Topic Linkages between Papers and Patents

Shuo Xu[1], Lijun Zhu[1], Xiaodong Qiao[1], Qingwei Shi[1,2], and Jie Gui[1]

[1] Information Technology Supporting Center, Institute of Scientific and Technical
Information of China, Beijing 100038, P.R. China
[2] School of Software, Liaoning Technical University, Huludao 125105, P.R. China
{xush, zhulj, qiaox, shiqw, guij}@istic.ac.cn

Abstract. The papers and patents are usually considered as the indicators of basic science studies and technologies, respectively. Previous linkage research between papers and patents mainly focus on the analysis of non-patent literature cited by patent from the viewpoint of citation analysis. Thus, one will miss many valuable scientific papers that are not cited by patents until now. This paper proposes a simple procedure for constructing topic linkages between papers and patents by analyzing these two kinds of information resources simultaneously. Experimental results on *new energy vehicles* indicate that our approach is feasible and efficient.

Keywords: Topic Linkages; Topic Models; Topic Similarity; Latent Dirichlet Allocation; Optimal Transportation Problem

1 Introduction

The papers and patents are usually considered as the indicators of basic science studies and technologies, respectively. Intuitively, there should be some interactive and exclusive relationships between papers and patents, and it can be advantageous to analyze these two kinds of information resources (corpora) simultaneously. This motivated the introduction of the linkage between papers and patents, which is pioneered by Narion and his co-works in 1976 [1]. There has been abundant literature on the linkage between papers and patents showing that the linkages are indeed helpful to understand the technology development trends [2], university-industry-government relations [3], to measure innovation [4], etc.

Previous linkage research between papers and patents mainly focus on the analysis of non-patent literature cited by patent from the viewpoint of citation analysis. Thus, one will miss many valuable scientific papers that are not cited by patents until now. The main problem, solved in the study, is to link the topics between these papers and patents. Fig. 1 gives a detailed illustration. A naïve approach assumes that the papers and patents are part of a single corpus, and are exchangeable within it. However, this assumption is not appropriate for many text analysis problems. Wang et al. [5] apply Gaussian (Markov) random fields to model the correlations of different corpora, and develop Markov topic models (MTMs). MTMs can capture both the internal topic strcture within each corpus and the relationships between topics across the corpus.

In fact, MTM is a joint model with Markov assumption. In this study, we propose a non-joint method for topic linkages between papers and patents. Specifically, our procedure has the following simple steps: (1) to discover the topics in the papers and patents corpus, respectively, with probabilistic topic models, in Section 2; (2) to calculate the topic similarity in Section 3; (3) to construct topic linkages between papers and patents in Section 4. In Section 5, an experimental evaluation is conducted, and Section 6 concludes this work.

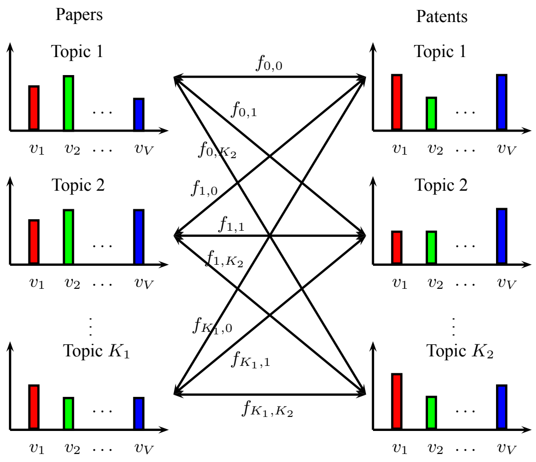

Fig. 1. The illustration of topic linkages between papers and patents

2 Latent Dirichlet Allocation (LDA)

The first step in topic linkages between papers and patents is to discover the topics in the papers and patents corpus, respectively. Probabilistic topic models, which are *generative model* for documents, are based upon the idea that documents are mixtures of topics, where a topic is a probability distribution over words. For more elaborate and detailed surveys we refer the readers to [6, 7].

In the study, Latent Dirichlet Allocation (LDA) [8], a widely used topic model, is utilized. In the generative process, for each document $m \in [1, M]$, a multinomial distribution $\boldsymbol{\vartheta}_m$ over topics is randomly sampled from the Dirichlet($\boldsymbol{\alpha}$), and then to generate each word, a topic $z_{m,n}(m \in [1, M], n \in [1, N_m])$ is chosen from this topic distribution, and a word $w_{m,n}$ is generated by randomly sampling from a topic-specific multinomial distribution $\boldsymbol{\varphi}_{z_{m,n}}$. A topic-specific multinomial distribution $\boldsymbol{\varphi}_k(k \in [1, K])$ is also randomly sampled from the Dirichlet($\boldsymbol{\beta}$).

Although LDA is still a relatively simple model, exact inference is generally intractable. A variety of algorithms have been used to estimate the parameters of topics models, such as variational EM (Expectation Maximization) [9, 8], expectation propagation [10, 11], belief propagation [12], and Gibbs sampling [13, 14], etc. In this paper, Gibbs sampling algorithm is used, since it provides a simple method for obtaining parameter estimates under Dirichlet priors and allows combination of estimates from several local maxima of the posterior distribution.

The Gibbs sampling procedure considers each word token in the text collection in turn, and estimates the probability of assigning the current word token to each topic, conditioned on the topic assignments to all other word tokens. The LDA model has two unknown multinomial parameter sets $\Theta = \{\boldsymbol{\vartheta}_m\}_{m=1}^{M}$ and $\Phi = \{\boldsymbol{\varphi}_k\}_{k=1}^{K}$ as well as the latent variables $\boldsymbol{z}$. The Gibbs sampling algorithm gives direct estimates of $\boldsymbol{z}$. Θ and Φ can be obtained from the count matrices as follows: $\varphi_{k,v} = \frac{n_k^{(v)}+\beta_v}{\sum_{v=1}^{V}(n_k^{(v)}+\beta_v)}, \vartheta_{m,k} = \frac{n_m^{(k)}+\alpha_k}{\sum_{k=1}^{K}(n_m^{(k)}+\alpha_k)}$, where V is the number of unique words, $n_k^{(v)}$ is the number of tokens of word v assigned to topic k, and $n_m^{(k)}$ represent the number of tokens in document m assigned to topic k.

3 Topic Similarity Measurement

To construct topic linkages between papers and patents, the similarity between a pair of topics, $\boldsymbol{\varphi}_1$ and $\boldsymbol{\varphi}_2$, should be measured. In the previous topic modeling research where topic similarity must be measured, symmetrized Kullback-Leibler (KL) divergence [15, 14], Jensen-Shannon (JS) divergence [16, 14] and cosine similarity [17] are frequently used without any formal validation.

In fact, apart from the three metrics above, there exist many alternatives, such as Spearman's rank order correlation coefficient (Spearman's ρ) [18], Kendall's τ [18], Jaccard's coefficient [19] and so on. Since as a multinomial distribution over the vocabulary, a topic, $\boldsymbol{\varphi}$, is also be seen as a V-dimensional vector, where each dimension i is a probability of v_i in $\boldsymbol{\varphi}$, or as a ranked list of words. Additionally, a topic can also be represented by a subset of topic words: words with a probability over a threshold or top words that contribute a cumulative probability mass over a threshold.

Symmetrized KL divergence is a symmetrized KL measure of the difference between two probability distributions. Formally, $\mathrm{symKL}(\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2) = \frac{1}{2}\left(\mathrm{KL}(\boldsymbol{\varphi}_1||\boldsymbol{\varphi}_2) + \mathrm{KL}(\boldsymbol{\varphi}_2||\boldsymbol{\varphi}_1)\right)$ with $\mathrm{KL}(\boldsymbol{\varphi}_1||\boldsymbol{\varphi}_2) = \sum_{v=1}^{V} \varphi_{1,v} \log \frac{\varphi_{1,v}}{\varphi_{2,v}}$.

JS divergence is another symmetric variation of KL divergence. Formally, $\mathrm{JS}(\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2) = \frac{1}{2}\left[\mathrm{KL}(\boldsymbol{\varphi}_1||\boldsymbol{\varphi}) + \mathrm{KL}(\boldsymbol{\varphi}_2||\boldsymbol{\varphi})\right]$, where $\boldsymbol{\varphi} = \frac{1}{2}(\boldsymbol{\varphi}_1 + \boldsymbol{\varphi}_2)$.

Cosine similarity measures the similarity between two vectors by finding the cosine of the angle between them, i.e., $\cos(\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2) = \frac{\boldsymbol{\varphi}_1 \cdot \boldsymbol{\varphi}_2}{\|\boldsymbol{\varphi}_1\| \times \|\boldsymbol{\varphi}_2\|}$.

Spearman's ρ is defined as the linear correlation coefficient of the ranks and is given by $\rho(\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2) = 1 - \frac{6\sum_{v=1}^{V}(\varphi_{1,v} - \varphi_{2,v})^2}{V(V^2-1)}$.

Kendall's τ measures the correlation between the relative ordering of ranks of the two ranked lists. It compares all the possible pairs of ranks $(\varphi_{1,i}, \varphi_{1,j})$ and $(\varphi_{2,i}, \varphi_{2,j})$ to determine the number of matching and non-matching pairs. A pair is matching or concordant if $\varphi_{1,i} > \varphi_{1,j} \Rightarrow \varphi_{2,i} > \varphi_{2,j}$ or $\varphi_{1,i} < \varphi_{1,j} \Rightarrow \varphi_{2,i} < \varphi_{2,j}$, and non-matching or discordant if $\varphi_{1,i} > \varphi_{1,j} \Rightarrow \varphi_{2,i} < \varphi_{2,j}$ or $\varphi_{1,i} < \varphi_{1,j} \Rightarrow \varphi_{2,i} > \varphi_{2,j}$. The correlation between the two ranked lists is defined as $\tau(\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2) = \frac{(n_c - n_d)}{\sqrt{n_c + n_d + n_1^t} \times \sqrt{n_c + n_d + n_2^t}}$, where n_c/n_d are the number of concordant/discordant pairs, n_1^t/n_2^t are the number of ties in $\boldsymbol{\varphi}_1/\boldsymbol{\varphi}_2$.

Jaccard's coefficient measures the similarity and diversity of two sets. Let A, B as a subset of topic words for $\boldsymbol{\varphi}_1$ and $\boldsymbol{\varphi}_2$. These subsets can be obtained in many ways. In the study, we let the subsets consist of top words that contribute a cumulative probability mass over a threshold ν $(= 0.5$ here). Then Jaccard's coefficient is can be easily calculated as $\text{Jaccard}(\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2) = \frac{|A \bigcap B|}{|A \bigcup B|}$.

Each metric looks at the relationship between two topics from different views. The symmetrized KL divergence and JS divergence consider the divergence of tow multinomial probabilities, and lower divergence would indicate higher similarity between two topics. Cosine similarity measures the angle of two vectors. Spearman's rank order correlation coefficient and Kendall's τ consider the ranks of words within a topic, and Jaccard's coefficient focuses on the association between two sets. Note that the range of both ρ and τ is $[-1, 1]$, which are transformed linearly into $[0, 1]$.

4 Topic Linkages Construction

If one can see topics in papers and patents as sources and sinks, respectively, or vise verse, and topic similarities as distances between sources and links (one can easily transform similarities into distances), the topic linkages construction problem can be transformed into the well-known *optimal transportation problem* [20, 21]. The question answered by the optimal transportation problem is: what is the cheapest way to move a set of masses from sources to sinks? Here cost is defined as the total *mass* × *distance* moved. For example, one can think of the sources as factories and the sinks as warehouses to make the problem concrete. We assume that the sources are shipping exactly as much mass as the sinks are expecting.

Formally, in the optimal transportation problem, we are given topic sets $\Phi_1 = \{\boldsymbol{\varphi}_1^1, \boldsymbol{\varphi}_2^1, \cdots, \boldsymbol{\varphi}_{K_1}^1\}$, $\Phi_2 = \{\boldsymbol{\varphi}_1^2, \boldsymbol{\varphi}_2^2, \cdots, \boldsymbol{\varphi}_{K_2}^2\}$ with respective associated nonnegative weights $\boldsymbol{p} = (p_1, p_2, \cdots, p_{K_1})$, $\boldsymbol{q} = (q_1, q_2, \cdots, q_{K_2})$ summing to one. Since there are no prior knowledge about the importance of topics, uniform weights, i.e. $\boldsymbol{p} = (\frac{1}{K_1}, \frac{1}{K_1}, \cdots, \frac{1}{K_1})$, $\boldsymbol{q} = (\frac{1}{K_2}, \frac{1}{K_2}, \cdots, \frac{1}{K_2})$, are utilized in the study. The optimal transportation distance between Φ_1 and Φ_2 is defined as $d(\Phi_1, \Phi_2) = \sum_{i=1}^{K_1} \sum_{j=1}^{K_2} f_{i,j}^* d(\boldsymbol{\varphi}_i^1, \boldsymbol{\varphi}_j^2)$, where the optimal flow (topic linkages) $F^* = [f_{i,j}^*]_{K_1 \times K_2}$ between Φ_1 and Φ_2 is the solution of the following linear programming, which guarantees an optimal solution.

$$\min_{F \in \mathbb{R}^{K_1 \times K_2}} \sum_{i=1}^{K_1} \sum_{j=1}^{K_2} f_{i,j} d(\boldsymbol{\varphi}_i^1, \boldsymbol{\varphi}_j^2) \tag{1}$$

$$\text{s.t.} \quad f_{i,j} > 0, 1 \leq i \leq K_1, 1 \leq j \leq K_2 \tag{2}$$

$$\sum_{j=1}^{K_2} f_{i,j} = p_i, 1 \leq i \leq K_1 \tag{3}$$

$$\sum_{i=1}^{K_1} f_{i,j} = q_j, 1 \leq j \leq K_2 \tag{4}$$

$$\sum_{i=1}^{K_1}\sum_{j=1}^{K_2} f_{i,j} = 1 \tag{5}$$

Example: Assume that papers and patents are mixtures of 4 and 5 topics, respectively. The similarity matrix between these topics is given in Figure 2 (a). In order to construct the topic linkages between papers and patents with the optimal transportation solving, the similarity matrix is transformed into the distance matrix simply by $d(\varphi_i^1, \varphi_j^2) = 1 - Sim(\varphi_i^1, \varphi_j^2)$. The optimal flow matrix F^*, which is our expected topic linages, is shown Figure 2 (b). Once the optimal flow matrix is known, it is very easy to construct the topic linkages. For instance, according to Figure 2 (b), one can link the Topic 1 in papers to Topic 1 and Topic 5 in patents with different linkage strength $f_{1,1} = 0.05$ and $f_{1,5} = 0.20$.

Patents Papers	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Topic 1	0.5	0.3	0.4	0.1	0.9
Topic 2	0.2	0.4	0.5	0.3	0.7
Topic 3	0.1	0.4	0.5	0.3	0.5
Topic 4	0.4	0.7	0.5	0.3	0.3

(a) Topic similarity matrix

Patents Papers	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Topic 1	0.05	0.00	0.00	0.00	0.20
Topic 2	0.10	0.00	0.15	0.00	0.00
Topic 3	0.00	0.00	0.05	0.20	0.00
Topic 4	0.05	0.20	0.00	0.00	0.00

(b) Optimal flow matrix

Fig. 2. The topic similarity and optimal flow matrices.

5 Experiments and Discussions

In this study, we choose the Derwent Innovation Index (DII) as the data source for patents, and National Science and Technology Library [1] as the data source for papers. The same query, described in [22], is used for two different data sources. The fields *title, keywords, and abstracts* are considered for the paper data set, and *TI, AB* for the patent data set. Totally, there are 39,827 papers and 79,104 patents. In addition to downcasing and removing stopwords and numbers, we also removed the words appearing only one kind of data set. In our experiments, the number of topics K_1, K_2 are fixed at 100, and the symmetric Dirichlet priors α, β are set at 0.5 and 0.1 respectively. Gibbs sampling is run for 2000 iterations.

In order to evaluate the performance of six metrics, the procedure proposed by Kim & Oh [23] is utilized. Specifically, starting from a set of topics extracted for papers/patents, we substitute five topics with the topics from patents/papers that are found to be most similar according to each of the six metrics to form six modified sets of topics. Then the normalized negative log likelihoods of the

[1] NSTL, http://www.nstl.gov.cn/

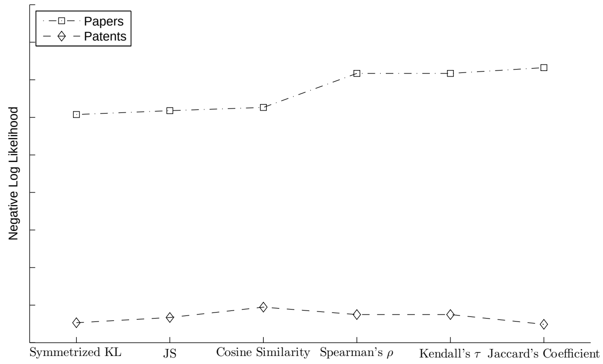

Fig. 3. Comparison of negative log likelihood for six similarity metrics.

respective corpus, which measures how well the model explains the corpus, are calculated using the modified sets of topics, and are given in Fig 3.

As shown in Fig 3, symmetrized KL divergence produces consistently the lowest normalized negative log likelihood scores. Another way to say this is symmetrized KL divergence performs the best among the six metrics, which is different from observations by Kim & Oh [23]. Fig. 4 gives 4 topics from papers and patents, respectively, in which each topic is shown with the 10 words. With our proposed approach, topic 88, 77, 83, 10 in papers are linked to topic 23, 80, 93, 76, respectively on the basis of the symmetric KL divergence. From Fig. 4, one can easily see that these topic linkages are quite intuitive and quite precise in the sense of conveying a semantic summary of *new energy vehicles* field, and LDA model can explain better patents than papers. This indicates that our approach is feasible and efficient.

6 Conclusions

The papers and patents are usually considered as the indicators of basic science studies and technologies, respectively. Previous linkage research between papers and patents mainly focus on the analysis of non-patent literature cited by patent from the viewpoint of citation analysis. Thus, one will miss many valuable scientific papers that are not cited by patents until now. The study tries to link the topics between in papers and patents, which enables knowledge navigation among heterogeneous data sources from underlying topic levels.

In order to construct the topic linkages between papers and patents, we propose a non-joint method, consisting of three simple steps: (1) to discover the topics in the papers and patents corpus, respectively, with probabilistic topic models; (2) to calculate the topic similarity; (3) to transform the topic linkage construction problem into the well-known optimal Finally, experimental results on *new energy vehicles* indicate that our approach is feasible and efficient.

Topic 88		Topic 77		Topic 83		Topic 10	
Word	Prop.	Word	Prop.	Word	Prop.	Word	Prop.
electron	0.360167	vehicle	0.196953	energized	0.708323	analysis	0.134864
field	0.102091	hybrid	0.177909	distribute	0.085559	loss	0.105088
emission	0.090633	electric	0.147691	source	0.019392	present	0.051836
function	0.065482	drive	0.038712	range	0.016205	component	0.041501
work	0.058801	power	0.017861	depend	0.015792	result	0.031639
ev	0.028510	hev	0.017530	cs	0.013460	analyzer	0.026125
hot	0.020922	brake	0.014130	primary	0.009063	part	0.023950
secondary	0.015511	economy	0.013488	discuss	0.006525	detail	0.022785
low	0.013667	powertrain	0.012264	total	0.005994	due	0.022249
enhance	0.008498	motor	0.010845	considered	0.003840	discuss	0.022154

(a) Topics in Papers

Topic 23		Topic 80		Topic 93		Topic 76	
Word	Prop.	Word	Prop.	Word	Prop.	Word	Prop.
device	0.893249	motor	0.979361	vehicle	0.532848	control	0.929425
draw	0.023259	stepper	0.000924	hybrid	0.370708	block	0.017018
show	0.013568	asynchronous	0.000545	decelerated	0.026651	diagram	0.015058
schematics	0.011200	directly	0.000370	show	0.016610	feedback	0.007267
function	0.006906	order	0.000156	draw	0.009068	function	0.004806
effect	0.006247	overheated	0.000156	improve	0.005725	regulator	0.003404
manner	0.004564	block	0.000127	schematics	0.005080	program	0.001894
simple	0.003528	relevant	0.000108	accept	0.004388	accordance	0.001361
design	0.001549	standard	0.000088	ensure	0.002945	perform	0.001060
configure	0.001509	conceptually	0.000088	mild	0.001526	instruction	0.000727

(b) Topics in Patents

Fig. 4. An illustration of 4 topic linkages between papers and patents.

Acknowledgments. We thank the financial support from "Key Technologies Research on Large Scale Semantic Computation for Foreign Scientific & Technical Knowledge Organization System", and "Application Demonstration of Knowledge Service based on STKOS", which are sponsored by Key Technologies R&D Program of Chinese 12th Five-Year Plan (2011–2015) under grant number 2011BAH10B04 and 2011BAH10B06, respectively, Social Science Foundation of Jiangsu Province: Study on Automatic Indexing of Digital Newspapers under grant 09TQC011 and MOE Project of Humanitites and Social Sciences: Research on Further Processing of e-Newspaper under grant 09YJC870014.

References

1. Narin, F., Hamilton, K.S., Olivastro, D.: The increasing linkage between U.S. technology and public science. Research Policy **26**(3) (1997) 317–330
2. Lee, M., Lee, S., Kim, J., Seo, D., Kim, P., Jung, H., Lee, J., Kim, T., Koo, H., Sung, W.K.: Decision-making support service based on technology opportunity discovery model. In Kim, T.h., Adeli, H., Ma, J., Fang, W.c., Kang, B.H., Park, B., Sandnes, F., Lee, K., eds.: FGIT-UNESST 2011. Volume 264 of Communications in Computer and Information Science. Springer Berlin Heidelberg (2011) 263–268
3. Leydesdoref, L., Meyer, M.: The scientometrics of a triple helix of university-industry-government relations. Scientometrics **70**(2) (2007) 207–222
4. Jibu, M.: An analysis of the achievements of JST operations through scientific patenting: Linkage between patents and scientific papers. In: Proceedings of the Conference on Science and Innovation Policy. (2011) 1–7

5. Wang, C., Thiesson, B., Meek, C., Blei, D.: Markov topic models. In: Proceedings of the 12th International Conference on Artificial Intelligence and Statistics. (2009) 583–590
6. Steyvers, M., Griffiths, T.: Probabilistic Topic Models. In: Handbook of Latent Semantic Analysis. Laurence Erlbaum (2007) 427–448
7. Blei, D.M.: Probabilistic topic models. Communications of the ACM **55**(4) (2012) 77–84
8. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of Machine Learning Research **3**(Jan) (2003) 993–1022
9. Winn, J.M.: Variational Message Passing and its Applications. PhD thesis, University of Cambridge (2004)
10. Minka, T.P.: Expectation propagation for approximate Bayesian inference. In: Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc. (2001) 362–369
11. Minka, T., Lafferty, J.: Expectation-propagation for the generative aspect model. In: Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence. (2002) 352–359
12. Zeng, J.: A topic modeling toolbox using belief propagation. Journal of Machine Learning Research **13**(Jul) (2012) 2233–2236
13. Griffiths, T.L., Steyvers, M.: Finding Scientific Topics. Proceedings of the National Academy of Sciences of the United States of America **101**(Suppl. 1) (2004) 5228–5235
14. Heinrich, G.: Parameter Estimation for Text Analysis. Technical report version 2.9, vsonix GmbH and University of Leipzig (2009)
15. Newman, D., Asuncion, A., Smyth, P., Welling, M.: Distributed algorithms for topic models. Journal of Machine Learning Research **10**(Aug) (2009) 1801–1828
16. Mimno, D., Wallach, H.M., Naradowsky, J., Smith, D.A., McCallum, A.: Polylingual topic models. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Stroudsburg, PA, USA, Association for Computational Linguistics (2009) 880–889
17. He, Q., Chen, B., Pei, J., Qiu, B., Mitra, P., Giles, C.L.: Detecting topic evolution in scientific literature: How can citations help? In: Proceedings of the 18th ACM International Conference on Information and Knowledge Management, New York, NY, USA, ACM (2009) 957–966
18. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: Numerical Recipes in C: The Art of Scientific Computing. 2nd edition edn. Cambridge University Press, New York, USA (1992)
19. Jain, A.K., Dubes, R.C.: Algorithms for Clustering Data. Prentice-Hall, Englewood Cliffs, New Jersey (1988)
20. Hillier, F., Lieberman, G.J., eds.: Introduction to Mathematical Programming. McGraw-Hill (1995)
21. Rachev, S.T., Ruschendorf, L., eds.: Mass Transportation Problems: Volume I: Theory (Probability and its Applications). Springer (1998)
22. Xu, S., Qiao, X., Zhu, L., Zhang, Y.: The service platform on monitoring and analyzing the content of science & technology information resource (in Chinese). Digital Library Forum **11** (2011) 38–44
23. Kim, D., Oh, A.: Topic chains for understanding a news corpus. In: Proceedings of the 12th International Conference on Intelligent Text Processing and Computational Linguistics. (2011)