Correlated Author-Topic: a Supervised Topic Model for Identifying Best Interactions in CQA

Siyang Wang,Tong Zhao,Chunping Li, and Ran Chen.

***Abstract*—As having become a very popular online platform for people to share professional knowledge, Community Question Answering (CQA) system stores multiple interactions between *askers, answerers, reviewers,* and *voters* into individual documents. Because most of these interactions do not apply with equal specificity across the whole text, thus they have multinomial social influences-make it difficult to identify individual authorities. In this paper, we present Correlated Author-Topic (C-AT), a supervised graphical probabilistic model based on Author-Topic (AT) and Author-Recipient-Topic (ART), by defining an authority-based mixture for asker/responder pairs in each document with a predefined prior knowledge. Associating a mixture with each pair, C-AT generates a topic distribution that is conditioned distinct on different interactions between correlated authors. We cast C-AT within a specific prior in order to obtain the best interaction. We demonstrate that C-AT's improved expressiveness and performances over AT and ART with visualizations of datasets from Yahoo! Answers. Experiments show that C-AT helps to build a novel way toward solving the topical experts identifying problem with interpreting in user interactions.**

***Index Terms*—topic model, probabilistic, content, social influence analysis.**

I. Introduction

AS flourishing of user-generated contents (UGC) services such as Wikipedia, Yahoo! Answers, YouTube and Flickr, a significant proportion of the world's textual data contains structural-type information. Embedded in these services, increasingly types of interaction between users have emerged as the developing of various user interfaces, such as posting, reviewing, voting, thumb upping, et al,. The collection of interactions from a single document means that an interaction between two users may have a unique social influence. For example, given a document in Community Question Answering (CQA) system, an asking/best-responding interaction generates a stronger influence than any other general asking/responding interactions do. A link from *asker* to the *best responder* denotes an endorsement for the quality of *best responder*'s answer. Now we define the "best interaction" for CQA as the most strongest interaction, which could reflect choices people make about what information is useful, interesting and authoritative.

One simple probabilistic-based approach used to model authors and topics can be found in Author-Topic (AT) [1], which explicitly identifies how users are interested in this topic by associating individual words in a text with their most appropriate authors. AT has derived from Latent Dirichlet Allocation [2]. Another promising approach is an AT-extended model, which we called as Author-Recipient-Topic [3]. Although ART models the sender-recipient structure over topics, it samples a recipient in at uniform manner as if each sender/recipient pair has the same authorities on text. Therefore, we manage to add some new attributions on ART - the authority distribution of asker and responders.

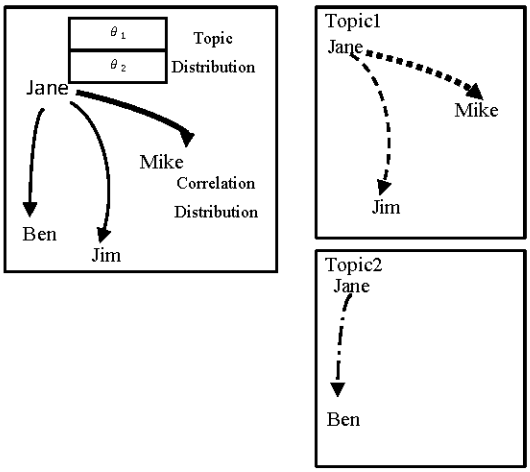

Fig. 1. A frame of correlated authors and topics in each document.

In order to achieve this goal, we put the information into a frame in Figure 1. The left figure illustrates generative process for each document: a generating of a coauthor network between four correlated users in a multinomial distribution with a specific prior, as well as a topic distribution. The right figure shows words' generating process, remaining similar with AT's. Thus, the crucial goal is how to effectively and efficiently obtain the multinomial distribution for correlated authors in each document in CQA.

In this paper, we present Correlated Author-Topic (C-AT) with modeling the correlated authors' authorities and interests into a probabilistic distribution. Extended from AT and ART, C-AT creates a mixture with a specific prior for correlated authors and marries the authority supervision common to modern texts with the author assignment ambiguity resolution of the AT-family models. In contrast to ART, C-AT mixes a probabilistic distribution for each asker/responder pairs in order to conditionally identify the authorities. C-AT is shown to be a natural extension of both topic model and social influence analysis as:

- incorporating an authority-based mixture,
- generating a user graph for each document.

The following sections are organized as: Section II formally formulates the previous works; Section III explains the proposed approach. Section IV presents experimental results that validate the computational efficiency of our methodology. Finally, Section V concludes.

Manuscript received April 3rd, 2012; revised April 16th, 2012.
S. Wang, T. Zhao, C. Li and R. Chen are with the School of Software, Tsinghua University, Beijing, 100084, China. e-mail: thu.wsy@gmail.com.

ISBN: 978-988-19251-3-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

II. RELATED WORKS

A. Modeling Authors and Topics

Employing Athor-Topic (AT) and Author-Recipient-Topic (ART) for CQA corpus, we could generate the underlying topics for each author (or each sender-recipient pair), as showed in Figure 2. First, we formulate the document from CQA corpus. A single document usually consists of three parts, i.e., the *subject* (a brief statement of the question), the *content* (additional detailed descriptions of the question), the *answers' content* posted by other users, and the *users' id*. We define the document as *associated text* [4], which is the concatenation of its subject, content, answers and users' id. Formally, we let D denote the collection of document, U denote the collection of users in CQA. When considering the associate text, each document $d \in D$ has a collection of authors $U^d \in U$. Each author $u^d \in U^d$ can be defined as a U-dimensional vector as:

$$u^d = \{u_1, ..., u_U\} \tag{1}$$

where $\{u_1, ..., u_U\} \in \{0, 1\}$. In AT's generative process for each document d, the set of d's authors, U^d, is observed. To generate each word, an author x is chosen at uniform from this set, then a topic z is selected from a topic distribution θ_x that is specific to the author, and then a word w is generated by sampling from a topic-specific multinomial distribution ϕ_z.

We employ ART on modeling sender/recipient pairs and topics for CQA: we treat each document d as a piece of message sent between an asker and corresponding responders. In its generative process for each document d, an asker a_d, and a set of responders, r_d, are observed. To generate word, a responder, x, is chosen at uniform from r_d, and then a topic z is chosen from a multinomial topic distribution $\theta_{(a_d,x)}$, where the distribution is specific to the asker/responder pair (a_d, x). Finally, the word w is generated by sampling from a topic-specific multinomial distribution ϕ_z.

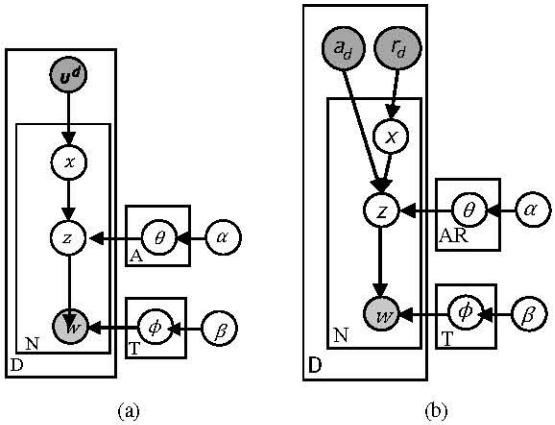

Fig. 2. Using AT and ART modeling authors and topics for CQA. The 2a goes to AT, the 2b goes to ART.

B. Social Influence Studies

Graph-theoretic approaches have been widely applied to social influence study, the majority of which are transferred to a link analysis problem, such as PageRank [5] and HITS [6]. They provide a score of the relative authority of each node in user graph in CQA [7] [8] [9] [10] [11]. The PageRank assumption is that a node transfers its PageRank values evenly to all the nodes it connects to. HITS algorithm introduces some special nodes that act as hubs in user graph. Askers can act as hubs and best responders can act as authorities.

Both of PageRank and HITS have some weakness and sparseness when applying to CQA [12]. A simplest method used to rate the authority of each user can be found in InDegree algorithm [13]. InDegree measures the authority of a node by the number of best nodes that link to this node. It is reasonable to assume that a user who answers 100 questions with 50 best answers among them is more professional than a user who answers 200 questions with 0 best answers. We employ InDegree algorithm to evaluate authorities between asker and responder. Let $I(y_i)$ denotes the InDegree value of a user u_i. We normalize the InDegree in such a way that $\sum_{i=1}^{U}(y_i)^2 = 1$. The normalized InDegree provides a relative score of the authority of each user. Let an InDegree value of B denotes the authority between asker A and responder B, as $A(a, r) = I(y_r)$.

III. OUR METHOD

A. Correlated Author-Topic Model

Correlated Author-Topic (C-AT) is a probabilistic model that describes a supervision for automatically generating an authority-based mixture into the process of author and topic mixtures for each document in CQA corpus. C-AT creates an authority-based mixture Π from each document, with a specific prior knowledge η. Like ART model, we use pair (a, r) to denote the correlate author. We assume that all junk responds have been eliminated. A document d is represented as a vector of words w_d, with N_d entries. A CQA corpus with D documents is represented as a concatenation of the document vectors, which we will denote w, having $N = \sum_{(d=1)}^{D} N_d$ entries. Let U be a set of users in CQA, U_a be a set of askers, and U_r be a set responders, where $U = U_a \cup U_r$. (There could be a large proportion of users that acted as voter or reviewer, although we did not find those to be necessary in our experiments.) We describe the generative process in Figure 3, and the notations in Table I.

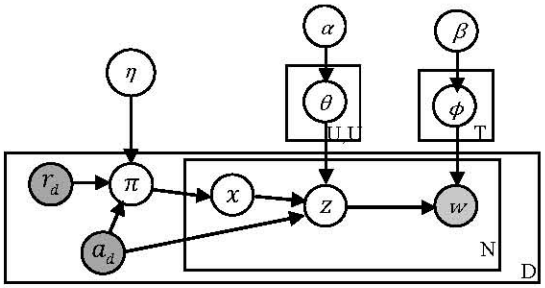

Fig. 3. A probabilistic graphical model of C-AT

1) Defining an Authority-Based Mixture: Given a single document d, a set of coauthors $u_d = a_d \cup r_d$ is observed, where d contains a unique asker a_d and a set of responders r_d. Discovering a definite asker a_d in document d, a multinomial variable π_{r_a} over responders is sampled from a Dirichlet prior η. Under generating a vector π_{r_a} over responders for an asker a, we could define a coauthor network for each

ISBN: 978-988-19251-3-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

TABLE I
NOTATIONS

Symbol	Description
C^{TAR}	Number of words assigned to correlation and topic
C^{WT}	Number of words assigned to topic
C^{RA}	Number of responders assigned to author
w_{di}	i^{th} word in the d^{th} document
x_{di}	Correlation assignment for word w_{di}
z_{di}	Topic assignment for word w_{di}
α, β, η	Dirichlet prior
π_{r_a}	Probabilities of responders r could influence an asker a
ϕ_t	Probabilities of words given topic t
θ_{ar}	Probabilities of topics given pair (a, r)

document. Taking authority-based mixture in background, C-AT subsequently conducts generative process for each author, each topic and each document. We describe the generative process of C-AT in Procedure 1.

Step 1 to 3 draw the generating of authority-based mixture. Remaining similar as ART, Step 4 to 9 draw the author and topic mixtures, from Dirichlet prior α and β. As a generative process for each document in Step 10 to 17, a variable x is chosen from π_{r_a} according to the unique asker a_d, then a topic z is selected from a topic distribution θ_x that is specific to that topic, finally, a word w is generated by sampling from a topic-specific multinomial distribution ϕ_z. Within an authority-based mixture, we could add some restrictions that ensure all topic assignments are limited to the correlated authors that have specific authorities.

Procedure 1 Generative Algorithm of C-AT

1: **for** each document d in $\{1, \ldots, D\}$ **do**
2: Generate $\pi_{r_a} \sim Dirichlet(\eta)$
3: **end for**
4: **for** each correlated author (a, r) in $\{1, \ldots, U\}$ **do**
5: Generate $\theta_{a,r} \sim Dirichlet(\alpha)$
6: **end for**
7: **for** each topic t in $\{1, \ldots, T\}$ **do**
8: Generate $\phi_t \sim Dirichlet(\beta)$
9: **end for**
10: **for** each document d in $\{1, \ldots, D\}$ **do**
11: Given Π
12: **for** each word i in $\{1, \ldots, N_d\}$ **do**
13: Generate $x_{di} \sim Multinomial(\pi_{r_a})$
14: Generate $z_{di} \sim Multinomial(\theta_{x_{di}})$
15: Generate $w_{di} \sim Multinomial(\phi_{z_{di}})$
16: **end for**
17: **end for**

2) Casting a Prior Knowledge: Considering the application of identifying best interactions, it would be nice if we establish the most authoritative interaction like asker/best-responder as a predefined facet and guide C-AT model to generate most relevant correlated authors over topics. We see that an authority-based mixture Π models the dependency between askers and responders. Each elements in π_{r_a} represents the responder that have the highest probability to influence asker a. Towards this objective, we define a feature function $l(a_d, r_d, b_d)$ for each document d, where $b_d \subset r_d$ denotes a best-responder of d. Let $l(a_d, r_d, b_d)$ be a multinomial as Equation 2:

$$l(a_d, r_d, b_d) = \begin{cases} l(a_d, r_d) = 1 \\ l(a_d, b_d) = 10 \\ otherwise \quad 0 \end{cases} \tag{2}$$

This function reflects the authority feature of correlated authors. We also define a global feature function $l_D(a_d, r_d, b_d)$ as:

$$l_D(a_d, r_d, b_d) = \sum_{d=1}^{D} l(a_d, r_d, b_d). \tag{3}$$

We use the Equation 3 to project the parameter vector of the Dirichlet prior η as:

$$\eta = l_d \times \eta = (\eta_{l_{D_1}}, \ldots, \eta_{l_{D_A}}). \tag{4}$$

Intuitively, $l_D(\cdot)$ constrains the model to bias towards the "best-responding" representative interactions.

B. Modeling Generative Process

In C-AT's generative process, each topic is drawn independently when conditioned on Θ, and each word is drawn independently when conditioned on Φ and z. The probability of the corpus w, conditioned on Θ, Φ and Π (implicitly on a fixed number of topics T), is: $P(w|\Theta, \Phi, \Pi, a_d, r_d) = \Pi_{d=1}^{D} P(w_d|\Theta, \Phi, \Pi, a_d, r_d)$. We can obtain the probability of the words in each document, w_d, by summing over the latent variables x and z, to give:

$$\begin{aligned} &P(w_d|\Theta, \Phi, \Pi, a_d, r_d) = \\ &\sum_{i=1}^{N_d} \sum_{r=1}^{R} \sum_{t=1}^{T} P(w_{di}, z_{di} = t, x_{di} = ra|\Theta, \Phi, \Pi, a_d, r_d) = \\ &\sum_{i=1}^{N_d} \sum_{r=1}^{R} \sum_{t=1}^{T} \phi_{w_{di}t} \theta_{tx_{di}} \pi_{x_{di}} \end{aligned} \tag{5}$$

Equation 5 expresses the probability of the words w in terms of the entries of the parameter matrices Θ, Φ and Π. The probability distribution over responders assignments in C-AT, $P(x_{di} = ra|\pi)$, is assumed to be multinomial of the elements of π, this multinomial distribution indicates the latent authorities that can influence the CQA. The probability distribution over topic assignments, $P(z_{di} = t|x_{di} = ra, \Theta)$ is the multinomial distribution θ_{ar} in Θ that corresponds to pair (a, r), and the probability of a word given a topic assignment, $P(w_{di}|z_{di} = t, \Phi)$ is the multinomial distribution ϕ_t in Φ that corresponds to topic t. We treat Θ, Φ, Π as random variables, and compute the marginal probability of a corpus by integrating them out:

$$\begin{aligned} &P(w|U, \alpha, \beta, \eta) = \\ &\int_{\Theta} \int_{\Phi} \int_{\Pi} P(w|a_d, r_d, \Theta, \Phi, \Pi) P(\Theta, \Phi, \Pi|\alpha, \beta, \eta) d\Theta d\Phi d\Pi \end{aligned} \tag{6}$$

C. Gibbs Sampling

C-AT contains three continuous random variables: Θ, Φ and Π. We use Markov Chain Monte Carlo (MCMC) algorithm or more specifically, Gibbs sampling, for its ease

ISBN: 978-988-19251-3-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

of implementation. Our inference scheme is based upon the observation that:

$$P(\Theta, \Phi, \Pi | D^{train}, \alpha, \beta, \eta) = \Sigma_{z,x} P(\Theta, \Phi, \Pi | z, x, D^{train}, \alpha, \beta, \eta) P(z, x, | D^{train}, \alpha, \beta, \eta) \quad (7)$$

We wish to construct a Markov chain that converges to the posterior distribution over x and z conditioned on training set D^{train}, α, β and η. Using Gibbs sampling, we can generate a sample from the joint distribution $P(z, x, | D^{train}, \alpha, \beta, \eta)$ by

- sampling a correlation assignment x_{di} and a topic assignment z_{di} for an individual word w_{di}, conditioned on a multinomial assignment of responders and a fixed assignment of topics for all other words in the corpus,
- repeating this process for each word.

A single Gibbs sampling iteration consists of sequentially performing this sampling of author and topic assignments for each individual word in the corpus. We see that joint distribution of the set of word tokens in the corpus w, the set of topic assignments z, the set of correlation assignments x, the set of author mixtures Θ, the set of topic mixtures Φ, and the set of authority-based mixtures Π, given the Dirichlet hyperparameters α, β and η, and the user set U, can be simplified as follows:

$$P(w, z, x, \Theta, \Phi, \Pi | \alpha, \beta, \eta, U) = C[\prod_{t=1}^{T} \prod_{w=1}^{W} \phi_{wt}^{C_{wt}^{WT} + \beta - 1}][\prod_{ar=1}^{A \times R} \prod_{t=1}^{T} \theta_{tar}^{C_{tar}^{TAR} + \alpha - 1}][\prod_{a=1}^{A} \prod_{r=1}^{R} \pi_{r_a}^{C_{ra}^{RA} + \eta - 1}] \quad (8)$$

where $C = [\frac{\Gamma(W\beta)}{\Gamma(\beta)^W}]^T [\frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T}]^{A \times R} [\frac{\Gamma(R\eta)}{\Gamma(\eta)^R}]^A$. Because we define a specific prior knowledge for η in Equation 4, as $\eta = l_D \times \eta = (\eta_{l_{D_1}}, \ldots, \eta_{l_{D_A}}) = \eta_{ra}$. So we get:

$$P(w, z, x, \Theta, \Phi, \Pi | \alpha, \beta, \eta, U) \propto [\prod_{t=1}^{T} \prod_{w=1}^{W} \phi_{wt}^{C_{wt}^{WT} + \beta - 1}][\prod_{ar=1}^{A \times R} \prod_{t=1}^{T} \theta_{tar}^{C_{tar}^{TAR} + \alpha - 1}][\prod_{a=1}^{A} \prod_{r=1}^{R} \pi_{r_a}^{C_{ra}^{RA} + \eta_{ra} - 1}] \quad (9)$$

We integrate out the random variables Θ, Φ, Π, making use of the fact that the Dirichlet distribution is a conjugate prior of multinomial distribution:

$$P(w, z, x | \alpha, \beta, \eta, U) \propto [\prod_{t=1}^{T} \frac{\prod_{w=1}^{W} \Gamma(C_{wt}^{WT} + \beta)}{\Gamma(\Sigma_{w'} C_{w't}^{WT} + W\beta)}][\prod_{ar=1}^{A \times R} \frac{\prod_{t=1}^{T} \Gamma(C_{tar}^{TAR} + \alpha)}{\Gamma(\Sigma_{t'} C_{t'ar}^{TAR} + T\alpha)}] [\prod_{a=1}^{A} \frac{\prod_{r=1}^{R} \Gamma(C_{ra}^{RA} + \eta_{ra})}{\Gamma(\Sigma_{r'} C_{r'a}^{RA} + R\eta_{ra})}] \quad (10)$$

Now setting w, z, x to $w_{-di}, z_{-di}, x_{-di}$ respectively, where $z_{-di}, x_{-di}, w_{-di}$ stand for the vector of topic assignments, vector of correlation assignments, and vector of word observations in all corpus except for the i^{th} word of the d^{th} document, we could obtain the following Gibbs sampling equation for the C-AT model:

$$P(x_{di} = ra, z_{di} = t | w_{di} = w, z_{-di}, x_{-di}, w_{-di}, U, \alpha, \beta, \eta) \propto \frac{(C_{wt,-di}^{WT} + \beta)}{(\Sigma_{w'} C_{w't,-di}^{WT} + W\beta)} \frac{(C_{tar,-di}^{TAR} + \alpha)}{(\Sigma_{t'} C_{t'ar,-di}^{TAR} + T\alpha)} \frac{(C_{ra,-di}^{RA} + \eta_{ra})}{(\Sigma_{r'} C_{r'a,-di}^{RA} + R\eta_{ra})} \quad (11)$$

IV. Experimental Results

A. Experiment Setup

We perform our experiments on a set of real-world data from Yahoo! Answers, which is one of the biggest question-answering services all over the world. We take 3539 documents within 11109 authors from the top four popular categories: Politic & Government (PG), Pets (P), Beauty & Style (BS) and Health (H). The datasets' statistics are reported in Table II.

TABLE II
Datasets' Statistics

Category	Doc	User	Term	Responder	Avr.Authors
PG	838	2441	16050	4903	7
P	814	2654	14948	3438	6
BS	839	2681	6767	3650	6
H	1048	3332	12812	2234	4

The basic modeling algorithm is implemented using JAVA with Eclipse Indigo and all experiments with it are performed on a Server running Windows 2003 with two Dual-Core Intel Xeon processors (2.0GHz) and 8GB memory.

B. Comparing Perplexity for Different Models

The perplexity score of a new unobserved document d that contains words w_d, and is conditioned on the known correlated authors of the document a_d, r_d, is defined as:

$$Perp(w_d | a_d, r_d, D^{train}) = exp\{-\frac{logP(w_d | a_d, r_d, D^{train})}{N_d}\} \quad (12)$$

It could be noted that the lower the perplexity, the better the performance of the model.

In our experiments, we trained the C-AT, AT and ART models separately. We held out 10 percent of the data for test purposes and trained the models on the remaining 90 percent. The experiments are over only word distributions, since AT cannot differentiate the type of authors. We trained all the hidden variable models using Gibbs sampler with exactly the same number of samples. Figure 4 shows the results of the perplexity. It shows that C-AT yields significantly lower perplexity on the test set than AT and ART, which indicates that C-AT is a better generative model. Figure 4 also shows that the models trained using the Gibbs sampler appear to stabilize rather quickly (after about 100 iterations), at least in terms of perplexity on test documents.

ISBN: 978-988-19251-3-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

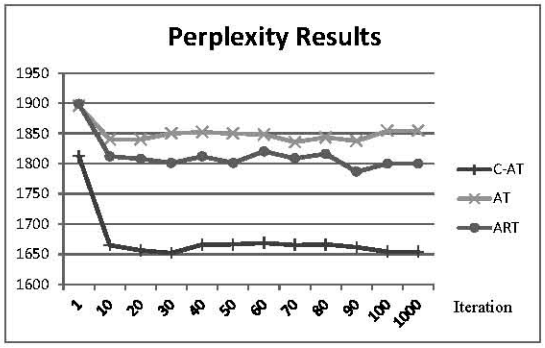

Fig. 4. Perplexity results of C-AT, ART and AT.

C. Authors Visualization

Figure 5 visualizes the C-AT's results. The clarity and specificity of topics are typically discovered by the model. Under the word distribution stands five correlated authors with the highest probability of correlated authors on discussing that topic (each pair taken up a raw, with the asker on the left of the responder.) We draw authors' nickname from user profile in Yahoo! Answers. The probabilities illustrate how strong the correlated authors contribute to that topic. For example, *(PB, Ahso)*, ranked No.1 in topic of Police & Law, was an asker/best-responder interaction. As being reported by Yahoo, the responder *Ahso* was chosen as a top contributor in the category of Politics & Government. He had answered 24304 questions and 43 percent of them were chosen as best answers. Implementing C-AT, we can find out all the best interactions over topics. Figure 5 also shows that the top 5 correlated authors over these 5 topics are all the real asker/best-responder pairs. Therefore, C-AT is able to gather out the authorities of best-responding correlations over topics.

POLICE & LAW		RELIGION		MILITARY SERVICE	
police help court gun life legal call laws law case guns order		Religion god Christian hate religious man called America church Muslim Christians		Military job army work school live force post days service months training	
PB--Ahso	0.622	Stev--Ranger	0.557	Nadine--NWIP	0.648
Meow--GFB	0.608	HDB--Destroyer	0.556	Mitch--Gonzo	0.627
John H--Bruce	0.579	Ame--Really	0.534	Nich M--Jessica	0.626
Soul Sbcnu	0.565	17--Ma G	0.524	Benjamin--Marine	0.61
Karl--Stuart	0.564	Domini--Bias	0.521	Gage--Tim	0.598
PUBLIC HEALTH		**WORLD WAR**		**ELECTION**	
Government country health care free system death private life movement		War Iran Israel Iraq military nuclear weapons country killed		Obama Santorum vote Paul Republican president Republicans Party	
Luke T--Susan	0.646	Scoo P--Trut M	0.890	Jim B--Billy B	0.683
Ethan--Mac K	0.628	Tom F--Wally	0.789	Sun S--Marian	0.563
Phoenix--Jason	0.542	Jaco B--C	0.776	T E--Tommy	0.551
Dusti--Lynn	0.526	Vincent--Herb	0.629	Alexa--Blue M	0.550
Gandhi--Ddd B	0.516	Siam V--Dylan	0.622	Kevin--Data U	0.506

Fig. 5. An illustration of C-AT results of 6 topics for the Politics & Government. Each topic is shown with the top 10 words. The titles are our own summary for the topics with following prominent best pairs for each topic.

D. Identifying Best Interactions

We now consider the case of identifying best interactions within a particular dataset (the time span is from February 2010 to March 2010). In our experiment, those documents that have at least 4 authors are considered. Specifically, the dataset contains 1023 documents, 5163 authors. We pick 143 documents as the test data, which contain 1082 authors. Recall the goal of best interactions finding is to identify user interactions with some expertise or experience on a specific topic (query) q. We define the baseline model as the combination of the ART $P(q|(a,r))$ and InDegree $A(a)$(Section II-B), as $\lambda A(a) + (1-\lambda)P(q|(a,r))$.

We compare C-AT with baseline method on how well it identifies the quality of the asker/best-responder pair. If the real best interaction is among the top N (say 1, 10 or 20) predicted pairs, the prediction for that topic is considered to be successful. We evaluate the performance of different top N in terms of $precision@1, precision@5, precision@10, precision@20$.

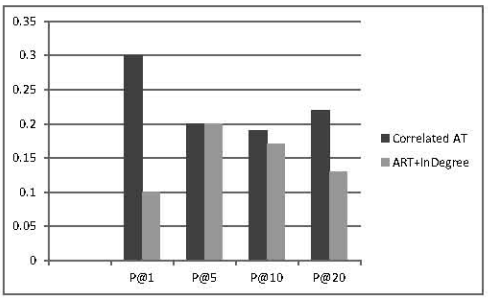

Fig. 6. Performance of identifying best interactions with different methods.

Figure 6 illustrates the result of identifying best interactions with different approaches. We see that the C-AT can indeed improve the accuracy, which confirms the effectiveness of the proposed approach.

V. Conclusion

In this paper, we conduct both authority and interest analysis of interactions in CQA. Considering interactions as a multinomial distribution, we create C-AT based on AT and ART models. C-AT extends ART along with a correlation mixture and a best-responder-biased prior. Experiments on Yahoo! Answers' datasets demonstrate that C-AT can effectively improve the performance over AT and ART models. C-AT has a good expressive visualization on user interactions and been demonstrated an improved accuracy on identifying best interactions over a competitive baseline method.

Extended from AT family, C-AT enables a range of natural extension for future investigation. The current model does not capture the labels of correlated pairs, which might be introduced by combining with Labeled LDA. C-AT lends itself naturally to modeling semi-supervised datasets where best-responding interactions are observed for some specific documents.

References

[1] M. Rosen-Zvi, T. L. Griffiths, M. Steyvers, and P. Smyth, "The Author-Topic Model for Authors and Documents," in *Uncertainty in Artificial Intelligence*, 2004, pp. 487–494.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[3] A. McCallum, A. . es Corrada-Emmanuel, and X. Wang, "The Author-Recipient-Topic Model for Topic and Role Discovery in Social Networks: Experiments with Enron and Academic Email."

[4] Y. Miao, C. Li, J. Tang, and L. Zhao, "Identifying new categories in community question answering archives: a topic modeling approach," in *International Conference on Information and Knowledge Management*, 2010, pp. 1673–1676.

ISBN: 978-988-19251-3-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

[5] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web," 1998.
[6] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of The ACM*, vol. 46, pp. 604–632, 1999.
[7] P. Jurczyk and E. Agichtein, "Discovering authorities in question answer communities by using link analysis," in *International Conference on Information and Knowledge Management*, 2007, pp. 919–922.
[8] M. Liu, Y. Liu, and Q. Yang, *Predicting Best Answerers for New Questions in Community Question Answering*, 2010.
[9] A. Pal, R. Farzan, J. A. Konstan, and R. E. Kraut, *Early Detection of Potential Experts in Question Answering Communities*.
[10] X. Liu, W. B. Croft, and M. B. Koll, "Finding experts in community-based question-answering services," in *International Conference on Information and Knowledge Management*, 2005, pp. 315–316.
[11] P. Jurczyk and E. Agichtein, "Hits on question answer portals: exploration of link analysis for author ranking," in *Research and Development in Information Retrieval*, 2007, pp. 845–846.
[12] M. Sannella, "Constraint satisfaction and debugging for interactive user interfaces," 1994.
[13] M. Bouguessa, B. Dumoulin, and S. Wang, "Identifying authoritative actors in question-answering forums: the case of Yahoo! answers," in *Knowledge Discovery and Data Mining*, 2008, pp. 866–874.

ISBN: 978-988-19251-3-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)