirichlet Allocation). By deploying topic modeling, documents can be modeled as multinomial distributions over topics instead of those over words. Topics are better features of documents than words due to its low dimension and good semantic interpretability.

Existing topic models do not explicitly discriminate the importance of documents on different topics, while in non-theoretical situations documents have different degrees of importance on different topics, thus viewing them as equally important may inherently hurt the performance of topic modeling. Ranking methods are initially proposed for the purpose of ranking web pages. Since concepts and entities in those domains are similar it can be also used to rank other kind of documents. To specify the importance of documents on different topics, topical ranking methods can be used which is an extension of basic ranking algorithms such as page rank and HITS (Hyperlink-Induced Topic Search). The proposed work incorporates link based importance into topic modeling.

Topical ranking method computes the importance scores of documents over topics which are then leveraged to aid the topic modeling process. The proposed framework, Link Importance Based Topic Model denoted as LIMTM for short. When comparing to existing topic models, LIMTM discriminates the importance of documents while accomplishing topic modeling. The ideology behind the methodology is that the more important documents are given more weights than the less important ones.

Corresponding author: Kanagadurga Natarajan

1. RELATED WORK:

Topic Models:

The prior link combined topic models can capture the topical correlations between linked documents but does not leverage the topical ranking of documents to aid the topic modeling process. The proposed framework closely resembles the Topic Flow model. The distinguished features of proposed work from Topic Flow lie in the following two folds. First, LIMTM provides a more flexible combination between link importance and topic modeling while Topic Flow couples flow network and topic modeling tightly. This feature makes LIMTM more extendable. Second, LIMTM builds a generalized relation between link importance and topic modeling rather than a hard relation like Topic Flow.

Ranking:

The proposed work is tightly related to ranking technology. The most customary link based ranking algorithms are Page Rank and HITS. The other commonly used ranking algorithms include FurtureRank, P-Rank and RankClus. When compared to RankClus which performs ranking based on hard clustering, LIMTM incorporates link based importance into topic modeling which is a soft clustering. Another difference is that RankClus is a clustering algorithm based on only links while LIMTM is a topic modeling framework based on both links and texts.

2. PRELIMINARIES:

Topic Modeling:

The objective of topic modeling is to extract conceptually coherent topics that are shared by a set of documents. LIMTM framework is built over PLSA topic model. LIMTM can be regarded as PLSA with informative prior.

Given a collection of N documents D, let V be the total number of unique words in the vocabulary and K denote the number of topics, the goal of PLSA is to maximize the likelihood of the collection of documents with respect to model parameters Θ and B.

$$P(D|\Theta,B)=\prod_{i=1}^{N}\prod_{w=1}^{V}\left(\sum_{z=1}^{K}\theta_{iz}\beta_{zw}\right)^{s_{iw}} \quad \text{... (1)}$$

Where $\Theta = \{\theta\}_{N*K}$ is the topic distribution of documents, $B = \{\beta\}_{K*V}$ is the word distribution of topics, and s_{iw} represents the times that word w occurs in document i.

After the conjecture of PLSA, each topic is represented as a distribution over words in which top probability words form a semantically coherent concept, and each document can be represented as a distribution over the discovered topics.

Topical Link Importance:

The documents' global importance computed based only on the link structure of the document network is given by Link importance. Topical page rank and topical HITS (Topical link analysis) are proposed by LIMTM.

As the input of topical page rank, each document i is associated with a topic distribution θ_i, which can be obtained via topic modeling methods. Taking θ_i into account, topical page rank produces an importance vector for each document, in which each element represents the importance score of the document on each topic. Letting γ_{zi} denote the importance of document i on topic z, topical page rank is formally expressed as

$$\gamma_{zi}^{(t)}=\lambda\sum_{j\in I_i}\frac{\alpha\gamma_{zj}^{(t-1)}+(1-\alpha)\theta_{jz}\gamma_{j}^{(t-1)}}{|O_j|}+(1-\lambda)\frac{\theta_{iz}}{M} \quad \text{... (2)}$$

Where α and λ are parameters that control the process of prorogating the ranking score, which are both empirically set to 0.85. $\gamma_{.j} = \sum_{z=1}^{K}\gamma_{zi}$ denotes the global importance of document j, I_i is the set of in-link neighbors of document i, $|O_j|$ denotes the number of out-link neighbors of document j, and θ_{jz} is the topic proportion of document j on topic z and M is the total number of documents.

3. LIMTM FRAMEWORK:

Relation between link importance and topic modeling:

The link importance γ_{zi} can be interpreted as the probability $P(i|z)$ of the node i involved in the topic z by normalizing the importance vector such that $\sum_{i=1}^{M}P(i|z)=1,\forall z$. By using the sum and product rules of the Bayesian theorem, the topic proportion $P(z|i)$ can be expressed in terms of γ_{zi}.

$$\theta_{iz}=P(z|i)=\frac{P(i|z)p(z)}{\sum_{i'=1}^{M}P(i|z)p(z)}=\frac{\gamma_{zi}\pi_z}{\sum_{z'=1}^{K}\gamma_{z'i}\pi_{z'}} \quad \text{... (3)}$$

Where $\pi_z = P(z)$ is the prior probability of topic z.

To reduce the effects of noise, the degree of belief on the link importance is modeled instead of removing the noise links. A parameter ξ is introduced ranging from 0 to 1 to indicate belief on the link importance.

$$\theta_{iz}=P(z|i)\alpha\left[\xi\gamma_{zi}+(1-\xi)\phi_{zi}\right]\pi_z \quad \text{... (4)}$$

Where $\phi_{zi} = p(i|z)$ has the same interpretation as γ_{zi}.

LIMTM Framework:

Different from the traditional topic models, the probability $p(i|z)$ of a document i involved in a topic z is governed by the weighted mixture of topical ranking γ_{zi} and the hidden variable ϕ_{zi} in the LIMTM model such that the effects of link importance on topic modeling is integrated.

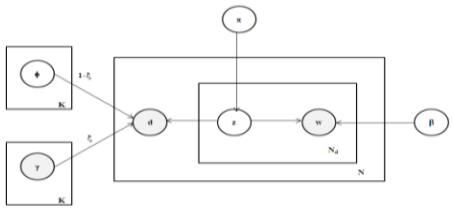

Figure 1: LIMTM Framework

In LIMPTM, the topical link importance γ of documents is labeled as observational variable since it can be obtained by the topical pagerank or topical HITS algorithm, although in an overall view topical link importance is in fact unknown. By incorporating topical link importance γ_{zi} into the topic modeling, the link information is naturally taken into account since the topical link analysis process is performed on the link structure.

In LIMTM Framework, the likelihood of a collection of documents D with respect to the model parameters is

$$P(D|\gamma,\pi,\phi,\beta) = \prod_{i=1}^{M}\prod_{w=1}^{V}\left(\sum_{z=1}^{K}\left[\xi\gamma_{zi} + (1-\xi)\phi_{zi}\right]\pi_z\beta_{zw}\right)^{s_{iw}} \quad \ldots (5)$$

Derivation of LIMTM:

The maximum likelihood estimation is adopted to derive the model parameters involved in LIMTM which is obtained by the expectation maximization (EM) algorithm.

The logarithm of the likelihood function is

$$L = \log P(D|\gamma,\pi,\phi,\beta)$$
$$= \sum_{i=1}^{M}\sum_{w=1}^{V} s_{iw}\log\sum_{z=1}^{K}\beta_{zw}\left[\xi\gamma_{zi} + (1-\xi)\phi_{zi}\right]\pi_z \quad \ldots (6)$$

In the E step the posterior distribution $P(z|i,w)$ of topics conditioned on each document-word pair (i,w) is computed by

$$\psi_{iwz}^{(t)} = P^{(t)}(z|i,w)\alpha\beta_{zw}^{(t)}\left[\xi\gamma_{zi} + (1-\xi)\phi_{zi}^{(t)}\right]\pi_z^{(t)} \quad \ldots (7)$$

Then, the lower bound of L can be derived by using Jensen inequality twice as follows,

$$L = \sum_{i=1}^{M}\sum_{w=1}^{V} s_{iw}\log\sum_{z=1}^{K}\psi_{iwz}^{(t)}\frac{\beta_{zw}\left[\xi\gamma_{zi} + (1-\xi)\phi_{zi}\right]\pi_z}{\psi_{iwz}^{(t)}}$$
$$\geq \sum_{i=1}^{M}\sum_{w=1}^{V} s_{iw}\sum_{z=1}^{K}\psi_{iwz}^{(t)}\log\beta_{zw}\left[\xi\gamma_{zi} + (1-\xi)\phi_{zi}\right]\pi_z - \sum_{i=1}^{M}\sum_{w=1}^{V} s_{iw}\sum_{z=1}^{K}\psi_{iwz}^{(t)}\log\psi_{iwz}^{(t)}$$
$$\geq \sum_{i=1}^{M}\sum_{w=1}^{V} s_{iw}\sum_{z=1}^{K}\left[\xi\psi_{iwz}^{(t)}\log\beta_{zw}\gamma_{zi}\pi_z + (1-\xi)\psi_{iwz}^{(t)}\log\beta_{zw}\phi_{zi}\pi_z\right] - \sum_{i=1}^{M}\sum_{w=1}^{V} s_{iw}\sum_{z=1}^{K}\psi_{iwz}^{(t)}\log\psi_{iwz}^{(t)} \quad \ldots (8)$$

In the M-step, the lower bound of L is maximized under the constraint $\sum_{w=1}^{V}\beta_{zw} = 1$, $\sum_{z=1}^{K}\pi_z = 1$ and $\sum_{i=1}^{M}\phi_{zi} = 1$. Through Lagrange multipliers, the constrained maximization problem is converted to the following.

$$max_{\theta,\pi}\sum_{i=1}^{M}\sum_{w=1}^{V} s_{iw}\sum_{z=1}^{K}\left[\xi\psi_{iwz}^{(t)}\log\beta_{zw}\gamma_{zi}\pi_z + (1-\xi)\psi_{iwz}^{(t)}\log\beta_{zw}\phi_{zi}\pi_z\right]$$

$$+\sum_{z=1}^{K}\lambda_z\left(\sum_{w=1}^{V}\beta_{zw}-1\right)+\lambda\left(\sum_{z=1}^{K}\pi_z-1\right)+$$
$$\sum_{z=1}^{K}\lambda'_z\left(\sum_{i=1}^{M}\phi_{zi}-1\right) \quad \ldots (9)$$

The maximization problem has a closed form which gives out the update rules that monotonically increase L.

$$\beta_{zw}^{(t+1)} \alpha \sum_{i=1}^{M} s_{iw}\psi_{iwz}^{(t)}$$
$$\pi_z^{(t+1)} \alpha \sum_{i=1}^{M}\sum_{w=1}^{V} s_{iw}\psi_{iwz}^{(t)}$$
$$\phi_{zi}^{(t+1)} \alpha \sum_{w=1}^{V} s_{iw}\psi_{iwz}^{(t)} \ldots (10)$$

As the parameter updating process converges, the topic proportion θ can be computed by removing noise.

4. CONCLUSION:

Thus the proposed LIMTM framework has incorporated Link based importance into Semantically coherent cluster of correlated words known as Topics as a unified framework. LIMTM framework distinguishes the importance of documents on different topics using Topical PageRanking algorithm. The performance of Topical PageRank lies in the features such as generalization performance, document clustering and classification, topic interpretability, and document network summarization performance. Topic modeling provides flexibility because it can be used according to specific application requirements.

5. REFERENCES:

1. Alper Kursat Uysal and Serkan Kunal, "The Impact of Preprocessing on Text Classification", IEEE Transactions on Information processing and management, Volume 50, Issue 1, January 2014.
2. Dongsheng Duan, Yuhua Li, Ruixuan Li, Rui Zhang and Aiming Wen, "Rank Topic: Ranking Based Topic Modeling", IEEE Transactions on Data Mining, Volume 12, December 2013.
3. Jia Zeng, William K. Cheung, Chun-hung Li and Jiming Liu, "Multi-relational Topic Models", IEEE Transactions on Data Mining, Volume 7, December 2012.
4. Mark A. Greenwood, "Implementing a Vector Space Document Retrieval System", Dept. of Computer Science, University of Sheffield, Electronic copy and resources available from http://www.dcs.shef.ac.uk /~ mark/ follow the PhD link.
5. Michal Rosen-Zvi, Thomas Griffi ths, Mark Steyvers and Padhraic Smyth, "The Author-Topic Model for Authors and Documents", published in UAI '12 Proceedings of the 20th conference on Uncertainty in artificial intelligence.