Second Order Probabilistic Models for Within-Document Novelty Detection in Academic Articles

Laurence A. F. Park and Simeon Simoff
School of Computing, Engineering and Mathematics
University of Western Sydney, Australia
{l.park,s.simoff}@uws.edu.au

ABSTRACT

It is becoming increasingly difficult to stay aware of the state-of-the-art in any research field due to the exponential increase in the number of academic publications. This problem effects authors and reviewers of submissions to academic journals and conferences, who must be able to identify which portions of an article are novel and which are not. Therefore, having a process to automatically judge the flow of novelty though a document would assist academics in their quest for truth. In this article, we propose the concept of Within Document Novelty Location, a method of identifying locations of novelty and non-novelty within a given document. In this preliminary investigation, we examine if a second order statistical model has any benefit, in terms of accuracy and confidence, over a simpler first order model. Experiments on 928 text sequences taken from three academic articles showed that the second order model provided a significant increase in novelty location accuracy for two of the three documents. There was no significant difference in accuracy for the remaining document, which is likely to be due to the absence of context analysis.

Categories and Subject Descriptors: H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing

Keywords: novelty detection; novelty location; academic articles

1. INTRODUCTION

Academic articles are written to communicate novel research results to the community. We assume that the author of the article is an expert in the field on which the article is written, knowing when research is novel, but we are not able to make the same assumption for the reader. Therefore, for the article to be effective, the author must provide a literature review, describing the state-of-the-art in the field, then describe the research question and present the experiments that lead to answering the question.

It is becoming increasingly difficult to stay aware of the state-of-the-art in any research field due to the exponential increase in the number of academic publications in journals, magazines, conference proceedings, workshops, and scientific blogs [3]. This difficulty to obtain coverage results in large numbers of articles with missing background information being sent to venues for review. This process results in one of two outcomes: 1) the reviewers are aware of the missing references and the article is rejected since its research question has been answered elsewhere, or 2) the reviewers are unaware of the missing references and duplicate redundant work is published. In both cases, the time of the reviewers is wasted, reading articles that need further work.

It is essential for authors and reviewers of academic work to know where novelty in their work lies and for them to be able to communicate this novelty to their readers. A useful tool for academic authors would allow them to identify where novelty lies in a document and also show where novelty is absent and why. Such a tool would increase the confidence of the author in their research, and also provide them with articles that may be useful to include in their review, hence assisting their quest for truth.

In this article, we examine the concept of Within-document Novelty Location for academic articles. Specifically, we investigate the benefit of using a second order generative model over a first order model, for within-document novelty location. The contributions of the article are:

- The definition of a generative first and second order model for within-document novelty location (Section 2).
- An analysis of the benefit of the second order model over the first order model (Section 3).

This article proceeds by describing the first and second order models in Section 2, then provides a set of experiments Section 3 to examine the benefit of the second order model.

2. LOCATING REGIONS OF NOVELTY

The goal of a within-document novelty location system is to automatically identify the flow of novelty throughout a single article; showing which portions are novel and which are not. To construct such a system, we can draw from the related fields of Novelty Detection and Plagiarism Detection.

Novelty Detection can be reduced to the problem of identifying objects that have not been seen before. An overview of statistical methods for detecting novelty in databases is found in Markou and Singh [5]. These statistical methods have been used on the TREC Novelty Track [6], a text retrieval task where participants must not only retrieve relevant information, but the results must not be redundant. Novelty detection has also been applied to whole documents [7], where the document is given a novelty score based on its content.

Plagiarism is the reproduction of work without attribution. Therefore, methods of plagiarism detection based on text and citations [2, 1] can be used to assist in finding non-novelty.

In this article we investigate if there is a benefit in using a second order generative model compared to a first order model. To minimise the complexity of the experiments and the number of variables, we treat each document as a sequence of terms and ignore the meaning of the terms. In this section, we present the first and second order generative models for within-document novelty location and the likelihood ratio with a general language model.

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.
Copyright 2014 ACM 978-1-4503-2257-7/14/07
http://dx.doi.org/10.1145/2600428.2609520.

2.1 A generative model for novelty detection

Context Free Grammars can be used to generate realistic research articles based on a set of rules. The articles are realistic, in that the sequence of words in the articles are grammatically correct, but there is no novelty in the ideas presented, making the articles useless as research articles[1]. The generated articles are formed by randomly choosing rules from the grammar. Given a large number of relations and terminals in a grammar, we can generate a large number of different articles. Therefore, there may be a small chance that an article with academic novelty can be produced using a CFG, but we can safely say that it is unlikely that this will happen.

We make the following proposition: *If it is likely that an article could have been generated by a machine, then it is likely that the article is not novel.* Note that the proposition states the quality of non-novelty, but does not mention the quality of a novel article. Based on this proposition, we can compute the non-novelty of an article by constructing a document model of an existing concept, then computing the likelihood that the article is generated from the document model. Evidence of novelty of an article is provided for each document model that shows not to be non-novel, but if as little as one document model shows that the article is non-novel, then we know the article is not novel. The likelihood of the text $\vec{w}$ being generated from document model $\vec{\theta}_j$ is given as:

$$\begin{aligned} L(\vec{\theta}_j|\vec{\omega} = \vec{w}) &= P(\vec{\omega} = \vec{w}) \\ &= \prod_{i=1}^{N} P(\omega_i = w_i|\vec{\omega}_{i-1..1} = \vec{w}_{i-1..1}) \end{aligned}$$

where $P(\vec{\omega} = \vec{w})$ is the probability of sampling the text sequence $\vec{w}$ from the document model $\vec{\theta}_j$, $\vec{\omega} = [\omega_1, \omega_2, \ldots, \omega_N]$ is a random variable representing a sequence of terms, N is the length of the text sequence $\vec{w}$, $\vec{w}$ is the text sequence we are examining for novelty, and $\vec{w}_{i-1..1}$ is the sequence of words from position 1 to position $i - 1$.

Therefore, to compute the likelihood, we must obtain $P(\omega_i = w_i|\omega_{i-1} = w_{i-1}, \ldots, \omega_1 = w_1)$. Unfortunately, this probability space is very large[2], so we are unlikely to obtain a good estimate of the conditional distribution. Instead, we will examine the document model under two separate assumptions: the independence assumption and the Markov assumption.

2.2 Independence assumption: 1st order model

A first order generative model for non-novelty detection is a categorical distribution of the words in the baseline set of documents. Therefore each term is sampled independently of each other term:

$$P(\omega_i = w_i|\vec{\omega}_{i-1..1} = \vec{w}_{i-1..1}) = P(\omega = w_i)$$

To generate a new document, we sample n words from the categorical distribution using Dirichlet smoothing:

$$P(\omega = w_i|\alpha) = \frac{\sum_d f_{d,t} + \alpha P(\omega_i = w_i)}{\sum_t \sum_d f_{d,t} + \alpha}$$

where $f_{d,t}$ is the frequency of term t in document d, $P(\omega = w_i)$ is background probability of word i, and $\alpha > 0$ is the smoothing parameter. Note that if we assume the background probability distribution over the words to be uniform, we obtain Additive smoothing.

		$L(\vec{\theta}_j|\vec{w})$	
		High	Low
$L(\vec{\theta}_0|\vec{w})$	High	Uncertain	Novel
	Low	Not novel	Uncertain

Table 1: The desired results when determining the novelty state of text from document model $\vec{\theta}_j$ and general language model $\vec{\theta}_0$.

2.3 Markov assumption: 2nd order model

A second order generative model for non-novelty detection is a conditional categorical distribution, where the distribution is conditioned based on the previous sample only. Therefore each term is sampled from a Markov chain:

$$P(\omega_i = w_i|\vec{\omega}_{i-1..1} = \vec{w}_{i-1..1}) = P(\omega = w_i|\omega_{-1} = w_{i-1})$$

where ω is the current word and ω_{-1} is the previous word. To generate a new document, we sample n words from the categorical distribution conditioned on the previous word, using Dirichlet smoothing:

$$P(\omega = w_i|\omega_{-1} = w_{i-1}, \alpha) = \frac{\sum_d f_{d,t_i \leftarrow t_j} + \alpha P(\omega = w_i)}{\sum_t \sum_d f_{d,t \leftarrow t_j} + \alpha}$$

where $f_{d,t_i \leftarrow t_j}$ is the frequency of term t_i following term t_j in document d, $P(\omega_i = w_i)$ is background probability of word i, and $\alpha > 0$ is the smoothing parameter.

2.4 Likelihood ratio

If a document model shows that it is likely, given a word sequence, it does not imply that that the word sequence is not novel. It may be that the text sequence uses language that is common but not specific to the text's concept. Therefore, to investigate if a piece of text is non-novel, we will examine the likelihood ratio of a document model θ_j to the general language model θ_0.

The general language model θ_0 is modelled on the language, regardless of concepts. Using the general language model, we can detect uncertainty in novelty. For example, even though the sequence "The experimental results showed our method provides high accuracy" is a high probability sequence for a given document model (implying that it is not novel), it does not provide evidence that the text is not novel, since it is a common phrase in academic articles. On the other hand, if we find a text sequence that is novel with respect to the document model and the general language model, it may be that the text is written in a unique style and is in fact not novel, making the novelty status uncertain. Therefore, the general language model becomes our Null model and we predict novelty using Table 1. To achieve these outcomes, we use the log likelihood ratio:

$$u(\vec{w}|\vec{\theta}_j) = \log\left(\frac{L(\theta_j|\tilde{w})}{L(\theta_0|\tilde{w})}\right) \tag{1}$$

Equation 1 provides us with the non-novelty score of the text $\vec{w}$ with respect to the document model $\vec{\theta}_j$ and the language model $\vec{\theta}_0$. A concept is novel if it cannot be found elsewhere, therefore a piece of text $\vec{w}$ contains a novel concept if its likelihood ratio is low when compared to all document models. Therefore, to determine if $\vec{w}$ is novel with respect to a set of document models, we use the function:

$$U(\vec{w}) = \max_{\vec{\theta}_j \in \Theta} \log\left(\frac{L(\theta_j|\tilde{w})}{L(\theta_0|\tilde{w})}\right) \tag{2}$$

where Θ is the set of document models and $U(\vec{w})$ is the non-novelty score for $\vec{w}$. A positive score for $U(\vec{w})$ implies that $\vec{w}$

[1] See `http://pdos.csail.mit.edu/scigen/`

[2] If we have 100,000 words and $\vec{\omega}$ is a sequence of 100 words, then $\vec{\omega}$ is a categorical distribution with $100^{100,000}$ elements.

is not novel; a zero score implies uncertainty and a negative score implies novelty with respect to Θ.

3. EXPERIMENTS

Our experiments are designed to determine if there is any benefit in using a second order model, when compared to a first order model, for within document novelty location in academic articles.

3.1 Method

To perform our experiment, we require: 1) a set of articles $\mathbb{N}$ to extract the text sequences $\vec{w}$ and examine for within-document novelty, 2) a set of articles $\mathbb{M}$ to construct the document models $\vec{\theta}_j$, and 3) a set of articles $\mathbb{G}$ to compute the general language model $\vec{\theta}_0$. The set of word sequences $\vec{w}$ were obtained by selecting one document d from $\mathbb{N}$, then extracting each 100 word sequence, separated by 50 words (so that each consecutive sequence overlapped half of the previous and next sequence). The length of the sequence was chosen as an acceptable length of text to contain one concept.

The set $\mathbb{M}$ was chosen as the set documents given in the references from each document in $\mathbb{N}$. The set $\mathbb{G}$ should be chosen to contain a good coverage of the known concepts, hence not focusing on a specific concept and producing a general language model. We chose the set $\mathbb{G}$ to be equal to $\mathbb{M}$, where all elements of the set $\mathbb{G}$ are used to compute θ_0. Our set $\mathbb{N}$ contained a Ph.D. thesis, a journal article and a conference article, all with known sections of non-novelty. The stop words were removed from each document and all words were stemmed using Porter's stemmer. Finally, we chose $\alpha = 0.001$ as the value that provided the greatest sampling variance from the set of document models.

3.2 Results

The non-novel likelihood ratio of the thesis, journal article and conference article are presented in Figures 1, 2 and 3. Each of these plots show the non-novelty likelihood ratio $U(\vec{w})$ for text spanning from the beginning to the end of the document. Each of the chapters/sections of the articles are shown using solid vertical lines.

We will first examine the results in Figure 1. The thesis contained 1) the front matter, 2) introduction, 3) literature review, 4-8) research chapters, 9) conclusion, and 10) bibliography and appendix. Of these, the concepts from chapters 4, 5 and 7 were published elsewhere, making them not novel to the thesis.

From examining the plot, both first and second order models seem equivalent for the first two chapters. They both show that non-novelty exists in the 4th, 5th and 7th chapters, but the second order model shows that there are portions of these chapters where novelty exists. On further examination, we found that there is additional information in chapters 5 and 7 that is not shown in the published articles, making these parts novel. We find in chapters 3, 6 and 8 that the first order model provides scores about 0, implying uncertainty, while the second order model provides negative scores implying novelty. Chapters 6 and 8 contained unpublished work and so are definitely novel, while chapter 3 contains the literature review. The literature review can be seen as non-novel since it is describing existing work, but it can also be seen as novel if these descriptions are provided in a unique way.

It seems that the second order model was able to detect novel and non-novel regions of the thesis, while the first order model was only able to detect the non-novel regions. The second order model provided more confident results (with larger magnitude).

The second document we will analyse is a journal article where a similar article was previously published at a conference by the same authors; the major difference being a new section on efficiency. We can see the non-novelty scores in Figure 2 containing nine sections, where the seventh section is the new addition to the article. We again see that the first order model has less confidence than the second order model (producing scores closer to zero). Both order models show that there is novelty in the seventh section, but the second order model drops in confidence as the section progresses. When examining the document model and null model score separately, we find that both the document model and null model drop in score for the seventh section, meaning that there is uncertainty due to the language not being seen before.

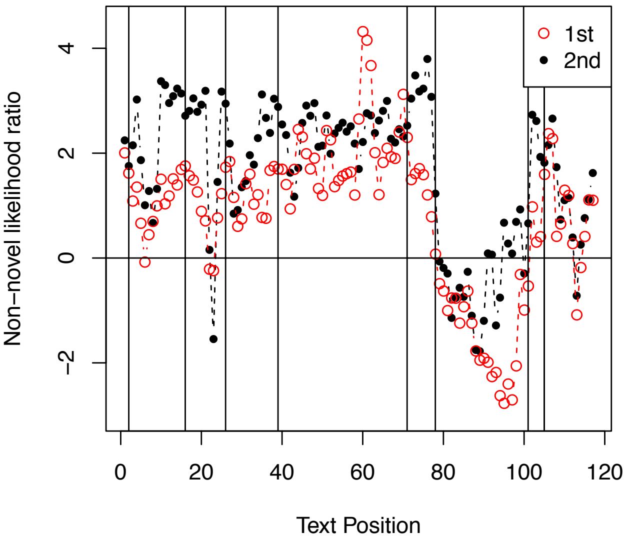

Figure 2: The non-novelty likelihood ratio $U(\vec{w})$ for a chosen journal article, where the beginning of each section has been marked with a vertical bar.

The last document we will examine is a conference article with scores shown in Figure 3. This article has 10 sections, where the 1st, 2nd, 6th, 7th, and 9th sections are novel. We can see that both the first and second order models provided similar results, not clearly showing the novelty of sections 6, 7 and 9. This article contained tables of text results that were given in other articles; the novelty was in the way the tables were ordered, which could not be determined by the first and second order models.

We examined the accuracy of the first and second order models by labelling each chapter/section of the three articles as either novel or not novel, based on our knowledge of prior work. We then computed the probability of each text sequence being novel by applying the logistic function to each non-novelty score. The probability p_i of a method providing the correct decision is given as $p_i = l_i^{(1-n)}(1-l_i)^n$ where l_i is the result of applying the logistic function to the non-novelty score and n is 1 if the text sequence is manually judged novel and 0 otherwise.

To determine the benefit of using a second order model, we computed the difference distribution of the probabilities p_i for the second order model minus the first order model. The number of text sequences (#Seq.) and the sample mean and standard deviation of this difference distribution are reported in Table 2, along with the p value from the testing if the mean of the difference distribution is greater than zero (implying that the second order model increases the novelty prediction accuracy).

We see from Table 2 that if using the second order model, we are expected to obtain a 7.7% increase in within-document novelty location on the thesis, 3.4% increase on the journal article and a 2.8% drop in accuracy on the conference article. We also found that the increase in novelty location for the thesis and journal article is significant, but there is not enough evidence to show a difference in the first and second order model means on the conference article.

Note that location of novelty was difficult in the conference article since tables of results were presented within the article, contain-

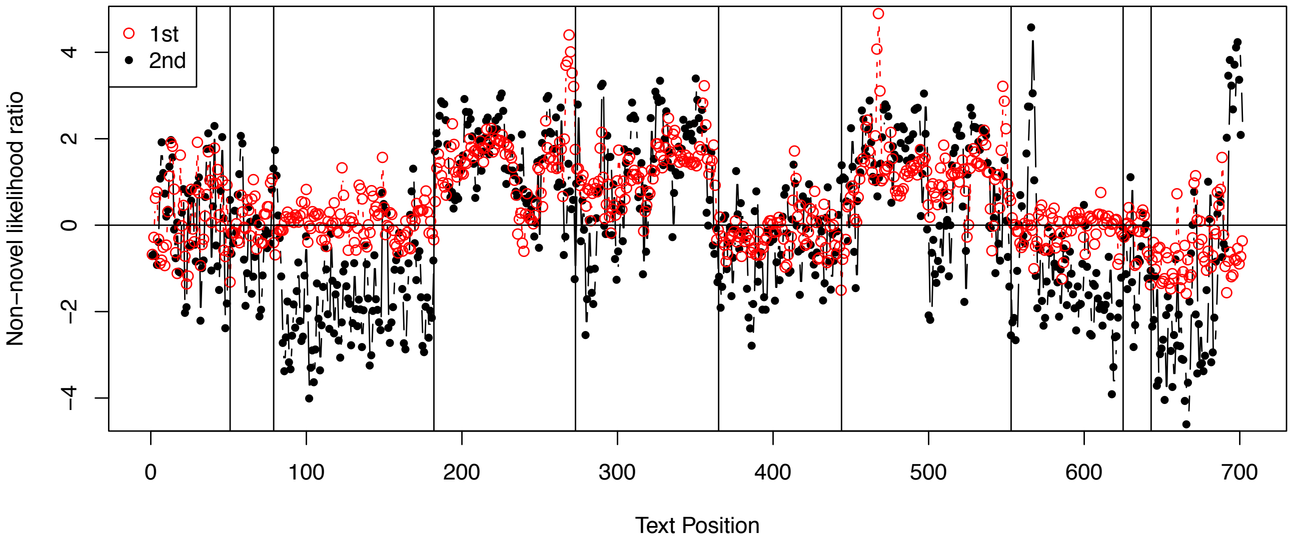

Figure 1: The non-novelty likelihood ratio $U(\vec{w})$ for a chosen thesis, where the beginning of each chapter has been marked with a vertical bar.

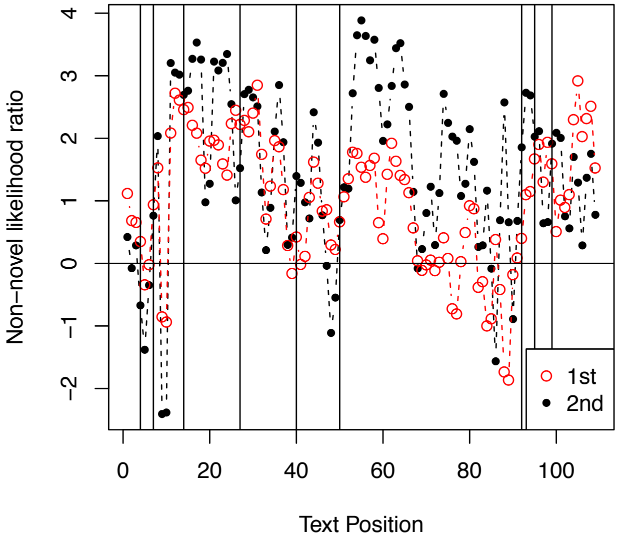

Figure 3: The non-novelty likelihood ratio $U(\vec{w})$ for a chosen conference article, where the beginning of each section has been marked with a vertical bar.

ing names that were reordered for each experiment. These tables were presented in a previous article, with a different ordering, making these results novel, but difficult to determine since the first and second order methods only examine the text. If the context of the tables were taken into account (e.g. by using methods from Li and Croft [4]), we should obtain a more thorough analysis.

4. CONCLUSION

Novelty detection is the process of automatically determining the novelty of a text sequence, based on a set of known text sequences, and has been thoroughly examined for document retrieval. Within-document novelty location is the classification of the novelty of text throughout a document, allowing us to locate novel and non-novel sections of the document.

Type	#Seq.	Mean	SD	p-value
Thesis	702	0.077	0.260	7.677×10^{-13}
Journal	117	0.034	0.172	8.631×10^{-5}
Conference	109	-0.028	0.206	0.7769

Table 2: Increase in probability of correctly classifying novelty due to using a second order, rather than a first order, model.

In this article we examined if there is a benefit in accuracy when using a second order probabilistic model, compared to a first order probabilistic model. We examined 928 text sequences from three documents and found that the first and second order models gave similar predictions of novelty location, but the second order predictions were more confident. By converting the confidence values to probabilities, we found that using the second order model provided a significant increase in novelty location for two of the three documents. There was no difference for one of the documents, which we discovered was due to lack of text context. Therefore, our evidence suggests that there may be benefit in using a second order model, but there are no drawbacks.

References

[1] J.-P. Bao and J. A. Malcolm. Text similarity in academic conference papers. In *2nd International Plagiarism Conference Proceedings*. Northumbria University Press, 2006.

[2] B. Gipp, N. Meuschke, C. Breitinger, M. Lipinski, and A. Nürnberger. Demonstration of citation pattern analysis for plagiarism detection. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 1119–1120. ACM, 2013.

[3] P. O. Larsen and M. von Ins. The rate of growth in scientific publication and the decline in coverage provided by science citation index. *Scientometrics*, 84(3): 575–603, 2010.

[4] X. Li and W. B. Croft. An information-pattern-based approach to novelty detection. *Inf. Process. Manage.*, 44(3):1159–1188, May 2008. ISSN 0306-4573. doi: 10.1016/j.ipm.2007.09.013.

[5] M. Markou and S. Singh. Novelty detection: a review-part 1: statistical approaches. *Signal processing*, 83(12):2481–2497, 2003.

[6] I. Soboroff and D. Harman. Novelty detection: The trec experience. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 105–112, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics. doi: 10.3115/1220575.1220589.

[7] F. S. Tsai and Y. Zhang. D2s: Document-to-sentence framework for novelty detection. *Knowledge and information systems*, 29(2):419–433, 2011.