What do you know?
A topic-model approach to authority identification

Alexandre Passos, Jacques Wainer
RECOD Lab
University of Campinas, Brazil
{tachard,wainer}@ic.unicamp.br

Aria Haghighi
Department of Computer Science
University of Massachusetts, Amherst
aria@cs.umass.edu

1 Introduction

A significant problem in the analysis of social media – e.g., web pages, discussion forums, blogs, product reviews, etc– is how to identify authoritative documents in a given domain. Broadly, authoritative documents are those which exhibit novel and relevant information relative to a document collection and generally demonstrate domain knowledge. Identifying authoritative documents and finding useful content will become a key natural language processing (NLP) and social computing problem as the amount of social media grows. In this work, we present a preliminary study of basic approaches to the problem, including a surprisingly strong and simple baseline. Our best performing approach is a hierarchical topic model [Blei et al., 2004] similar to Haghighi and Vanderwende 2009.

Previous work have explored graph-theoretical approaches to authority in networks where formal links are present between documents, whether they be in the form of paper citations [Salton and Bergmark, 1979, Kleinberg, 1999] or hyperlinks [Page et al., 1998]. Textual analysis of authority has focused on recognizing text of high linguistic register [Harper et al., 2008, Kim et al., 2006], e.g., text which is well-written and cogent.

In this paper we focus on identifying authoritativeness from the textual content of social media documents relative to other documents on the same topic. This approach borrows many of the ideas and techniques from the multi-document summarization research. Specifically, we focus on product (book and restaurant) reviews (see Section 2) and utilize user votes as a proxy for the helpfulness and authoritativeness of a review. In this context we treat authoritativeness identification as a ranking problem over reviews for a given product. Interestingly, many of our strongest performing baselines are *unsupervised* and do not require having user helpfulness ratings at training time.

2 Problem setting

In this paper we assume that user votes on online reviews correlate with review authoritativeness. We then focus on predicting the helpfulness of online product reviews as a proxy for identifying authoritative documents.

Our experimental setting is as follows: we assume each model is given a set of products, and for each product there is a description and a set of reviews, which are just free-text documents. We ignore star ratings, author identity, and review age, even though all of these factors are potentially relevant to predict the helpful votes, and focus only on the textual content of these reviews. Each model will then rank these reviews according to some model-specific criterion, and we will evaluate how well these ranks correlate with the true ranks (as measured by the user votes), according to standard information retrieval metrics.

For the evaluation we focused on two online reviews datasets: the **goodreads** dataset, extracted from `goodreads.com` book reviews and the **yelp** dataset, extracted from `yelp.com` restaurant

reviews. For the **goodreads** dataset we collected, from the first 326 books in the "Best Books Ever" on the `goodreads.com` website, the first 60 or so reviews from each book. For the **yelp** dataset we collected all the reviews for the 283 most reviewed restaurants in the Boston/Cambridge area. For both datasets we use the number of "helpful" votes as the true ranking function.

3 Models

In this section we present several models for ranking product reviews. We look broadly at three classes of baseline models: heuristic, summarization algorithms, and a discriminative classifier.

Heuristic Approaches: The simple heuristic models we explore are:

- **random**: Sort reviews randomly.
- **nwords**: Sort reviews by number of words; this predicts that a review with more tokens is more authoritative.
- **unique**: For each word w, let g_w be its count across all documents for all products. Let p_w be its count amongst documents of a given product. Rank a review d of this product by the number of words unique amongst the document collection. Specifically, the score associated with a document is, $\sum_{w \in d \text{ s.t. } p_w=1} \log(g_w + 1)$

random's purpose to serve as a sanity check. **nwords** as a baseline is advocated in [Ghose and Ipeirotis, 2007] and has been shown to weakly correlate with helpfulness. **unique** is, to the best of our knowledge, novel to this work. It also is by a large margin the best performing of the simple heuristic baseline (see Section 4). The intuition behind it is that a good review should introduce relevant words to a discussion that wouldn't be there otherwise. An authoritative review adds 'context.'

Summarization-based Approaches: We consider two summarization-related baseline models, because one aspect of a helpful or authoritative document is that it provides a complete picture of a given topic [Liu et al., 2007]. In this sense, the document which is the best summary of user sentiment about a topic is the most helpful document in a collection.

- **sumbasic**: Rank documents by the sum-basic criterion [Nenkova and Vanderwende, 2005], ordering reviews of the same product by how many high-frequency words they have relative to the product document collection. The score of a document D is $\sum_{w \in D} P(w)$.
- **klsum**: rank by the kl-sum criterion proposed by Haghighi and Vanderwende 2009. It ranks by the unigram KL divergence $KL(P_p||P_r)$, where P_p is a smoothed distribution for all reviews of the same product and P_r is a smoothed distribution for each review. We assume both distributions are drawn from a symmetric Dirichlet with hyper-parameter 0.01. Performance did not depend much on smoothing.

Discriminative Approach: Finally, we consider a simple discriminative classifier baseline.

- **logreg**: A regularized logistic regression classifier, trained to pick the best review for each single product versus the bottom 30%. We used L_2 regularization with $\sigma^2 = 5$ and the L-BFGS-B optimizer [Byrd et al., 1995].
  We used features on each review for word identity, punctuation, textual structure (average word length, average sentence length, number of words, number of sentences, number of paragraphs), and problem-specific features (number of words in common with other reviews of the same product, number of words in common with the product description, specific words in common with other reviews of the same product). This feature set extends the ones used in Agichtein et al. 2008 and Kim et al. 2006 with problem-specific features. For this model only the dataset was divided into training and test sets, and the reported performance numbers refer only to the test set.

3.1 Topic model

Our topic model is a variation of hierarchical LDA [Blei et al., 2004] with a fixed tree structure, similar to the **TopicSum** model proposed in Haghighi and Vanderwende 2009. This model assumes

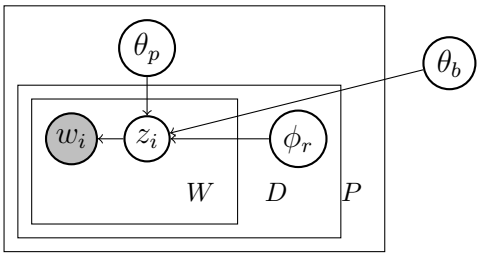

Figure 1: The **topic** model. Each word w_i comes either from a global background topic θ_b or a product-specific topic θ_p. Observed variables are shaded.

	precision@K			NDCG@K			Averaged	Normalized
Model	$K=1$	5	10	$K=1$	5	10	rank	Averaged rank
random	0.08	0.10	0.18	0.03	0.06	0.09	29	0.50
nwords	0.21	0.14	0.19	0.09	0.23	0.28	15	0.25
sumbasic	0.04	0.10	0.18	0.01	0.02	0.04	33	0.57
klsum	0.23	0.14	0.20	0.10	0.24	0.29	16	0.27
unique	0.22	0.14	0.20	0.09	0.22	0.28	14	0.25
logreg	0.24	0.11	0.19	**0.15**	0.23	0.29	17	0.30
topic	**0.28**	**0.20**	**0.24**	**0.15**	**0.28**	**0.36**	**10**	**0.18**

Table 1: Results for the **goodreads** dataset. The best results for each metric are in bold face.

that each word in each review either comes from a background distribution common to all reviews or from a product-specific content distribution common to all reviews of the same product. We rank the reviews by something that roughly represents its number of rare content words, $\sum_{w \in p} \frac{1}{df_w}$, where df_w is the number of reviews of the product that used the word w. For the **yelp** dataset, this sum was over words assigned to the product topic, while for the goodreads dataset it was over words assigned to the global topic. This model follows the intuition that authoritative reviews introduce good content words. Figure 1 shows a graphical depiction of this model.

We estimate this model using 10 iterations of collapsed Gibbs sampling. We experimented with many variants of this model, including changing the number of topics, ranking using a KL criterion, considering a review-specific topic, ranking by number of words in any of the topics, and using a decision variable to separate the topic proportions of the relevant and irrelevant documents, but they all significantly underperformed the version we present here. This model also performed better than the content model of Barzilay and Lee 2004 and than a discriminative classifier with access to Barzilay and Lee content topics as features.

4 Evaluation

We evaluate the review ranking induced by each model using multiple error metrics and report results averaged over all product rankings. The specific metrics we use are: precision@K ($\frac{K}{c}$, where c is how many reviews do you have to look at to find the K best reviews), NDCG@K (normalized discounted cumulative gain of the first K elements of the ranked list), average rank of the best review (from 1 to the number of reviews for that product), and normalized average rank of the best review (the average rank divided by the number of reviews for that product, from 0 to 1).

Table 1 shows the results for the **goodreads** dataset and Table 2 shows the results for the **yelp** dataset. For both datasets, the best unsupervised baseline models are **unique** and **klsum**, closely followed by **nwords**. The **topic** model was the best model in all settings.

5 Conclusions and future work

The performance of the **unique** baseline and the **topic** model are impressive. However, there are two important concerns:

Model	precision@K $K=1$	5	10	NDCG@K $K=1$	5	10	Averaged rank	Normalized Averaged rank
random	0.03	0.03	0.05	0.01	0.02	0.03	109	0.48
nwords	0.10	0.05	0.06	0.05	0.10	0.13	47	0.21
sumbasic	0.01	0.02	0.05	0.00	0.00	0.01	141	0.62
klsum	0.09	0.04	0.06	0.05	0.09	0.11	58	0.25
unique	0.13	0.05	0.07	0.08	0.13	0.17	40	0.17
logreg	0.16	0.04	0.05	0.12	0.14	0.16	60	0.26
topic	**0.18**	**0.06**	**0.07**	**0.13**	**0.18**	**0.22**	**31**	**0.13**

Table 2: Results for the **yelp** dataset. The best results for each metric are in bold face.

- A review that has nothing in common with the topic at hand (such as spam) will tend to be very well ranked, as its unusual words will be assigned to the product distribution and will have a low product count. The models assume all documents are indeed relevant.
- The **topic** model, while not without motivation, is still relatively heuristic: the mix between a topic model to select words and a tf-idf-like heuristic to rank documents is unusual.

We are currently extending the ideas in this model to better capture the phenomena involved without falling in the obvious failure mode of favoring irrelevant or spam content. We are also experimenting with models that aim to directly capture review content specificity; we believe that by explicitly considering such information we can avoid the failure modes of these models.

Acknowledgments

Alexandre Passos is supported financially by a CAPES scholarship. The authors would like to thank Joseph Turian, Alais de Hoogh, Iara Malbouisson, and Daniel Cason for reading preliminary versions of this paper and making suggestions.

References

[Agichtein et al. 2008] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. 2008. Finding high-quality content in social media. In *Proceedings of the international conference on Web search and web data mining*, pages 183–194. ACM.

[Barzilay and Lee 2004] R. Barzilay and L. Lee. 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *Proceedings of HLT-NAACL 2004*, pages 113–120.

[Blei et al. 2004] D. Blei, T.L. Griffiths, M.I. Jordan, and J.B. Tenenbaum. 2004. Hierarchical topic models and the nested Chinese restaurant process. *Advances in neural information processing systems*, 16:106.

[Byrd et al. 1995] R.H. Byrd, P. Lu, J. Nocedal, and C. Zhu. 1995. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208.

[Ghose and Ipeirotis 2007] A. Ghose and P.G. Ipeirotis. 2007. Designing novel review ranking systems: predicting the usefulness and impact of reviews. In *Proceedings of the ninth international conference on Electronic commerce*, page 310. ACM.

[Haghighi and Vanderwende 2009] A. Haghighi and L. Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proceedings of HLT-NAACL 2009*, pages 362–370. Association for Computational Linguistics.

[Harper et al. 2008] F.M. Harper, D. Raban, S. Rafaeli, and J.A. Konstan. 2008. Predictors of answer quality in online Q&A sites. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 865–874. ACM.

[Kim et al. 2006] S.M. Kim, P. Pantel, T. Chklovski, and M. Pennacchiotti. 2006. Automatically assessing review helpfulness. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 423–430. Association for Computational Linguistics.

[Kleinberg 1999] J.M. Kleinberg. 1999. Hubs, authorities, and communities. *ACM Computing Surveys (CSUR)*, 31(4es):5.

[Liu et al. 2007] J. Liu, Y. Cao, C.Y. Lin, Y. Huang, and M. Zhou. 2007. Low-quality product review detection in opinion summarization. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 334–342.

[Nenkova and Vanderwende 2005] A. Nenkova and L. Vanderwende. 2005. The impact of frequency on summarization. *Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005-101*.

[Page et al. 1998] L. Page, S. Brin, R. Motwani, and T. Winograd. 1998. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford University.

[Salton and Bergmark 1979] G. Salton and D. Bergmark. 1979. A citation study of the computer science literature. Technical report, Cornell University.