Tackling Sparsity, the Achilles Heel of Social Networks: Language Model Smoothing via Social Regularization

Rui Yan[1], **Xiang Li**[1,2], **Mengwen Liu**[3] **and Xiaohua Hu**[3]
[1]Baidu Research, Baidu Inc., Beijing, China
[2]Dept. of Computer Science & Technology, Peking University, Beijing, China
[3]College of Information Science & Technology, Drexel University, Philadelphia, USA
{yanrui02,lixiang32}@baidu.com, {ml943,xh29}@drexel.edu

Abstract

Online social networks nowadays have the worldwide prosperity, as they have revolutionized the way for people to discover, to share, and to diffuse information. Social networks are powerful, yet they still have Achilles Heel: extreme data sparsity. Individual posting documents, (e.g., a microblog less than 140 characters), seem to be too sparse to make a difference under various scenarios, while in fact they are quite different. We propose to tackle this specific weakness of social networks by smoothing the posting document language model based on social regularization. We formulate an optimization framework with a social regularizer. Experimental results on the *Twitter* dataset validate the effectiveness and efficiency of our proposed model.

1 Introduction

Along with Web 2.0 online social networks have revolutionized the way for people to discover, to share and to propagate information via peer-to-peer interactions (Kwak et al., 2010). Although powerful as social networks are, they still suffer from a severe weakness: extreme sparsity. Due to the special characteristics of real-time propagation, the postings on social networks are either officially limited within a limit length (140 characters on Twitter), or generally quite short due to user preference. Given limited text data sampling, a language model estimation usually encounters with zero count problem when facing with data sparsity, which is not reliable. Therefore, *sparsity* is regarded as the Achilles Heel of social networks and now we aim at tackling the bottleneck (Yan et al., 2015).

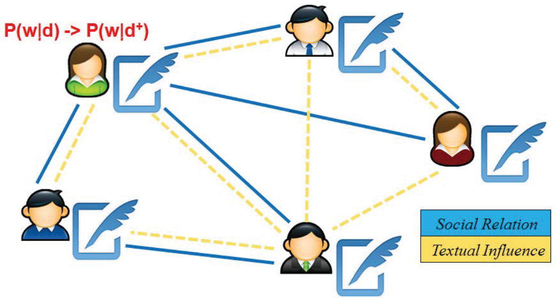

Figure 1: 2 different sources to smooth document language models: texts (colored in yellow) and social contacts (colored in blue). Each piece of texts is authored by a particular social network user.

Statistical language models have attracted much attention in research communities. Till now much work on language model smoothing has been investigated based on textual characteristics (Lafferty and Zhai, 2001; Yan et al., 2013; Liu and Croft, 2004; Tao et al., 2006; Lavrenko and Croft, 2001; Song and Croft, 1999). However, for social networks, texts are actually associated with users (as illustrated in Figure 1). We propose that social factors should be utilized as an augmentation to better smooth language models.

Here we propose an optimization framework with regularization for language model smoothing on social networks, using both textual information and the social structure. We believe the social factor is fundamental to smooth language models on social networks. Our framework optimizes the smoothed language model to be closer to social neighbors in the online network, while avoid deviating too much from the original user language models. Our contributions are as follows:

• We have proposed a balanced language model smoothing framework with optimization, using text information with social structure as a regularizer;

• We have investigated an effective and efficient strategy to model the social information among social network users.

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 623–629,
Beijing, China, July 26-31, 2015. ©2015 Association for Computational Linguistics

We evaluate the effect of our proposed language model smoothing model using datasets from Twitter. Experimental results show that language model smoothing with social regularization is effective and efficient in terms of intrinsic evaluation by perplexity and running time: we show that the Achilles Heel of social networks could be to some extent tackled.

The rest of the paper is organized as follows. We start by reviewing previous works. Then we introduce the language model smoothing with social regularization and its optimization. We describe the experiments and evaluation in the next section and finally draw the conclusions.

2 Related Work

Language models have been paid high attention to during recent years (Ponte and Croft, 1998). Many different ways of language modeling have been proposed to solve different tasks. Better estimation of query language models (Lafferty and Zhai, 2001; Lavrenko and Croft, 2001) and more accurate estimation of document language models (Liu and Croft, 2004; Tao et al., 2006) have long been proved to be of great significance in information retrieval and text mining, etc. Language models are typically implemented based on retrieval models, e.g., text weighting and normalization (Zhai and Lafferty, 2001), but with more elegant mathematical and statistical foundations (Song and Croft, 1999).

There is one problem for language models. Given limited data sampling, a language model estimation sometimes encounters with the zero count problem: the maximum likelihood estimator would assign unseen terms a zero probability, which is not reliable. Language model enrichment is proposed to address this problem, and has been demonstrated to be of great significance (Zhai and Lafferty, 2001; Lafferty and Zhai, 2001).

There are many ways to enrich the original language model. The information of background corpus has been incorporated using linear combination (Ponte and Croft, 1998; Zhai and Lafferty, 2001). In contrast to the simple strategy which smooths all documents with the same background, recently corpus structures have been exploited for more accurate smoothing. The basic idea is to smooth a document language model with the documents similar to the document under consideration through clustering (Liu and Croft, 2004; Tao et al., 2006). Position information has also been used to enrich language model smoothing (Zhao and Yun, 2009; Lv and Zhai, 2009) and has been used in the combination of both enrichment of position and semantic (Yan et al., 2013). Beyond the semantic and/or position related smoothing intuitions, document structure based language model smoothing is another direction to investigate (Duan and Zhai, 2011). Mei *et al.* have proposed to smooth language model utilizing structural adjacency (2008). None of these methods incorporates social factors in language model smoothing.

There is a study in (Lin et al., 2011) which smooths document language models of tweets for topic tracking in online text streams. Basically, it applies general smoothing strategies (e.g., Jelinek-Mercer, Dirichlet, Absolute Discounting, etc.) on the specific tracking task. Social information is incorporated into a factor graph model as features (Huang et al., 2014; Yan et al., 2015). These factor graph model based methods are less efficient so as to better handle *cold-start* situations with little training data. In contrast with these works, we have proposed a language model smoothing framework which incorporates social factors as a regularizer. According to the experimental results, our method is effective with social information and as well much more efficient.

3 Smoothing with Social Regularization

To motivate the model, we briefly discuss the intuitions of proposed language model smoothing. Generally, given a non-smoothed document language model $P(w|d)$, which indicates a word distribution for a term w in document d, we attempt to generate a smoothed language model $P(w|d^+)$ that could better estimate the text contents of a document d as d^+ to avoid zero probabilities for those words not seen in d. Arbitrary assignment of pseudo word counts such as add-λ to every unseen words once was a major improvement for language model smoothing (Chen and Goodman, 1996). However, the purpose of smoothing is to estimate language model more accurately. One of the most useful resources to smooth is the documents similar to d: documents with the larger textual similarity indicate the smaller distance and the better smoothing effects.

Moreover, the author information of the posting documents is easily accessible on social networks. We hence have information related to social factors, which could be used to better estimate the document language model. Through our observation, people are more likely to inherit language habits and usages from their contacts on the social networks. This social factor is important and unique for language model smoothing on social networks. It should be not surprising that smoothing with social factors will be a better optimum. Previously, the pure similarity based smoothing without social factors indicates equal distance for every document from any user on the networks, which is not a fair assumption and presumably leads to a weaker performance.

Yet, with the objective of textual similarity based smoothing with social factors, the smoothed language model might possibly deviate from the original posting documents of a specific user dramatically. It is intuitive that we ought to keep the original representation of document language models of the particular user, and in the meanwhile the postings could be distinguished from one another. Therefore, the combination of the original language model with the social factor as a regularizer ensures the optimum smoothing effects with proper optimization to balance both the textual and social components.

3.1 Problem Formulation

Now we give a formal definition as follows:

Input. Given the entire document set D, and the social network of users U, we aim to smooth the language model of the target document, denoted as $P(w|d_0)$, based on the influence from all other documents d where $\{d|d \in D\}$, and d is authored by $u_d \in U$.

Output. The smoothed language model of $P(w|d_0^+)$ for the original document d_0.

3.2 Methodology Framework

We frame social language smoothing as the interpolation of document representation from the original user and the social factor regularization. Regularization has been cast as an optimization problem in machine learning literature (Zhou and Schölkopf, 2005), and we could form the language model smoothing under this optimization framework. Formally, we propose the smoothing framework for language models with the regularized social factor as follows:

$$O(d_0) = \lambda \sum_{u_{d_i}=u_0} \phi_{d_i} |P(w|d_0^+) - P(w|d_i)|^2 + \\ (1-\lambda) \sum_{u \in U \backslash u_0} \pi_u \sum_{u_{d_j} \neq u_0} \phi_{d_j} |P(w|d_0^+) - P(w|d_j)|^2 \quad (1)$$

where $u_0 = u_{d_0}$, which means the author of d_0 to smooth. Function π_u indicates the social relationship between user u and u_0. Function ϕ_d measures the textual similarity between document d and the document d_0 to smooth. The smoothed document language model is denoted as $P(w|d_0^+)$, and the unsmoothed document language model for d is written as $P(w|d)$.

The objective function of $O(.)$ implement two intuitions: 1) the first component guarantees the smoothed language model would not deviate too much from the language habits of the user of u_0, controlled by the similarity between all the documents from the author of d_0; 2) the second term, namely a harmonic function in semi-supervised learning, incorporating the influence from contacts on the social networks. The framework is general since the functions could be initiated in different instances. Different initiations of functions indicate different features or factors to be taken into account. In this paper, we formulate the textual similarity of ϕ_d, and the social relationship π_u based on the social network dimension. Eventually, we can find the flexibility to extend features and factors in future work.

Firstly, we will define the correlation ϕ_d between document pairs. It is intuitive to measure the relationship among documents based on the textual similarity. In this paper, we utilize the standard cosine metric to measure the similarity between posting document in vector space model representations (Salton et al., 1975). Vector components are set to their *tf.idf* values (Manning et al., 2008). *tf* is the term frequency and *idf* is the inverse document frequency. Next we continue to define the social factor among users.

For π_u, the most intuitive way is to calculate the contacts similarity of the social network users, i.e., friends or followees in common. We first apply the Jaccard distance (Jaccard, 1912; Pang-Ning et al., 2006) on the social contact sets for the two network users (i.e., between u_0 and another particular user u) as follows:

$$\pi_u = \frac{|\{nb(u_0)\} \cap \{nb(u)\}|}{|\{nb(u_0)\} \cup \{nb(u)\}|} \quad (2)$$

#User	#Docs	#Link
9,449,542	364,287,744	596,777,491

Clusters	#Docs	Notes
1. apple	42,528	Tech: apple products
2. nfl	40,340	Sport: American football
3. travel	38,345	General interst

Table 1: Statistics of dataset and topic clusters.

where $\{nb(u)\}$ indicates the set of all neighbor contacts of node u, each of which shares an edge to u.

Now we have finished modeling the language model smoothing with social factors as regularization, and have defined the context correlation between documents and user social relationships. By plugging in Equation (2) into Equation (1), we could compute the smoothed language model of $P(w|d_0^+)$. All the definitions for $\pi(.)$ result in a range which varies from 0 to 1. Particularly, the ego user similarity π_{u_0} = 1, which would be a natural and intuitive answer.

4 Experiments and Evaluation

4.1 Datasets and Experimental Setups

Utilizing the data in (Yan et al., 2012), we establish the dataset of microblogs and the corresponding users from 9/29/2012 to 11/30/2012. We use roughly one month as the training set and the rest as testing set. Based on this dataset, we group the posting documents with the same hashtag '#' into clusters as different datasets to evaluate (Lin et al., 2011; Yan et al., 2015; Yan et al., 2011). We manually selected top-3 topics based on popularity (measured in the number of postings within the cluster) and to obtain broad coverage of different types: sports, technology, and general interests, as listed in Table 1.

Pre-processing. Basically, the social network graph can be established from all posting documents and all users. However, the data is noisy. We first pre-filter the pointless babbles (Analytics, 2009) by applying the linguistic quality judgments (e.g., OOV ratio) (Pitler et al., 2010), and then remove inactive users that have less than one follower or followee and remove the users without any linkage to the remaining posting documents. We remove stopwords and URLs, perform stemming, and build the graph after filtering. We establish the language model smoothed with both text information and social factors.

4.2 Algorithms for Comparison

The first baseline is based on the traditional language model: **LM** is the language model without smoothing at all. We include the plain smoothing of **Additive** (also known as Add-δ) smoothing and **Absolute Discounting** decrease the probability of seen words by subtracting a constant (Ney et al., 1995). We also implement several classic strategies smoothed from the whole collection as background information: **Jelinek-Mercer (J-M)** applies a linear interpolation, and **Dirichlet** employs a prior on collection influence (Zhai and Lafferty, 2001; Lafferty and Zhai, 2001).

Beyond these simple heuristics, we also examine a series of semantic based language model smoothing. The most representative two semantic smoothing methods are the Cluster-Based Document Model (**CBDM**) proposed in (Liu and Croft, 2004), and the Document Expansion Language Model (**DELM**) in (Tao et al., 2006). Both methods use semantically similar documents as a smoothing corpus for a particular document. We also include Positional Language Model (**PLM**) proposed in (Lv and Zhai, 2009), which is the state-of-art positional proximity based language smoothing. PLM mainly utilizes positional information without semantic information. We implemented the best reported PLM configuration. We also include the Factor Graph Model (**FGM**) method to make a full comparison with our proposed social regularized smoothing (**SRS**).

4.3 Evaluation Metric

We apply language *perplexity* to evaluate the smoothed language models. The experimental procedure is as follows: given the topic clusters shown in Table 1, we remove the hashtags and compute its *perplexity* with respect to the current topic cluster, defined as a power function:

$$\text{pow}\left[2, -\frac{1}{N}\sum_{w_i \in V} \log P(w_i)\right]$$

Perplexity is actually an entropy based evaluation. In this sense, the lower perplexity within the same topic cluster, the better performance in purity the topic cluster would have.

Topic	#apple	#nfl	#travel
LM	15851	11356	10676
Additive	15195	10035	10342
Absolute	15323	10123	10379
J-M	14115	10011	10185
Dirichlet	13892	9516	10138
PLM	13730	9925	10426
CBDM	12931	9845	9311
DELM	11853	9820	9513
FGM	10788	9539	8408
SRS	11808	9888	9403

Table 2: Perplexity in hashtag clusters.

4.4 Overall Performance

We compare the performance of all methods of language model smoothing on the Twitter datasets. In Table 2 we list the overall results against all baseline methods. We have an average of -7.28% improvement in terms of language perplexity in hashtag topic clusters against all baselines without social information.

The language model without any smoothing strategy performs worst as expected, and once again demonstrates the Achilles Heel of data sparsity on social networks! Simple intuition based methods such as additive smoothing does not help a lot, since it only arbitrarily modifies the given term counts straightforward to avoid zero occurrence, which is proved to be insufficient. Absolute smoothing performs slightly better, due to the idea to incorporate the collection information by term counts. Jelinek-Mercer (J-M) and Dirichlet methods are more useful since they include the information from the whole collection as background language models, but they fail to distinguish documents from documents and use all of them equally into smoothing. PLM offers a strengthened language model smoothing strategy within each posting document based on positions, and smooth the terms outside of the posting document formulating the background collection into a Dirichlet prior. The performance of CBDM and DELM indicates a prominent improvement, and proves that semantic attributes included into the smoothing process really make a difference. Both of the smoothing methods cluster documents, and use the clustered documents as a better background. However, none of these methods has made use of the social factors during the language model smoothing, while both FGM and SRS suggests social factors do have an impact on language model smoothing.

We make a further comparison between FGM and SRS: both are using social information. An interesting phenomenon is that FGM slightly outperforms SRS. The proposed SRS has more efficiency than FGM. It is quite intuitive that FGM is a complicated model based on propagation via linkage while our proposed SRS is a lightweight model using linear combination. Hence SRS is proved to be both effective due to the comparable performance with FGM, and more efficient as the result of simple interpolation.

5 Conclusions

We present a language model smoothing method based on text correlation with social factors as regularization to solve the zero count phenomenon (sparsity!) for short postings on social networks. We smooth the extremely sparse language model based on texts and social connections in optimization. We evaluate the performance of our proposed smoothing method. In general, the social factor is proved to have a meaningful contribution. Our model outperforms all baseline smoothing methods without social information while takes less time to run: the lightweight method balances effectiveness and efficiency best.

Acknowledgments

This work was partially done when the first author was at University of Pennsylvania. We thank all the anonymous reviewers for their valuable and constructive comments in ACL short paper track[1].

References

Pear Analytics. 2009. Twitter study–august 2009. 15.

Stanley F. Chen and Joshua Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*, ACL '96, pages 310–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Huizhong Duan and Chengxiang Zhai. 2011. Exploiting thread structures to improve smoothing of language models for forum post retrieval. In *Advances in Information Retrieval*, pages 350–361. Springer.

[1] This paper was at first submitted to the ACL long paper track. One reviewer insisted his/her (*perhaps disputable*) opinions and the other two reviewers were outvoted. If interested, we would welcome this reviewer to write emails to us and to discuss his/her *very quick* review offered initially before the author response period.

Yu-Yang Huang, Rui Yan, Tsung-Ting Kuo, and Shou-De Lin. 2014. Enriching cold start personalized language model using social network information. In *Proceedings of the 52nd Annual Meeting on Association for Computational Linguistics*, ACL '14, pages 611–617.

Paul Jaccard. 1912. The distribution of the flora in the alpine zone. *New phytologist*, 11(2):37–50.

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 591–600, New York, NY, USA. ACM.

John Lafferty and Chengxiang Zhai. 2001. Document language models, query models, and risk minimization for information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 111–119, New York, NY, USA. ACM.

Victor Lavrenko and W. Bruce Croft. 2001. Relevance based language models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 120–127, New York, NY, USA. ACM.

Jimmy Lin, Rion Snow, and William Morgan. 2011. Smoothing techniques for adaptive online language models: Topic tracking in tweet streams. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 422–429, New York, NY, USA. ACM.

Xiaoyong Liu and W. Bruce Croft. 2004. Cluster-based retrieval using language models. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, pages 186–193, New York, NY, USA. ACM.

Yuanhua Lv and ChengXiang Zhai. 2009. Positional language models for information retrieval. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 299–306, New York, NY, USA. ACM.

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*, volume 1.

Qiaozhu Mei, Duo Zhang, and ChengXiang Zhai. 2008. A general optimization framework for smoothing language models on graph structures. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 611–618, New York, NY, USA. ACM.

Hermann Ney, Ute Essen, and Reinhard Kneser. 1995. On the estimation ofsmall'probabilities by leaving-one-out. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 17(12):1202–1212.

Tan Pang-Ning, Michael Steinbach, Vipin Kumar, et al. 2006. Introduction to data mining. In *Library of Congress*, page 74.

Emily Pitler, Annie Louis, and Ani Nenkova. 2010. Automatic evaluation of linguistic quality in multi-document summarization. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 544–554, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jay M. Ponte and W. Bruce Croft. 1998. A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 275–281, New York, NY, USA. ACM.

G. Salton, A. Wong, and C. S. Yang. 1975. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, November.

Fei Song and W. Bruce Croft. 1999. A general language model for information retrieval. In *Proceedings of the Eighth International Conference on Information and Knowledge Management*, CIKM '99, pages 316–321, New York, NY, USA. ACM.

Tao Tao, Xuanhui Wang, Qiaozhu Mei, and ChengXiang Zhai. 2006. Language model information retrieval with document expansion. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, pages 407–414, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rui Yan, Xiaojun Wan, Jahna Otterbacher, Liang Kong, Xiaoming Li, and Yan Zhang. 2011. Evolutionary timeline summarization: A balanced optimization framework via iterative substitution. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 745–754, New York, NY, USA. ACM.

Rui Yan, Mirella Lapata, and Xiaoming Li. 2012. Tweet recommendation with graph co-ranking. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 516–525, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rui Yan, Han Jiang, Mirella Lapata, Shou-De Lin, Xueqiang Lv, and Xiaoming Li. 2013. Semantic v.s. positions: Utilizing balanced proximity in language model smoothing for information retrieval. In

Proceedings of the 6th International Joint Conference on Natural Language Processing, IJCNLP'13, pages 507–515.

Rui Yan, Ian E.H. Yen, Cheng-Te Li, Shiqi Zhao, and Xiaohua Hu. 2015. Tackling the achilles heel of social networks: Influence propagation based language model smoothing. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 1318–1328, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Chengxiang Zhai and John Lafferty. 2001. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 334–342, New York, NY, USA. ACM.

Jinglei Zhao and Yeogirl Yun. 2009. A proximity language model for information retrieval. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 291–298, New York, NY, USA. ACM.

Dengyong Zhou and Bernhard Schölkopf. 2005. Regularization on discrete spaces. In *Pattern Recognition*, pages 361–368. Springer.