Combining the Missing Link: An Incremental Topic Model of Document Content and Hyperlink

Huifang Ma, Zhixin Li, Zhongzhi Shi

Combining the Missing Link: an Incremental Topic Model of Document Content and Hyperlink

Huifang Ma[1,2], Zhixin Li[1,2], Zhongzhi Shi[1]

[1] Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences,100080,Beijing,China
[2] Graduate School of the Chinese Academy of Sciences,100039,Beijing,China
{mahf, lizhixin, shizz}@ics.ict.ac.cn

Abstract. The content and structure of linked information such as sets of web pages or research paper archives are dynamic and keep on changing. Even though different methods are proposed to exploit both the link structure and the content information, no existing approach can effectively deal with this evolution. We propose a novel joint model, called Link-IPLSI, to combine texts and links in a topic modeling framework incrementally. The model takes advantage of a novel link updating technique that can cope with dynamic changes of online document streams in a faster and scalable way. Furthermore, an adaptive asymmetric learning method is adopted to freely control the assignment of weights to terms and citations. Experimental results on two different sources of online information demonstrate the time saving strength of our method and indicate that our model leads to systematic improvements in the quality of classification.

Keywords: Topic model; Link-IPLSI; Incremental Learning; Adaptive Asymmetric learning

1 Introduction

Obtaining multi-side semantic information from a topic report containing dynamic online data streams is useful both from a theoretical point of view, as there are many complex phenomena to be addressed, and from purely practical applications such as topic modeling. A variety of techniques for automatically extracting thematic content of a set of documents are proposed, such as latent semantic indexing(LSI)[1], probabilistic latent semantic indexing (PLSI)[2]. The topics learned by a topic model can be regarded as themes discovered from documents sets, while the topic-term distributions focus on the high probability words that are relevant to a theme.

With lots of electronic documents connected with hyperlinks/citations can be easily and readily acquired through the Internet, scholars demonstrate an increasing academic interest concerning how to effectively construct models for these correlated hypertexts. Automatic techniques to analyze and mine these document collections are at the intersection of the work in link analysis [3, 4], hypertext and Web mining [5, 6]. The most well known algorithms in link mining are PageRank [7] and HITS [8]. Both algorithms exploit the hyperlinks of the Web to rank pages based on their levels of "prestige" or "authority". Link mining encompasses a wide range of tasks [9] and we focus on the core challenges addressed by a majority of ongoing research in the field of topic modeling.

There are many noteworthy works. Cohn and Chang [10] introduced PHITS as a probabilistic analogue of the HITS algorithm, attempting to explain the link structure in terms of a set of latent factors. One of the first efforts in applying topic models to modeling both citation and content came from Cohn and Hoffman [11], they constructed Link-PLSI to integrate content and connectivity together. Erosheva et al. [12] defined a generative model for hyperlinks and text and thereby modeled topic specific influence of documents. We refer to this model as Link-LDA. Nallapati et al. [13] addressed the problem of joint modeling of text and citations in the topic modeling framework and presented two different models. Gruber et al. [14] recently presented a probabilistic generative model LTHM for hypertext document collections that explicitly models the generation of links.

These methods, however, are not suitable to be applied to the changing situation as the links and documents are probably only valid at a certain time. When new documents and a set of inter-connections are added, the existing model should be updated correspondingly. A similar situation happens when part of old documents and citations are deleted. A naïve approach to catch the update of links and contents is to rerun the batch algorithm from scratch on all existing data each time new data comes in, which is computationally expensive. Another obvious shortcoming for the naïve approach is that after re-running of batch algorithm, changes to the links and contents themselves can not be captured with the content of latent topics maintained.

As for incremental learning of topic modeling, Chien et al. [15] proposed an incremental PLSI learning algorithm which efficiently updates PLSI parameters using the maximum a posterior. Chou et al. [16] introduced another Incremental PLSI (IPLSI), aiming to address the problem of online event detection. Although these models capture the basic concept of incremental learning for PLSI, their weakness is that they do not take additional link information into consideration. Even so, these models offer excellent foundations on which to build our model.

In this paper, we present a new model Link-IPLSI, which extends the existing Link-PLSI by modeling interactions between document and link structure incrementally. In contrast to PLSI and Link-PLSI, the new model processes incoming online documents incrementally for each time period, discards out-of-date documents, terms and links not used recently, and folds in new terms, links and documents with the latent semantic indices preserved from one time period to the next. To the best of our knowledge, there is no previous work constructing the interconnected documents incrementally.

This paper has the following technical contributions:

- Present an incremental Link-PLSI model, owning the ability to identify meaningful topics while reducing the amount of computations by maintaining latent topics incrementally.
- Extend Link-PLSI model for updating two modalities simultaneously, which supports addition/deletion for both terms and citations.
- By means of integrating link information, our incremental model takes advantage of adaptive asymmetric learning method to weigh terms and links respectively.

In a word, this paper presents an incremental topic model that is applicable to a set of dynamic interconnected data. We have applied this method to both cited and hyperlinked data sets. Experiments show that our method is effective for topic modeling and works more efficiently than the corresponding batch method.

The remainder of this paper is organized as follows: In Section 2, we introduce the Link-PLSI model and its principles. In Section 3, we give detailed information on our proposed Link-IPLSI model. Section 4 describes the test corpora, the performance measures and the baseline method together with the experiment results. We conclude and discuss future work in Section 5.

$P(c_r|z_k)$ c

$P(d_i)$ $P(z_k|d_i)$

d z

$P(w_j|z_k)$ w

Note that in the rest of the paper, we use the terms “citation” and “hyperlink” interchangeably. Likewise, the term “citing” is synonymous to “linking” and so is “cited” to “linked” [13].

2 Link-PLSI Model

Link-PLSI [11] is based on the assumption that similar decomposition of the document term co-occurrence matrix can be applied to the cite-document co-occurrence matrix in which each entry is a count of appearances of a linked-document (or citation) in a source document. Under this assumption, a document is modeled as a mixture of latent topics that generates both terms and citations. A representation of Link-PLSI model is depicted in Fig. 1.

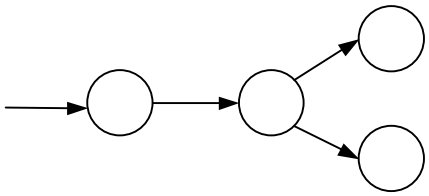


Fig. 1 Representation of Link-PLSI model

The model is defined as a generative process: a document d_i is generated with some probability $P(d_i)$, a latent topic z_k associated with documents, terms and citations is chosen probabilistically so that their association can be represented as conditional probabilities $P(w_j|z_k)$, $P(c_r|z_k)$ and $P(z_k|d_i)$. The following joint model for predicting citations/links and terms in documents is defined as:

$$P(c_r \mid d_i) = \sum_k P(c_r \mid z_k) P(z_k \mid d_i), \tag{1}$$

$$P(w_j \mid d_i) = \sum_k P(w_j \mid z_k) P(z_k \mid d_i), \tag{2}$$

where w_j represents one term and c_r indicates one citation, c and d both refer to document in the corpus and they may be identical. They are kept separate notationally to reinforce different roles they play in the model, c is conveyed by being cited and d is conveyed by citing [11].

An EM algorithm is used to compute the parameters $P(w_j|z_k)$, $P(c_r|z_k)$ and $P(z_k|d_i)$ through maximizing the following log-likelihood function with a relative weight α of the observed data:

$$\begin{aligned} L = \sum_i [\alpha \sum_j \frac{N_{ji}}{\sum_{j'} N_{j'i}} \log \sum_k P(w_j \mid z_k) P(z_k \mid d_i) \\ + (1-\alpha) \sum_r \frac{A_{ri}}{\sum_{r'} A_{r'i}} \log \sum_k P(c_r \mid z_k) P(z_k \mid d_i)], \end{aligned} \tag{3}$$

where N_{ji} is the count of term w_j in document d_i and A_{ri} is the count of citation c_r from document d_i. The steps of the EM algorithm are described as follows:

E-step. The conditional distributions $P(z_k|d_i, w_j)$ and $P(z_k|d_i, c_r)$ are computed from the previous estimate value of the parameters $P(w_j|z_k)$, $P(c_r|z_k)$ and $P(z_k|d_i)$:

$$P(z_k \mid d_i, w_j) = \frac{P(z_k \mid d_i)P(w_j \mid z_k)}{\sum_k P(z_k \mid d_i)P(w_j \mid z_k)}, \tag{4}$$

$$P(z_k \mid d_i, c_r) = \frac{P(z_k \mid d_i)P(c_r \mid z_k)}{\sum_k P(z_k \mid d_i)P(c_r \mid z_k)}. \tag{5}$$

M-step. The parameters $P(w_j|z_k)$, $P(c_r|z_k)$ and $P(z_k|d_i)$ are updated with the ne w expected values $P(z_k|d_i, w_j)$ and $P(z_k|d_i, c_r)$:

$$P(w_j \mid z_k) = \sum_i \frac{N_{ji}}{\sum_{j'} N_{j'i}} P(z_k \mid d_i, w_j), \tag{6}$$

$$P(c_r \mid z_k) = \sum_i \frac{A_{ri}}{\sum_{r'} A_{r'i}} P(z_k \mid d_i, c_r), \tag{7}$$

along with the mixing proportions

$$\begin{aligned} P(z_k \mid d_i) \propto \alpha \sum_j \frac{N_{ji}}{\sum_{j'} N_{j'i}} P(z_k \mid d_i, w_j) \\ + (1-\alpha) \sum_r \frac{A_{ri}}{\sum_{r'} A_{r'i}} P(z_k \mid d_i, c_r). \end{aligned} \tag{8}$$

3 A Joint Incremental Link-PLSI For Content And Hyperlink

The topic modeling process often requires simultaneous model construction and testing in an environment which constantly evolves over time. It is assumed that the most effective topic model to be used under such environment does not stay constant over time, but varies with progression of the data stream.

For the effective update of contents and links when new documents are added or old linked-data are removed, we propose an incremental approach to Link-PLSI technique, which is referred to as Link-IPLSI. The basic idea of our updating algorithm is straightforward: the Link-IPLSI model is performed on the initial linked-documents at the beginning. When a set of new documents are added introducing new terms and citations, a cycle will be created for folding in these documents, terms and citations and the model is then updated during the cycle. For new adding of documents or removing of old ones, our model adjusts both term-topic and link-topic probabilities at the lowest cost.

3.1 Preprocessing

The preprocessing phase is the first step for the incremental learning, involving elimination of out-of-date documents, terms and hyperlinks. The corresponding

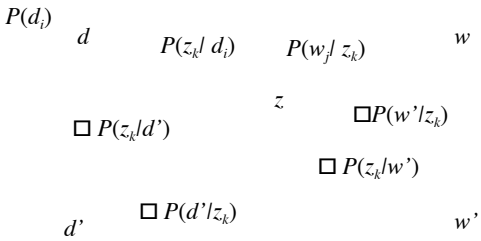


Fig. 2. Illustration of Link-IPLSI model

parameters $P(w_{out}/z)$, $P(c_{out}/z)$ and $P(z|d_{out})$ are removed. (d_{out} is an out-of-date document, and so are w_{out} and c_{out}) We can not simply augment the model directly, as the basic principle of probability that the total probability will be equal to one should be observed, the remaining parameters need to be renormalized proportionally:

$$P(w \mid z) = \frac{P_0(w \mid z)}{\sum_{w' \in W_0} P_0(w' \mid z)}, \tag{9}$$

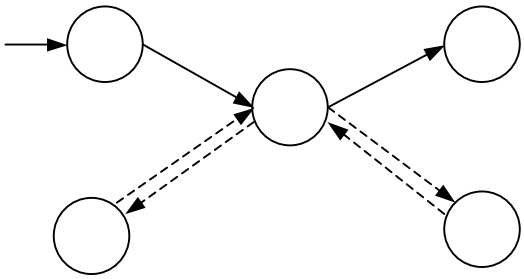
$$P(c \mid z) = \frac{P_0(c \mid z)}{\sum_{c' \in C_0} P_0(c' \mid z)}, \tag{10}$$

where $P_0(w/z)$ and $P_0(c/z)$ stand for the probabilities of the remaining terms and citations, whereas W_0 and C_0 are the respective sets of remaining terms and citations.

3.1 Incremental Link-PLSI technique

In this section, we give a detailed illustration of Link-IPLSI. The novelty of our model is that it takes advantage of the existing information to handle the streaming data without retraining the model. Therefore, the model is much faster and more scalable which makes model construction easier than that of the batch system. Fig. 2 is an illustration of sequences for updating related information of Link-IPLSI, where *d'* and *w'* indicate new documents and new terms respectively.

As the figure shows, new documents should first be folded in with old terms and links fixed, and then $P(d'/z_k)$ are calculated which sets a foundation for folding in new terms and links in the followings. In this way, $P(w_{all}/z_k)$, $P(c_{all}/z_k)$ and $P(z_k/d_{all})$ are updated as better initial values for the final EM algorithm, which guarantees a faster convergence. (d_{all} is a final document in the entire document set, and so are w_{all} and c_{all}). A specified illustration is given below.



Fold in new document. There is a need to realize how many data have already been well explained by the existing model in order to integrate the streaming data into the model effectively. Using a partial version of EM algorithm, folding-in, we can update the unknown parameters with the known parameters fixed so as to maximize the likelihood with respect to the previously trained parameters. Obviously, documents should first be folded in, since old terms/links are well trained and the arriving documents contain old terms/links while old documents convey no corresponding information to aid the folding in of new terms and links.

For new documents d_{new}, we first randomize and normalize $P(z|d_{new})$. Thereafter $P(z|d_{new})$ are updated through fusion of $P(w|z)$ and $P(c|z)$ in the follows:

E-step. The conditional distributions $P(z|d_{new}, w)$ and $P(z|d_{new}, c)$ are obtained from the previous estimate value of the parameters $P(z|d_{new})$:

$$P(z \mid d_{new}, w) = \frac{P(w \mid z)P(z \mid d_{new})}{\sum_k P(w \mid z_k)P(z_k \mid d_{new})}, \tag{11}$$

$$P(z \mid d_{new}, c) = \frac{P(c \mid z)P(z \mid d_{new})}{\sum_k P(c \mid z_k)P(z_k \mid d_{new})}. \tag{12}$$

M-step. The parameters $P(z|d_{new})$ are updated with the new expected values $P(z|d_{new}, w)$ and $P(z|d_{new}, c)$:

$$\begin{aligned} P(z \mid d_{new}) \propto \alpha \sum_{j \in d_{new}} \frac{N_{j,new}}{\sum_{j' \in d_{new}} N_{j',new}} P(z \mid d_{new}, w_j) \\ + (1-\alpha) \sum_{r \in d_{new}} \frac{A_{r,new}}{\sum_{r' \in d_{new}} A_{r',new}} P(z \mid d_{new}, c_r). \end{aligned} \tag{13}$$

Link-PLSI assumes that terms and citations make different contributions in defining the latent topic. The only potential imbalance could result from the mix parameter α between different terms and citations while these values cannot be freely controlled. Unlike Link-PLSI, our method allows to assign weights to each modality according to the average amount of information it offers. This concretely allows modeling of a document as a mixture of latent topics that is defined by the relative importance of its terms and its citation patterns, resulting in different mixtures.

Specifically, we use entropy as a criterion for weight assignment as entropy is useful for evaluating the average amount of information needed to specify the state of a random variable. The idea is quite straightforward as the distributions over terms in each document can be good indications for their informativeness. Distributions of terms in each document that are sharply peaked around a few values will have relatively low entropy, whereas those that are spread more evenly across different values will have higher entropy. The entropy of term feature distribution of a specific document is defined as:

$$H(d_i) = -\sum_j \frac{N_{ij}}{N_i} \log \frac{N_{ij}}{N_i}, \tag{14}$$

where N_i is the total number of feature terms in document d_i. The fusion parameter α is then defined according to our empirical formula:

$$\alpha = \begin{cases} 1 & H(d_i) \le \theta, \\ \exp(\theta - H(d_i)) & H(d_i) > \theta. \end{cases} \tag{15}$$

where θ equals to the average entropy of the entire document set.

Fold in new terms and hyperlinks. In this step we consider the problem of how to fold in new terms and citations. It is a pity that they can not be folded in by using

$P(z|d_{new})$ directly, it is because the sum of all probabilities of terms/citations in old term/citation sets under z already equals to one, which means $P(w|z)$ and $P(c|z)$ have been well trained and normalized. If we randomize and normalize all $P(w_{new}|z)$ and $P(c_{new}|z)$ when new documents arrive, the sum of the probabilities of all terms/citations under z will be larger than one. This restriction makes it inapplicable to update new terms and citations directly. To avoid this, we first derive $P(d_{new}|z)$ in the following way:

$$P(z \mid d_{new}, w) = \frac{P(w \mid z)P(z \mid d_{new})}{\sum_k P(w \mid z_k)P(z_k \mid d_{new})}, \tag{16}$$

$$P(z \mid d_{new}, c) = \frac{P(c \mid z)P(z \mid d_{new})}{\sum_k P(c \mid z_k)P(z_k \mid d_{new})}, \tag{17}$$

$$\begin{aligned} P(d_{new} \mid z) \propto \alpha \sum_{j \in d_{new}} \frac{N_{j,new}}{\sum_{i' \in D_{new}} N_{ji'}} P(z \mid d_{new}, w_j) \\ + (1-\alpha) \sum_{r \in d_{new}} \frac{A_{r,new}}{\sum_{i' \in D_{new}} A_{ri'}} P(z \mid d_{new}, c_r), \end{aligned} \tag{18}$$

where D_{new} is the set of new documents. Thereafter we need to develop a mechanism for new terms/citations update that satisfies the basic principle of topics under new terms/citations equal to one. $P(z|w_{new})$ and $P(z|c_{new})$ are randomly initialized and normalized. We then update $P(z|w_{new})$ and $P(z|c_{new})$ with the above $P(d_{new}|z)$ fixed.

E-step. The conditional distributions $P(z|d_{new}, w_{new})$ and $P(z|d_{new}, c_{new})$ are calculated from the previous estimate value of the parameters $P(z|w_{new})$ and $P(z|c_{new})$:

$$P(z \mid d_{new}, w_{new}) = \frac{P(z \mid w_{new})P(d_{new} \mid z)}{\sum_k P(z_k \mid w_{new})P(d_{new} \mid z_k)}, \tag{19}$$

$$P(z \mid d_{new}, c_{new}) = \frac{P(z \mid c_{new})P(d_{new} \mid z)}{\sum_k P(z_k \mid c_{new})P(d_{new} \mid z_k)}. \tag{20}$$

M-step. The parameters $P(z|w_{new})$ and $P(z|c_{new})$ are updated with the new expected values $P(z|d_{new}, w_{new})$ and $P(z|d_{new}, c_{new})$:

$$P(z \mid w_{new}) = \sum_{i \in D_{new}} \frac{N_{new,i} P(z \mid d_i, w_{new})}{\sum_{i' \in D_{new}} N_{new,i'}}, \tag{21}$$

$$P(z \mid c_{new}) = \sum_{i \in D_{new}} \frac{A_{new,i} P(z \mid d_i, c_{new})}{\sum_{i' \in D_{new}} A_{new,i'}}. \tag{22}$$

We can add the corresponding parameters of w_{new}, c_{new} and d_{new} reasonably at different times in this way.

Update the parameters. The third step of our incremental algorithm deals with the issues of how to calculate $P(w_{new}|z)$ and $P(c_{new}|z)$ and how to get the final normalized $P(w_{all}|z)$ and $P(c_{all}|z)$ by means of adjusting $P(w|z)$ and $P(c|z)$. Following the basic

principle of the total probability of terms/citations in the entire terms/citations sets under z should equal to one, we normalize $P(w_{all}|z)$ and $P(c_{all}|z)$ as:

$$P(w_{all} \mid z) = \sum_{i \in D \cup D_{new}} \frac{N_{all,i}}{\sum_{j' \in d_i} N_{j'i}} P(z \mid d_i, w_{all}), \tag{23}$$

$$P(c_{all} \mid z) = \sum_{i \in D \cup D_{new}} \frac{A_{all,i}}{\sum_{r' \in d_i} A_{r'i}} P(z \mid d_i, c_{all}). \tag{24}$$

For new terms w_{new} and new citations c_{new}, $P(z/d, w)$ and $P(z/d, c)$ are calculated according to formula (19) and (20) while for old terms and citations, we use formula (4) and (5) to get $P(z/d, w)$ and $P(z/d, c)$.

In the last step, we use the above parameters to execute the original Link-PLSI model for updating the model. As new documents arrive and old documents disappear, the above Link-IPLSI model can preserve the probability and continuity of the latent parameters during each revision of the model in a fast way.

4 Experimental Results

In this section, our empirical evaluation on the performance of our approach is presented. In all experiments, we used a PC with an Intel core2 duo p8400 3GHz CPU, 2G bytes of memory on the Windows XP Professional SP2 platform. We designed three experiments to test the viability of our model: time expenditure by comparing execution time with the Naïve Link-IPLSI; some preliminary results to demonstrate the performance of our algorithm in classification, which indicates the increased power of our adaptive learning of fusion parameter.

Data description. The performance of our model was evaluated using two different types of linked data: scientific literature from Citeseer which is connected with citations, Wikipedia webpages and WebKB dataset containing hyperlinks. We first adjust the link structure to include the incoming links and outgoing links only within each corpus, and then take advantage of these dataset for our model construction with adding new documents and citations and deleting out-of-date information.

The Citeseer data can be obtained from Citeseer collection that was made publicly available by Lise Getoor's research group at University of Maryland. There are altogether 3312 documents using abstract, title and citation information in the corpus with the vocabulary of 3703 unique words. The Citeseer dataset only includes articles that cite or are cited by at least two other documents. Thereafter the corpus size is limited to 1168 documents, of which only 168 documents have both incoming and outgoing links. The WebKB dataset contains approximately 6000 html pages from computer science departments of four schools (Cornell, Texas, Washington, and Wisconsin). The dictionary contains 2800 words in the WebKB domain and 9843 links. The dataset of Wikipedia webpages is downloaded from Wikipedia by crawling within the Wikipedia domain, starting from the "Artificial Intelligence" Wikipedia page and the dataset is composed of 1042 documents and 4912 links with the vocabulary of 3072 words.

Experiments on time cost To evaluate time efficiency of Naïve Link-IPLSI and Link-IPLSI, we ran these two algorithms on the subset of each database consisting of 90% of its entire documents respectively. We constructed five perturbed versions of

the databases, containing a randomly deleted 10% subset of the original documents and adding of the same amount of data. Remind that the time cost depends highly on the number of topics k, we examined the impact of k in the experiment. For each k, we ran the Naïve Link-IPLSI and Link-IPLSI on each dataset mentioned above, the average number of iterations to reach convergence and the total time spent on model construction are recorded. Table 1 gives a detailed illustration on the total time and the number of iterations required to achieve convergence. (The total time of Link-IPLSI is divided into two parts: Link-PLSI time and folding time).

As seen in this table, the Link-IPLSI method can save a large amount of time compared with the naïve method. In general, the computation time of the Naïve Link-IPLSI approach is much longer than that of our model. With k =30 on WebKB dataset, Link-IPLSI can reduce the time cost by more than 13 times. The reason is that the

Table1: Execution time (in seconds) of Naïve Link-IPLSI and Link-IPLSI

k	WebKB				Citeseer				Wiki			
	NLI		LI		NLI		LI		NLI		LI	
	Aver Iter	Total Time	Aver Iter	Total Time	Aver Iter	Total Time	Aver Iter	Total Time	Aver Iter	Total Time	Aver Iter	Total Time
10	42.11	7290	8.42	728	38.12	3072	7.87	310	42.97	3019	8.34	298
15	41.32	8598	8.13	818	42.31	3912	8.23	324	41.38	3718	8.37	307
20	43.46	11921	7.91	862	42.19	6184	8.12	373	45.23	5615	8.11	352
25	39.12	13063	8.61	943	47.81	7045	9.12	418	41.22	6412	8.01	384
30	41.32	15532	8.11	1131	55.39	8280	10.03	523	45.87	7123	8.77	472

Note: NLI stands for Naïve Link-IPLSI, LI stands for Link-IPLSI,
Aver.Iter stands for Average Iterations; k indicates number of latent topics

Naïve Link-IPLSI approach uses new random initial settings to re-estimate all relevant parameters of EM algorithm each time and requires a large number of iterations to converge to a different local optimum while Link-IPLSI has preserved a better starting point and can therefore converge much faster than the Naïve Link-IPLSI approach. The larger the dataset is, the more time our model can save. This is because the key point of Link-IPLSI is to reduce the EM iteration cost on more estimated parameters. Furthermore, when k increases, time cost increases as well, these results are consistent with our intuition.

Classification. Apart from its superior performance in time saving, another attractive feature of our model is its stability of latent topics. In this section, we use three Link-IPLSI variant models, that is, Link-PLSI, Naïve Link-IPLSI and the Link-IPLSI without learning of fusion parameter, together with Link-LDA and LTHM as baseline for comparison. Besides, we use Adaptive Link-IPLSI to denote our model for using adaptive asymmetric learning mechanism.

We perform classification on the WebKB dataset and Citeseer dataset. The task of this experiment is to classify the data based on their content information and link structure. From the original datasets, five perturbed versions of the datasets were created. We randomly split each dataset into five folds and repeat the experiment for five times, for each time we use one fold for test, four other folds for training incrementally. To give these models more advantage, we set the number of latent topics to be seven and six on WebKB and Citeseer respectively which correspond to the exact number of clusters. Classification accuracy is adopted as the evaluation metric, which is defined as the percentage of the number of correct classified documents in the entire data set. We demonstrate the average classification accuracies and its standard deviation over the five repeats as results. Since the accuracy of the

Link-PLSI model depends on the parameter α, we use the average classification accuracies for Link-IPLSI.

Table 2 shows the average classification accuracies on different datasets using different methods. From Table 2 we can see that the accuracies of Naïve Link-IPLSI and Link-PLSI are worse than that of Link-IPLSI and our model. Specifically, Though Link-IPLSI performs slightly better than other variant models of Link-PLSI, our method of Adaptive Link-IPLSI clearly outperform all other models and receives the highest accuracy among all these approaches. As described above, the latent

Table2: Classification accuracy (mean ± std-err %) on WebKB data set and Citeseer data set

Method	WebKB	Citeseer
Naïve Link-IPLSI	0.332± 0.90	0.453± 0.68
Link-PLSI	0.358± 0.88	0.478± 0.75
Link-IPLSI	0.371±0.87	0.481± 0.83
Link-LDA	0.382±0.77	0.501± 0.90
LTHM	0.411±0.67	0.534± 0.52
Adaptive Link-IPLSI	0.431±0.71	0.562± 0.81

variables generated by the Naïve- Link-IPLSI algorithm are discontinuous, whereas the latent variables generated by our algorithm are continuous. This shows that latent continuity can improve the performance of classification. The difference between the results of Link-LDA, LTHM and Adaptive Link-IPLSI is significant. This indicates that the enhanced classification performance is largely attributed to the adaptive weighing mechanism, i.e. the automatically obtained reasonable parameter α plays an important role in the improvement of classification.

5 Conclusion

Existing topic model cannot effectively update itself when changes occur. In this paper, we developed an incremental technique to update the hyperlinked information in a dynamic environment. Our technique computes and updates corresponding parameters by analyzing changes to linked documents and by re-using the results from previous Link-PLSI computation. Besides, our model can assign weights to terms and citations by means of adaptive asymmetric learning mechanism. In addition, we have demonstrated its faster and scalable performance on three distinctive dataset and illustrated preliminary results of our model in classification. However, our model learns the asymmetric fusion parameter through empirical formula hence further theoretical analysis is needed. Meanwhile, the number of latent topics of our model is fixed which is inconsistent with human intuition. Extending the model to grow or shrink as needed that permits easier model selection is our future work.

Acknowledgments. This work is supported by the National Science Foundation of China (No. 60933004, 60903141), the National Basic Research Priorities Programme

(No. 2007CB311004), 863 National High-Tech Program (No.2007AA01Z132), and National Science and Technology Support Plan (No.2006BAC08B06).

References

[1] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer and R. Harshman, "Indexing by latent semantic Analysis", Journal of the American Society for Information Science, 1990, 41, pp. 391-407.
[2] T. Hofmann, "Unsupervised Learning by Probabilistic Latent Semantic Analysis", Machine Learning, 2001, pp. 177-196.
[3] D. Jensen and H. Goldberg, "AAAI Fall Symposium on AI and Link Analysis", AAAI Press, 1998.
[4] R. Feldman, "Link analysis: Current state of the art", In: Proceedings of International Conference on Knowledge Discovery and Data Mining, 2002, pp. 23-26.
[5] S. Chakrabarti, "Mining the Web", Morgan Kaufman, 2002.
[6] Z.Z Shi, H.F Ma and Q. He, "Web Mining: Extracting Knowledge from the World Wide Web", Data Mining for Business Applications, Springer, 2008, pp. 197-209.
[7] L. Page, S. Brin, R. Motwani and T. Winograd, "The PageRank citation ranking: Bringing order to the web", Technical report, Stanford University, 1998.
[8] J. Kleinberg, "Authoritative sources in a hyperlinked environment", Journal of the ACM, 1999, 46, pp. 604-632.
[9] L. Getoor and C. P. Diehl, "Link Mining: A Survey", In: ACM SIGKDD Explorations Newsletter, 2005, 7: pp. 3-12.
[10] D. Cohn and H. Chang, "Learning to Probabilistically Identify Authoritative Documents", In: Proceedings of International Conference on Machine Learning, 2000, pp. 167-174.
[11] D. Cohn and T. Hofmann, "The missing link-A probabilistic model of document content and hypertext connectivity", In: Advances in Neural Information Processing Systems, 2001, pp. 430-436.
[12] E. Erosheva, S. Fienberg and J. Lafferty, "Mixed-membership models of scientific publications", In: Proceedings of the National Academy of Sciences, 2004, pp. 5220-5227.
[13] R. Nallapati, A. Ahmed, E. P. Xing and W. W. Cohen, "Joint Latent Topic Models for Text and Citations", In: Proceedings of International Conference on Knowledge Discovery and Data Mining, 2008, pp. 542-550.
[14] A. Gruber, M. Rosen-Zvi and Y. Weiss, "Latent Topic Models for Hypertext", In: Conference on Uncertainty in Artificial Intelligence (UAI), 2008, pp. 230-240.
[15] J. T. Chien and M. S. Wu, "Adaptive Bayesian Latent Semantic Analysis", IEEE Transactions on Audio, Speech, and Language, 2008, 16, pp. 198-207.
[16] T. C. Chou and M.C Chen, "Using Incremental PLSA for Threshold Resilient Online Event Analysis", IEEE Transaction on Knowledge and Data Engineering, 2008, 20, pp. 289-299