TaxoPhrase: Exploring Knowledge Base via Joint Learning of Taxonomy and Topical Phrases

Weijing Huang
Key Laboratory of High Confidence Software Technologies (Ministry of Education), EECS, Peking University
huangwaleking@gmail.com

Wei Chen*
Key Laboratory of High Confidence Software Technologies (Ministry of Education), EECS, Peking University
pekingchenwei@pku.edu.cn

Tengjiao Wang
Key Laboratory of High Confidence Software Technologies (Ministry of Education), EECS, Peking University
tjwang@pku.edu.cn

Shibo Tao
SPCCTA, School of Electronics and Computer Engineering, Peking University
expo.tao@gmail.com

ABSTRACT

Knowledge bases restore many facts about the world. But due to the big size of knowledge bases, it is not easy to take a quick overview onto their restored knowledge. In favor of the taxonomy structure and the phrases in the content of entities, this paper proposes an exploratory tool TaxoPhrase on the knowledge base. TaxoPhrase (1) is a novel Markov Random Field based topic model to learn the taxonomy structure and topical phrases jointly; (2) extracts the topics over subcategories, entities, and phrases, and represents the extracted topics as the overview information for a given category in the knowledge base. The experiments on the example categories *Mathematics*, *Chemistry*, and *Argentina* in the English Wikipedia demonstrate that our proposed TaxoPhrase provides an effective tool to explore the knowledge base.

KEYWORDS

Knowledge Base, Exploratory Tool, Topical Phrases, Taxonomy Structure, Topic Model, Markov Random Field

1 INTRODUCTION

Knowledge bases[5][3][10][13] are constructed elaborately to restore the information representing the facts about the world. And due to the big size of knowledge bases, it's necessary to provide an exploratory tool to take a quick overview on them. For example, there are more than 5.3 million articles in Wikipedia (March 2017)[1], far beyond the scale that human can read all. A suitable exploratory tool benefits the users of the knowledge base to have an overall perspective of the restored knowledge.

We attempt to achieve this purpose by answering the following three questions: (1) Q1, what are the main subtopics related to a given topic in knowledge base; (2) Q2, what are the related entities for each subtopic; (3) Q3, what are the key summarization corresponding to these subtopics. As many knowledge bases are constructed based on the Wikipedia, such as YAGO[10] and DBPeida[13], we answer the above three questions on Wikipedia without loss of generality. For this reason, in this paper the terms *page* and *entity* are used interchangeably.

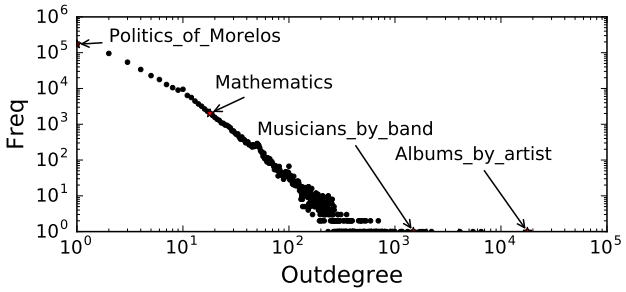

Figure 1: Distribution of categories' out degrees to subcategories, according to the study on the English Wikipedia.

Since categories in the knowledge base are used to group entities into similar subjects and are further organized hierarchically into the taxonomy, it seems straightforward to utilize the taxonomy to answer Q1 and Q2. But the size of taxonomy is too large to provide a quick overview for the whole knowledge base. Taking the English Wikipedia's taxonomy as an example, there are about 1.5 million category nodes. And the distribution of these categories' degrees is unbalanced, as shown in Figure 1. It means some categories contain too many subcategories, such as *Albums_by_artist*, whose out degree is 17,749; while most categories have very few subcategories, such as *Politics_of_Morelos* containing only one subcategory *Morelos_elections*. These characteristics indicate that the taxonomy structure is a large scale-free network[1]. So it's not easy to answer Q1 and Q2 directly only by the taxonomy structure.

Besides, the topic model[2], especially its extension on phrases, such as [6][9], is developed for the exploratory analysis on text corpora, and suitable for answering Q3. Usually the most frequent words or phrases in topics are used for summarizing the corpus[8]. However the meanings of the learned topics need to be manually interpreted[4], which may limit the usability of existing methods on Q3.

In this paper, we propose a novel exploratory tool TaxoPhrase, which learns the taxonomy structure and topical phrases in knowledge base jointly, and makes the questions Q1, Q2, Q3 tractable in a unified framework. The joint learning algorithm of TaxoPhrase is

*Corresponding author.
[1]https://stats.wikimedia.org/EN/TablesWikipediaEN.htm

OKBQA 2017, Tokyo,
2017. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

inspired by the complementary relation among the three parts in knowledge base: the categories in the taxonomy, the entities, and the phrases in entities' contents.

We take two examples to illustrate this kind of complementary relation, as shown in Figure 2. For the example ①, the Wikipedia page `Zeroth-order logic` directly belongs to two categories *Propositional_calculus* and *Systems_of_formal_logic*, which are descendant subcategories of the category *Mathematics*. The content of this page contains the phrases such as "zeroth-order logic", "first-order logic", "propositional calculus", and etc.. For the example ②, the Wikipedia page `Propositional calculus` belongs to five categories *Logical_calculi*, *Classical_logic*, *Propositional_calculus*, *Systems_of_formal_logic*, and *Boolean_algebra*, and contains the phrases "propositional calculus", "propositional logic", "logical connectives", and etc.. Obviously, these two pages share the similar categories in the taxonomy and the phrases in the content. Therefore, these two pages are more likely to correspond to the same subtopic *Mathematical_logic* under the category *Mathematics*. This fact is beneficial to answer Q1 and Q2. Meanwhile, the phrases shared by these two pages, e.g. "propositional calculus" and "propositional logic", are more likely to be grouped together as the topical phrases for the subtopic *Mathematical_logic*. These topical phrases are further used to give the answer of Q3.

To utilize the complementary relation among the three parts in the knowledge base, we extract the phrases and the related categories for each entity, and model them together in our proposed topic model TaxoPhrase.

To sum up, the contribution of our proposed TaxoPhrase is mainly in two aspects. (1) It is a novel Markov Random Field based topic model to learn the taxonomy structure and topical phrases jointly. (2) It extracts the topics over subcategories, entities, and phrases, and the extracted topics function as the overview information for a given category in the knowledge base. Furthermore, the experiments on example categories *Mathematics*, *Chemistry*, and *Argentina* in English Wikipedia demonstrate that our proposed method TaxoPhrase provides an effective tool to explore the knowledge base.

2 RELATED WORKS

To the best of our knowledge, there's few work on providing an explorative tool for the knowledge base. The most closest work to our motivation is Holloway's analyzing and visualizing the semantic coverage of Wikipedia[16], which visualizes the category network in two dimensions by the layout algorithm DrL (used to be VxOrd)[14]. However the layout doesn't provide the overview information of the knowledge base directly. The other works related to our approach can be grouped into two groups, which are taxonomy related and topical phrases related.

The taxonomy related works mainly focus on how to use the taxonomy of the knowledge base to enhance the quality of text mining tasks, such as Twixonomy[7], LGSA[11], and TransDetector[12].

The phrase related works extend the topic model to phrase level, such as ToPMine[6] and TPM[9]. ToPMine firstly extracts phrases by the frequent pattern mining on corpus, and secondly mine topical phrases on the "bag of phrases", in which the single word is treated as the shortest phrase. TPM reuses the phrases generated from

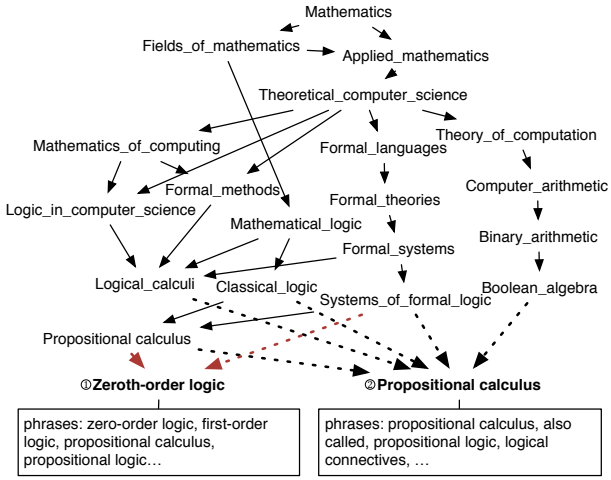

Figure 2: The illustration on the complementary relation among the three parts of knowledge base: the categories in the taxonomy, the entities, and the phrases in contents.

ToPMine. We follow it and use the phrases as the input of our tool. Considering the big size of the knowledge base, we run the LDA[2] on phrases as a baseline, rather than running the full ToPMine.

3 PROPOSED METHOD

3.1 Preprocessing

There are two parts to be extracted for each entity. The first are phrases, that are generated from ToPMine[6]. The second are the category-information, a.k.a., the set of *category* → *subcategory* edges related to the entity, defined formally in Definition 1.

Definition 1 (Category-Information). Given the entity d, the category information is a set of edges $s_{dn'} \rightarrow t_{dn'}$, where $t_{dn'}$ is the direct subcategory of $s_{dn'}$ and they are both the ancestors of the entity d. And we denote d's category-information as $\boldsymbol{c}_d = \{(s_{dn'}, t_{dn'})\}_{n'=1}^{N'_d}$.

For instance, all the categories in Figure 2 are the ancestors of the entity `Propositional calculus`, so all the 25 edges in Figure 2 are included in its category-information. As suggested by [12], we prune the cycles according to nodes' PageRank score to make the taxonomy as a Directed Acyclic Graph. We extract the category-information for a given entity on the taxonomy DAG.

3.2 TaxoPhrase

In this subsection, we present the model TaxoPhrase, which is illustrated in Figure 3. Since the phrases and category-information for each entity are already generated in the preprocessing phase, we treat them as the input of the TaxoPhrase model.

Joint Learning. Same as the traditional probabilistic topic models such as LDA[2], we assume that there are K topics in TaxoPhrase, and for each entity d we use the K-dimension vector θ_d to represent its latent topic distribution. The difference is that we model the categories and the phrases jointly. We denote the topics as $\{(\phi_k^{(\tau)}, \phi_k)\}_{k=1}^{K}$, where $\phi_k^{(\tau)}$ is the category distribution on the k-th topic, and ϕ_k is the word distribution on the k-th topic.

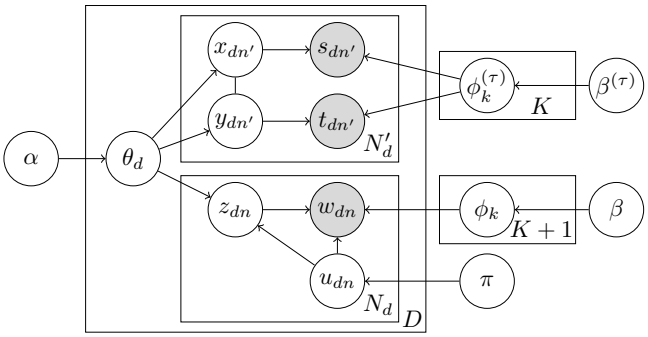

Figure 3: Illustration of the model TaxoPhrase

We connect the generation process of the phrases and the category-information by the entity-topic distribution θ_d. For each entity d, the input data include the category-information $\boldsymbol{c}_d = \{(s_{dn'}, t_{dn'})\}_{n'=1}^{N'_d}$ and the phrases $\boldsymbol{w}_d = \{w_{dn}\}_{n=1}^{N_d}$. We use the discrete values $\boldsymbol{x}_d = \{x_{dn'}\}_{n'=1}^{N'_d}$, $\boldsymbol{y}_d = \{y_{dn'}\}_{n'=1}^{N'_d}$ to represent the hidden topics of the category nodes $\boldsymbol{s}_d$ and $\boldsymbol{t}_d$ respectively. Correspondingly, we use the discrete values $\boldsymbol{z}_d$ for the phrases $\boldsymbol{w}_d$. The discrete topic assignments $\boldsymbol{x}_d$, $\boldsymbol{y}_d$, and $\boldsymbol{z}_d$ are all drawn from the same distribution $Multinomial(\theta_d)$, and are impacted by each other. Thus, the topic-category distribution $\phi_k^{(\tau)}$ and the topic-phrase distribution ϕ_k are aligned with each other.

Additionally, we use the background topic ϕ_0 to model the high frequent background phrases to enhance the topic modeling quality. The switcher variable u_{dn} is introduced for determining whether the phrase w_{dn} belongs to the background topic.

Markov Random Field on the taxonomy. As mentioned in Section 1, categories connected via an edge in the taxonomy tend to share the similar topic. Given the category edges $(s_{dn'}, t_{dn'})$, we put their latent topic assignments $x_{dn'}$ and $y_{dn'}$ into a Markov Random Field to capture this tendency. Specifically, we define the binary potential $\exp(\mathbb{I}(x_{dn'} = y_{dn'}))$ to encourage $x_{dn'}$ to have the same topic as $y_{dn'}$, where $\mathbb{I}(.)$ is the indicator function. And we use the unary potential $p(x_{dn'}|\theta_d)$ to link θ_d and $x_{dn'}$, which is defined by the multinomial distribution with the parameter θ_d as $p(x_{dn'}|\theta_d) = \prod_{k=1}^{K} \theta_{dk}^{\mathbb{I}(x_{dn'}=k)}$. The unary potential $p(y_{dn'}|\theta_d)$ is defined in the same way to link the entity's topic distribution θ_d and the topic assignment $y_{dn'}$ of the category $t_{dn'}$.

Because of the joint learning for categories and phrases, the entity's topic distribution θ_d also links with the phrases' topic assignments $\boldsymbol{z}_d$, which are generated with the probability $p(z_{dn}|\theta_d, u_{dn} = 1) = \prod_{k=1}^{K} \theta_{dk}^{\mathbb{I}(z_{dn}=k)\mathbb{I}(u_{dn}=1)}$. Therefore, the topic assignments share the following joint distribution.

$$p(\boldsymbol{x}_d, \boldsymbol{y}_d, \boldsymbol{z}_d|\theta_d, \boldsymbol{u}_d) = \frac{1}{A_d(\theta_d)} \prod_{n=1}^{N_d} p(z_{dn}|\theta_d, u_{dn}) \prod_{n'=1}^{N'_d} p(x_{dn'}|\theta_d) \cdot \prod_{n'=1}^{N'_d} p(y_{dn'}|\theta_d) \exp\{\sum_{n'=1}^{N'_d} \mathbb{I}(x_{dn'} = y_{dn'})\} \tag{1}$$

In Equation (1), A_d is the partition function to normalize the joint distribution. According to the Equation (1), the topic assignments do not only depends on the entity's topic distribution θ_d, but also depends on the other topic assignments in the Markov Random Field.

Generation Process. To sum up, given the hyper parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\beta}^{(\tau)}$, π, and the number of topics K, the generation process of the entities in the knowledge base can be described as follows.

1. Draw phrases' background topic $\boldsymbol{\phi_0} \sim Dir(\boldsymbol{\beta})$.
2. For each topic $k \in \{1, \cdots, K\}$,
   (a) draw phrase distribution on the topic $\boldsymbol{\phi_k} \sim Dir(\boldsymbol{\beta})$,
   (b) draw category distribution on the topic $\boldsymbol{\phi_k^{(\tau)}} \sim Dir(\boldsymbol{\beta^{(\tau)}})$.
3. For each entity index $d \in \{1, \cdots, D\}$,
   (a) draw the topic distribution on the entity $\boldsymbol{\theta_d} \sim Dir(\boldsymbol{\alpha})$,
   (b) for each phrase index $n \in \{1, \cdots, N_d\}$,
      (i.) draw the switcher $u_{dn} \sim Bernoulli(\pi)$,
   (c) draw topic assignments $\boldsymbol{x}_d$, $\boldsymbol{y}_d$, and $\boldsymbol{z}_d$ according to the Equation (1),
   (d) for each phrase index $n \in \{1, \cdots, N_d\}$,
      (i.) if $u_{dn} = 0$ draw the phrase $w_{dn} \sim Multinomial(\boldsymbol{\phi_0})$, else draw the phrase $w_{dn} \sim Multinomial(\boldsymbol{\phi_{z_{dn}}})$,
   (e) for $\boldsymbol{c}_d$'s each category edge index $n' \in \{1, \cdots, N'_d\}$,
      (i.) draw the category $s_{dn'} \sim Multinomial(\boldsymbol{\phi^{(\tau)}_{x_{dn'}}})$,
      (ii.) draw the category $t_{dn'} \sim Multinomial(\boldsymbol{\phi^{(\tau)}_{y_{dn'}}})$.

Inference. Firstly, we joint sample for z_{dn} and u_{dn} together according to the Equation (2) and (3), as z_{dn} is meaningful only when the switcher variable u_{dn} is set to 1. The sampling result of z_{dn} and u_{dn} depends on three parts, the entity's topic distribution $n_{d,k}$, the phrase's topic distribution $n_{k,v}$ and $n_{B,v}$, and the coin toss π. The sampling result also takes the impact from the categories into consideration, because $n_{d,k} = \sum_{n=1}^{N_d} \mathbb{I}(z_{dn} = k)\mathbb{I}(u_{dn} = 1) + \sum_{n'=1}^{N'_d} \mathbb{I}(x_{dn'} = k) + \sum_{n'=1}^{N'_d} \mathbb{I}(y_{dn'} = k)$.

$$p(z_{dn} = k, u_{dn} = 1|\boldsymbol{z_{\neg dn}}, \boldsymbol{u_{\neg dn}}, \boldsymbol{x}, \boldsymbol{y}, w_{dn} = v, \boldsymbol{w_{\neg dn}}, \boldsymbol{s}, \boldsymbol{t}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\beta^{(\tau)}}, \pi) \propto \frac{n_{d,k} + \alpha_k}{\sum_{k=1}^{K} n_{d,k} + \sum_{k=1}^{K} \alpha_k} \cdot \frac{n_{k,v} + \beta_v}{\sum_{v=1}^{V} n_{k,v} + \sum_{v=1}^{V} \beta_v} \cdot \pi \tag{2}$$

$$p(u_{dn} = 0|\boldsymbol{z_{\neg dn}}, \boldsymbol{u_{\neg dn}}, \boldsymbol{x}, \boldsymbol{y}, w_{dn} = v, \boldsymbol{w_{\neg dn}}, \boldsymbol{s}, \boldsymbol{t}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\beta^{(\tau)}}, \pi) \propto \frac{n_{B,v} + \beta_v}{\sum_{v=1}^{V} n_{B,v} + \sum_{v=1}^{V} \beta_v} \cdot (1 - \pi) \tag{3}$$

Secondly, we sample for x_{dn} and y_{dn} sequentially as the Equation (4). The exponential part $\exp\{\mathbb{I}(y_{d,n} = k)\}$ encourages that x_{dn} is sampled with the same topic as y_{dn}. That's where the Markov Random Field plays the role on the taxonomy structure.

$$p(x_{dn} = k|\boldsymbol{x_{\neg dn}}, \boldsymbol{y}, s_{dn} = v^{(\tau)}, \boldsymbol{s_{\neg dn}}, \boldsymbol{t}, \boldsymbol{z}, \boldsymbol{u}, \boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\beta^{(\tau)}}, \pi) \propto (n_{d,k} + \alpha_k) \frac{n^{(\tau)}_{k,v^{(\tau)}} + \beta^{(\tau)}_{v^{(\tau)}}}{\sum_{v^{(\tau)}=1}^{V^{(\tau)}} n^{(\tau)}_{k,v^{(\tau)}} + \sum_{v^{(\tau)}=1}^{V^{(\tau)}} \beta^{(\tau)}_{v^{(\tau)}}} \exp\{\mathbb{I}(y_{d,n} = k)\} \tag{4}$$

The sampling equation for y_{dn} is symmetric with x_{dn}'s as they are symmetric in the Markov Random Field.

Table 1: Top 5 topics learned by TaxoPhrase. The line in the *italic font* indicates the categories.

Topic 1	*Mathematics_awards*, *Mathematicians_by_award*, *Mathematicians_by_nationality*, *Mathematicians_by_field*
(Entities)	`John Cedric Griffiths Teaching Award`, `Santosh Vempala`, `Aisenstadt Prize`, `Subhash Suri`, `David P. Dobkin`
(Phrases)	university of california, american mathematical society, professor of mathematics, princeton university, computer science, harvard university, american mathematician, stanford university, massachusetts institute of technology, columbia university
Topic 2	*Geometry_stubs*, *Differential_geometry_stubs*, *Elementary_geometry_stubs*, *Polyhedron_stubs*
(Entities)	`Enneadecahedron`, `Icosahedral pyramid`, `Expanded icosidodecahedron`, `Pentadecahedron`, `Cubic cupola`
(Phrases)	three dimensional, platonic solids, johnson solids, uniform polyhedron compound, symmetry group, regular dodecahedron, triangular faces, vertex figure, nonconvex uniform polyhedron, four dimensional
Topic 3	*Topology_stubs*, *Knot_theory_stubs*, *Theorems_in_topology*, *Theorems_in_algebraic_topology*
(Entities)	`Knot operation`, `Chromatic homotopy theory`, `Infinite loop space machine`, `Simple space`, `Base change map`
(Phrases)	topological space, algebraic topology, category theory, topological spaces, fundamental group, simply connected, homotopy theory, 3 manifold, 3 manifolds, knot theory
Topic 4	*Cryptography_stubs*, *Cryptography*, *Combinatorics_stubs*, *Number_stubs*
(Entities)	`PC1 cipher`, `PKCS 8`, `KR advantage`, `Ccrypt`, `BEAR and LION ciphers`
(Phrases)	dual ec drbg, block cipher, sha 1, public key, hash function, stream cipher, escape wheel, balance wheel, secret key, private key
Topic 5	*Algebra_stubs*, *Abstract_algebra_stubs*, *Linear_algebra_stubs*, *Theorems_in_algebra*
(Entities)	`C-closed subgroup`, `Torsion abelian group`, `Fixed-point subgroup`, `Change of rings`, `Acceptable ring`
(Phrases)	algebraic geometry, group theory, abstract algebra, finite group, finitely generated, abelian group, finite groups, galois group, commutative ring, normal subgroup

4 EXPERIMENT RESULTS

In this section, we demonstrate the effectiveness of our proposed model TaxoPhrase, by evaluating the quality of the learned topics.

Dataset. We extract the taxonomy graph, the category-page graph, and pages' content from the latest dump of the English Wikipedia[2][3]. We choose *Mathematics*[4], *Chemistry*[5], and *Argentina*[6] to construct the datasets. The resulted datasets are described in the Table 2.

Baselines and Settings. We compare our learned topics on phrases with LDA, and compare the learned topics on categories with SSN-LDA[17]. SSN-LDA utilizes the co-occurrence relation of users in the network to discover the communities. We apply it on the category-entity graph to learn the topics on categories. We set $\beta = 0.01$ for LDA and SSN-LDA, $\beta^{(\tau)} = \beta = 0.01$ for TaxoPhrase, set $\alpha = 0.1$ and do the hyper-parameter optimization every 50 sampling iterations for all methods as [15]. All the algorithms are implemented in Mallet[7] with 1000 iterations. And the topic number is set to 100 for all algorithms and datasets.

Evaluating Metric. We choose the point-wise mutual information (PMI) as the measure of the topic coherence. For each topic, the PMI are computed among all pairs of top-30 topical phrases/categories. Specifically, $PMI{-}Score(z) = \frac{1}{435}\sum_{i<j} PMI(w_{z,i}, w_{z,j}), i, j \in \{1...30\}$, where $PMI(w_{z,i}, w_{z,j})$ are computed on the reference corpus. To make the evaluation result robust, we use the whole English Wikipedia as the reference corpus for computing PMI. The final PMI is the average score over all the topics.

Effectiveness. The results are shown in Table 2. Considering the quality of the topics on phrases and categories, our proposed method TaxoPhrase both achieve the optimum scores. Also shown in the Table 1, it's easy to confirm that the joint learning on categories and phrases provide more interpretable topics. Overall, TaxoPhrase provides an effective tool to explore the knowledge base.

Table 2: The statistics of the datasets, and the evaluation result on the learned topics on phrases/categories.

		Maths	Chemistry	Argentina
#Entities		27,947	60,375	8,617
#Category Types		1,391	3,038	1,479
#Phrase Types		116,013	248,769	21,183
on phrases	LDA	4.55	4.30	3.52
	TaxoPhrase	4.67	4.55	3.81
on categories	SSN-LDA	4.01	3.97	3.06
	TaxoPhrase	4.51	4.48	3.73

5 CONCLUSION

To provide an overview information for Knowledge Base, we joint model the taxonomy structure and phrases in the entity's content. Specifically, we propose the novel model TaxoPhrase. TaxoPhrase encourages that: the category nodes in the same edge tend to share the same topic with each other; the category nodes in the same category-information tend to have very few but coherent topics; and the category nodes and the phrases are more likely to have semantically coherent topics.

The experiments on three datasets, which are *Mathematics*, *Chemistry* and *Argentina* extracted from English Wikipedia, verify the effectiveness of TaxoPhrase on exploring the Knowledge Base.

6 ACKNOWLEDGEMENTS

This research is supported by the Natural Science Foundation of China (Grant No.61572043) and National Key Research and Development Program (Project Number: 2016YFB1000704).

[2]https://dumps.wikimedia.org/enwiki/latest/enwiki-latest-categorylinks.sql.gz
[3]https://dumps.wikimedia.org/enwiki/latest/enwiki-latest-pages-articles.xml.bz2
[4]https://en.wikipedia.org/wiki/Category:Mathematics
[5]https://en.wikipedia.org/wiki/Category:Chemistry
[6]https://en.wikipedia.org/wiki/Category:Argentina
[7]http://mallet.cs.umass.edu/

REFERENCES

[1] Albert-László Barabási and Eric Bonabeau. 2003. Scale-free networks. *Scientific American* 288 (2003), 50–59.
[2] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *JMLR* (2003).
[3] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *SIGMOD*.
[4] Jonathan Chang, Jordan L Boyd-Graber, Sean Gerrish, Chong Wang, and David M Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Nips*, Vol. 31. 1–9.
[5] Ludovic Denoyer and Patrick Gallinari. 2006. The wikipedia xml corpus. In *International Workshop of the Initiative for the Evaluation of XML Retrieval*. Springer, 12–19.
[6] Ahmed El-Kishky, Yanglei Song, Chi Wang, Clare R Voss, and Jiawei Han. 2014. Scalable topical phrase mining from text corpora. *Proceedings of the VLDB Endowment* 8, 3 (2014), 305–316.
[7] Stefano Faralli, Giovanni Stilo, and Paola Velardi. 2015. Large Scale Homophily Analysis in Twitter Using a Twixonomy.. In *IJCAI*. 2334–2340.
[8] Lauren A Hannah and Hanna M Wallach. Summarizing topics: From word lists to phrases. In *NIPS 2014 Workshop on Modern Machine Learning and Natural Language Processing*. 1–5.
[9] Yulan He. 2016. Extracting Topical Phrases from Clinical Documents. In *AAAI*. 2957–2963.
[10] Johannes Hoffart, Fabian M Suchanek, Klaus Berberich, and Gerhard Weikum. 2013. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence* 194 (2013), 28–61.
[11] Zhiting Hu, Gang Luo, Mrinmaya Sachan, Eric Xing, and Zaiqing Nie. 2016. Grounding topic models with knowledge bases. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*.
[12] Weijing Huang, Tengjiao Wang, Wei Chen, and Yazhou Wang. 2017. Category-Level Transfer Learning from Knowledge Base to Microblog Stream for Accurate Event Detection. In *International Conference on Database Systems for Advanced Applications*. Springer, 50–67.
[13] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, and others. 2015. DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web* 6, 2 (2015), 167–195.
[14] Shawn Martin, W Michael Brown, and Brian N Wylie. 2007. *Dr. L: Distributed Recursive (Graph) Layout*. Technical Report. Sandia National Laboratories.
[15] Hanna M Wallach, David M Mimno, and Andrew McCallum. 2009. Rethinking LDA: Why priors matter. In *Advances in neural information processing systems*. 1973–1981.
[16] Torsten Zesch, Iryna Gurevych, and Max Mühlhäuser. 2007. Analyzing and accessing Wikipedia as a lexical semantic resource. *Data Structures for Linguistic Resources and Applications* (2007), 197–205.
[17] Haizheng Zhang, Baojun Qiu, C Lee Giles, Henry C Foley, and John Yen. 2007. An LDA-based community structure discovery approach for large-scale social networks. In *Intelligence and Security Informatics, 2007 IEEE*. IEEE, 200–207.