Discovery of topic flows of authors

Young-Seob Jeong¹ · Sang-Hun Lee² ·
Gahgene Gweon³ · Ho-Jin Choi⁴

Published online: 6 May 2017

© The Author(s) 2017. This article is an open access publication

Abstract With an increase in the number of Web documents, the number of proposed methods for knowledge discovery on Web documents have been increased as well. The documents do not always provide keywords or categories, so unsupervised approaches are desirable, and topic modeling is such an approach for knowledge discovery without using labels. Further, Web documents usually have time information such as publish years, so knowledge patterns over time can be captured by incorporating the time information. The temporal patterns of knowledge can be used to develop useful services such as a graph of research trends, finding similar authors (potential co-authors) to a particular author, or finding top researchers about a specific research domain. In this paper, we propose a new topic model, Author Topic-Flow (ATF) model, whose objective is to capture temporal patterns of research interests of authors over time, where each topic is associated with a research domain. The state-of-the-art model, namely Temporal Author Topic model, has the same objective as ours, where it com-

✉ Ho-Jin Choi
hojinc@kaist.ac.kr

Young-Seob Jeong
bytecell@sch.ac.kr

Sang-Hun Lee
xhoon01@add.re.kr

Gahgene Gweon
ggweon@snu.ac.kr

¹ Soonchunhyang University, Asan-si, South Korea

² Agency for Defense Development, Daejeon, South Korea

³ Seoul National University, Seoul, South Korea

⁴ Korea Advanced Institute of Science and Technology, 291 Daehak-ro, Yuseong-gu, Daejeon, South Korea

putes the temporal patterns of authors by combining the patterns of topics. We believe that such ‘indirect’ temporal patterns will be poor than the ‘direct’ temporal patterns of our proposed model. The ATF model allows each author to have a separated variable which models the temporal patterns, so we denote it as ‘direct’ topic flow. The design of the ATF model is based on the hypothesis that ‘direct’ topic flows will be better than the ‘indirect’ topic flows. We prove the hypothesis is true by a structural comparison between the two models and show the effectiveness of the ATF model by empirical results.

Keywords Probabilistic topic model · Topic flow · Knowledge discovery

1 Introduction

As the number of Web documents is increasing exponentially, it becomes important to develop methods to extract useful information or knowledge from the documents. There are many knowledge discovery problems, one of which is the discovery of academic research interests. The discovery of research interests may give an insight into research trends according to a particular period and further may help researchers to make wise decisions for their future research topics.

It is important to note the difference between the discovery of academic interests and identifying experts [7]. The task of discovery of academic interests is to find people who write about particular topics, and the identifying experts task is to find the most skilled people for particular research topics. The discovery of academic interests, therefore, assumes that an author writes much about a particular topic when the author is interested in the topic. If the academic interests are plotted along the time sequence, then it will be a trend of academic interests. The trend of academic interests can be used to develop useful services such as finding similar authors (potential co-authors) to a particular author, finding top researchers about a specific research domain, or a graph of research trends. For example, the research trend graph is provided by Microsoft academic research, as shown in Fig. 1. When we pick the research topic *Networks and Communications*, then the dominant author list is shown in the bottom. By navigating such research trend graph, we can get useful information about ‘Which research area is hot, nowadays?’, ‘Who is the expert in this area?’, or ‘Will this research area rise again?’.

There are three practical factors to consider for the task of discovering research interests from academic Web documents. First, the academic documents typically do not have consistent labels or categories. Although each academic document provides keywords and categories that it belongs to, the keywords and categories are not standardized with other documents, conferences, or journals. The inconsistent keywords and categories make the discovery of academic interests more difficult. Second, it is necessary to capture latent semantic structures (i.e., research topics) that are common across all the documents. Each academic document may contain multiple research topics, so every topic is shared by all the documents. Third, academic Web documents are dynamic, so the research topics rise or fall as time goes by. The time factor conveys temporal patterns of research interests, which is important information.

Fig. 1 Sample of research trends obtained from Microsoft academic research

To address the above factors, clustering techniques have been investigated to overcome this obstacle [15]. In this paper, we adopt a topic modeling approach which is one of such techniques, where each topic is regarded as a specific research domain. When we use a word as a feature, then each document can be represented as a sequence of words. This representation, however, is not practical because the number of possible sequences of unique words grows exponentially as the size of each document increases. Therefore, the sequence of words is usually ignored, and only the word frequencies are utilized, which is called bag-of-words (BOW) method. Although this method misses location information of words, it has shown good performances with various types of data. The topic modeling approach also takes the BOW method. Each topic or cluster obtained from the topic modeling approach conveys latent semantic meaning across the documents. When the number of unique words, called the vocabulary size, is V , then a topic is represented as a V -dimensional vector of weighted unique words. For each topic, each unique word has a weight, and unique words that have greater weights are more representative of the topic. By looking at the top representative words of each topic, human can interpret and give a name to the topic. The sum of word weights in each topic is 1, so a topic is a distribution across the set of unique words. Thus, each topic is a distribution across the set of unique words, and different topics represent different semantic meaning by having different weights of the unique words. For example, if there are only three unique words {music, film, rock} and the top two words of a particular topic is {film, music}, then we can interpret the topic is *OST music*. If the top two words of another topic are {music, rock}, then we might interpret the topic is *rock music*. In this paper, we regard each topic as a research domain, so the topic proportion can be interpreted as the interest of the domains. Further, if the topic proportions are listed according to time information, then it will be a temporal pattern or a trend of research interests. Our proposed new topic model captures such temporal pattern of each author from academic Web documents.

The rest of this paper is organized as follows. Section 2 provides preliminary background for understanding the topic modeling approaches and related studies. Section 3 describes our motivation and approach in detail, and Section 4 presents experiments and results. Finally, Sect. 5 brings this paper to conclusions and future work.

2 Background

The past efforts for discovering research interests can be divided into three categories as described in Daud [6]. First, from a piece of text, research topics are identified by considering predefined features such as sentence length, forensic linguistics, and authorship attributions [8, 12]. The predefined features have tremendous impact on the performance, so these approaches heavily rely on the features. One big limitation of this approach is that it will require huge effort to develop different feature sets for different types of data. Second, from explicit connections between authors (e.g., co-writing), arbitrary relationships between the authors can be extracted in a form of graph structure [11, 21, 30]. This approach usually does not utilize the content of documents, so it is not likely to exactly capture the research interests of the authors. For example, if two authors A and B wrote a paper together, then the two authors might have similar research interests. However, this does not help to answer the question ‘What is the most favorite research topic of A ?’. Moreover, this approach is not applicable to the authors who usually write papers alone. Third, from the contents of documents, arbitrary semantic structures can be extracted. As the number of documents without annotations (i.e., unstructured texts) is growing exponentially, it is preferred to take unsupervised methods [5, 23, 31]. Topic modeling is one of such methods, and it captures the latent semantic structures across the documents. Many topic models and related studies have been proposed [19, 26, 28], where they are mainly motivated from the probabilistic latent semantic analysis (PLSA) model [14] or the latent Dirichlet allocation (LDA) model [3]. For instance, some models extract topics from the perspectives of authors Mimno and McCallum [20], Rosen-Zvi et al [24], Steyvers et al [25]. These models commonly assume that authors have topic distributions. In Chang et al [4], Newman et al [22], the topics are extracted from the perspectives of entities, where the entities are commonly supposed to have topic distributions. These models, however, do not consider the time dimension, so they suffer from a topic exchangeability problem. Assume that one uses AT model [24] to identify academic trends of a given author for two years. The AT model simply assumes that each author has a topic distribution which does not have any time information or temporal patterns. Thus, the model should be applied to each year separately. In other words, there will be two AT models for two years, respectively, and there will be two sets of topics $\phi^{\text{first year}}$ and $\phi^{\text{second year}}$. It should be noted that the t th topic $\phi_t^{\text{first year}}$ of the first year will not be equivalent to the t th topic $\phi_t^{\text{second year}}$ of the second year. In the worst case, the t th topic of the first year may be out of existence in the second year. This inconsistency of topics is called as the exchangeability problem [6].

We propose a new topic model to obtain the research interests of authors without the drawbacks of the exchangeability problem. As each topic model has its unique

generative structure that represents its own hypothesis, it will be helpful to discuss the existing salient models related to our proposed model.

2.1 Latent Dirichlet allocation (LDA) model and Author Topic (AT) model

Figure 2 shows the graphical representations of the LDA model [3] and the Author Topic (AT) model [24]. Given D documents, the LDA model hypothesizes that each document is written by N_d steps of an iterative process, where each step generates a single word. As the N_d -dimensional word sequence is observable, it is represented by the shaded node w . For each n th word, the topic variable z is sampled from the topic distribution θ of the d th document, where θ is derived from the hyperparameter α . The n th word is also derived from T topics ϕ , where the topics are derived from the hyperparameter β . The hyperparameters α and β are usually given by human. Assume that there are V unique words in the documents, then every topic ϕ is V -dimensional distribution over the vocabulary words. In other words, every topic is a vector of the same length, but different topics will have different distributions. It is important to note that the topic ϕ is a probability distribution over the unique words while the topic distribution θ , also called topic distribution, is the probability distribution of the topics. This implies that each document can be represented as the topic distribution θ rather than the distribution ϕ over the unique words. As the number of topics T is usually smaller than the number of documents D , the topic distribution θ makes it easy to compute distance or similarity between documents.

The LDA model hypothesizes that the word sequence is generated from the topics, where each document has a topic distribution. While the LDA model captures some latent patterns from the document's point of view, the AT model extracts some patterns from the authors' point of view. That is, as shown in Fig. 2, the AT model hypothesizes that there is a topic distribution θ_a for every a th author rather than d th document. If the total number of authors is A , then there are A topic distributions of the authors. For each document, the author list a_d and the word sequence w are observable, so they are represented as the shaded nodes. For each n th word, the author variable a is sampled from the author list a_d . The topic variable z is then sampled from the topic distribution

Fig. 2 (Left) a graphical representation of the latent Dirichlet allocation (LDA) model [3]; (right) a graphical representation of the Author Topic (AT) model [24]

AUTHOR = Jordan_M		
PROB.	TOPIC	WORDS
.1389	37	MIXTURE, EM, LIKELIHOOD, EXPERTS, MIXTURES, EXPERT, GATING, PARAMETERS, LOG, JORDAN
.1221	60	BELIEF, FIELD, STATE, APPROXIMATION, MODELS, VARIABLES, FACTOR, JORDAN, NETWORKS, PARAMETERS
.0598	52	ALGORITHM, ALGORITHMS, PROBLEM, STEP, PROBLEMS, LINEAR, UPDATE, FIND, LINE, ITERATIONS
.0449	77	MOTOR, TRAJECTORY, ARM, INVERSE, HAND, CONTROL, MOVEMENT, JOINT, DYNAMICS, FORWARD

AUTHOR = Koch_C		
PROB.	TOPIC	WORDS
2518	4	LIGHT, RESPONSE, INTENSITY, RETINA, OPTICAL, KOCH, BACKGROUND, CONTRAST, CENTER, FEEDBACK
.0992	45	VISUAL, STIMULUS, CORTEX, SPATIAL, ORIENTATION, RESPONSE, CORTICAL, RECEPTIVE, TUNING, STIMULI
.0882	84	SPIKE, FIRING, SYNAPTIC, SYNAPSES, MEMBRANE, POTENTIAL, CURRENT, SPIKES, RATE, SYNAPSE
.0504	64	CIRCUIT, CURRENT, VOLTAGE, ANALOG, CHIP, VLSI, CIRCUITS, SILICON, PULSE, MEAD

AUTHOR = LeCun_Y		
PROB.	TOPIC	WORDS
2298	13	RECOGNITION, CHARACTER, TANGENT, CHARACTERS, DISTANCE, HANDWRITTEN, DIGITS, SEGMENTATION, DIGIT, IMAGE
.0930	53	GRADIENT, FUNCTION, DESCENT, ERROR, VECTOR, DERIVATIVE, DERIVATIVES, OPTIMIZATION, PARAMETERS, LOCAL
.0930	69	LAYER, WEIGHTS, PROPAGATION, BACK, OUTPUT, LAYERS, INPUT, NUMBER, WEIGHT, FORWARD
.0762	36	INPUT, OUTPUT, INPUTS, OUTPUTS, VALUES, ARCHITECTURE, SUM, ADAPTIVE, PREVIOUS, PROCESSING

AUTHOR = Sejnowski_T		
PROB.	TOPIC	WORDS
.0927	9	SOURCE, INDEPENDENT, SOURCES, SEPARATION, INFORMATION, ICA, BLIND, COMPONENT, SEJNOWSKI, NATURAL
.0852	45	VISUAL, STIMULUS, CORTEX, SPATIAL, ORIENTATION, RESPONSE, CORTICAL, RECEPTIVE, TUNING, STIMULI
.0495	36	INPUT, OUTPUT, INPUTS, OUTPUTS, VALUES, ARCHITECTURE, SUM, ADAPTIVE, PREVIOUS, PROCESSING
.0439	74	MOTION, FIELD, DIRECTION, RECEPTIVE, FIELDS, VELOCITY, MOVING, FLOW, DIRECTIONS, ORDER

AUTHOR = Vapnik_V		
PROB.	TOPIC	WORDS
.3374	28	KERNEL, VECTOR, SUPPORT, MARGIN, SVM, DATA, SPACE, KERNELS, SET, MACHINES
.1243	44	LOSS, ESTIMATION, METHOD, ESTIMATE, PARAMETER, INFORMATION, ENTROPY, BASED, LOG, NEURAL
.0943	72	BOUND, BOUNDS, THEOREM, EXAMPLES, DIMENSION, FUNCTIONS, CLASS, PROBABILITY, NUMBER, RESULTS
.0669	92	ERROR, TRAINING, GENERALIZATION, EXAMPLES, SET, ENSEMBLE, TEST, FUNCTION, LINEAR, ERRORS

Fig. 3 Sample topics of five authors obtained from AT model

θ of the sampled author a . The n th word is generated in the process as same as the process of the LDA model. The biggest difference between the LDA model and the AT model is the position of the topic distribution θ . Thus, the hypothesis of the AT model is different from that of the LDA model, and such difference makes them to be the best for different types of data. To discover the research interests of authors, it is obvious that the AT model will be better than the LDA model. Figure 3 shows the sample topics of five authors obtained from the AT model. For example, the most prominent topic of the author Jordan can be interpreted as *mixture of experts*.

2.2 Dynamic Topic (DT) model, Sequential LDA (S-LDA) model, and Sequential Entity Group Topic (SEGT) model

Although the LDA model and AT model are capable of capturing latent semantics across the documents, both of them have a limitation that they do not consider temporal patterns. The temporal patterns can be important to some tasks such as a trend analysis and action recognition. There are some topic models that capture temporal patterns of topics [2, 9, 10, 16]. Blei and Lafferty proposed the Dynamic Topic (DT) model [2], which captures the evolution of topics by employing a Gaussian kernel. The topics are obtained from sequentially organized data, and the topics change over time, e.g.,

years. This implies that the research topics of different years will be different, which implies that it requires human to interpret the topics of every year. It also means that it is not possible to obtain the rise and fall of particular topics (i.e., a research trend).

If the topics are not changed, then a flow of a particular topic can be obtained by concatenating the proportions of the corresponding topic of every time span (i.e., year). Assume that there are documents of three time tags 2010, 2011, and 2012, where each document has only one time tag. T -dimensional topic distribution will be obtained for every year, where the topics are same for every year. If the topic distributions are aligned or ordered by the time tags, then it will be $T \times 3$ (time span) matrix, where each t th row is a temporal pattern of t th topic. We call the temporal pattern as *topic flow* in this paper.

Sequential LDA (S-LDA) model [10] was proposed to capture the topic flows. This model is designed based on a hypothesis that each segment (time span) is influenced by its previous segment. The hypothesis is represented by a nested extension of the two-parameter Poisson–Dirichlet Process (PDP) which basically gives a kind of smoothing effect to adjacent segments. In other words, it provides a smoothed version of topic flows in the assumption that the topic distributions do not rise and fall rapidly. However, the topics obtained by the S-LDA model are not from authors' point of view.

Sequential Entity Group Topic (SEGT) model [16] was proposed to capture the topic flows from the entity's (or entity group's) point of view. It is basically an extension of the S-LDA model, so it uses the nested extension of the two-parameter PDP to model the temporal patterns. The biggest difference between the SEGT model and the S-LDA model is that the position of the topic distributions. The SEGT model allows the entities or entity groups to have topic distributions, while the documents have the topic distributions in the S-LDA model. Although the results of S-LDA model and SEGT model were quite impressive, they have a common limitation that they are designed to capture the topic flows within only one document. That is, the authors assume that the smoothing effects between adjacent segments are valid only within a single document.

2.3 Topics Over Time (TOT) model and Temporal Author Topic (TAT) model

To get topic flows from a set of documents, Wang and McCallum [29] proposed the Topics Over Time (TOT) model. The model has a time stamp variable y which is generated from topic-specific multinomial distributions ϕ and beta distribution ψ . It has a limitation that it generates the topics from the perspective of the documents. To get topic flows of authors from a set of documents, the Temporal Author Topic (TAT) model [6] was appeared, where the TAT model is a variation of the Author Conference Topic (ACT1) model [27].

The graphical representations of the TOT model and the TAT model are depicted in Fig. 4. The biggest difference between the two models is the position of the topic distribution θ . In the TOT model, each document has a topic distribution, which implies that the topics are obtained from the perspective of the documents. On the other hand, each author has a topic distribution in the TAT model, which means that the topics are generated from the authors' point of view. The TAT model is basically an extension of the AT model. The process of word sequence generation of the TAT model is same

Fig. 4 (Left) a graphical representation of the Topics Over Time (TOT) model [29]; (right) a graphical representation of the Temporal Author Topic (TAT) model [6]

Table 1 Meaning of notations for graphical representations of the TAT model

Notation	Meaning of notation
D	Number of documents
N	Number of words
T	Number of topics
A	Number of authors
V	Size of vocabulary
Y	Number of time spans, e.g., years, in this paper
α	A T -dimensional Dirichlet prior vector for θ of the authors
β	A V -dimensional Dirichlet prior vector for ϕ of the topics
γ	A Y -dimensional Dirichlet prior vector for ψ of the topics
θ	T -dimensional topic distributions of the authors
ψ	Y -dimensional time span distributions of the topics
ϕ	The topics (the V -dimensional probability distributions over the vocabulary)

as the process of the AT model, except for the additional variable ψ . The meaning of notations of the TAT model is summarized in Table 1. Note that the number of variable ϕ is same as the number of variable ψ . That is, there are T topics, each of which is a distribution over V unique words, while each time span distribution ψ is a distribution over time spans. Thus, ψ represents the flow (i.e., rise and fall) of every t th topic. The TAT model is the model most similar to ours, and its objective is the same as ours.

We believe that the TAT model has a limitation that it uses only one variable ψ to represent the flows of every topic. This implies that the topic flows of every author are shared, and this will lead to wrong results for a particular authors whose topic flows are much different from many other authors. Moreover, to compute the topic flows of authors, it is required to combine the topic patterns. Such topic flows are called as *indirect* topic flows in this paper. On the other hand, our proposed model allows each author to have a topic distribution for each time span, so the topic flows of authors

are directly represented by the topic distribution. We call this topic flow as *direct* topic flows in this paper. The proposed model is designed based on the hypothesis that modeling directly topic flows will be more effective than the indirect topic flows.

3 Finding topic flows of authors over time

In this section, we clarify the definition of the problem to be solved and offer our contributions. Then we describe the new proposed topic model in detail.

3.1 Problem definition and contributions

The problem is to discover a research trend of authors. If we regard the research field as a topic of the topic modeling, then the problem becomes the task of finding the topic flows of authors. If an author a writes many papers on a particular topic t for year y , then the author will have a bigger topic proportion of the topic t for the year. Likewise, if the author writes only few papers on the topic t for the year $y + 1$, then the topic proportion for the year will decrease. The two consecutive topic proportions can be concatenated and will be a flow the topic t .

The topic flows of authors can be used to answer to the following questions: “Who are the authors who wrote most on topic Z in year Y ?”, “Which topics did author P write most about in year Y ?”, and “Can we see the temporal patterns of interests, e.g., topic flows of author P for the past 5 years?” The TAT model [6] solved the same problem as ours, but we demonstrate that the proposed model is more effectively used to answer to the above questions by showing many experimental results. The contributions of our work are as follows: (1) A new topic model to get topic flows of authors without suffering from the exchangeability problem; (2) Experimental verification of the effectiveness of the proposed model by comparison with the TAT model on a real-world dataset.

3.2 Author Topic-Flow (ATF) model

Motivated by the hypothesis that direct topic flows will be more effective than indirect topic flows, we designed a new topic model called the Author Topic-Flow (ATF) model. The graphical representation of the ATF model is shown in Fig. 5 and the meaning of notations is given in Table 2.

To effectively explain the structure of the ATF model, we here clarify the three big differences between the ATF model and the TAT model. First, the position of time span distribution ψ is different. Both models utilize the time span distribution to capture the temporal patterns of topics, so the position of ψ determines the way of representing temporal patterns of topics. The TAT model says that each topic has a time span distribution, which means that each topic has a flow or a temporal pattern. A higher proportion ψ_t to the t th topic in the y th time span implies that there are more documents about the t th topic in the y th time span, and vice versa. It is worth noting that the topic distributions of every author contribute to the global time span

Fig. 5 A graphical representation of the Author Topic-Flow (ATF) model

Table 2 Meaning of notations for graphical representations and probabilistic equations of the ATF model

Notation	Meaning of notation
Y	Number of time spans, e.g., years, in this paper
D	Number of documents
N	Number of words
T	Number of topics
A	Number of authors
V	Size of vocabulary
α_a	A T -dimensional Dirichlet prior vector for θ of the a th author
β	A V -dimensional Dirichlet prior vector for ϕ
γ_a	A Y -dimensional Dirichlet prior vector for ψ of the a th author
θ_{ay}	A T -dimensional topic distribution of the a th author for the y th year
ψ_a	A Y -dimensional time span distribution of the a th author
ϕ_r	The r th topic (the V -dimensional probability distribution over the vocabulary)
a_d	An observed author list of the d th document
y_d	An observed year, e.g., the time tag of the d th document
a_{dn}	An author assignment of the n th word in the d th document
z_{dn}	A topic assignment of the n th word in the d th document
$\mathbf{z}_{-dn}$	A vector of topic assignments for all words except the n th word of the d th document
w_{dn}	The observed n th word in the d th document
$\mathbf{w}$	A vector of all the observed words
C_{ayt}^{AYT}	The number of words that are assigned to the r th topic and the a th author, within the y th year
C_{ay}^{AY}	The number of words that are assigned to the a th author within the y th year
C_{tv}^{TY}	The number of the v th unique words that are assigned to the r th topic in every document

distribution, so the time span distribution ψ is globally shared by all the authors. That is, it will make many authors to have similar topic flows to each other, which is obviously wrong results. Moreover, it is required to combine the topic distribution and temporal patterns of the topics, in order to indirectly obtain the topic flow of a particular author. On the other hand, the ATF model says that each author has a time span distribution ψ and topic distributions θ for every time span (e.g., year). The ATF model allows each author to have a different topic flow which can be directly used as a topic flow of the author.

Second, the number of topic distributions θ is different. In the TAT model, each author has only one topic distribution, while each author has topic distributions for every time span in the ATF model. By joining with the first difference, this second difference allows each author to have a different topic flow.

Third, in the ATF model, every author has the prior variables α_a and γ_a , which are prior information about the topic distributions and time span distribution, respectively. This helps the model to avoid an overfitting problem that stems from the restriction that every future document has a topic distribution the same as was seen in the training documents [3].

One may argue that why do not move the topic ϕ into each author. If each author has the topic ϕ , then different authors will have different topics. That is, the exchangeability problem appears between the authors, so it will be impossible to measure a similarity or a distance between the authors. To avoid the exchangeability problem, the topic ϕ should be shared by all the authors and all the time spans. The ATF model can be explained in a more formal generative way as follows.

1. For the t th topic,
 - (a) Draw a word distribution ϕ_t from $Dirichlet(\beta)$.
2. For the a th author,
 - (a) Draw a time span distribution ψ_a from $Dirichlet(\gamma_a)$.
 - (b) For the y th time span,
 - i. Draw a topic distribution θ_{ay} from $Dirichlet(\alpha_a)$.
3. For the d th document,
 - (a) Generate a list of authors a_d and a time span y_d from $Multinomial(\psi)$.
 - (b) For the n th word,
 - i. Choose an author a_{dn} from $Uniform(a_d)$.
 - ii. Given a_{dn} and y_d , choose a topic z_{dn} from $Multinomial(\theta_{a_{dn}y_d})$.
 - iii. Given z_{dn} , generate a word w_{dn} from $Multinomial(\phi_{z_{dn}})$.

In the ATF model, each document has the observable time span y , which intuitively makes sense because each research article likely has only one time tag. The TAT model, in contrast, has a time span for every word, which is not natural. The probability of generating the n th word w_{dn} with year j given an author g of the d th document is represented as:

$$P(w_{dn}, y_d | g, d, \theta, \phi, \psi) = P(a_d, y_d | \psi_{gy_d}) \times \sum_{z_{dn}=1}^T P(w_{dn} | z_{dn}, \phi_{z_{dn}}) P(z_{dn} | g, y_d, \theta_{gy_d}) P(g | a_d) \quad (1)$$

where the meaning of the notations is described in Table 2. The exact computation of the parameters intractable, we use a collapsed Gibbs sampling algorithm for approximating the parameters [1, 13], which is one of the MCMC algorithms. For each step of the Markov chain, a latent topic assignment z_{dn} and author assignment a_{dn} of the n th word m in the d th document are drawn from the conditional posterior distribution:

$$P(z_{dn} = k, a_{dn} = g | \mathbf{z}_{-dn}, \mathbf{w}_{-dn}, w_{dn} = m, y_d = j) \propto \frac{\alpha_{ak} + C_{ajk}^{AYT}}{\sum_{t=1}^T \alpha_{at} + C_{ajt}^{AYT}} \frac{\beta_m + C_{km}^{TV}}{\sum_{v=1}^V \beta_v + C_{kv}^{TV}} \frac{\gamma_{aj} + C_{aj}^{AY}}{\sum_{y=1}^Y \gamma_{ay} + C_{ay}^{AY}}. \quad (2)$$

The meaning of the notations is represented in Table 2, with a minor exceptional use of the notation that C_{ayt}^{AYT} , C_{ay}^{AY} , and C_{tv}^{TV} in this expression exclude the n th word.

The ATF model has three parameters, the topic distribution θ , the word distribution (topic) ϕ , and the time span distribution ψ . To obtain the parameters, we need to keep track of $V * T$ (word by topic) and $T * Y * A$ (topic by time span of the author) count matrices. From the count matrices, the three parameters can be obtained as

$$\theta_{ajk} = \frac{\alpha_{ak} + C_{ajk}^{AYT}}{\sum_{t=1}^T \alpha_{at} + C_{ajt}^{AYT}}, \quad (3)$$

$$\phi_{km} = \frac{\beta_m + C_{km}^{TV}}{\sum_{v=1}^V \beta_v + C_{kv}^{TV}}, \quad (4)$$

$$\psi_{aj} = \frac{\gamma_{aj} + C_{aj}^{AY}}{\sum_{y=1}^Y \gamma_{ay} + C_{ay}^{AY}}. \quad (5)$$

The time span distribution ψ is not associated with the two random variables a and z , so it can be exactly computed without an iterative approximation. This implies that the time span distribution in Eq. (2) is not changed during the approximation, so it can make each step of the Markov chain faster by using a fixed time span distribution.

The probability of the a th author given the t th topic and the y th year can be obtained using a joint conditional probability as follows:

$$P(a|t, y) = \frac{P(a, t, y)}{P(t, y)}. \quad (6)$$

In terms of the joint probability, we compare the ATF model with the TAT model in order to emphasize the structural differences between them, which result in performance gaps that we will show by experiments. The joint probability of the TAT model is

$$P(a, t, y) = P(t, y|a)P(a) = P(t|a)P(y|a)P(a), \quad (7)$$

where $P(y|a) = \sum_{z=1}^T P(y|z)P(z|a)$, $P(y|z) = \psi_{zy}$, $P(z|a) = \theta_{az}$, and $P(a) = C_{ay}^{AY} / \sum_{a'=1}^A C_{a'y}^{AY}$. The joint probability of the ATF model is

$$P(a, t, y) = P(t|a, y)P(a, y) = P(t|a, y)P(y|a)P(a), \quad (8)$$

Fig. 6 Graphical representations of three models (e.g., AT model, TAT model, and ATF model)

where $P(y|a) = \psi_{ay}$, $P(t|a, y) = \theta_{ayt}$, and $P(a) = C_{ay}^{AY} / \sum_{a'=1}^A C_{a'y}^{AY}$.

In the TAT model, a conditional probability $P(t|a, y)$ of the t th topic given the a th author and the y th year can be computed from the joint probability $P(a, t, y)$. However, to compute the joint probability, it is required that the time span distribution of each author $P(y|a)$ should be ‘indirectly’ obtained by combining the time span distribution of every topic and the topic distribution of the author. This causes the worse performance of the TAT model about the task of discovery of topic flows of authors. Moreover, the topic flow of a th author is generated by combining the set of the conditional probabilities, which means that the topic flow of the author will be smoothed by the global time span distribution of the topics. This also means that every author tends to have similar temporal patterns of research interests, which will be obviously wrong results.

In contrast, the ATF model has the parameter θ_{ayt} which itself is the conditional probability $P(t|a, y)$, and the parameter ψ_{ay} is the time span distribution of the a th author. Thus, the topic flow of each author can be ‘directly’ obtained by combining θ_{ayt} . We believe that this difference between the two models will make a huge impact on the results. To effectively demonstrate structural differences, the graphical representation of three models (e.g., AT model, TAT model, and ATF model) is depicted in Fig. 6, where different colored lines denote different distributions. The generative process of AT model can be seen as two steps. The first step is to sample the author variable z followed by the second step of sampling the word. The generative process of TAT model has additional step of sampling the year (time span variable). The difference from the AT model is highlighted using the red dotted lines. Note that the variable γ , which represents the temporal patterns of topics, is a global variable. This implies that the temporal patterns of topics are shared by every author. On the other hand,

in the generative process of the ATF model, each author has a separated variable γ . This eventually allows the temporal patterns of topics of a particular author to be independent from the temporal patterns of other authors. We believe that such structural difference makes our proposed model will be better for capturing the topic flow of authors, and proved it by experimental results.

4 Experiments

4.1 Description of dataset and environment

The dataset is a set of 816 research articles on computer science, which was collected from a Microsoft academic research site. The dataset is collected by searching articles of some key authors and articles of the co-authors of the key authors. The dataset covers artificial intelligence, algorithms, databases, the natural language process, information retrieval, machine learning, networks, real-time systems, and so on. The time spans of the dataset were unique publish years between 2007 and 2011. The articles contained only abstracts, not the full texts. We removed stop-words, punctuations and numbers. All the words were lowercase, and we performed stemming by Porter stemmer. Words and authors that appeared less than three times in a dataset were removed. The sentences were recognized by '.', '?', '!', and "newline". The size of vocabulary was $V = 8300$, the number of total words was 70,373, and the number of authors was $A = 1127$.

For fair comparison, the parameters of the two models were equally initialized. We symmetrically set $\alpha = 0.1$, $\beta = 0.1$, and $\gamma = 0.1$. We varied the number of topics T and found that various topics were obtained with few redundancies when $T = 50$. Therefore, for all experiments except the author prediction test, we set $T = 50$. The number of Gibbs sampler iterations for the models was 1000.

4.2 Topic discovery

The purpose of this experiment was to show the quality of topics obtained from the ATF model by comparing the topics of the ATF model with the topics of the TAT model. We used the three criteria of Jo and Oh [17] for measuring the quality. First, the discovered topics should be coherent. In other words, they should be meaningful or comprehensible to people. Second, they should be specific enough to capture the details of research interests in the dataset. Third, they should be those that are discussed the most in the dataset. In other words, the topics should sufficiently reflect the dataset.

Two sample topics are shown in Table 3, where the top 10 representative words were obtained according to their weights in each topic. The topic names were manually determined by people in accordance with the top representative words. The topics obtained from both models were similar. Specifically, for each topic, a list of top representative words of the ATF model was similar to that of the TAT model. Not only these topics, but also the overall topics obtained from the ATF model were similar to the topics obtained from the TAT model. The topics obtained from both models were comprehensible to people, and specific enough to capture the details of research interests. For example, with respect to the topic *Document searching*, it was trivial

Table 3 Sample topics obtained from the two models

Topic Models	Descriptive languages		Robot group control	
	TAT	ATF	TAT	ATF
Top 10 words	Ontology	Query	Robot	Robot
	Owl	Ontology	Group	Group
	Logic	Owl	Robots	Control
	Reason	Logic	Individual	System
	Query	DI	Behavior	Task
	DI	Reason	Collect	Individual
	Language	Description	Evolution	Robots
	Description	Language	Task	Collect
	Express	Answer	Swarm	Swarm
	Semantic	Express	Fault	Communication

for people to determine the name of the topic as *Document searching* in accordance with the top representative words. Further, the topics of both models were generally discussed the most in the dataset. The topics *Descriptive languages*, *Robot group control*, and *Document searching* were discussed in the research fields of artificial intelligence, real-time systems, and information retrieval, respectively.

4.3 Topic-wise research interest over years

In this subsection, we demonstrate that the ATF model is effective in obtaining topic-wise research interests of authors over years, which can be obtained using the probabilities of authors given a particular topic and a year, where the probabilities are computed through Eq. (6). In other words, given a particular topic t , it determines how much each author is interested in the topic t over years. Therefore, the topic-wise research interest of authors over years can be thought of as temporal patterns of research interests of authors from a particular topics point of view.

A sample of research interest patterns of five authors for the topic *Descriptive languages* is depicted in Fig. 7, where research interest patterns of different authors are plotted with different colors. Although the two graphs in the figure are obtained from the same dataset, they are significantly different. For instance, with respect to the author Ian Horrocks, the TAT model says that his interest in *Descriptive languages* increased in 2008 but rapidly decreased in 2010, while the ATF model says that his interest decreased slowly between 2008 and 2010. For the author Bernardo Cuenca, the TAT model says that his research interest slowly decreased from 2008 to 2011, while the ATF model says that his research interest was highest in 2009. These differences imply that one of the models is better than the other model for representing the topic-wise research interest over years. To know which model is more effective, we manually counted the number of papers written by the authors for each year, as described in Table 4, where each digit between brackets represents the number of papers about the topic *Descriptive language*. The plot obtained from the ATF model is more consistent

Fig. 7 Topic-wise research interests for the topic Descriptive language of the TAT model (*left*) and of the ATF model (*right*)

Table 4 Number of total papers written by five authors for each year

Years	2007	2008	2009	2010	2011
Ian Horrocks	10 (8)	10 (8)	10 (7)	10 (6)	10 (8)
Bernardo Cuenca	4 (4)	5 (4)	7 (6)	6 (3)	3 (2)
Yevgeny Kazakov	4 (4)	1 (1)	3 (2)	4 (2)	2 (1)
Ulrike Sattler	8 (8)	6 (6)	1 (1)	2 (2)	2 (1)
Birte Glimm	3 (3)	2 (2)	1 (1)	2 (2)	3 (3)

The digits within brackets are the number of papers written by the authors about the topic Descriptive language

Fig. 8 Topic-wise research interests for the topic Robot group control of the TAT model (*left*) and of the ATF model (*right*)

with the table, as the plot of the ATF model reflects the digits between brackets in the table. For example, with respect to the author Bernardo Cuenca, the number of papers about the topic was at its peak in 2009. The ATF model, therefore, shows a consistent plot for him, while the TAT model's plot says that his interest decreased from 2008.

Another sample for the topic *Robot group control* is shown in Fig. 8, and the topic flows obtained from the two models are again different. For example, the TAT model says that the research interest of the author Marco Dorigo is significantly decreased from between 2009 and 2011. However, as shown in Table 5, the author Marco Dorigo consistently wrote papers about the topic *Robot group control* between 2007 and 2011. The plot of ATF model is more consistent with the Table 5, because the ATF model allows each author to have a unique topic flow while the TAT model pushes every author to share the topic flow.

Table 5 Number of total papers written by five authors for each year

Years	2007	2008	2009	2010	2011
Marco Dorigo	10 (7)	10 (9)	10 (8)	10 (7)	10 (7)
Stefano Nolfi	10 (9)	4 (3)	7 (5)	8 (5)	2 (2)
Anders Lyhne Christensen	4 (4)	3 (3)	3 (3)	6 (4)	4 (2)
Rehan O'Grady	3 (3)	3 (3)	2 (2)	4 (3)	3 (3)
Christos Ampatzis	2 (2)	1 (1)	3 (1)	4 (2)	3 (3)

The digits within brackets are the number of papers written by the authors about the topic Robot group control

Fig. 9 A time span distribution of the topic Descriptive language over years in the TAT model

Such results mainly come from the structural difference between the models. The topic flows of every author are shared in the TAT model, which means that the topic flow of each author is affected by the other authors. On the other hand, the ATF model directly models the topic flows of authors, so the research interest of each author is captured well. To be more specific, the reason for the poor topic-wise plots of the TAT model is that its joint probability was computed using the time span distributions of topics ϕ_{zy} , as shown in Eq. (7). Specifically, there was no term that exactly represents topic flow of each author, so it could not capture the exact patterns of research interests of the authors. As the topic-wise plot can be regarded as research interest patterns of authors from a particular topics point of view, the lack of a term for topic flows of authors caused poor representation of research interest patterns of authors. In Fig. 9, the time span distribution $\psi^{Descriptive\ language}$ of the TAT model is depicted. The research interest patterns of authors obtained from the TAT model in Fig. 7 generally show dependency upon the plot of the time span distribution ψ in Fig. 9. That is, for every topic t , research interest patterns of authors are smoothed by the time span distribution of the topic t , because time span distributions of topics were used to compute the research interest patterns of the authors. In contrast, the joint probability of the ATF model is computed using the term θ_{ay} which directly represents the research interest patterns of authors, as shown in Eq. (8). Its topic-wise plot, therefore, represents more exact research interest patterns of authors given a particular topic.

4.4 Author-wise research interest over years

In this subsection, we demonstrate that the ATF model is effective in obtaining author-wise research interests over years, which can be thought of as topic flows or research

Fig. 10 Research interest patterns of two topics, Robot group control and Networks, for the author Christos Ampatzis, which were obtained from the TAT model (*left*) and the ATF model (*right*)

Table 6 Number of papers, written by Christos Ampatzis, about two topics, Robot group control and Networks, for each year

Years	2007	2008	2009	2010	2011
Robot group control	2	1	1	2	3
Network	0	0	0	2	0

interest patterns from a particular authors point of view. Samples of research interest patterns for two topics, *Robot group control* and *Networks*, of the author Christos Ampatzis are shown in Fig. 10, where the left figure is obtained from the TAT model and the right figure is obtained from the ATF model. The plots of the two models are significantly different. The ATF model says that the research interest about the topic *Robot group control* increases from 2008, while the TAT model says that the research interest decreases from 2009. To know which model gives the right result, we again manually counted the number of papers written by the author for each year, as described in Table 6. It is obvious that the plot of the ATF model is more consistent with the Table 6. For instance, his research interest about the topic *Robot group control* grew from 2008, which is well depicted in the plot of the ATF model. The reason for the difference in the plots of the two models is the same as the reason from the experiment in the previous subsection. That is, in the TAT model, each author has only a topic proportion without time span distribution, so its representation of interest patterns is poor than the ATF model in which each author has research interest patterns.

4.5 Author ranking

The purpose of this experiment was to show that the ATF model is more effective in finding authors who are most interested in a particular topic for each year. For the topic *Descriptive language*, the top five authors obtained from the TAT model and from the ATF model are shown Tables 7, 8, respectively. The result of the TAT model puts Ian Horrocks in the top rank for every year, and the top five authors in 2007 are also top ranked in all the other years. In contrast, in the result of the ATF model, the set of the top five authors of each year is different from the set of the top ranked authors of the other years. To identify which result is more consistent with the dataset, we manually

Table 7 Top five authors who are interested in the topic Descriptive language, obtained from the TAT model

Years	2007	2008	2009	2010	2011
Rank 1	Ian	Ian	Ian	Ian	Ian
	Horrocks	Horrocks	Horrocks	Horrocks	Horrocks
Rank 2	Ulrike	Bernardo	Bernardo	Yevgeny	Carsten
	Sattler	Cuenca	Cuenca	Kazakov	Lutz
Rank 3	Birte	Ulrike	Carsten	Bernardo	Birte
	Glimm	Sattler	Lutz	Cuenca	Glimm
Rank 4	Bernardo	Birte	Yevgeny	Birte	Bernardo
	Cuenca	Glimm	Kazakov	Glimm	Cuenca
Rank 5	Yevgeny	Carsten	Birte	Carsten	Yevgeny
	Kazakov	Lutz	Glimm	Lutz	Kazakov

Table 8 Top five authors who are interested in the topic Descriptive language, obtained from the ATF model

Years	2007	2008	2009	2010	2011
Rank 1	Ulrike	Ian	Boris	Ian	Ian
	Sattler	Horrocks	Motik	Horrocks	Horrocks
Rank 2	Ian	Ulrike	Bernardo	Boris	Frank
	Horrocks	Sattler	Cuenca	Motik	Wolter
Rank 3	Yevgeny	Bernardo	Ian	Bernardo	Ilianna
	Kazakov	Cuenca	Horrocks	Cuenca	Kollia
Rank 4	Birte	Boris	Carsten	Yevgeny	Carsten
	Glimm	Motik	Lutz	Kazakov	Lutz
Rank 5	Bernardo	Birte	Yevgeny	Giorgos	Birte
	Cuenca	Glimm	Kazakov	Stoilos	Glimm

made a table of the top five authors according to the number of papers written by the authors for each year, as shown in Table 9. The top ranked author in Table 9 is Ian Horrocks every year, which may seem consistent with the result of the TAT model. If we see, however, the set of authors for each year, then Table 9 is more consistent with the result of the ATF model as there are more common authors between Table 8 and Table 9 for every year. As each author has only a topic proportion covering all the years in the TAT model, the research interest of each author for a particular year is smoothed by the temporal patterns of topics, as described in Eq. (7). In other words, every author is more likely to share similar temporal patterns of research interests. This makes the ranked authors in 2007 are also ranked in other years, finally resulting in poor results. The ATF model allows each author to have a unique temporal pattern of research interest, so it is more sensitive to the variations of research interests of each author. For example, in the year 2011, the author Frank Wolter appears in the

Table 9 Top five authors who are interested in the topic Descriptive language, obtained by counting the number of papers

2007		2008		2009		2010		2011	
Ian	10 (8)	Ian	10 (8)	Ian	10 (7)	Ian	10 (6)	Ian	10 (8)
Horrocks		Horrocks		Horrocks		Horrocks		Horrocks	
Ulrike	8 (8)	Ulrike	6 (6)	Bernardo	7 (6)	Boris	5 (4)	Frank	4 (4)
Sattler		Sattler		Cuenca		Motik		Wolter	
Yevgeny	4 (4)	Bernardo	5 (4)	Boris	7 (6)	Bernardo	6 (3)	Carsten	4 (4)
Kazakov		Cuenca		Motik		Cuenca		Lutz	
Bernardo	4 (4)	Boris	4 (3)	Carsten	4 (3)	Giorgos	3 (3)	Ilianna	3 (3)
Cuenca		Motik		Lutz		Stoilos		Kollia	
Birte	3 (3)	Birte	2 (2)	Frank	3 (3)	Yevgeny	4 (2)	Birte	3 (3)
Glimm		Glimm		Wolter		Kazakov		Glimm	

The digits are the number of papers written by the authors each year, and the digits between brackets are the number of papers about the topic written by the authors

top ranked author list although he is not ranked in all the other years, which is more consistent with the dataset as described in Table 9.

4.6 Finding authors similar to a particular author

In this experiment, we find a list of authors who are similar to a particular author in a given year. We assume that the similar authors should have similar research interests. The similarity between authors is measured by symmetric KullbackLeibler (sKL) divergence between topic proportions of authors of each year, where the sKL divergence(a, b) = $KL(a, b) + KL(b, a)$. As sample results, the top three authors similar to Ian Horrocks obtained from the TAT model and the ATF model are described in Tables 10, 11, respectively. The digits in the tables are the number of papers co-authored with Ian Horrocks, so more similar authors should have bigger digits. The similar authors obtained from the ATF model have bigger digits than the authors obtained from the TAT model, which implies that the ATF model is more effective in finding authors similar to a particular author. For example, the TAT model says that the most similar author to Ian Horrocks in 2007 should be Yevgeny Kazakov whose number of co-authored papers is four. On the other hand, the ATF model says that the most similar author is Ulrike Sattler who wrote eight papers with Ian Horrocks in 2007. With authors other than Ian Horrocks, we also observed similar results. The reason is that the ATF model directly captures the temporal patterns of each author's research interest, while the TAT model does not.

4.7 Author prediction on unseen documents

As the objective was to get temporal patterns of research interests of authors, we can measure the model by the performance of author prediction task on unseen documents.

Table 10 Top three authors similar to Ian Horrocks obtained from the TAT model for each year

2007		2008		2009		2010		2011	
Yevgeny	4	Bernardo	5	Bernardo	6	Bernardo	4	Carsten	0
Kazakov		Cuenca		Cuenca		Cuenca		Lutz	
Bernardo	4	Birte	4	Carsten	0	Yevgeny	3	Birte	3
Cuenca		Glimm		Lutz		Kazakov		Glimm	
Birte	3	Ulrike	6	Yevgeny	0	Carsten	0	Bernardo	3
Glimm		Sattler		Kazakov		Lutz		Cuenca	

The digits are the number of papers co-authored with Ian Horrocks

Table 11 Top three authors similar to Ian Horrocks obtained from the ATF model for each year

2007		2008		2009		2010		2011	
Ulrike	8	Ulrike	6	Bernardo	6	Boris	4	Frank	4
Sattler		Sattler		Cuenca		Motik		Wolter	
Yevgeny	4	Bernardo	5	Carsten	0	Yevgeny	3	Birte	3
Kazakov		Cuenca		Lutz		Kazakov		Glimm	
Birte	3	Birte	4	Hector	2	Bernardo	4	Bernardo	3
Glimm		Glimm		Perez		Cuenca		Cuenca	

The digits are the number of papers co-authored with Ian Horrocks

The prediction process of unseen documents is described in Fig. 11. We divided the dataset into 10 subsets, and performed tenfold cross-validation. We assumed that unseen documents were written by the known authors of the training dataset, so we ignored the authors who appeared only in unseen documents. We again compared the ATF model with the TAT model, and each model was used to rank all the known authors for each unseen document for a given publish year of the document. The ranking process is done using the symmetric KL divergence between the topic distribution of unseen document and the topic distribution of authors. For each unseen document, we collected the gap between the predicted author ranks and the ground-truth authors who are assumed to be the first rank. The mean and standard deviation of the collected gaps are depicted in Fig. 12 with various settings for the number of topics from 10 to 90, where the lower mean and standard deviation should be better. The performances increased as the number of topics T increased from 10 to 50, because the latent research interests were well discovered when $T = 50$. The performances declined and the gap between the two models decreased when the number of topics T was greater than 50, because redundant topics were generated when $T > 50$. Both models had their best performances when $T = 50$, and the ATF model generally exhibited better performances than the TAT model. It is worth noting that the purpose of ATF model is not the task of author prediction. The ATF model is designed to grasp the temporal patterns of research interests of authors.

Fig. 11 Prediction process of unseen documents

Fig. 12 Author prediction results from the TAT model and the ATFM model on unseen documents: the mean of predicted ranks (*left*) and the standard deviation of predicted ranks (*right*). The horizontal axis represents the number of topics

4.8 Efficiency of models

We compared training time of three models (AT model, TAT model, and ATF model). For every model, we use the same machine of dual core 3.20GHz CPU, 8GB RAM, 500GB HDD with Microsoft windows 7. We used Java and Eclipse Juno for implementation and experiments. From 10 to 90 topics, parameter approximation time with a fixed number of iterations (e.g., 1000 iterations) using a collapsed Gibbs sampling is measured and averaged. The result is shown in Table 12, which shows that AT model takes the shortest time because the sampling function of parameter approximation is simplest among the three models. The TAT model takes little more time than the ATF model. The reason is that the TAT model has to update the topic-year distribution for

Table 12 Spent time for training three models (seconds)

# of topics	10	30	50	70	90	Average
AT model	126.8	372.9	568.6	856.3	1013.1	587.54
TAT model	147.2	520.9	682.3	981.3	1212.1	708.76
ATF model	132.7	418.8	617.9	941.7	1105.1	1105.1

each iteration, while the time span distribution ψ of the ATF model can be exactly computed without an iterative approximation. Thus, our proposed ATF model is not only more effective than the TAT model, but also efficient in terms of the training time.

5 Conclusion

The amount of Web documents increases exponentially, which makes it necessary to develop systems or models to automatically capture some latent patterns among the documents. The latent patterns typically change over time, which compose topical flows or trends. In this paper, we propose Author Topic-Flow (ATF) model, which effectively discovers research trends in each author's point of view. The state-of-the-art model, namely Temporal Author Topic (TAT) model, combines the temporal patterns of topics to compute the research trends of the authors, so we denote it as *indirect topic flow*. On the other hand, our proposed model has a variable to directly represent the research trends of each author, so we denote it as *direct topic flow*. That is, it allows each author to directly have a temporal pattern of research interests, while each author in the TAT model has only a topic proportion covering all the time spans. We performed empirical comparisons with the TAT model with a real-world dataset, and we proved that the ATF model is more effective and efficient in capturing the temporal patterns of research interests of authors. We hope that this study would be helpful in some other research areas such as user-adapted Web content recommendation on mobile platform [18] or out-of-domain detection of intelligent dialog systems.

Acknowledgements This work was supported by Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korean Government (MSIP) (No. 2013-0-00131, Development of Knowledge Evolutionary WiseQA Platform Technology for Human Knowledge Augmented Services). This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP; Ministry of Science, ICT & Future Planning) (No. 2017017836).

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

1. Andrieu C, de Freitas N, Doucet A, Jordan MI (2003) An introduction to MCMC for machine learning. *J Mach Learn* 50(1):5–43

2. Blei DM, Lafferty JD (2006) Dynamic topic models. In: *Proceedings of the 23rd International Conference on Machine Learning*. Pittsburgh, USA, pp 113–120
3. Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. *J Mach Learn Res* 3:993–1022
4. Chang J, Boyd-Graber J, Blei DM (2009) Connections between the lines: augmenting social networks with text. In: *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. France, Paris, pp 169–178
5. Cui X, Zhu P, Yang X, Li K, Ji C (2014) Optimized big data k-means clustering using mapreduce. *J Supercomput* 70(3):1249–1259
6. Daud A (2012) Using time topic modeling for semantics-based dynamic research interest finding. *Knowledge-Based Syst* 26:154–163
7. Daud A, Li J, Zhou L, Muhammad F (2010) Temporal expert finding through generalized time topic modeling. *Knowledge-Based Syst* 23(6):615–625
8. Diederich J, Kindermann J, Leopold E, Paass G (2003) Authorship attribution with support vector machines. *Appl Intell* 19(1):109–123
9. Du L, Buntine W, Jin H (2010a) A segmented topic model based on the two-parameter poisson-dirichlet process. *Mach Learn* 81(1):5–19
10. Du L, Buntine WL, Jin H (2010) Sequential latent dirichlet allocation: discover underlying topic structures within a document. In: *Proceedings of 2010 IEEE International Conference on Data Mining*. Australia, Sydney, pp 148–157
11. Erten C, Harding PJ, Kobourov SG, Wampler K, Lee G (2004) Exploring the computing literature using temporal graph visualization. In: *Conference on Visualization and Data Analysis 2004*, San Jose, USA, pp 45–56
12. Gray A, Sallis P, Macdonell S (1997) Software forensics: Extending authorship analysis techniques to computer programs. In: *Proceedings of the 3rd Biannual Conference of the International Association of Forensic Linguists*. Durham, USA, pp 1–8
13. Griffiths TL, Steyvers M (2004) Finding scientific topics. In: *Proceedings of the National Academy of Sciences of the United States of America*, pp 5228–5235
14. Hofmann T (1999) Probabilistic latent semantic analysis. In: *Proceedings of Uncertainty in Artificial Intelligence*. Stockholm, Sweden, pp 289–296
15. Jain AK, Dubes RC (1988) *Algorithms for clustering data*. Prentice Hall, Englewood Cliffs
16. Jeong Y-S, Choi H-J (2012) Sequential entity group topic model for getting topic flows of entity groups within one document. In: *Proceedings of the 16th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*. Kuala Lumpur, Malaysia, pp 366–378
17. Jo Y, Oh AH (2011) Aspect and sentiment unification model for online review analysis. In: *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, HongKong, pp 815–824
18. Lee W, Leung CKS, Lee JJH (2011) Mobile web navigation in digital ecosystems using rooted directed trees. *IEEE Trans Ind Electron* 58(6):2154–2162
19. Liu L, Tang L, Dong W, Yao S, Zhou W (2016) An overview of topic modeling and its current applications in bioinformatics. *Springerplus* 5(1):1–22
20. Mimno D, McCallum A (2007) Expertise modeling for matching papers with reviewers. In: *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Jose, USA, pp 500–509
21. Mutschke P (2003) Mining networks and central entities in digital libraries. A graph theoretic approach applied to co-author networks. In: *Proceedings of the 5th International Symposium on Intelligent Data Analysis (IDA)*. Berlin, Germany, pp 155–166
22. Newman D, Chemudugunta C, Smyth P (2006) Statistical entity-topic models. In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Philadelphia, USA, pp 680–686
23. Poon H, Domingos P (2010) Unsupervised ontology induction from text. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden, pp 296–305
24. Rosen-Zvi M, Chemudugunta C, Griffiths T, Smyth P, Steyvers M (2010) Learning author-topic models from text corpora. *ACM Trans Inf Syst* 28(1):1–38
25. Steyvers M, Smyth P, Rosen-Zvi M, Griffiths T (2004) Probabilistic author-topic models for information discovery. In: *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Seattle, USA, pp 306–315
26. Sun X, Yung NHC, Lam EY (2016) Unsupervised tracking with the doubly stochastic dirichlet process mixture model. *IEEE Trans Intell Transp Syst* 17(9):2594–2599

27. Tang J, Zhang J, Yao L, Li J, Zhang L, Su Z (2008) Arnetminer: extraction and mining of academic social networks. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas, USA, pp 990–998
28. Vulic I, Smet WD, Tang J, Moens MF (2015) Probabilistic topic modeling in multilingual settings: an overview of its methodology and applications. *Inf Process Manag* 51(1):111–147
29. Wang X, McCallum A (2006) Topics over time: a non-markov continuous-time model of topical trends. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, USA, pp 424–433
30. White S, Smyth P (2003) Algorithms for estimating relative importance in networks. In: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, USA, pp 266–275
31. Yasami Y, Mozaffari SP (2010) A novel unsupervised classification approach for network anomaly detection by k-means clustering and id3 decision tree learning methods. *J Supercomput* 53(1):231–245