2015 IEEE 15th International Conference on Data Mining Workshops

Constructing Topic Hierarchies from Social Media Data

Yuhao Zhang Wenji Mao Daniel Zeng
State Key Laboratory of Management and Control for Complex Systems
Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
{zhangyuhao2012, wenji.mao, dajun.zeng}@ia.ac.cn

***Abstract*—Constructing topic hierarchies from the data automatically can help us better understand the contents and structure of information and benefit many applications in security informatics. The existing topic hierarchy construction methods either need to specify the structure manually, or are not robust enough for sparse and noisy social media data such as microblog. In this paper, we propose an approach to automatically construct topic hierarchies from microblog data in a bottom up manner. We detect topics first and then build the topic structure based on a tree combination method. We conduct a preliminary empirical study based on the Weibo data. The experimental results show that the topic hierarchies generated by our method provide meaningful results.**

Keywords—topic hierarchies; topic detection; social media

I. INTRODUCTION

The analysis of social media data is increasingly important in recent years, especially for various security-related applications. Constructing hierarchical topic structure from online social media can help us better understand the contents and structure of information and facilitate decision making, emergency response and management, and many other applications in security informatics.

Many works have been devoted to constructing topic hierarchies. Ontology construction methods can be used for topic hierarchy construction. These methods use a single term to represent a topic which restrict their ability to handle complex topics. Hierarchical topic models like nCRP [1], nCRF [2] can detect and organize the topics into the tree structure automatically, but the results of these methods are not very interpretable sometimes for sparse and noisy data. PAM [3] and hPAM [4] use the directed acyclic graph to organize the topics. Recent work CATHY [5] use a top-down, recursive way to construct the topic hierarchies. But these methods need to specify the topic structure manually.

In this paper, we propose an approach that can automatically construct topic hierarchies from microblog data in a bottom up manner. To avoid the limitations of the related work, our approach extracts topics first and then builds topic hierarchies using a tree combination method. We also conduct an empirical study to compare our approach with the related works and show its advantage in topic hierarchy construction.

II. RELATED WORK

Topic analysis is widely used in text processing and is an essential technique for security-related event detection and monitoring. Traditional topic detection methods like pLSI [6], LDA [7] and HDP [8] model a document as a mixture of topics and a topic as a distribution over words. However, these methods can only provide a pool of flat topics without inferring the relationships among them.

Ontology construction methods can be used to organize topics into a hierarchical structure. These methods [9], [10] find the topic terms from text and identify the topic relations using statistical method and additional information sources. However, using merely one term as a topic restrict the ability of these approaches to represent complex topics.

Hierarchical topic models have also been proposed for this task. nCRP [1] organizes the topics into a tree structure using a non-parametric Bayesian approach. Each document is generated by topics along a path of the topic tree. nCRF [2] extend nCRP by modeling document as a distribution over all the nodes of the hierarchy. But these methods sometimes could not provide good results on social media data like tweets as the data is often sparse and noisy.

Another kind of generative methods adopt a directed acyclic graph model to construct the topic hierarchies, including PAM [3] and hPAM [4]. In PAM, each internal node represents a distribution over its child nodes and each leaf node represents a distribution over the words. hPAM extends the PAM by allowing the internal nodes also represent distributions over words. A more recent work CATHY [5] adopts a top-down, recursive way to construct the topic hierarchies. It constructs the topic hierarchies by repeating a graph partition process. However, the methods mentioned above all need to specify the topic number at each level manually.

To address these challenges, we propose a bottom-up approach that can construct the topic hierarchies automatically from the data. We first detect topics with a widely used topic detection method nonnegative matrix factorization (NMF) [11] and get the fine grained topics including small topics as well as different aspects of the large topics. Then we use a tree combination method to construct the topic hierarchies. The method is similar to Bayesian rose tree (BRT) [12], a hierarchical clustering method that choose the tree structure based on the marginal data likelihood. But instead of handling data in a generative way as BRT, we construct the topic hierarchical structure based on topic similarities. Our approach chooses the proper combination mode for topic trees by comparing three topic similarity measures which reflect the closeness of the subtopics within trees and the topic similarity between two trees. The experimental results show that our method can provide a meaningful topic hierarchical structure.

III. PROBLEM FORMULATION

We propose a method that organize topics into a tree structure, in which each node represents a topic and the non-

978-1-4673-8493-3/15 $31.00 © 2015 IEEE
DOI 10.1109/ICDMW.2015.146

IEEE computer society

leaf topic covers the semantics of its children. Here we give formal definitions of some concepts involved in the problem.

DEFINITION 1 (TOPIC). *Given a lexicon **l** ranked in the lexicographic order, a topic **t** is represented as a vector* $\mathbf{v_t}$, *whose elements are weights of the corresponding words in **l**.*

Each topic has a topic weight which reflect the proportion of the topic in the corpora.

DEFINITION 2 (TOPIC WEIGHT). *The topic weight* $\mathbf{w_t}$ *is defined as the sum of the proportion of topic **t** of each document in the corpora.*

We use topics as basic units to constructing the topic tree, which reflect the topic hierarchies.

DEFINITION 3 (TOPIC TREE). *A topic tree is a tree in which each node represents a topic. For each non-leaf topic, its subtopics comprise the children. Let **Tree(r)** represents a topic tree. It either contains one node **r** or a root node **r** with sub-trees **Tree(r.1)**, **Tree(r.1)**, ⋯, **Tree(r.n***$_r$***)** connected to **r**, where* n_r *is the number of the children of **r**.*

We define inter-tree similarity to reflect the topic similarity between two topic trees. As the root topic can cover the semantics of the entire topic tree, we define this measure based on the root topics of the trees.

DEFINITION 4 (INTER-TREE SIMILARITY). *Given Tree(a) and Tree(b), their inter-tree similarity* $\mathbf{P_{a,b}}$ *is defined as the similarity of their root topics.*

We define the intra-tree similarity to reflect the closeness of the subtopics within a tree.

DEFINITION 5 (INTRA-TREE SIMILARITY). *The intra-tree similarity is the average subtopic similarities of a topic tree. For a topic tree **Tree(r)** with* n_r *sub-trees connected to its root, its intra-tree similarity* $\mathbf{I_r}$ *is defined as:*

$$I_r = \frac{n_r(n_r-1)}{2} \sum_{i,j\in[1,n_r],\, i<j} sim(Tree(r.i), Tree(r.j))$$

For the topic tree contains one node, we set its intra-tree similarity as infinity.

Our problem is constructing a topic tree that can reflect the topic hierarchies of the given corpora.

IV. PROPOSED METHOD

To solve the problem, we propose an approach that constructs the topic hierarchies in a bottom up manner. We first use traditional topic detection method to extract the topics from the document collection. Then we construct the topic hierarchies using a multi-branch hierarchical construction method based on tree combination.

A. Topic Detection

We use NMF to extract topics from the given documents. We turn the document collection into the term-document matrix V and use the *tf-idf* to represent the terms. Let k be the topic number. We need to decompose V into two nonnegative matrix, namely term-topic matrix $W_{m\times k}$ and topic-document matrix $H_{k\times n}$. We add L2 norm to the optimization objective to avoid over fitting. The object function is:

$$\min_{W,H} f(W,H) = \| WH - V \|_F^2 + \alpha \| W \|_F^2 + \beta \| H \|_F^2, \quad s.t. W, H \ge 0$$

The $\| \bullet \|_F$ is the Frobenius norm. α and β are regularization parameters. We use alternating nonnegative least square (ANLS) [13] algorithm to solve it. Then we can get the topics and topic weights based on W and H separately. Although we adopt nonnegative matrix factorization for this paper, other topic detection methods can also be used.

B. Topic Similarity Calculation

As our topic hierarchical construction method is based on the topic similarities, choosing a good similarity measure which could reflect the closeness of topics is important. Traditional measure cosine similarity do not perform very well when the topics are sparse. So we propose a new method for calculating the similarity between sparse topics based on the positive point mutual information (PPMI) [14], which can reflect the co-occurrence of two words. As the sparse topics often adopt a narrow range of terms, we design the topic similarity measure based on the weights of the top words of the topics and their PPMI values. Re-rank the terms of the topic vector in the descending order of the term weight. Let $e_{a,i}$ be the weight of the term $t_{a,i}$. The topic similarity measure using the top m terms is defined as follows:

$$sim(t_a, t_b) = \sum_{i,j\in[1,\,m]} e_{a,i} e_{b,j} PPMI(t_{a,i}, t_{b,j})$$

C. Topic Hierarchy Construction

We propose a topic hierarchy construction method based on tree combination. We take the detected topics as basic topic trees. By iteratively combining these trees, we get the entire topic hierarchies. At each round, we choose the two topic trees with maximal inter-tree similarity and combine them in a proper way which can best reflect the relationship between the topics within the two trees. Suppose Tree(a) and Tree(b) are the two trees prepared for combination as follows:

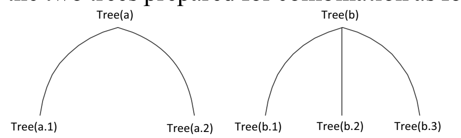


Fig.1 Tree(a) and Tree(b)

We choose the combination mode based on three measures: the intra-tree similarities I_a and I_b and the inter-tree similarity $P_{a,b}$, which represent the closeness of the subtopics of the two trees and the closeness of the two topic groups separately. We combine the topic trees by making the topics which are close enough as siblings and maintaining the structural integrity of the tree that has relative high intra-tree similarity.

So we get four combination modes of the topic trees as shown in Fig. 2. Let γ valued between 0-1 be the threshold for the comparison of I_a, I_b and $P_{a,b}$. We choose the combination mode based on the following rules and considerations:

1) If $P_{a,b} > \gamma \times I_a$ *and* $P_{a,b} > \gamma \times I_b$, *choose mode (a):* If $P_{a,b}$ is close enough to I_a and I_b, the subtopics of Tree(a) and Tree(b) are very similar. We put the subtopics of the two trees together and give them a common root as shown in Fig. 2(a).

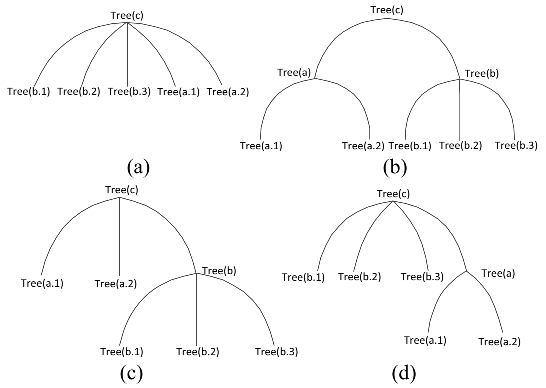

Tree(c)
Tree(b.1) Tree(b.2) Tree(b.3) Tree(a.1) Tree(a.2)
(a)
Tree(c)
Tree(a)
Tree(b)
Tree(a.1)
Tree(a.2) Tree(b.1) Tree(b.2) Tree(b.3)
(b)
Tree(c)
Tree(a.1)
Tree(a.2)
Tree(b)
Tree(b.1) Tree(b.2) Tree(b.3)
(c)
Tree(c)
Tree(b.1) Tree(b.2) Tree(b.3)
Tree(a)
Tree(a.1) Tree(a.2)
(d)

Fig.2 Combination modes of two topic trees

2) If $P_{a,b} \le \gamma \times I_a$ and $P_{a,b} \le \gamma \times I_b$, choose mode (b): If $P_{a,b}$ is much smaller than I_a and I_b, the topics of the two trees are relatively independent. So we construct the new tree by keeping Tree(a) and Tree(b) as subtrees as shown in Fig. 2(b).

3) If $P_{a,b} > \gamma \times I_a$ and $P_{a,b} \le \gamma \times I_b$, choose mode (c): If $P_{a,b}$ is close to I_a and much smaller than I_b, it means that Tree(b) has a higher intra-tree similarity than Tree(a) and its root topic is similar to the subtopics of Tree(a). We can maintain the structural integrity of Tree(b) and make Tree(b) and the subtopics of Tree(a) as siblings as shown in Fig. 2(c).

4) If $P_{a,b} \le \gamma \times I_a$ and $P_{a,b} > \gamma \times I_b$, choose mode (d): The situation $P_{a,b}$ is close to I_b and much smaller than I_a is similar to the combination mode (c) except for exchanging the places of Tree(a) and Tree(b) as shown in Fig. 2(d).

D. Topic Information Update

After the tree combination at each round, we need to produce the topic vector and topic weight for the new node. We define the non-leaf topic as the weighted sum of its children. And the weight of a non-leaf topic is the sum of the weights of its children. So the topic vector v_t and topic weight w_t for the new non-leaf node t can be computed as follows:

$$v_t = \frac{\sum_{i\in[1,n_t]} w_{t.i} v_{t.i}}{w_t}, \quad w_t = \sum_{i\in[1,n_t]} w_{t.i}$$

V. EXPERIMENT

A. Dataset and Preprocessing

We evaluate our method using data from Weibo, a microblog site in China. The dataset is about the 26th Asia-Pacific Economic Cooperation (APEC) summit, which took place in Beijing, China from Nov. 5, 2014 to Nov. 11, 2014. We collect tweets that related to this event in Weibo from Nov. 1, 2014 to Nov. 18, 2014 and remove the short tweets which have less than 5 words. We get nearly 8000 tweets in total. Then we remove the stop words and the words whose inverse document frequency are less than 5. We use the data after the preprocessing to construct the topic hierarchies.

B. Results and Discussions

We evaluate our method and compare it with two typical topic hierarchies construction approaches hPAM and nCRP. We first demonstrate the topic hierarchical results produced by these methods qualitatively. Then we provide a quantitative analysis based on the 'topic intrusion' [15], a human evaluation method for topic models.

For our method, we use the top 50 words of each topic for topic similarity calculation, which can cover the most important words in the topics. And we empirically set the coefficient γ to 0.7. We detect 50 topics from the tweets collection and construct the topic hierarchies based on these topics. The result shows that the topic hierarchies our method constructed make sense as the unpopular topics are usually represented by sub-trees with few topic nodes while the hot ones are represented by sub-trees with multiple branches. Fig. 3(a) shows a sub-tree generated by our method as an example.

The hPAM construct the topic hierarchical structure of three levels. The first level is the root topic. We set the model contains 10 topics at the second level and 50 topics at the third level. The topics hPAM produced are interpretable, but the parent-child relationship of topics is not very clear sometimes. Fig. 3(b) shows part of the constructed topic hierarchies.

The nCRP produces a relatively large-scale topic hierarchical structure. The method put many key words of the event on the root topic while many subtopics are just noise of the Weibo data. So many topic branches produced are not helpful for understanding the topic structure of the event.

In order to measure the topic hierarchical structure quantitatively, we conduct a topic intruder task, which can be used to evaluate the parent-child relationship of the topic hierarchies based on human judgments [15]. Evaluators are shown a parent topic and N candidate child topics for each question. N-1 candidates are true child topics generated by the models while the other one is chosen randomly from the rest part of the generated hierarchical structure. Each topic is represented by the top 5 words. The evaluators either pick the intruder topic out of the candidates or choose not to answer the question if they could not make the choice. The answer rate of the questionnaire could reflect the distinguishability of the topics and the correct answer rate could reflect the quality of the parent-child relationship of the generated topics.

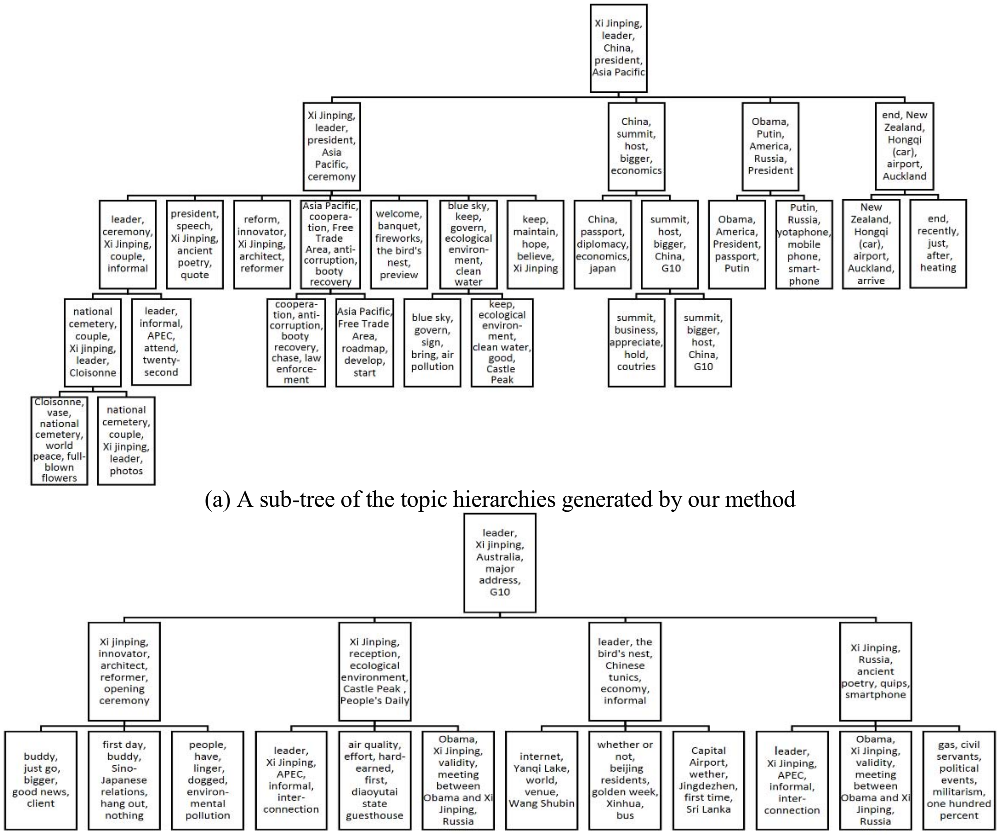
We compare our method only with hPAM in this task as topics generated by nCRP contain too much noise. As the hPAM is a directed acyclic graph model, we take the three strongest subtopics of the non-leaf topic as the true child topics. We invite three participants and show them 30 topic intrusion questions. The results are shown in TABLE I. We can see the hierarchies our method constructed reflect the parent-child relationship of the topics better than the hPAM.

TABLE I. TOPIC INTRUSION RESULTS

	Our Method		hPAM	
	Correct	Answered	Correct	Answered
Rater 1	50.0%	80.0%	30.0%	60.0%
Rater 2	55.0%	75.0%	40.0%	70.0%
Rater 3	45.0%	80.0%	30.0%	50.0%
Average	50.0%	76.7%	33.3%	60.0%

VI. CONCLUSION

In this paper, we propose an approach to automatically


Xi Jinping, leader, China, president, Asia Pacific
Xi Jinping, leader, president, Asia Pacific, ceremony
China, summit, host, bigger, economics
Obama, Putin, America, Russia, President
end, New Zealand, Hongqi (car), airport, Auckland
leader, ceremony, Xi Jinping, couple, informal
president, speech, Xi Jinping, ancient poetry, quote
reform, innovator, Xi Jinping, architect, reformer
Asia Pacific, coopera-tion, Free Trade Area, anti-corruption, booty recovery
welcome, banquet, fireworks, the bird's nest, preview
blue sky, keep, govern, ecological environ-ment, clean water
keep, maintain, hope, believe, Xi Jinping
China, passport, diplomacy, economics, japan
summit, host, bigger, China, G10
Obama, America, President, passport, Putin
Putin, Russia, yotaphone, mobile phone, smart-phone
New Zealand, Hongqi (car), airport, Auckland, arrive
end, recently, just, after, heating
national cemetery, couple, Xi jinping, leader, Cloisonne
leader, informal, APEC, attend, twenty-second
coopera-tion, anti-corruption, booty recovery, chase, law enforce-ment
Asia Pacific, Free Trade Area, roadmap, develop, start
blue sky, govern, sign, bring, air pollution
keep, ecological environ-ment, clean water, good, Castle Peak
summit, business, appreciate, hold, coutries
summit, bigger, host, China, G10
Cloisonne, vase, national cemetery, world peace, full-blown flowers
national cemetery, couple, Xi jinping, leader, photos

(a) A sub-tree of the topic hierarchies generated by our method

leader, Xi jinping, Australia, major address, G10
Xi jinping, innovator, architect, reformer, opening ceremony
Xi Jinping, reception, ecological environment, Castle Peak , People's Daily
leader, the bird's nest, Chinese tunics, economy, informal
Xi Jinping, Russia, ancient poetry, quips, smartphone
buddy, just go, bigger, good news, client
first day, buddy, Sino-Japanese relations, hang out, nothing
people, have, linger, dogged, environ-mental pollution
leader, Xi Jinping, APEC, informal, inter-connection
air quality, effort, hard-earned, first, diaoyutai state guesthouse
Obama, Xi Jinping, validity, meeting between Obama and Xi Jinping, Russia
internet, Yanqi Lake, world, venue, Wang Shubin
whether or not, beijing residents, golden week, Xinhua, bus
Capital Airport, wether, Jingdezhen, first time, Sri Lanka
leader, Xi Jinping, APEC, informal, inter-connection
Obama, Xi Jinping, validity, meeting between Obama and Xi Jinping, Russia
gas, civil servants, political events, militarism, one hundred percent

(b) Part of the topic hierarchies generated by hPAM (We show three strongest subtopics for each topic in the second level)

Fig.3 The generated topic hierarchical structures (We translate Chinese into English)

construct the topic hierarchies in a bottom up manner. We detect topics by NMF and build the topic hierarchies using a multi-branch hierarchical construction method. We conduct a preliminary empirical study using the Weibo dataset and compare our method with the related work. The experimental results show the advantage of our method.

ACKNOWLEDGEMENTS

This work is supported in part by NNSFC grants #61175040, #91224008 and #U1435221, Ministry of Health grant #2013ZX10004218, and grant #2015A069.

REFERENCES

[1] D. M. Blei, T. L. Griffiths, M. I. Jordan, and J. B. Tenenbaum, “Hierarchical Topic Models and the Nested Chinese Restaurant Process.,” *NIPS*, 2003, pp. 17–24.

[2] A. Ahmed, L. Hong, and A. Smola, “Nested Chinese Restaurant Franchise Process: Applications to User Tracking and Document Modeling,” *Proceedings of the 30th International Conference on Machine Learning*, 2013, pp. 1426–1434.

[3] W. Li and A. McCallum, “Pachinko Allocation: DAG-structured Mixture Models of Topic Correlations,” *Proceedings of the 23rd International Conference on Machine Learning*, 2006, pp. 577–584.

[4] D. Mimno, W. Li, and A. McCallum, “Mixtures of Hierarchical Topics with Pachinko Allocation,” *Proceedings of the 24th International Conference on Machine Learning*, 2007, pp. 633–640.

[5] C. Wang, M. Danilevsky, N. Desai, Y. Zhang, P. Nguyen, T. Taula, and J. Han, “A Phrase Mining Framework for Recursive Construction of a Topical Hierarchy,” *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2013, pp. 437–445.

[6] T. Hofmann, “Probabilistic Latent Semantic Indexing,” *Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999, pp. 50–57.

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[8] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, “Hierarchical dirichlet processes,” *J. Am. Stat. Assoc.*, vol. 101, no. 476, 2006.

[9] X. Zhu, Z.-Y. Ming, X. Zhu, and T.-S. Chua, “Topic Hierarchy Construction for the Organization of Multi-source User Generated Contents,” *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2013, pp. 233–242.

[10] X. Zhu, Z.-Y. Ming, Y. Hao, X. Zhu, and T.-S. Chua, “Customized Organization of Social Media Contents Using Focused Topic Hierarchy,” *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, 2014, pp. 1509–1518.

[11] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, vol. 401, no. 6755, pp. 788–791, Oct. 1999.

[12] C. Blundell, Y. W. Teh, and K. A. Heller, “Bayesian Rose Trees,” *ArXiv12033468 Cs Stat*, Mar. 2012.

[13] A. N. Langville, C. D. Meyer, R. Albright, J. Cox, and D. Duling, “Algorithms, Initializations, and Convergence for the Nonnegative Matrix Factorization,” *ArXiv14077299 Cs Stat*, Jul. 2014.

[14] P. D. Turney and P. Pantel, “From Frequency to Meaning: Vector Space Models of Semantics,” *J Artif Int Res*, vol. 37, no. 1, pp. 141–188, Jan. 2010.

[15] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-graber, and D. M. Blei, “Reading Tea Leaves: How Humans Interpret Topic Models,” *NIPS* 2009, pp. 288–296.