Identifying the Original Contribution of a Document via Language Modeling

Benyah Shaparenko
Cornell University
4130 Upson Hall
Ithaca NY 14853
benyah@cs.cornell.edu

Thorsten Joachims
Cornell University
4153 Upson Hall
Ithaca NY 14853
benyah@cs.cornell.edu

ABSTRACT

One goal of text mining is to provide readers with automatic methods for quickly finding the key ideas in individual documents and whole corpora. To this effect, we propose a statistically well-founded method for identifying the original ideas that a document contributes to a corpus, focusing on self-referential diachronic corpora such as research publications, blogs, email, and news articles. Our statistical model of passage impact defines (interesting) original content through a combination of impact and novelty, and it can be used to identify the most original passages in a document. Unlike heuristic approaches, this statistical model is extensible and open to analysis. We evaluate the approach on both synthetic and real data, showing that the passage impact model outperforms a heuristic baseline method.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms, Experimentation

Keywords

Language Modeling, Topic Modeling, Original Contributions

1. INTRODUCTION

For diachronic text corpora like research publications, web discussions, and news, identifying the origin and flow of ideas is a crucial step towards understanding their content and structure. To automatically detect the origin of ideas, we propose a generative probabilistic model that allows inference for originality in an unsupervised way. Unlike methods for Topic Detection and Tracking [2], our model provides an operational definition of originality by combining novelty and impact, and we show how it can find text passages that best summarize the original contribution of a document.

We define original ideas as being both novel and having impact. Anybody can write novel nonsense, but adding impact focuses on new ideas that interest and are important to many people. Applications range from automatically extracting snippets to summarize important developments in news corpora, to attributing ideas to their originators in web discussions, and to making the ideas in the large body of research literature easily accessible to humans by summarizing the original contribution of each document.

Our originality-detection methods are text-based and do not require hyperlinks. While heuristic approaches for identifying important words or passages exist (e.g., based on TFIDF [4]), the statistical methods we propose have a concise probabilistic semantic which affords easy analysis and extensibility. Furthermore, we show that using the language-modeling approach empirically outperforms simple heuristics for novelty detection on a discussion-group corpus.

2. METHODS

We take a language modeling approach and define a generative model for diachronic corpora. An author writes a new document using a mixture of novel ideas and ideas "copied" from earlier documents. An idea has impact, if it is copied (i.e., discussed, elaborated on) by future documents. This picture is one of idea flows, originating in documents with impact, and "flowing" to documents based on idea development. We directly model idea flows between documents, without an extra level of the topic as in topic models [3]. Identifying the original contribution of a document means separating novel ideas from old ideas, and simultaneously assessing impact. We assume that documents generally contain a key paragraph or sentence(s) that succinctly summarize the new idea, and we aim to identify this piece of original text as a summary. This task differs from summarization, however, because our method focuses on originality [1]: While all documents can be summarized, some do not contribute original ideas (e.g., a textbook). The task also differs from the TREC novelty track [6], since we detect impact, not relevance. The following gives more detail on our probabilistic model and inference method.

2.1 Passage Impact Model

We propose a generative model of a diachronic corpus that extends the model in [5]. A document $D^{(i)}$ of length n_i is modeled as a vector of n_i random variables $W^{(i)} = (W_1^{(i)}...W_{n_i}^{(i)})'$, one per word. The $W^{(i)}$ can be partitioned into two sets: $Z^{(i)} \subseteq \{1...n_i\}$ for the word indices in $D^{(i)}$ that are original and have impact, while $\bar{Z}^{(i)} = \{1...n_i\} - Z^{(i)}$ contains the rest of $D^{(i)}$. Words in the original portion

k_F	0	1	2
% Err	37.89 ± 3.23	2.95 ± 0.78	0.26 ± 0.16

Table 1: Percentage of misranked non-original passages. Passage length $L = 100$, $\delta = 0.2$, $\pi_n^{(i)} = 0.5$, $\pi_i^{(l)} = 0.05$, and $\pi_n^{(l)} = 0.6$. Two documents $d^{(l)}$, with k_F used in inference. One standard error is shown.

δ	0.2	0.3	0.4
% Err	0.00 ± 0.00	4.74 ± 1.21	24.89 ± 2.80

Table 2: Percentage of misranked non-original passages for $k_F = 2$, $L = 100$, $\pi_n^{(i)} = 0.5$, $\pi_i^{(l)} = 0.05$, and $\pi_n^{(l)} = 0.6$. One standard error is shown.

$Z^{(i)}$ are drawn from a unigram language model with word probabilities $\theta^{(i)}$. The remaining words $\bar{Z}^{(i)}$ come from a mixture of a novel unigram model $\bar{\theta}^{(i)}$ (new, but without impact) and words copied from the original sections of existing documents. Words are copied uniformly and independently so that copying uses a unigram model with parameters $\hat{\theta}^{(k)}$ for each prior document $D^{(k)}$. The document-specific mixing weights $\pi^{(i)}$ are $(\pi_n^{(i)}, \pi_k^{(i)})$ for $\bar{\theta}^{(i)}$ and $\hat{\theta}^{(k)}$, respectively. Formally, we model a diachronic corpus as follows:

MODEL 1. (PASSAGE IMPACT MODEL)
A corpus $\mathcal{C} = (D^{(1)}...D^{(n)})$ of temporally-sorted documents $D^{(i)} = (W^{(i)}, Z^{(i)})$, each having parameters $(\theta^{(i)}, \bar{\theta}^{(i)}, \pi^{(i)})$, has probability $P(\mathcal{C}) = \prod_{i=1}^{n} P(D^{(i)}|D^{(1)}...D^{(i-1)})$ where

$$P(D^{(i)}|D^{(1)}...D^{(i-1)}) = \prod_{j \in z^{(i)}} \left(\theta^{(i)}_{w_j^{(i)}}\right) \prod_{j \in \bar{z}^{(i)}} \left(\pi_n^{(i)} \bar{\theta}^{(i)}_{w_j^{(i)}} + \sum_{k=1}^{i-1} \pi_k^{(i)} \hat{\theta}^{(k)}_{w_j^{(i)}}\right) P(Z^{(i)})$$

and where $\hat{\theta}_w^{(k)}$ is the probability of uniformly drawing word w from the words in the original section $z^{(k)}$ of document $D^{(k)}$. Note that $\pi_n^{(i)} + \sum_k \pi_k^{(i)} = 1$, $\sum_j \theta_j^{(i)} = 1$, and $\sum_j \bar{\theta}_j^{(i)} = 1$.

2.2 Inference

Using the PIM, we want to infer the section $Z^{(i)}$ in $D^{(i)}$ that most succinctly summarizes the original idea. Only document text is observed. After a few justifiable approximations, we use a MAP inference procedure based on Model 1 to infer $Z^{(i)}$ by maximizing $P(D^{(i)}...D^{(n)}|D^{(1)}...D^{(i-1)})$ w.r.t. $Z^{(i)}$, $\theta^{(i)}$, $\bar{\theta}^{(i)}$, and $\pi^{(i)}$. This problem reduces to a sequence of convex programs that can be solved efficiently.

3. EXPERIMENTS

We tested this method on synthetically-generated data based on the Neural Information Processing Systems proceedings and on web documents from Slashdot discussions.

3.1 Synthetically-Generated Data

Synthetic data is based on language models $\theta^{(i)}$ from the MLEs of NIPS documents. Document $d^{(i)}$ has 20 passages. One is $Z^{(i)}$, and the others form $\bar{Z}^{(i)}$. Since real documents may not have exactly one $Z^{(i)}$ passage, we test robustness by diffusing δ original content through $\bar{Z}^{(i)}$. Future documents are $d^{(l)}$. The evaluation measure is the average percentage of $\bar{Z}^{(i)}$ passages that have a higher likelihood than $Z^{(i)}$.

Table 1 shows that impact is critical to define originality. Using just one future document can differentiate real original ideas and novel noise. Table 2 shows that the method is robust for diffused original content, up to $\delta = 0.3$.

	Prec@2 ± One Std Err	Rec@10 ± One Std Err
PIM	22.13 ± 3.38	36.09 ± 3.61
TFIDF	9.84 ± 3.03	25.01 ± 4.04
RAND	10.63 ± 1.10	23.92 ± 2.27

Table 3: Prec@2 and Rec@10 use the predicted sentence ranking by likelihood and TFIDF sum. Original sentences are those quoted word-for-word from the article. Results are for $\pi_o^{(i)} = 0.2$ and $\pi_n^{(l)} = 0.001$.

3.2 Slashdot Discussions

We also evaluate on Slashdot, where users post and discuss articles. Sometimes the first post links to and directly quotes a web document. We collect such linked-to web documents and discussions, treating human-selected direct quotations as labeled original contributions. For inference, we rank all sentences in the linked-to web document $d^{(i)}$ by likelihood. The documents $d^{(1)}...d^{(i-1)}$ are the ones that have been linked to in earlier discussions. The future "document" $d^{(i+1)}$ is the user discussion of $d^{(i)}$, except with direct article quotations removed. We evaluated the method on identifying human-selected sentences on 61 articles (7 months) matching these criteria from the Games subtopic. The baseline uses TFIDF, with sentences scored by the sum its words' IDF values. Evaluation is by Precision and Recall.

The Prec@2 statistics in Table 3 show that the Passage Impact Model outperforms the TFIDF heuristic baseline for predicting original sentences at the top of the ranking. In addition, Rec@10 shows that the PIM is also significantly better when trying to find a large subset of original content.

4. CONCLUSIONS

We proposed a generative model for diachronic corpora and an inference procedure to identify original ideas. This method significantly beats a heuristic baseline for selecting a document section to summarize its original contribution.

5. ACKNOWLEDGMENTS

We acknowledge Adam Siepel, Art Munson, Yisong Yue, Nikos Karampatziakis, and the ML Discussion Group. This work was supported in part by NSF Grant IIS-0812091.

6. REFERENCES

[1] Document understanding conferences. http://duc.nist.gov/.

[2] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic detection and tracking pilot study: Final report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop-1998*, 1998.

[3] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(5):993–1022, 2003.

[4] G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988.

[5] B. Shaparenko and T. Joachims. Information genealogy: Uncovering the flow of ideas in non-hyperlinked document databases. In *Proceedings of KDD-07*, pages 619–628, New York, 2007. ACM Press.

[6] I. Soboroff and D. Harman. Overview of the TREC 2003 novelty track. In *Proceedings of TREC-2003*, 2003.