

# Identifying the Original Contribution of a Document via Language Modeling

Benyah Shaparenko and Thorsten Joachims

Department of Computer Science, Cornell University, Ithaca NY 14853, USA

**Abstract.** One major goal of text mining is to provide automatic methods to help humans grasp the key ideas in ever-increasing text 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 the model is used to identify each document’s most original passages. Unlike heuristic approaches, the statistical model is extensible and open to analysis. We evaluate the approach both on synthetic data and on real data in the domains of research publications and news, showing that the passage impact model outperforms a heuristic baseline method.

## 1 Introduction

With the rapid proliferation of large text corpora, it is especially relevant to provide automatic methods to support users in understanding global aspects of a corpus without requiring them to read it in full. In diachronic corpora that grow over time, one such global aspect is the dependency structure between ideas and documents throughout the corpus. In particular, what is the original contribution that a given document makes, and how does this idea further “flow” through the corpus? In this paper, we focus on the first half of this question and develop methods that automatically identify the original ideas that a document contributes. Our methods leverage the diachronic nature of many text corpora, where ideas originate in some documents and get discussed and refined in later documents. Such corpora include research publications, email, news articles, Wikipedia content, discussion boards, and blogs. Useful applications include the visualization of corpora, the detection of important developments in news corpora, or the attribution of ideas in blogs or email discussions.

When identifying the original ideas expressed in a document, we are most interested in ideas that ultimately had impact. Anybody can write some spam on a discussion board, which would likely be novel to the discussion (at least the first time), but not particularly interesting. In addition to novelty, measuring the impact of an idea lets us focus on those ideas that are important, or that at least are interesting to a large number of people. Therefore, our operational definition of an original contribution combines both novelty and impact.

Unlike methods that rely on explicit citations that must be localizable in each document [1], our methods require only the text of the documents. This makes them more broadly applicable than citation-based measures (e.g., for email, news). Furthermore, unlike novelty detection methods [2] (e.g., based on TFIDF-style measures), our methods combine novelty with impact, which provides a way of measuring the importance of novel ideas. The originality-detection methods we propose are derived from a probabilistic language model of diachronic corpora – called the Passage Impact Model (PIM), which makes them theoretically well-founded and more extensible than heuristic approaches. The method is evaluated on a corpus of Slashdot discussions, as well as through a blind experiment with human judges on a collection of NIPS research articles. In both experiments, the language modeling approach was found to outperform a heuristic that focuses on novelty detection alone.

## 2 Related Work

The task of succinctly describing the original contribution of a document relates to several existing research areas, including document summarization, topic detection, topic modeling, and language modeling.

The largest body of related work is in document summarization (see e.g. [3]). Document summarization methods provide the user with a summary of the entire document, including both original and existing ideas, without explicitly making a distinction. The difference between summarization and originality detection is most apparent for documents that do not necessarily contain original content (e.g., textbooks, review articles). While such documents have a summary, their original contribution can be quite different or even non-existent.

Another area of related work lies in novelty detection for Topic Detection and Tracking [4, 5] in news streams. There, the task is to identify new topics and events as they appear in the news. One major difference is that the Passage Impact Model segments the document to identify a single passage that best describes that document’s original contribution. Thus the inference method can actually find a text description within the document, instead of just marking that the document contains a novel topic. A second difference is that the Passage Impact Model combines novelty with impact, focusing on ideas that not only are novel but also affect the rest of the corpus. The TREC Novelty track [2] solves a different problem, combining novelty and relevance, not novelty and impact.

One previous paper has tackled the problem of making “impact-based summaries” [1]. Their method is based on citation contexts for explicit citations to a document  $d$ . The task is to select the sentence  $s$  in document  $d$  that best describes the contribution of  $d$  that had impact in these citation contexts. That work followed a KL-divergence-based information retrieval framework where the document  $d$  stands for the corpus, the sentences  $s$  stand for the documents to be retrieved, and the citation context is descriptive of the “query.” The Passage Impact Model is quite different in model and inference, since it does not require citations. Instead, our method is based on an extensible generative and unsu-

pervised language-modeling framework. We start from a generative model of the corpus and derive an inference method to identify the most densely-concentrated original contribution in the document  $d$ . We do not need to use a citation context, as the method is completely text-based.

On a higher level, topic models and other language models also provide generative models of corpora. In topic models, however, the focus is on discovering underlying topics, without any explicit notion of originality or impact. Typically, topics are inferred by fitting graphical models with topics as the latent variables. Latent Dirichlet Allocation (LDA) [6, 7] and its extensions [8, 9] are the most well-known, but there is much other work in topic modeling [10–20]. In this sense, topic models describe the relationship between topics and documents, but not the relationships between individual documents. Our Passage Impact Model directly models relationships between documents via a copy process. In this sense it builds on the models in [21, 15], extending them to recognizing document substructure. We use simple unigram language models in the PIM, but one could also use more complex language models [22, 10, 6, 23–26].

### 3 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 [6]. 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 describe the new idea, and we aim to identify this piece of original text. The following gives more detail on our probabilistic model and inference method.

#### 3.1 Passage Impact Model

We propose a generative model of a diachronic corpus that extends the model in [21] with respect to modeling originality. We model a document  $D^{(i)}$  containing  $n_i$  words as a vector of  $n_i$  random variables  $W^{(i)} = (W_1^{(i)} \dots W_{n_i}^{(i)})'$ , one per word. Considering the process by which authors write documents, the text can be split into several types: original content that will have impact on following documents, novel content that will not have impact, and content “copied” from already-existing ideas in the corpus. The location of the original content in  $D^{(i)}$  is denoted by  $Z^{(i)}$ , where  $Z^{(i)} \subseteq \{1 \dots n_i\}$ . More concretely, the random variables  $W^{(i)}$  are partitioned into two sets:  $Z^{(i)} \subseteq \{1 \dots n_i\}$  for the indices of the words of  $D^{(i)}$  that are original and have impact, while  $\bar{Z}^{(i)} = \{1 \dots n_i\} - Z^{(i)}$  contains

the rest of  $D^{(i)}$  (i.e., the copied content and the novel content without impact). With these definitions, the document is described by the tuple

$$D^{(i)} = (W^{(i)}, Z^{(i)}) \quad (1)$$

and we will now define a probabilistic model of a document  $P(D^{(i)}|D^{(1)}\dots D^{(i-1)})$ . Each document  $D^{(i)}$  can draw on the ideas already expressed in the existing documents  $D^{(1)}\dots D^{(i-1)}$  in the corpus. The probability of an entire corpus  $\mathcal{C}$  consisting of documents  $D^{(1)}\dots D^{(n)}$ , can be decomposed as

$$P(\mathcal{C}) = \prod_{i=1}^n P(D^{(i)}|D^{(1)}\dots D^{(i-1)}). \quad (2)$$

We decompose the probability for a single document  $D^{(i)}$  into

$$\begin{aligned} P(D^{(i)}|D^{(1)}\dots D^{(i-1)}) &= P(W^{(i)}, Z^{(i)}|D^{(1)}\dots D^{(i-1)}) \\ &= P(W^{(i)}|Z^{(i)}, D^{(1)}\dots D^{(i-1)})P(Z^{(i)}) \end{aligned}$$

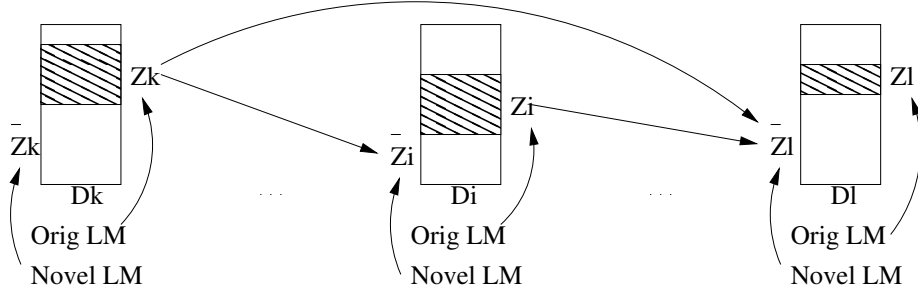
since the document text  $W^{(i)}$  depends on the previous documents, but the author's selection of placement of original content is independent of previous documents. Prior information about the placement of  $Z^{(i)}$  in the document can be encoded in  $P(Z^{(i)})$ . Furthermore, in the inference described below, the quantity  $P(Z^{(i)})$  can be used to encode constraints on the form of original content summary that is desirable (e.g., a single sentence or a single paragraph).

Words in the original portion  $Z^{(i)}$  are generated from a unigram language model with word probabilities  $\theta^{(i)}$ . The rest of the document (i.e. the words indexed by  $\bar{Z}^{(i)}$ ) comes from a mixture of existing ideas and text that is novel but without impact. That is, the words indexed by  $\bar{Z}^{(i)}$  are drawn from a mixture of a novel unigram model  $\bar{\theta}^{(i)}$  (new but without impact) and words copied from the original sections of prior documents. Words are drawn uniformly and independently in this copy process so that it can also be described by 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.

With the assumption that text is generated from these unigram multinomial language models, the generative model of the text given  $Z^{(i)}$  and the existing corpus at time  $i$  is

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

Figure 1 illustrates the generative process at document  $d^{(i)}$ , showing how  $d^{(i)}$  copies content from the original part  $Z^{(k)}$  of earlier documents  $d^{(k)}$  and showing how terms indexed by  $Z^{(i)}$  are copied by later documents  $d^{(l)}$ . We summarize this generative process of a diachronic corpus in the Passage Impact Model.



**Fig. 1.** The generative process for a corpus. Document  $d^{(i)}$  is the current document, while  $d^{(k)}$  precede  $d^{(i)}$  in time and  $d^{(l)}$  follow  $d^{(i)}$ . The shaded boxes are original content  $Z^{(\cdot)}$ , while the rest of the documents form  $\bar{Z}^{(\cdot)}$ . The arrows depict the copy process.

### Model 1 (PASSAGE IMPACT MODEL)

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

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

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  $\sum_j \bar{\theta}_j^{(i)} = 1$ .

### 3.2 Inference

Using the Passage Impact Model, we are primarily interested in inferring the subset  $Z^{(i)}$  of words in  $D^{(i)}$  where the original contribution is most succinctly contained. The only observed quantity is the text  $w^{(1)} \dots w^{(n)}$  of all documents. We use maximum-likelihood inference based on Model 1 for inferring  $Z^{(1)} \dots Z^{(n)}$  by maximizing  $P(D^{(1)} \dots D^{(n)})$  given  $w^{(1)} \dots w^{(n)}$  w.r.t.  $Z^{(i)}$ ,  $\theta^{(i)}$ ,  $\bar{\theta}^{(i)}$ , and  $\pi^{(i)}$ . Applying Bayes rule and independence assumptions involving the placement of original content  $Z^{(\cdot)}$  in different documents  $D^{(i)} \dots D^{(n)}$ , the inferred original content  $Z^{(i)*}$  is given by the following:

$$\begin{aligned} (Z^{(1)*} \dots Z^{(n)*}) &= \operatorname{argmax}_{Z^{(1)} \dots Z^{(n)}} \max_{(\theta, \bar{\theta}, \pi)} P(w^{(1)} \dots w^{(n)} | Z^{(1)} \dots Z^{(n)}) P(Z^{(1)} \dots Z^{(n)}) \\ &= \operatorname{argmax}_{Z^{(1)} \dots Z^{(n)}} \max_{(\theta, \bar{\theta}, \pi)} P(w^{(1)} \dots w^{(n)} | Z^{(1)} \dots Z^{(n)}) P(Z^{(1)}) \dots P(Z^{(n)}) \end{aligned}$$

Note that we do not explicitly include the parameters  $\theta$ ,  $\bar{\theta}$ , and  $\pi$  in the notation for improved readability, since their dependence is straightforward. To

avoid the intractable simultaneous maximization over all  $(Z^{(i)} \dots Z^{(n)})$ , we introduce some simplifying assumptions that allow independent optimization for each  $Z^{(i)}$ . First, we assume that for all prior documents  $d^{(1)} \dots d^{(i-1)}$ , the copy probabilities  $\hat{\theta}^{(1)} \dots \hat{\theta}^{(i-1)}$  can be approximately estimated from the full set of words  $w^{(1)} \dots w^{(i-1)}$ , respectively, not merely the words indexed by the original markers  $z^{(1)} \dots z^{(i-1)}$ . In practice, this assumption can be expected to have only minor impact<sup>1</sup>, and it can be removed if  $z^{(1)} \dots z^{(i-1)}$  are already known. With this assumption, we have that for any  $i$

$$\begin{aligned} (Z^{(i)*} \dots Z^{(n)*}) &= \operatorname{argmax}_{Z^{(i)}, Z^{(n)}} \max_{Z^{(1)}, \dots, Z^{(i-1)}} \max_{(\theta, \bar{\theta}, \pi)} P(w^{(1)} \dots w^{(n)} | Z^{(1)} \dots Z^{(n)}) P(Z^{(1)}) \dots P(Z^{(n)}) \\ &= \operatorname{argmax}_{Z^{(i)}, Z^{(n)}} \max_{(\theta, \bar{\theta}, \pi)} P(w^{(i)} \dots w^{(n)} | Z^{(i)} \dots Z^{(n)}, \hat{\theta}^{(1)} \dots \hat{\theta}^{(i-1)}) P(Z^{(i)}) \dots P(Z^{(n)}) \end{aligned}$$

Second, we introduce a simplified model for the future documents  $D^{(i+1)} \dots D^{(n)}$  so that one can maximize over  $Z^{(i)}$  independently. When inferring  $Z^{(i)}$ , modeling exactly how future documents  $D^{(l)}$ ,  $l > i$ , had impact on each other is of minor importance, so that we do not model their  $Z^{(l)}$ . Instead, we assume that the original and novel content of future documents comes from a multinomial mixture, which can be captured by a single multinomial language model  $\bar{\theta}^{(l)}$ . Thus, each  $D^{(l)}$  depends only on the documents  $D^{(1)} \dots D^{(i)}$ , and

$$P(w^{(i+1)} \dots w^{(n)} | Z^{(i)} \dots Z^{(n)}, \hat{\theta}^{(1)} \dots \hat{\theta}^{(i-1)}) = \prod_{l=i+1}^n P(w^{(l)} | Z^{(i)}, w^{(i)}, \hat{\theta}^{(1)} \dots \hat{\theta}^{(i-1)})$$

Putting all of these assumptions together, we can rewrite the objective function as the likelihood of the documents in the corpus starting from  $D^{(i)}$ , given all the documents that precede  $D^{(i)}$ , which is  $P(D^{(i)} \dots D^{(n)} | D^{(1)} \dots D^{(i-1)})$ . We express this likelihood using the parameters  $(\theta^{(i)}, \bar{\theta}^{(i)}, \pi^{(i)})$  as follows:

$$\begin{aligned} Z^{(i)*} &= \operatorname{argmax}_{Z^{(i)}} \max_{(\theta, \bar{\theta}, \pi)} P(Z^{(i)}) P(w^{(i)} | Z^{(i)}, \hat{\theta}^{(1)} \dots \hat{\theta}^{(i-1)}) \prod_{l=i+1}^n P(w^{(l)} | Z^{(i)}, w^{(i)}, \hat{\theta}^{(1)} \dots \hat{\theta}^{(i-1)}) \\ &= \operatorname{argmax}_{Z^{(i)}} \max_{(\theta, \bar{\theta}, \pi)} \left[ P(Z^{(i)}) \prod_{j \in z^{(i)}} \binom{\theta^{(i)}}{w_j^{(i)}} \prod_{j \in \bar{z}^{(i)}} \left( \pi_n^{(i)} \bar{\theta}_{w_j^{(i)}}^{(i)} + \sum_{k=1}^{i-1} \pi_k^{(i)} \hat{\theta}_{w_j^{(i)}}^{(k)} \right) \right. \\ &\quad \left. \prod_{l=i+1}^n \prod_{j=1}^{n_l} \left( \pi_n^{(l)} \bar{\theta}_{w_j^{(l)}}^{(l)} + \sum_{k=1}^i \pi_k^{(l)} \hat{\theta}_{w_j^{(l)}}^{(k)} \right) \right] \quad (3) \end{aligned}$$

Note that the various  $\pi^{(\cdot)}$  and  $\bar{\theta}^{(\cdot)}$ , as well as  $\theta^{(i)}$ , are linearly constrained to form proper probability distributions, and that  $\hat{\theta}^{(i)}$  can be computed in closed

<sup>1</sup> Since each document can be a mixture of original content and previous content, when estimating  $\hat{\theta}^{(\cdot)}$  from the entire document, it is equal to the true  $\hat{\theta}^{(\cdot)}$ , mixed with some previous content that would have come from the  $\hat{\theta}^{(\cdot)}$  of even earlier documents in the corpus. This assumption means that the  $\hat{\theta}^{(\cdot)}$  also could include some content from  $\bar{\theta}^{(i)}$ . However, if this portion's mixture component is relatively small, the  $\hat{\theta}^{(i)}$  will still be quite faithful to the Passage Impact Model's definition.

form for a given  $z^{(i)}$ . For a fixed  $z^{(i)}$ , the above optimization problem is convex and has no local optima. The prior  $P(Z^{(i)})$  can be used to enforce a particular form of original content description (e.g., that the algorithm has to select a whole paragraph or a single sentence).

### 3.3 Implementation Details

When solving the optimization problem, the method can efficiently find the maximum likelihood if given a specific  $z^{(i)}$ . In the following, we therefore give non-zero prior  $P(Z^{(i)})$  only to a fairly small number of  $z^{(i)}$  that can be enumerated explicitly. This allows us to find the globally optimal solution of Eq. 3. In particular, we break documents into consecutive passages of equal length, which we denote  $s^1 \dots s^K$ . We set  $P(Z^{(i)} = s^k)$  to be uniform for each  $k = 1 \dots K$ , with all other  $P(z^{(i)}) = 0$ . One could also define a non-uniform prior over the candidate passages  $z^{(i)}$  to encode additional knowledge (e.g., bias toward the beginning or end of the document). With this particular assumption on  $z^{(i)}$ , the entire likelihood maximization can now be reduced to a sequence of convex problems, one per  $s^k$ . The solution to this sequence of optimizations is the global maximum likelihood across the passages. We use the general software optimization tool MOSEK to solve these convex optimizations [27].

While the individual problems are convex, for efficiency reasons, we have to consider the number of parameters in the Passage Impact Model. Therefore, when performing inference on document  $d^{(i)}$ , instead of using the full set of previous documents  $\{d^{(1)} \dots d^{(i-1)}\}$ , we choose the set of  $k_P$  nearest neighbors from these documents according to cosine similarity. The document indices for these  $k_P$  documents are given in the set  $\mathcal{P}$ . Besides  $d^{(i)}$ , the optimization also uses the likelihood of generating the documents  $d^{(i+1)} \dots d^{(n)}$ . Each of these “future” documents  $d^{(l)}$  has its own set of mixing weights and set of previous documents, again chosen from the documents  $\{d^{(1)} \dots d^{(i-1)}\}$  nearest to  $d^{(l)}$  by cosine similarity. While we do not use the following strategies for improving efficiency, one could further reduce the size of the optimization problem. For example, it is possible to consider a Passage Aggregated Impact Model, wherein all future text is “lumped” together into one single “document” for inference. Then, there would only be a single set of future document parameters. Equivalently, we could constrain all future documents to have the same mixing weights and choose the set of previous neighbors as those most similar to the concatenation of all future documents. There is a tradeoff between using more information in more future documents vs. using more parameters for a specific set of interesting previous documents.

## 4 Experiments

We conducted experiments to test the Passage Impact Model on both synthetic and real data from research publications and news articles.

## 4.1 Experiment 1: Synthetic Data

We use synthetic data to explore the range of problems and parameters under which the methods work effectively and robustly. The synthetic data is generated with underlying language models from documents in the full-text proceedings of the Neural Information Processing Systems (NIPS) conference [28] between 1987-2000. NIPS has 1955 documents with text obtained by OCR, resulting in 74731 unique words (multi-character alphabetic strings), except without stopwords.

To generate a document  $d^{(i)}$ , we selected a NIPS document  $d$  randomly and set the original language model  $\theta^{(i)}$  for  $d^{(i)}$  to be the distribution of words in  $d$ . The words indexed in  $Z^{(i)}$  are then generated according to  $\theta^{(i)}$ . For  $\bar{Z}^{(i)}$ , we set the novel language models  $\bar{\theta}^{(i)}$  and each  $\bar{\theta}^{(l)}$  similarly, with each document selected for  $\bar{\theta}^{(l)}$  following NIPS document  $d$  in time. The mixing weights  $\pi_k^{(i)}$  are selected uniformly at random, except for explicitly exploring  $\pi_i^{(l)}$ ,  $l > i$ , (how much future documents  $d^{(l)}$  copy from  $d^{(i)}$ ) and  $\pi_n^{(i)}$  (how much novel but not original content  $d^{(i)}$  has) according to the values they might take in practice.

The structure of  $Z^{(i)}$  and  $\bar{Z}^{(i)}$  takes the form of  $K = 20$  passages with  $L$  words per passage. In the simplest case,  $Z^{(i)}$  marks exactly one passage as original. In addition, we test scenarios where the original content is more diffused through the document, which poses a challenge in inference. One crucial assumption of our method is that the prior  $P(Z^{(i)})$  used during inference matches the data-generating process. However, the inference procedure as implemented above aims to find a single passage containing all the original content, while the true  $Z^{(i)}$  might diffuse it over other passages. To test the robustness of inference w.r.t. the degree of diffusion, we include a fraction  $\delta$  of original content in the (mostly) non-original passages in data generation, but not during inference.

Evaluation on the synthetic data uses the percentage of (mostly) non-original passages with a greater likelihood than the original passage likelihood. Random performance would be that half of the non-original passages are misranked, resulting in a score of 50%. The error values show one standard error.

**Impact Is Critical** In the first experiment, we explore the difference between pure novelty detection vs. the additional use of impact when identifying  $Z^{(i)}$ . When not using any future documents, our method might still be able to identify  $Z^{(i)}$  merely by fitting the mixture model and detecting that  $Z^{(i)}$  cannot be expressed as a mixture of previous documents. In this setting, our method becomes a pure novelty detection method. However, Table 1 shows that the signal from novelty alone is much weaker than novelty combined with impact. While the performance is better than random when no future documents are used ( $k_F = 0$ ), detection accuracy substantially improves when future documents and impact are considered by the method. The table shows that two future documents that copy 5% of their content from  $d^{(i)}$  already provide a robust signal.

**More Information in Longer Passages** We would like to determine the size of the original passage for which the Passage Impact Model can perform well.



**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  $\pi_n^{(l)} = 0.6$ . 10 future documents  $d^{(l)}$  were generated, and inference used the  $k_F$  documents  $d^{(l)}$  most (cosine) similar to  $d^{(i)}$ .

$k_F$	% Err $\pm$ One Std Err
0	37.89 $\pm$ 3.23
1	2.95 $\pm$ 0.78
2	0.26 $\pm$ 0.16
5	0.00 $\pm$ 0.00
10	0.16 $\pm$ 0.16

**Table 2.** Percentage of misranked non-original passages with  $k_F = 2$  future documents. The data was generated with  $\delta = 0.2$ ,  $\pi_n^{(i)} = 0.5$ ,  $\pi_i^{(l)} = 0.05$ , and  $\pi_n^{(l)} = 0.6$ .

Length	% Err $\pm$ One Std Err
25	8.16 $\pm$ 1.61
50	2.26 $\pm$ 0.65
100	1.00 $\pm$ 0.99
400	2.26 $\pm$ 0.74

Users may be interested in descriptions anywhere from one or more sentences to paragraphs. Table 2 shows that, in general, when performing inference on longer passages, the method is able to perform more accurately. The method performs very well for passages as short as 50 words. However, for very short passages of length 25 words, there is some drop in accuracy. Longer passages – and therefore longer documents – provide more observations, and it is less likely that the method will overfit to a few random draws.

**Diffusiveness of Original Content in  $d^{(i)}$**  The inference method searches for a single passage that contains the original contribution, but realistic documents will have original content spread throughout all passages. How much original content in other passages can our inference method tolerate? Table 3 shows that the method is very robust towards small to moderate diffusion. Even as  $\delta$  increases to 0.3 (i.e. 30% of each of the other passages is original content), the method is still quite accurate. After that, performance degrades rather quickly, at least when only two future documents are used.

**How Much Copying Is Necessary?** As shown above, the Passage Impact Model relies on future documents copying the ideas expressed in the original contribution of  $d^{(i)}$ . How much must each future document copy to provide a sufficient signal? Table 4 shows that the method performs with minimal errors for many values of  $\pi_i^{(l)}$ , even in the situation where future documents copy only 5% of their content (i.e. 100 words) from  $d^{(i)}$ . At lower values for copying, the percentage of correctly ranked passages smoothly decreases. As  $\pi_i^{(l)}$  approaches

**Table 3.** Percentage of misranked non-original passages.  $k_F = 2$  future documents, passages length  $L = 100$  words,  $\pi_n^{(i)} = 0.5$ ,  $\pi_i^{(l)} = 0.05$ , and  $\pi_n^{(l)} = 0.6$ .

$\delta$	% Err $\pm$ One Std Err
0.1	0.00 $\pm$ 0.00
0.2	0.00 $\pm$ 0.00
0.3	4.74 $\pm$ 1.21
0.4	24.89 $\pm$ 2.80
0.5	45.26 $\pm$ 3.38

**Table 4.** Percentage of misranked non-original passages.  $k_F = 2$  future documents, passage length  $L = 100$  words,  $\delta = 0.2$ ,  $\pi_n^{(i)} = 0.5$ , and  $\pi_n^{(l)} = 0.6$ .

$\pi_i^{(l)}$	% Err $\pm$ One Std Err
0.005	34.37 $\pm$ 3.16
0.01	28.58 $\pm$ 2.97
0.02	9.16 $\pm$ 1.41
0.05	0.11 $\pm$ 0.07
0.1	0.00 $\pm$ 0.00
0.2	0.00 $\pm$ 0.00


0, the method becomes essentially equivalent to a novelty detection method that does not using any future documents.

## 4.2 Experiment 2: Slashdot

Besides synthetic data, we also evaluate on the real world dataset of news articles linked to on Slashdot under the Games topic. When users post an entry, they often link to some article on the Web, and sometimes quote directly from it. Then other users read and respond to these postings in a discussion board format. We collect linked-to web documents and discussions from the Games topic where the original poster directly quotes from a linked-to document. We regard the sentences in the human-selected direct quotations as the label for the original content  $z^{(i)}$  of the web document  $d^{(i)}$ .

We collected a set of 61 documents from the Games topic of Slashdot. These are the entries posted from August 2008 through February 2009, inclusive, where the initial entry quotes a portion of the referenced article. The documents are the referenced articles. In addition, we collect the first page of the user discussion on this topic, as selected by Slashdot. Figure 2 shows a screenshot of Slashdot that depicts the data we collected.

**Experiment Setup** To do inference on Slashdot data, we sort the fulltext, linked-to news articles by their posting date. For each article, we use the Passage Impact Method to rank all the sentences in the linked-to web content  $d^{(i)}$  by

An original post including a quotation	Part of the discussion
<p><b>Simulating Emotions Within Games</b></p> <p>Posted by Soulskill on Thu Jan 29, 2009 08:06 AM from the dreams-of-electric-sheep dept.</p> <p>Gamasutra is running an opinion piece about <a href="#">the way video games handle simulated emotions</a>. Most often, an non-player character's emotional state is used to either tell a story or to drive gameplay. The author suggests that as both concepts become more complex in modern games, the simulation of emotions must also become more dynamic to remain interesting. Quoting:</p> <p>"Most of our emotional simulations use a simple sensation/calculation/behavior loop. Someone says or does something to a character; this influences his emotional state; he acts upon his feelings. His emotional state then reverts to a more neutral state over time (I was angry half an hour ago, but I've calmed down now), or changes again in response to another sensation. If these systems are really simple they produce absurd results: a character is furious one moment and cheerful a second later, like a Warner Brothers cartoon character. This is the kind of thing you get with finite state machines. This approach doesn't take into account the fact that behavior itself changes emotions. Behavior is not merely an output to be exhibited; it also affects how we feel. It feeds back into our emotional state."</p> 	<p>1. Finite state machines will be unrealistically simple when simulating emotional responses. 2. Behavioural-feedback is a necessary condition for realistic emotional displays.</p> <p>Point number 1 is unwarranted. Finite state machines may elaborate their input at an arbitrarily <i>-finite</i> may still be <i>very large</i>. Part of such an elaboration, of course, may be inner transitions be amount to behavioural-feedback. There is nothing intrinsically <i>un-dynamic</i> to FSM.</p> <p><a href="#">7 hidden comments</a></p> <p><b>Re: Don't hurt the feelings of FSMs (Score: 4, Interesting)</b> by <a href="#">Yvanhoe (564977)</a> on Thursday January 29, @10:10AM (#26652669) <a href="#">Journal</a></p> <p>Dwarf Fortress uses ASCII characters to display the actors and their various states. It is. It is not about the graphical feedback, it is about the behavior : once you see some throwing everything around, you know that something is wrong with him. When you see sleeping side by side in the same room, you suspect that something is going on. It is behaviors.</p>

**Fig. 2.** Left: A post that quotes from article  $d^{(i)}$  by the link “the way video games handle simulated emotions.” The label for the original content  $z^{(i)}$  in  $d^{(i)}$  is the quotation text. Right: Part of the discussion to be used as the future document  $d^{(l)}$ .

their likelihood under the model. The previous documents  $d^{(1)} \dots d^{(i-1)}$  in this setting are the web content that have been linked to in earlier discussions. The future content  $d^{(i+1)}$  in this experiment is the user discussion on this posting, except that any direct quotations from the fulltext article have been removed. The user discussion may not contain all the comments, but only those that have been voted up enough to be selected to appear with the posting. We collected seven months (August 2008 to February 2009, inclusive) of articles that satisfy these criteria from the Games subtopic of Slashdot, which netted a corpus of 61 web documents with their associated discussions.

**Evaluation Method** For evaluation, we rank the sentences in the fulltext article in decreasing order of likelihood. The user quotations typically contain no more than a handful of sentences, but often more than one. Thus, this implementation differs from the model where we assume that there is a single original contribution marked in the passage  $Z^{(i)}$ . As a baseline, we compare against a simple heuristic that identifies novelty. In particular, we rank the sentences by a TFIDF score given by the sum of each sentence term’s IDF value. Then, since we have the labels of the true original sentences, we evaluate using the standard metrics of precision and recall at certain points in the ranking. Precision at a point in a ranking is defined to be the number of original sentences at that position in the ranking divided by the total number of sentences up to that point. For a point near the top of the ranking, precision measures whether the sentences that the method most confidently predicts as original are indeed original. Thus we report results for Prec@2. Recall at a point in the ranking is defined to be the number of original sentences at that position in the ranking divided by the total number of original sentences in the document. Recall measures how well the method can find all the original content in the document. Since each labeled quotation typically contains several sentences, we report results for Rec@10.

**Table 5.** Prec@2 and Rec@10 are based on the predicted ranking of sentences by likelihood and TFIDF sum. Original sentences are the ones quoted word-for-word from the article. Results are for  $\pi_n^{(i)} = 0.2$  and  $\pi_n^{(l)} = 0.001$ .

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

**Table 6.** Comparing the PIM with future documents, and PIM as a novelty detection method (without future documents). Results are for  $\pi_n^{(i)} = 0.2$  and  $\pi_n^{(l)} = 0.001$ .

	Prec@2 $\pm$ One Std Err	Rec@10 $\pm$ One Std Err
PIM Impact	22.13 $\pm$ 3.38	36.09 $\pm$ 3.61
PIM Novelty	9.84 $\pm$ 3.03	28.04 $\pm$ 4.24

**Results** The Prec@2 results in Table 5 show that the Passage Impact Model outperforms the TFIDF heuristic baseline for predicting the human-selected sentences at the very top of the ranking. For the task of finding a description consisting of a few good sentences that succinctly describe the original content of a news article, the Passage Impact Model is better than the baseline. The PIM also significantly outperforms the baseline when trying to find most of the original content, as measured by Rec@10.

**Importance of Impact Component** Similar to the experiment with synthetic data, the use of impact substantially improves the performance over pure novelty detection. Table 6 compares the results when using the discussion for detecting impact with the results when no future documents are used. Using the discussion significantly improves the precision of the method.

**Robustness with respect to amount of novel content in  $d^{(i)}$**  During inference, the method needs to assume a mixture weight for the novel content in the non-original text  $\bar{Z}^{(i)}$ . How sensitive is the method to the selection of this parameter? Table 7 shows that the method is robust and provides good results for a wide range of values for  $\pi_n^{(i)}$ .

**Minor Effect of Novel Language Model in Future Documents** Similarly, since Slashdot discussions are somewhat notorious for getting off topic at times, we evaluated whether changing the amount of novel content in the “future document,” i.e., the discussion makes a difference. As it turns out, Table 8 shows that for a wide range of novel content mixing weights  $\pi_n^{(l)}$ , the method is quite robust. The model is able to focus on the portions that the discussion derives from the underlying linked article.

**Table 7.** Prec@2 and Rec@10 for various amounts of assumed novel content  $\pi_n^{(i)}$  in  $d^{(i)}$ . Sentences are marked as original if they appear word-for-word as in the linked article. Results are for  $\pi_n^{(l)} = 0.001$ .

$\pi_n^{(i)}$	Prec@2 $\pm$ One Std Err	Rec@10 $\pm$ One Std Err
0.01	18.85 $\pm$ 3.10	35.51 $\pm$ 3.55
0.05	20.49 $\pm$ 3.15	36.03 $\pm$ 3.57
0.2	22.13 $\pm$ 3.38	36.09 $\pm$ 3.61
0.8	22.95 $\pm$ 3.39	36.45 $\pm$ 3.63
0.9	22.95 $\pm$ 3.39	36.45 $\pm$ 3.63

**Table 8.** Prec@2 and Rec@10 for various mixing weights  $\pi_n^{(l)}$  for the noise model in fitting future documents. Sentences are marked as original if they appear word-for-word as in the linked article. The results are reported for  $\pi_n^{(i)} = 0.2$ .

$\pi_n^{(l)}$	Prec@2 $\pm$ One Std Err	Rec@10 $\pm$ One Std Err
0.0001	20.49 $\pm$ 3.15	36.77 $\pm$ 3.58
0.001	22.13 $\pm$ 3.38	36.09 $\pm$ 3.61
0.01	16.39 $\pm$ 3.42	34.55 $\pm$ 3.73
0.1	18.03 $\pm$ 3.29	30.34 $\pm$ 3.47
0.5	20.49 $\pm$ 3.73	31.04 $\pm$ 3.47

### 4.3 Experiment 3: Evaluation based on Human Judgments

While the Slashdot data provided a reasonable mechanism for inferring ground-truth labels, the most direct evaluation is by explicit human judgment. Therefore, we conducted an experiment with human judges to evaluate the Passage Impact Model on a corpus containing all 1955 papers from the NIPS conference [28] between 1987-2000. In a blind experiment, we asked judges to compare passages extracted by the PIM to those extracted by the TFIDF heuristic regarding how well they summarize the original contribution of a NIPS paper.

**Experiment Setup** Since breaking documents into paragraphs is non-trivial, especially when they are OCR-ed and have many math equations, we arbitrarily defined passages as consecutive blocks of text of length  $L = 100$  (non-stopword) words. On average, there are 14 passages per document.

For inference using the Passage Impact Model, we constrained the novel  $\bar{\theta}^{(i)}$  and original  $\theta^{(i)}$  language models to be equal because research publications typically discuss original contributions at length. Ideally, the identified passage should list the paper’s contributions or conclusions. (Although the abstract has original content, it mostly focuses on placing the paper with the context of existing ideas.) The future document novelty mixing weight of  $\pi_n^{(l)} = 0.01$  is small to force the model to “explain” the content of future documents  $d^{(l)}$  by identifying copied ideas. For efficiency, we used  $k_F = 5$  future documents. We

compare against the TFIDF heuristic baseline. Each paper’s passages predicted by the PIM and the baseline were highlighted, and three judges selected which passage better summarized the paper’s original contribution. The annotators are machine learning graduate students familiar with the corpus and do not include the authors of this paper.

Since the judgment process is time-consuming, we selected a subset of NIPS publications for evaluation. We ranked all NIPS publications by their number of intra-corpus citations and selected the top 50 most-cited documents. The first publication is “Optimal Brain Damage” by Le Cun, Denker, and Solla, with 27 citations. The entire set of 50 documents includes documents down to those with only 5 intra-1987-to-2000 NIPS citations. The PIM and the baseline selected the same passage on two documents, so we use the remaining 48 for evaluation.

**Results** On these 48 documents, the human judges preferred the Passage Impact Method over the baseline 58.33% of the time, with one standard error of 3.54%. Thus the judges significantly prefer the PIM over the baseline. To analyze the results more closely, we separated the 48 evaluation documents into two sets. On 20 documents, all three annotators (independently) agreed on a single passage. For these, they preferred the PIM 70% of the time. On the other 28 documents, two annotators preferred one passage, while the third annotator preferred the other passage. Here, the preferences for PIM and baseline were exactly 50%. This suggests that sometimes identifying a passage that summarizes the original contribution is quite difficult. When this is not the case, however, the PIM outperforms the baseline quite substantially with 70% preference.

## 5 Discussion and Future Work

While the Passage Impact Model provides a generative model of diachronic corpora and the relationships between individual documents, the model is still quite simple. For example, it is based on unigram models of text production. In modeling the probability of  $W^{(i)}$ , one could instead use a more sophisticated sequence model, or at least  $n$ -gram language models. Such information may help to identify coherent original ideas. Another limitation is that the model is constrained to evaluate only a small number of candidate  $Z^{(i)}$  for efficiency reasons. Developing pruning criteria is a promising direction for substantially increasing the scope of  $Z^{(i)}$  in hopes of finding better descriptions of original contributions.

Other information available for some corpora could be integrated into the model as well. For example, if citation information is available, it could provide additional constraints on the parameters during inference. Citations could be used as priors for mixing weights, modeling that documents copy primarily from those documents they cite. This could improve the accuracy of the model, and it could improve efficiency of the optimization since many mixing weights could be fixed at zero.

A more general direction for further work lies in the integration of originality detection with models for idea flow. The goal is to have a unified probabilistic

model that identifies the dependency structure of the corpus, with ideas originating in some documents and then flowing through the corpus. Treating these inference problems separately seems suboptimal.

## 6 Conclusions

We have proposed an unsupervised generative model for diachronic text corpora that provides a formal structure for the process by which authors form new ideas and build on existing ideas. The model captures both novelty and impact, defining an (important) original contribution as a combination of both. For this Passage Impact Model, we have proposed an inference procedure to identify the most original passage of a document. Under reasonable approximations, the inference procedure reduces to multiple convex programs that can be solved efficiently. The method is evaluated on synthetic and real data, and it is shown to significantly outperform a heuristic baseline for selecting a passage describing the original contribution in the domains of online discussions and research articles.

## 7 Acknowledgments

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

## References

1. Mei, Q., Zhai, C.: Generating impact-based summaries for scientific literature. In: Proceedings of the Association for Computational Linguistics (ACL). (2008) 816–824
2. Soboroff, I., Harman, D.: Overview of the TREC 2003 novelty track. In: Proceedings of the Text Retrieval Conference (TREC). (2003)
3. NIST: Document Understanding Conferences (DUC) <http://duc.nist.gov/>.
4. Allan, J., Carbonell, J., Doddington, G., Yamron, J., Yang, Y.: Topic detection and tracking pilot study: Final report. In: Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop. (1998)
5. Allan, J., Papka, R., Lavrenko, V.: On-line new event detection and tracking. In: Proceedings of the SIGIR Conference on Research and Development in Information Retrieval. (1998) 37–45
6. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. *Journal of Machine Learning Research (JMLR)* **3**(5) (2003) 993–1022
7. Blei, D., Griffiths, T., Jordan, M., Tenenbaum, J.: Hierarchical topic models and the nested chinese restaurant process. In: Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS). (2003)
8. Blei, D., Lafferty, J.: Correlated topic models. In: Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS). (2005)
9. Blei, D.M., Lafferty, J.D.: Dynamic topic models. In: Proceedings of the International Conference on Machine Learning (ICML). (2006) 113–120

10. Hofmann, T.: Probabilistic latent semantic analysis. In: Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI). (1999)
11. Mann, G., Mimno, D., McCallum, A.: Bibliometric impact measures leveraging topic analysis. In: Proceedings of the Joint Conference on Digital Libraries (JCDL). (2006)
12. Wang, X., McCallum, A.: Topics over time: A non-markov continuous-time model of topical trends. In: Proceedings of the Conference on Knowledge Discovery and Data Mining (KDD). (2006) 424–433
13. Steyvers, M., Smyth, P., Rosen-Zvi, M., Griffiths, T.: Probabilistic author-topic models for information discovery. In: Proceedings of the Conference on Knowledge Discovery and Data Mining (KDD). (2004) 306–315
14. Griffiths, T., Steyvers, M.: A probabilistic approach to semantic representation. In: Proceedings of the Annual Conference of the Cognitive Science Society. (2002)
15. Dietz, L., Bickel, S., Scheffer, T.: Unsupervised prediction of citation influences. In: Proceedings of the International Conference on Machine Learning (ICML). (2007) 233–240
16. McCallum, A., Corrada-Emanuel, A., Wang, X.: Topic and role discovery in social networks. In: Proceedings of International Joint Conference on Artificial Intelligence (IJCAI). (2005)
17. Mei, Q., Ling, X., Wondra, M., Su, H., Zhai, C.: Topic sentiment mixture: Modeling facets and opinions in weblogs. In: Proceedings of the World Wide Web Conference (WWW). (2007) 171–180
18. Li, W., McCallum, A.: Pachinko allocation: Dag-structured mixture models of topic correlations. In: Proceedings of the International Conference on Machine Learning (ICML). (2006)
19. Wang, X., Li, W., McCallum, A.: A continuous-time model of topic co-occurrence trends. In: AAAI Workshop on Event Detection. (2006)
20. Griffiths, T., Steyvers, M., Blei, D., Tenenbaum, J.: Integrating topics and syntax. In: Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS). (2004)
21. Shaparenko, B., Joachims, T.: Information genealogy: Uncovering the flow of ideas in non-hyperlinked document databases. In: Proceedings of the Conference on Knowledge Discovery and Data Mining (KDD). (2007) 619–628
22. Manning, C.D., Schuetze, H.: Foundations of Statistical Natural Language Processing. MIT Press (1999)
23. Jelinek, F.: Basic Language Modeling. In: Statistical Methods for Speech Recognition. MIT Press (1998) 57–78
24. Zhai, C.: Risk Minimization and Language Modeling in Information Retrieval. PhD thesis, Carnegie Mellon University (2002)
25. Kurland, O., Lee, L.: Corpus structure, language models, and ad hoc information retrieval. In: Proceedings of the SIGIR Conference on Research and Development in Information Retrieval. (2004) 194–201
26. Kurland, O., Lee, L.: Respect my authority! hits without hyperlinks, utilizing cluster-based language models. In: Proceedings of the SIGIR Conference on Research and Development in Information Retrieval. (2006) 83–90
27. MOSEK: <http://www.mosek.com/index.html>.
28. NIPS Online: The Text Repository. <http://nips.djvuzone.org/txt.html>.