Learning Topic Hierarchies for Wikipedia Categories

Linmei Hu†, Xuzhong Wang§, Mengdi Zhang†, Juanzi Li†, Xiaoli Li‡,
Chao Shao†, Jie Tang†, Yongbin Liu†
† Dept. of Computer Sci. and Tech. Tsinghua University, China
§ State Key Laboratory of Math. Eng. and Advanced Computing, China
‡ Institute for Infocomm Research(I2R), A*STAR, Singapore
{hulinmei1991,koodoneko,mdzhangmd,lijuanzi2008}@gmail.com
xlli@i2r.a-star.edu.sg, birdlinux@gmail.com
jietang@tsinghua.edu.cn, yongbinliu03@gmail.com

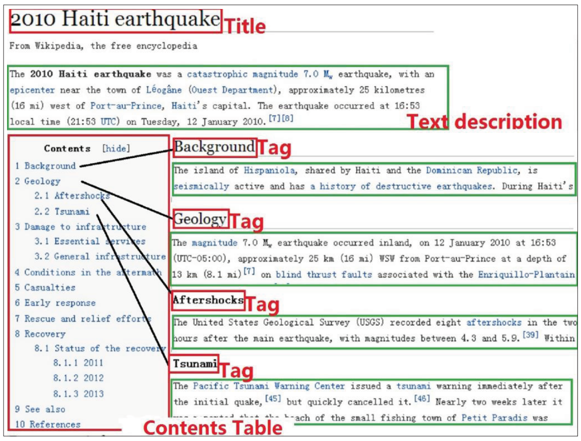
Abstract

Existing studies have utilized Wikipedia for various knowledge acquisition tasks. However, no attempts have been made to explore multi-level topic knowledge contained in Wikipedia articles' *Contents* tables. The articles with similar subjects are grouped together into Wikipedia *categories*. In this work, we propose novel methods to automatically construct *comprehensive* topic hierarchies for given categories based on the structured *Contents* tables as well as corresponding unstructured *text* descriptions. Such a hierarchy is important for information browsing, document organization and topic prediction. Experimental results show our proposed approach, incorporating both the structural and textual information, achieves high quality category topic hierarchies.

1 Introduction

As a free-access online encyclopedia, written collaboratively by people all over the world, Wikipedia (abbr. to Wiki) offers a surplus of rich information. Millions of articles cover various concepts and instances [1]. Wiki has been widely used for various knowledge discovery tasks. Some good examples include knowledge mining from Wiki infoboxes (Lin et al., 2011; Wang et al., 2013), and *taxonomy deriving* from Wiki category system (Zesch and Gurevych, 2007).

We observe that, in addition to Wiki's infoboxes and category system, Wiki articles' Contents tables (CT for short) also provide valuable structured topic knowledge with different levels of granularity. For example, in the article "*2010 Haiti Earthquake*", shown in Fig.1, the left Contents zone is a CT formed in a topic hierarchy format. If we view "*2010 Haiti earthquake*" as the root topic, the first-level "*Geology*" and "*Damage to infrastructure*" tags can be viewed as its subtopics, and the second-level "*Tsunami*" and "*Aftershocks*" tags underneath "*Geology*" are the subtopics of "*Geology*". Clicking any of the tags in Contents, we can jump to the corresponding text description. Wiki articles contain a wealth of this kind of structured and unstructured information. However, to our best knowledge, little work has been done to leverage the knowledge in CT.



2010 Haiti earthquake

From Wikipedia, the free encyclopedia

The **2010 Haiti earthquake** was a catastrophic magnitude 7.0 M_w earthquake, with an epicenter near the town of Léogâne (Ouest Department), approximately 25 kilometres (16 mi) west of Port-au-Prince, Haiti's capital. The earthquake occurred at 16:53 local time (21:53 UTC) on Tuesday, 12 January 2010.[7][8]

Contents [hide]

1 Background
2 Geology
2.1 Aftershocks
2.2 Tsunami
3 Damage to infrastructure
3.1 Essential services
3.2 General infrastructure
4 Conditions in the aftermath
5 Casualties
6 Early response
7 Rescue and relief efforts
8 Recovery
8.1 Status of the recovery
8.1.1 2011
8.1.2 2012
8.1.3 2013
9 See also
10 References

Background

The island of Hispaniola, shared by Haiti and the Dominican Republic, is seismically active and has a history of destructive earthquakes. During Haiti's

Geology

The magnitude 7.0 M_w earthquake occurred inland, on 12 January 2010 at 16:53 (UTC-05:00), approximately 25 km (16 mi) WSW from Port-au-Prince at a depth of 13 km (8.1 mi)[7] on blind thrust faults associated with the Enriquillo-Plantain

Aftershocks

The United States Geological Survey (USGS) recorded eight aftershocks in the two hours after the main earthquake, with magnitudes between 4.3 and 5.9.[39] Within

Tsunami

The Pacific Tsunami Warning Center issued a tsunami warning immediately after the initial quake,[45] but quickly cancelled it.[46] Nearly two weeks later it was reported that the beach of the small fishing town of Petit Paradis was

Figure 1: The Wiki article "2010 Haiti Earthquake" with structured *Contents* table and corresponding unstructured *text* descriptions.

In Wiki, similar articles (each with their own CT) belonging to the same subject are grouped together into a *Wiki category*. We aim to integrate multiple topic hierarchies represented by CT (from the articles under the same *Wiki category*) into a comprehensive *category topic hierarchy* (CTH). While there also exist manually built CTH represented by CT in corresponding Wiki articles, they are still too high-level and incomplete. Take the "*Earthquake*" category as an example, its corresponding Wiki article [2] only contains

[1] http://en.wikipedia.org/wiki/Encyclopedia

[2] http://en.wikipedia.org/wiki/Earthquake

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), pages 346–351,
Beijing, China, July 26-31, 2015. ©2015 Association for Computational Linguistics

some major and common topics. It does not include the subtopic "*nuclear power plant*", which is an important subtopic of the "*2011 Japan earthquake*". A comprehensive CTH is believed to be more useful for information browsing, document organization and topic extraction in new text corpus (Veeramachaneni et al., 2005). Thus, we propose to investigate the Wiki articles of the same category to *automatically* build a comprehensive CTH to enhance the manually built CTH.

Clearly, it is very challenging to learn a CTH from multiple topic hierarchies in different articles due to the following 3 reasons: 1) A topic can be denoted by a variety of tags in different articles (e.g., "*foreign aids*" and "*aids from other countries*"); 2) Structural/hierarchical information can be inconsistent (or even opposite) across different articles (e.g., "response *subtopicOf* aftermath" and "aftermath *subtopicOf* response" in different earthquake event articles); 3) Intuitively, text descriptions of the topics in Wiki articles are supposed to be able to help determine *subtopic* relations between topics. However, how can we model the textual correlation?

In this study, we propose a novel approach to build a high-quality CTH for any given Wiki category. We use a Bayesian network to model a CTH, and map the CTH learning problem as a structure learning problem. We leverage both structural and textual information of topics in the articles to induce the optimal tree structure. Experiments on 3 category data demonstrate the effectiveness of our approach for CTH learning.

2 Preliminaries

Our problem is to learn a CTH for a Wiki category from multiple topic hierarchies represented by CT in the Wiki articles of the category. For example, consider the category "earthquake". There are a lot of Wikipedia articles about earthquake events which are manually created by human experts. In these articles, the CTs imply hierarchical topic knowledge in the events. However, due to crowdsourcing nature, these knowledge is heterogeneous across different articles. We want to integrate these knowledge represented by CTs in different earthquake event articles to form a comprehensive understanding of the category "earthquake" (CTH).

Specifically, our input consists of a set of Wiki articles $A_c = \{a\}$, belonging to a Wiki category c. As shown in Fig.1, each article $a \in A_c$ contains a CT (topic tree hierarchy) $H_a = \{T_a, R_a\}$, where T_a is a set of topics, each denoted by a tag g and associated with a text description d_g, and $R_a = \{(g_i, g_j)\}, g_i, g_j \in T_a$ is a set of subtopic relations (g_j is a subtopic of g_i). The output is an integrated comprehensive CTH $H_c = \{T_c, R_c\}$ where $T_c = \{t\}$ is a set of topics, each denoted by *a set of tags* $t = \{g\}$ and associated by a text description d_t *aggregated by* $\{d_g\}_{g \in t}$, and $R_c = \{(t_i, t_j)\}, t_i, t_j \in T_c$ is a set of subtopic relations (t_j is a subtopic of t_i).

We map the problem of learning the output H_c from the input $\{H_a\}, a \in A_c$, as a structure learning problem. We first find clusters of similar tags T_c (each cluster represents a topic) and then derive hierarchical relations R_c among these clusters.

Particularly, given a category c, we first collect relevant Wiki articles $A_c = \{a\}$. This can be done automatically since each Wiki article has links to its categories. We can also manually find the Wikipage which summarizes the links of A_c (e.g., `http://en.wikipedia.org/wiki/Lists_of_earthquakes`) and then collect A_c according to the links.

Then we can get a global tag set $G = \{g\}$ containing all the tags including titles in the articles A_c. We cluster the same or similar tags from different articles using single-pass incremental clustering (Hammouda and Kamel, 2003) to construct the topic set T_c, with cosine similarity computed based on the names of tags g and their text descriptions d_g. Note that titles of all the articles belonging to the same cluster corresponds to a root topic.

Next, the issue is how to induce a CTH $H_c = \{T_c, R_c\}$ from a set of topics T_c.

3 Topic Hierarchy Construction

We first present a basic method to learn H_c and then describe a principled probabilistic model incorporating both structural and textual information for CTH learning.

3.1 Basic Method

After replacing the tags in a CT (see Fig.1) with the topics they belong to, we can then get a topic hierarchy $H_a = \{T_a, R_a\}$ for each article a. For each subtopic relation $(t_i, t_j) \in R_a$, we can calculate a count/weight $n(t_i, t_j)$, representing the number of articles in A_c containing the relation. We then construct a directed complete graph with a weight $w(t_i, t_j) = n(t_i, t_j)$ on each edge (t_i, t_j). Finally, we apply the widely used Chu-Liu/Edmonds algorithm (Chu and Liu, 1965; Edmonds, 1967) to find an optimal tree with the largest sum of weights from our constructed graph, meaning that the overall subtopic relations in the tree is best supported by all the CT/articles. The Chu-Liu/Edmonds algorithm works as follows. First, it selects, for each node, the maximum-weight incoming edge. Next, it recursively breaks cycles with the following idea: nodes in a cycle are collapsed into a pseudo-node and the maximum-weight edge entering the pseudo-node is selected to replace the other incoming edge in the cycle. During backtracking, pseudo-nodes are expanded into an acyclic directed graph, i.e., our final category topic hierarchy H_c.

However, the basic method has a problem. Consider if *n*("*earthquake*", "*damages to hospitals*")=10 and *n*("*earthquake*") =100, while *n*("*damages*", "*damages to hospitals*")=5 and *n*("*damages*")=5. We would prefer "*damages*" to be the parent topic of "*damages to hospitals*" with a higher *confidence* level (5/5=1 vs 10/100=0.1). However, the above basic method will choose "*earthquake*" which maximizes the weight sum. An intuitive solution is to normalize the weights. In Subsection 3.2, we present our proposed principled probabilistic model which can derive normalized structure based weights. In addition, it can be easily used to incorporate and combine textual information of topics into CTH learning.

3.2 Probabilistic Model for CTH Learning

We first describe the principled probabilistic model for a CTH. Then we present how to encode structural dependency and textual correlation between topics. Last, we present our final approach combining both structural dependency and textual correlation for CTH construction.

3.2.1 Modeling a Category Topic Hierarchy

In a topic hierarchy, each node represents a topic. We consider each node as a variable and the topic hierarchy as a Bayesian network. Then the joint probability distribution of nodes N given a particular tree H is

$$P(N|H) = P(root) \prod_{n \in N \setminus root} P(n|par_H(n)) \ ,$$

where $P(n|par_H(n))$ is the conditional probability of node n given its parent node $par_H(n)$ in H. Given the nodes, this is actually the likelihood of H. Maximizing the likelihood with respect to the tree structure gives the optimal tree:

$$\begin{aligned} H^* &= \text{argmax}_H P(N|H) \\ &= \text{argmax}_H P(root) \prod_{n \in N \setminus root} P(n|par_H(n)) \\ &= \text{argmax}_H \sum_{n \in N} logP(n|par_H(n)) \end{aligned} \tag{1}$$

Encoding Structural Dependency. Considering t_j is a subtopic of t_i, we define the structural conditional probability:

$$P_{struc}(t_j|t_i) = \frac{n(t_i, t_j) + \alpha}{n(t_i) + \alpha \cdot |T_c - 1|} \ , \tag{2}$$

where $n(t_i, t_j)$ is the count of articles containing relation (t_i, t_j) and $n(t_i)$ is the count of articles containing topic t_i. The parameter $\alpha = 1.0$ is the Laplace smoothing factor, and $|T_c - 1|$ is the total number of possible relations taking t_i as their parent topic.

Encoding Textual Correlation. Considering a topic text description d_t as a bag of words, we use the normalized word frequencies $\phi_t = \{\phi_{t,w}\}_{w \in V} s.t. \sum_{w \in V} \phi_{t,w} = 1$ to represent a topic t. To capture the subtopic relationship (t_i, t_j), we prefer a model where the expectation of the distribution for the child is exactly same with the distribution for its parent, i.e., $E(\phi_{t_j}) = \phi_{t_i}$. This naturally leads to the hierarchical Dirichlet model (Wang et al., 2014; Veeramachaneni et al., 2005), formally, $\phi_{t_j}|\phi_{t_i} \sim Dir(\beta\phi_{t_i})$ in which β [3] is the concentration parameter which determines how concentrated the probability density is likely to be. Thus we have:

$$P_{text}(t_j|t_i) = \frac{1}{Z} \prod_{w \in V} \phi_{t_j,v}^{\beta\phi_{t_i,w}-1} \ , \tag{3}$$

where $Z = \frac{\prod_{w \in V} \Gamma(\alpha\phi_{t_i,w})}{\Gamma(\sum_{w \in V} \alpha\phi_{t_i,w})}$ is a normalization factor and $\Gamma(\cdot)$ is the standard Gamma distribution. We note that for the root node we have the uniform prior instead of the prior coming from the parent.

3.2.2 Combining Structural and Textual Information

Substituting Eq.2 into Eq.1, we can solve the optimal tree structure by applying Chu-

[3] Experimental results are insensitive to β, we set β=5

Liu/Edmonds algorithm to the directed complete graph with structure based weights w_{struc}=$log(P_{struc}(t_j|t_i)$ $=$ $log\frac{n(t_i,t_j)+\alpha}{n(t_i)+\alpha\cdot|T_c-1|}$ on the edges (t_i, t_j). While this solves the problem of the basic method, it only considers structural dependency and does not consider textual correlation which is supposed to be useful.

Therefore, we calculate text based weights w_{text}=$log(P_{text}(t_j|t_i)$ $=$ $\sum_{w\in V} log\phi_{t_j,v}^{\alpha\phi_{t_i,w}-1}$ $-$ $logZ$ similarly. Then we combine structural information and textual information by defining the weights $w(t_i, t_j)$ of the edges (t_i, t_j) as a simple weighted average of $w_{struc}(t_i, t_j)$ and $w_{text}(t_i, t_j)$. Specifically, we define:

$$w(t_i, t_j) = \lambda w_{text}(t_i, t_j) + (1 - \lambda) w_{struc}(t_i, t_j) \ ,$$

where λ controls the impacts of text correlation and structure dependency in optimal structure learning. Note that w_{text} and w_{struc} should be scaled [4] first before applying Chu-Liu/Edmonds algorithm to find an optimal topic hierarchy.

4 Experiments

We evaluate the CTH automatically generated by our proposed methods via comparing it with a manually constructed ground-truth CTH.

4.1 Data and Evaluation Metric

Data. We evaluate our methods on 3 categories, i.e., English "earthquake" and "election" categories containing 293 and 60 articles, and Chinese "earthquake" category containing 48 articles [5]. After removing noisy tags such as "references" and "see also", they contain 463, 79 and 426 unique tags respectively. After tag clustering [6], we can get 176, 57 and 112 topics for each category.

Evaluation Metric. We employ the *precision* measure to evaluate the performance of our methods. Let **R** and $\mathbf{R}_s$ be subtopic relation sets of our generated result and ground-truth result respectively, then *precison*=$|\mathbf{R} \cap \mathbf{R}_s|/|\mathbf{R}|$. Due to the number of relations $|\mathbf{R}|$=$|\mathbf{R}_s|$ $=$ $|T_c - 1|$, we have *precison=recall=F1*=$|\mathbf{R} \cap \mathbf{R}_s|/|\mathbf{R}|$.

We compare three methods, including our basic method (Basic) which uses only non-normalized structural information, our proposed probabilistic method considering only structural information ($\lambda = 0$) (Pro+S), and considering both structural and textual information ($0 < \lambda < 1$) (Pro+ST).

4.2 Results and Analysis

Quantitative Analysis. From Table 1, we observe that our approach Pro+ST (with best λ values as shown in Fig.2) significantly outperforms Basic and Pro+S which only utilize the structural information (+24.3% and +5.1% on average, p <0.025 with *t-test*). Pro+S which normalizes structural information also achieves significant higher precision than Basic (+19.2% on average, p <0.025).

	Earth.(En)	Elect.(En)	Earth.(Ch)
Basic	0.5965	0.7719	0.7143
Pro+S	0.8971	0.8596	0.9017
Pro+ST	0.9543	0.9298	0.9286

Table 1: Precision of different methods on 3 categories

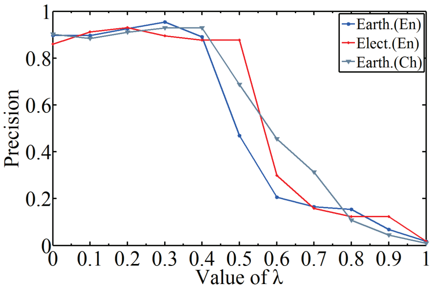


Figure 2: The precision of CTH with different λ values

To examine the influence of λ, we show the performance of our approach Pro+ST with different λ values on 3 categories in Fig.2. All the curves grow up first and then decrease dramatically as we emphasize more on textual information. They can always get consistent better results when 0.2$\leq$ λ $\leq$0.3. When λ approaches 1, the precision declines fast to near 0. The reason is that the topics with short (or null) text descriptions are likely to be a parent node of all other nodes and influence the results dramatically, but if we rely mostly on structural information and use the textual information as auxiliary for correcting minor errors in some ambiguous cases, we can improve the precision of the resultant topic hierarchy.

Qualitative Analysis. Due to space limitation, we only show the topic hierarchy for "Election" with smaller topic size in Fig.3. As we can see,

[4] We use min-max normalization $x^* = \frac{x-min}{max-min}$

[5] We filter articles with little information in Contents.

[6] We use an incremental clustering algorithm

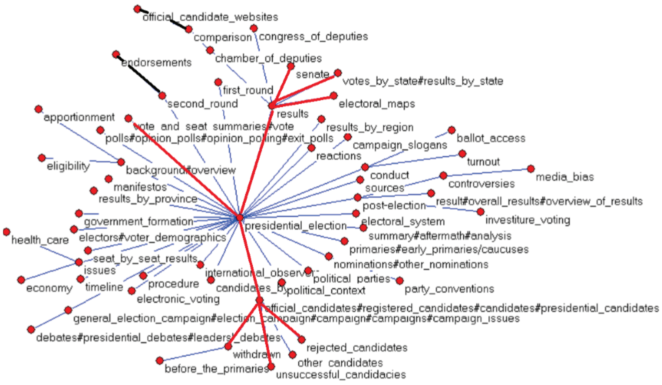


Figure 3: The category topic hierarchy for presidential elections. Topics are labeled by tags separated by "#".

the root topic "*presidential elections*" includes subtopics "*results*", "*vote*", "*official candidates*", etc. Furthermore, '*official candidates*" contains subtopics "*debates*, "*rejected candidates*", "*unsuccessful candidates*", etc. The above mentioned examples are shown with red edges. However, there are also a few (7%) mistaken relations (black edges) such as "*comparison*" (should be "*official candidates*" instead) → "*official candidate websites*". Overall, the above hierarchy aligns well with human knowledge.

5 Related Work

To our best knowledge, our overall problem setting is novel and there is no previous work using Wiki articles' *contents* tables to learn topic hierarchies for categories. Existing work mainly focused on learning topic hierarchies from *texts* only and used traditional hierarchical clustering methods (Chuang and Chien, 2004) or topic models such as HLDA (Griffiths and Tenenbaum, 2004), HPAM (Mimno et al., 2007), hHDP (Zavitsanos et al., 2011), and HETM (Hu et al., 2015). Differently, we focus on structured contents tables with corresponding text descriptions.

Our work is also different from ontology (taxonomy) construction (Li et al., 2007; Tang et al., 2009; Zhu et al., 2013; Navigli et al., 2011; Wu et al., 2012) as their focus is concept hierarchies (e.g. *isA* relation) rather than thematic topic hierarchies. For example, given the "*animals*" category, they may derive "*cats*" and "*dogs*", etc. as subcategories, while our work aims to derive thematic topics "*animal protection*" and "*animal extinction*", etc. as subtopics. Our work enables a fresher to quickly familiarize himself/herself with any new category, and is very useful for information browsing, organization and topic extraction.

6 Conclusion

In this paper, we propose an innovative problem, i.e., to construct high quality comprehensive topic hierarchies for different Wiki categories using their associated Wiki articles. Our novel approach is able to model a topic hierarchy and to incorporate both structural dependencies and text correlations into the optimal tree learning. Experimental results demonstrate the effectiveness of our proposed approach. In future work, we will investigate how to update the category topic hierarchy incrementally with the creation of new related articles.

Acknowledgments

The work is supported by 973 Program (No. 2014CB340504), NSFC-ANR (No. 61261130588), Tsinghua University Initiative Scientific Research Program (No. 20131089256), Science and Technology Support Program (No. 2014BAK04B00), China Postdoctoral Science Foundation (No. 2014M550733), National Natural Science Foundation of China (No. 61309007) and THU-NUS NExT Co-Lab.

References

Yoeng-Jin Chu and Tseng-Hong Liu. 1965. On shortest arborescence of a directed graph. *Scientia Sinica*, 14(10):1396.

Shui-Lung Chuang and Lee-Feng Chien. 2004. A practical web-based approach to generating topic hierarchy for text segments. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 127–136. ACM.

Jack Edmonds. 1967. Optimum branchings. *Journal of Research of the National Bureau of Standards B*, 71(4):233–240.

DMBTL Griffiths and MIJJB Tenenbaum. 2004. Hierarchical topic models and the nested chinese restaurant process. *Advances in NIPS*, 16:17.

Khaled M Hammouda and Mohamed S Kamel. 2003. Incremental document clustering using cluster similarity histograms. In *Web Intelligence, 2003. WI 2003. Proceedings. IEEE/WIC International Conference on*, pages 597–601. IEEE.

Linmei Hu, Juanzi Li, Jing Zhang, and Chao Shao. 2015. o-hetm: An online hierarchical entity topic model for news streams. In *Advances in Knowledge Discovery and Data Mining - 19th Pacific-Asia Conference, PAKDD 2015, Proceedings, Part I*, pages 696–707.

Rui Li, Shenghua Bao, Yong Yu, Ben Fei, and Zhong Su. 2007. Towards effective browsing of large scale social annotations. In *Proceedings of the 16th international conference on World Wide Web*, pages 943–952. ACM.

Wen-Pin Lin, Matthew Snover, and Heng Ji. 2011. Unsupervised language-independent name translation mining from wikipedia infoboxes. In *Proceedings of the First workshop on Unsupervised Learning in NLP*, pages 43–52. Association for Computational Linguistics.

David Mimno, Wei Li, and Andrew McCallum. 2007. Mixtures of hierarchical topics with pachinko allocation. In *Proceedings of the 24th ICML*, pages 633–640. ACM.

Roberto Navigli, Paola Velardi, and Stefano Faralli. 2011. A graph-based algorithm for inducing lexical taxonomies from scratch. In *IJCAI*, pages 1872–1877.

Jie Tang, Ho-fung Leung, Qiong Luo, Dewei Chen, and Jibin Gong. 2009. Towards ontology learning from folksonomies. In *IJCAI*, volume 9, pages 2089–2094.

Sriharsha Veeramachaneni, Diego Sona, and Paolo Avesani. 2005. Hierarchical dirichlet model for document classification. In *Proceedings of the 22nd ICML*, pages 928–935. ACM.

Zhigang Wang, Zhixing Li, Juanzi Li, Jie Tang, and Jeff Z Pan. 2013. Transfer learning based cross-lingual knowledge extraction for wikipedia. In *ACL (1)*, pages 641–650.

Jingjing Wang, Changsung Kang, Yi Chang, and Jiawei Han. 2014. A hierarchical dirichlet model for taxonomy expansion for search engines. In *Proceedings of the 23rd international conference on WWW*, pages 961–970. International World Wide Web Conferences Steering Committee.

Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Q Zhu. 2012. Probase: A probabilistic taxonomy for text understanding. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pages 481–492. ACM.

Elias Zavitsanos, Georgios Paliouras, and George A Vouros. 2011. Non-parametric estimation of topic hierarchies from texts with hierarchical dirichlet processes. *The Journal of Machine Learning Research*, 12:2749–2775.

Torsten Zesch and Iryna Gurevych. 2007. Analysis of the wikipedia category graph for nlp applications. In *Proceedings of the TextGraphs-2 Workshop (NAACL-HLT 2007)*, pages 1–8.

Xingwei Zhu, Zhao-Yan Ming, Xiaoyan Zhu, and Tat-Seng Chua. 2013. Topic hierarchy construction for the organization of multi-source user generated contents. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 233–242. ACM.