Entropy-Based Authorship Search in Large Document Collections

Ying Zhao and Justin Zobel

School of Computer Science and Information Technology, RMIT University
GPO Box 2476V, Melbourne, Australia
{yizhao,jz}@cs.rmit.edu.au

Abstract. The purpose of authorship search is to identify documents written by a particular author in large document collections. Standard search engines match documents to queries based on topic, and are not applicable to authorship search. In this paper we propose an approach to authorship search based on information theory. We propose relative entropy of style markers for ranking, inspired by the language models used in information retrieval. Our experiments on collections of newswire texts show that, with simple style markers and sufficient training data, documents by a particular author can be accurately found from within large collections. Although effectiveness does degrade as collection size is increased, with even 500,000 documents nearly half of the top-ranked documents are correct matches. We have also found that the authorship search approach can be used for authorship attribution, and is much more scalable than state-of-art approaches in terms of the collection size and the number of candidate authors.

1 Introduction

The purpose of authorship search (AS) is to find within a large collection the documents that appear to have been written by a given author. That is, given documents of known authorship, the task is to find other documents by the same author. AS has not previously been widely investigated, but is related to authorship attribution (AA), the task of identifying the authorship of unknown documents given a corpus of known authorship. AA and AS are valuable in applications such as plagiarism detection, literary analysis, and forensics. However, none of the AA approaches has been scaled to large document collections. For example, for the Federalist Papers, the authorship of 65 documents is explored [17]. Holmes et al. used 17 journal articles [13], while Koppel et al. used 21 English books [19]. Diederich et al. used a collection with seven authors and around 100 texts for each author [8]. Hoover's collection had 50 documents [14]. We previously used 4900 documents, so far the largest AA collection [28,30].

Our method for AS is motivated by information retrieval (IR) techniques, where matching is determined by computing the similarity between queries and documents, but there are significant differences. A key difference is choice of index terms; IR techniques make use of content-bearing words, while in AS it is necessary to identify style markers. We explore use of function words and part-of-speech (POS) tags. Another

G. Amati, C. Carpineto, and G. Romano (Eds.): ECIR 2007, LNCS 4425, pp. 381–392, 2007.

potential difference is choice of similarity measure. We propose relative entropy as the similarity measure for AS, inspired by the language models used in IR.

For data, we use collections of 10,700 to 500,700 newswire articles from the TREC collections. Our results show that, with sufficiently large queries, matching documents can be found from within large collections. While results are mixed, with even the largest document collection precision in the top 10 is 44.2%. The best results were achieved with stop words; use of style markers was less effective, while standard similarity measures were, compared to relative entropy, nearly useless. We also investigated the applicability of the proposed AS approach to AA, finding that, with a large volume of text by an unattributed author, authorship could be identified with reasonable reliability, greatly improving on previous methods. With less training text and on larger collections, the accuracy of attribution was not as good. Overall, however, we have demonstrated for the first time the feasibility of AA and AS on large document collections. As we have since shown for a small collection, the method is also highly effective for literature [29].

2 Document Search

IR systems are used to search for documents that satisfy users' information needs [2]. Current IR systems usually deal with large and heterogeneous collections and typically take as input queries of a few words, returning as a response a list of documents deemed most likely to be relevant. Search involves two stages, index term extraction and similarity computation. For documents in English, indexing involves separating the text into words, case-folding, stopping, and stemming [32].

Various models have been proposed as bases for measurement of similarity between documents and queries. One is the vector-space model [2], where items are represented as vectors. The assumption is that similar documents should be separated by a relatively small angle. Plausibly a similar assumption would apply to AS: given appropriate style markers, the distribution of style markers in the documents by an author should be similar. An alternative are models used to derive estimates for the probability that a document is relevant to a query. The BM25 model is one of the most successful probabilistic models in IR [24]. Whether such a model is suitable for AS is, intuitively, not clear, but given the success of BM25 in IR it is reasonable to consider its use for AS.

Language models, also based on probability theory, were originally motivated by tasks such as speech recognition. These models are used to estimate the probability distributions of words or word sequences. In IR, language models are used to estimate the likelihood that a given document and query could have been generated by the same model [7]. Given a document d and a model Θ_d inferred from d, language models estimate the probability that model Θ_d could have generated the query q. Smoothing techniques are applied to assign probabilities for missing terms [6,12,27].

Although language models have elements that are counter-intuitive (suggesting, for example, that queries comprised of common words are more likely than queries comprised of words that are specific to a given document), they are a high effective approach to IR. In this paper a key contribution is exploration of whether language models are suitable for AS.

3 Authorship Attribution

The purpose of AA is to identify documents that are written by a particular author. A range of AA methods have been proposed in recent research. Despite the differences amongst all these approaches, the framework of an AA technique involves two stages in general: extraction of document representations and making attribution decisions.

Document representations are comprised of style makers. Both lexical and grammatical style markers have been proposed for AA. A simple approach is to use lexical markers, that is, function words and punctuation symbols [8,13,15]. More sophisticated syntactic and grammatical components can be extracted by natural language processing (NLP) techniques, such as part-of-speech (POS) tags [16,25]. However, these more advanced style markers do not necessarily produce better performance for AA [30]. A particular issue is that the idiosyncratic grammatical patterns that are particular to an author may not be identified due to the lack of observations of such patterns in the process of training the parser. That is, NLP is not only error prone, but is likely to make errors on the most significant elements of the data.

In the attribution stage a variety of classification methods have been investigated. Principle component analysis (PCA) has been used in several approaches to AA [5,13]. Hoover [14] examined the scalability of PCA to large corpora or multi-class AA, in which the number of author candidates is greater than 2, finding only 25% accuracy given 50 samples by a total of 27 authors, suggesting that PCA would not scale to large numbers of authors.

Machine learning approaches such as support vector machines (SVMs) [8,18] are considered to be competitive alternatives for AA. SVMs are effective when provided with sufficient samples and features. However SVMs are not always superior to other methods when given small number of samples for training, which is often the case in AA. Computational cost is another issue of SVMs. Bayesian networks are less effective and more computationally expensive than SVMs [11,23].

Language models have also been proposed for AA. Benedetto et al. used a compression-based language model [4], based on a standard compression suite; however, Goodman [9] was unable to reproduce the result, and the method is not plausible. A Markov model has also been applied to AA by Khmelev et al. [17], in which the features are individual characters. Good accuracy was achieved on data collected from the Gutenberg project, but the accuracy may be overestimated, due to the duplicate texts provided by Gutenberg. For example the number of distinct texts by Burroughs is only 9, but Khmelev et al. included 25 of his works in their experiments.

Most of these AA methods are not directly applicable to search tasks. In a search system a query is evaluated by ranking the similarities measured between the query and each document individually in the collection. The result is a list of top-ranked documents. In contrast to search, there is no ranking required for AA; instead, an explicit decision is made for each unknown document individually. Documents are required for training to learn a model for a particular author in AA. There is no document-by-document calculation involved. AA techniques have not been applied to search problems. In this paper we propose what we believe is the first AS mechanism.

4 Relative Entropy for Authorship Search

In AA and AS, the underlying assumption is that there are patterns or characteristics of an author's writing that can be automatically extracted and then used to distinguish their work from that of others. Given appropriate style markers, the distribution with which they are observed should be similar in all of the author's documents, regardless of topic. Distributions can be compared via their entropy, and we therefore propose use of the Kullback-Leibler divergence (KLD, or relative entropy) as a similarity measure for AA. The distributions need to be estimated, and we propose use of the language models that have been successfully applied in IR [6,20,27]. We used a similar approach for AA [30], but it was not clear that such an approach could be used for AS.

Entropy measures the uncertainty of a random variable X, where, in this application, each $x \in X$ could be a token such as a word or other lexical feature, and $p(x)$ is the probability mass function of X. The KLD quantifies the dissimilarity between two distributions. In the context of AS, we can build entropy models for the queries and the documents. The differences between query models and document models can be measured by relative entropy as:

$$KLD(d\|q) = \sum_{x \in X} p_d(x) \log_2 \frac{p_d(x)}{p_q(x)} \tag{1}$$

The divergence is calculated between the query and every document in the collection. The documents whose entropy has the lowest divergence from the query are the most likely to share authorship and thus should be the highest ranked. However, if $p(x)$ is zero for some symbol the divergence is undefined. To address this issue, we use Dirichlet smoothing to estimate probabilities [27]:

$$\hat{p}_d(x) = \frac{f_{x,d}}{\mu + |d|} + \frac{\mu}{\mu + |d|} p_B(x) \tag{2}$$

Here x are the style markers used for document representations and $f_{x,d}$ is the frequency of token x in document d. The notation $|d| = \sum_{x \in d} f_{x,d}$ represents the number of token occurrences in d, and $p_B(x)$ is the probability of the token x in the *background model*, which provides statistics on the tokens. The parameter μ controls the mixture of the document model and the background model. The background probabilities dominate for short documents, in which the evidence for the in-document probabilities is weak; when the document is longer then the influence of the background model is reduced. In principle the background model could be any source of typical statistics for token occurrences.

Additionally, a set of style markers is required. Some researchers have found that function words are effective [1,8,13,15,18], while use of part-of-speech tags has also been considered [30]. We make use of both kinds of marker, but, in agreement with our earlier work [30], find that function words are superior.

5 Experiments

As data, we use collections of documents extracted from the TREC corpus [10]. There are large numbers of documents included in TREC that can be used to evaluate the proposed search technique, as the author is identified. We believe that this data presents a difficult challenge for AA or AS, as, compared to novelists or poets, journalists do not necessarily have a strong authorial style, and the work may have been edited to make it consistent with a publication standard.

We develop three collections of documents to evaluate the proposed AS system, which consist of 10,700, 100,700, and 500,700 documents respectively. We call these the 10k-collection, 100k-collection, and 500k-collection. The documents in the 10k-collection and 100k-collection are from the AP subcollection of TREC; the 500k-collection consists of documents from AP, WSJ, and SJM. Metadata, including author names is discarded. As authors, we select the seven[1] that we earlier used for AA [28,30]. These authors are regular contributors to AP and each of them has over 800 documents in the TREC corpus. We randomly select 100 documents of each author and include them as part of each of the collections, giving in total the extra 700 documents in each case. All queries and documents are pre-processed to obtain the style markers that are applied to the system; query construction is discussed below. The background models of different types of style markers used in all experiments are derived from the AP collection of over 250,000 documents; an alternative would have been to use the collection as the background model in each case, but we decided to hold the background model constant across all experiments.

We evaluate our proposed authorship search system from several perspectives. Scalability is examined by considering effectiveness on collections of different sizes. We run the experiments with different kinds of style marker. The differences between KLD-based search and other retrieval techniques are tested. Finally we explore use of the AS approach as an AA method.

Feasibility and scale in size. In the first experiment we examine whether AS is feasible for small and large collections. The first seven queries used in this experiment are generated by concatenating 500 randomly selected documents written by each of the seven authors. These documents are distinct from the 100 documents that are included as part of each collection. We call these the 500-document queries. The style markers are function words. The next seven queries are formed by concatenating the 100 documents that are included in the collection; we call these the 100-included queries.

The numbers of correct matches in the top-100 ranked documents are in Table 1, for the 10k-collection. Amongst the 500-document queries, those based on the documents of Currier and Dishneau are the most effective, while the query based on the documents of Beamish is much less effective. The 100-included queries are slightly better than 500-document queries in most cases, despite being based on less text, and are highly consistent with the 500-document queries, suggesting that the style of some authors is easier to identify than that of others.

Overall precision and recall is plotted in Figure 1, based on the 500-document queries on all three collections. We achieve average $p@10$ (precision at 10 documents retrieved) of 84.2% on the 10k-collection, 74.2% on the 100k-collection, and 30.0% on the 500k-collection. Thus, while the density of correct matches falls from 1% to 0.02%,

[1] The authors are Barry Schweid, Chet Currier, Dave Skidmore, David Dishneau, Don Kendall, Martin Crutsinger, and Rita Beamish.

Table 1. The number of correct matches in the top 100 documents in response to each query, on the 10k-collection

	Number of correct answers in top 100						
	Schweid	Currier	Skidmore	Dishneau	Kendall	Crutsinger	Beamish
500-document	48	61	35	61	44	52	30
100-included	59	58	49	61	46	56	37

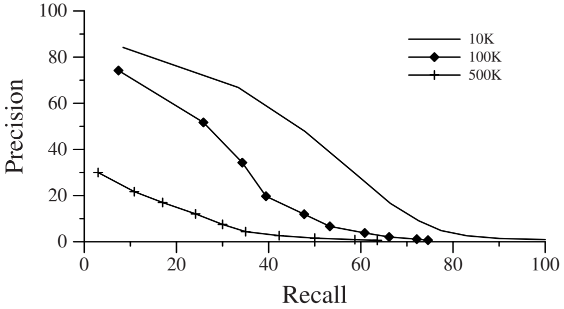

Fig. 1. Precision versus recall for 500-document queries on each of the three collections: 10k, 100k, and 500k

effectiveness drops more slowly. Achievement of high recall is much more difficult with the largest collection, but the results show that, with a large query, AS is indeed feasible on even half a million documents.

Another dimension of scale is the volume of training data available. In the experiments above we had a large volume of text per author. With less text, effectiveness may decline. For each author we constructed 5 100-document queries and 25 20-document queries; average results are shown in Figure 2. It can be seen that reducing the amount of training data does indeed reduce effectiveness. For low levels of recall, queries of 100 documents (whether 100-included or queries comprised of another 100 documents) lead to reasonable effectiveness; indeed, whether or not the documents are included has a surprisingly low effect on the results, demonstrating that style as measured by function words must be moderately consistent within the work of an author. However, queries of 20 documents are much less effective. While reasonable numbers of correct documents are still found in the top 10 to 50 answers, subsequent results are poor.

Style markers. In text categorization, documents are usually indexed or represented by topic words occurred in the documents [3,21,22,26]. However, in AA whether topic words are appropriate style markers is controversial; some researchers have used them, but most have not. In this experiment we contrasted use of function words and topic words for AA, using the 10k-collection. Results are shown in Figure 3. In this figure, the uppermost curve uses the 500-document queries and is the same as in Figure 1; the dashed line is the comparable results for queries of topic-words; and the solid line is based on topic words and the 100-included queries. As can be seen, AS with topic

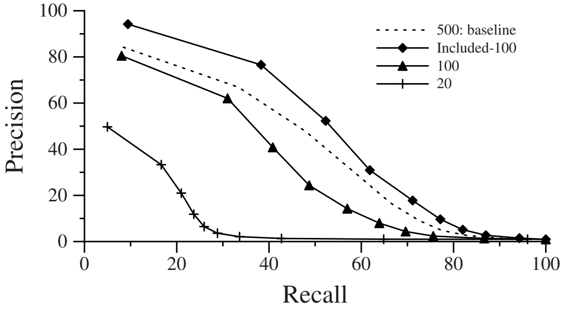

Fig. 2. Effectiveness for queries composed of 20–500 documents, on the 10k-collection

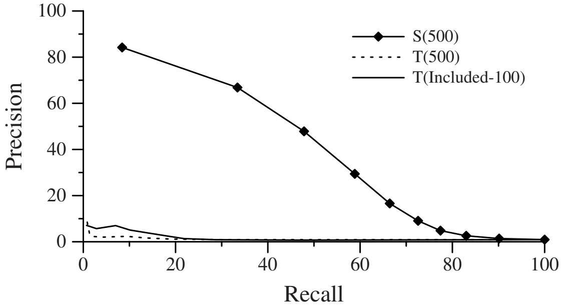

Fig. 3. Comparison of using different indexing methods: function words versus topic words on 10k-collection

words completely failed for authorship search; results are little better than random. The results show that the topic words are misleading in characterizing authors' writing style.

Other kinds of style marker are more plausible. For the next experiment, we used NLTK (a natural language toolKit)[2] has been applied to extract part-of-speech (POS) tags from documents. That is, in this approach, each document is represented by a stream of POS tags. We used a list of 183 POS tags, then indexed documents with function words, POS tags, and both combined. Results are shown in Figure 4.

Function words consistently lead to greater effectiveness than POS tags, which is consistent with our previous work in AA [30]. The combination of function words and POS tags leads to even greater effectiveness. With the smallest 10k-collection, function words are almost as good as the combined features, and both of them achieve the same $p@10$ of 84.2%. However, with larger collections the advantage of combination increases. On the 500k collection, function words achieve 30.0% $p@10$; addition of POS tags increases this to 44.2%. These results show that, even though POS tags by themselves do not yield good effectiveness, they are helpful additional evidence of style.

[2] Available from `http://nltk.sourceforge.net/index.html`

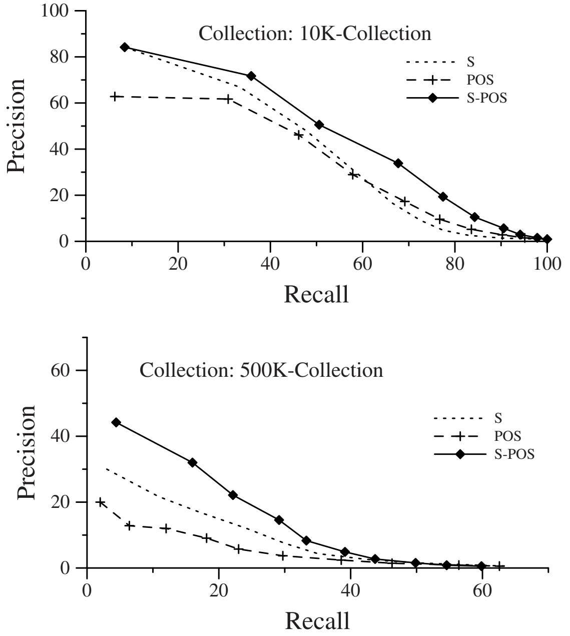

Fig. 4. Effectiveness of different style markers on the 10k (upper) and 500k (lower) collections, using the 500-included queries

KLD ranking versus other measures. In this experiment we compare similarity measures. In addition to KLD we used three measures that have been successfully used in IR, including BM25 and the vector-space measures BB-BCI-BCA and BB-ACB-BCA [31,32].

Results are in Figure 5. The IR similarity measures are surprisingly poor — none has proved suitable for AS. The BM25 measure is slightly better than the other two vector space models but none is usable. The reason why these measures are ineffective for AS is unclear and needs further investigation.

Applicability to authorship attribution. In this experiment we examine whether our AS approach can be used for AA. Instead of returning a list of documents that are judged likely to have the same authorship as to the query, an explicit authorship is returned corresponding to the query.

The proposed AA approach is as follows. We have a query for which authorship is unknown. Using search, a list of l top-ranked documents is returned. These are of known authorship, with k distinct authors and for each author a a count f_a of the number of documents by a in the list; thus $l = \sum_a f_a$.

A simple way to attribute authorship is to select a with the largest f_a. More strictly, a threshold t where $0 \leq t \leq 1$ can be selected so that the query can be assigned to a particular author a if $a = argmax_a(f_a)$ and $f_a/l > t$. Increasing t should reduce the likelihood of incorrect attribution.

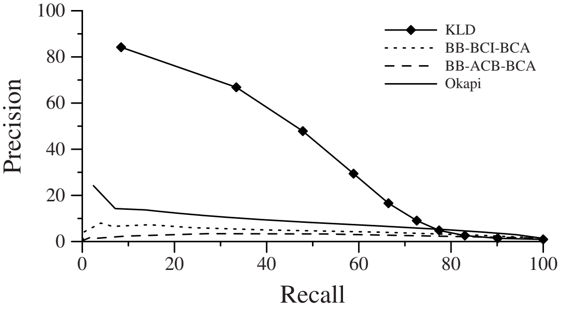

Fig. 5. Effectiveness of different similarity measures on 10k-collection, using the 500-document queries

To test these methods we built two collections from the AP data. The 10k-vote collection includes 10,000 documents from 342 authors, and the 100k-vote collection consists of 100,000 documents by 2229 authors. In both collections, 100 documents of each of the seven test authors are included. Overall the number of texts per author varies from 1 to 835. In both collections more than 10% of the distinct authors have written over 100 documents each. All documents in 10k-vote have identified authorship, while in the 100k-vote collection more than 90% of the texts have identified authorship. As style markers we use the combination of function words and POS tags.

Results from previous experiments show that it is feasible to search for documents written by the same author as that of the query, given a group of documents of known authorship as the query. In this experiment the authorship of the query is unknown and is to be identified. In this experiment, 500-document queries are unreasonably large. We experimented with queries that are formed from individual documents and from 10-document sets; none of the query documents are in the collections.

Results are shown in the Table 2, using the threshold $t = 0$ so that attribution is made to the authorship of the biggest f_a. Evaluation is based on the top l ranked documents, for l from 10 to 100. As can be seen, queries can be effectively attributed using the 10k-vote collection using only the top 10 documents retrieved; with both 1-document and 10-document queries, increasing l is not helpful.

With 1-document queries, the overall correctness of attribution is 51.0%. Previous methods achieve this accuracy only on small collections. Greater attribution effectiveness is achieved with 10-document queries, giving overall 74.3% correct attribution.

Table 2. Voting results for authorship attribution, showing the number of queries (1-document and 10-document queries) correctly attributed, on the 10k-vote collection, in the top 10, 20, 40, 60, 80, and 100 answers retrieved. There were 700 1-document queries and 70 10-document queries.

		Number of answers retrieved					
Queries	N_q	10	20	40	60	80	100
1-doc	700	357	343	334	346	335	337
10-doc	70	52	55	58	56	55	56

Table 3. Voting-based AA results for each author; for each author there are 100 1-document queries and 10 10-document queries on 10k-vote and 20 1-document queries and 5 10-document queries on 100k-vote. On the 100k-vote collection, for some authors only negligible numbers of correct documents were found; these are shown as *negl.*

		Number correctly attributed / Average correct in top 10						
Collection		Schweid	Currier	Skidmore	Dishneau	Kendall	Crutsinger	Beamish
10k-vote	Q_{1-doc}/100	39/3.2	69/9.2	36/4.4	76/9.8	58/4.8	54/5.5	25/2.7
	Q_{10-doc}/10	8/3.6	10/8.0	1/2.0	10/10.0	10/7.4	10/6.3	3/3.0
100k-vote	Q_{1-doc}/20	*negl.*	14/4.8	*negl.*	15/5.2	8/2.9	*negl.*	*negl.*
	Q_{10-doc}/5	*negl.*	3/7.0	*negl.*	5/7.4	5/4.4	*negl.*	*negl.*

There has been no previous attempt at multi-class AA with more than a few authors. Both the number of authors and the size of the collection are much more substantial than in all previous AA work.

We have observed strong inconsistencies amongst queries based on the work of different authors. Results extracted from top-10 lists are shown in Table 3. As can be observed, queries using documents by Currier and Dishneau are more effective than other queries, not only in accuracy of AA but also in confidence. This observation is consistent with results from previous search experiments.

The confidence is indicated by the average number of correct documents in the top-k ranked list. For instance, on the 10k-vote collection, the 100 1-document queries of Dishneau can be correctly attributed at 76% accuracy, providing around 98% confidence. Note that, unsurprisingly, the effectiveness of attribution for the 10-document queries is generally better than for the 1-document queries.

We also tested the proposed method on the 100k-vote collection, which has over 2000 known authors. This experiment is much less successful, with near-zero accuracy in four of the seven cases. Interestingly, these failures correspond to the results of lower confidence on the 10k-vote collection. For queries based on documents by Currier and Dishneau, the attribution accuracies are respectively 70% and 75%, suggesting 48% and 52% confidence. Again, use of 10-document queries leads to greater effectiveness. However, it can be seen that AA on large collections with large numbers of authors remains a challenge.

6 Conclusion

We have explored the novel task of authorship search. Our proposal is that simple entropy-based statistics and characterization of documents by distributions of style markers can be used to find documents by an author, given some training documents by that author.

Our experiments show that such a method can be highly successful for collections of moderate size. The proposed similarity measure, the Kullback-Leibler divergence, which is used to compute relative entropy, is far more effective than standard measures drawn from information retrieval. As style markers, both function words and part-of-speech tags are effective; for large collections, combined use of both kinds of marker led to even better results. Reasonable effectiveness can be achieved on collections of even half a million documents.

To our knowledge our approach is the first that is able to search a large collection for documents written by a particular author. The success of the method is highlighted by the fact that we have used experimental data, newswire articles, that we regard as challenging for this task: in contrast to material drawn from sources such as literature, we would not expect human readers to be aware of strong stylistic differences between the authors.

The proposed search approach can also be applied to author attribution. Previous methods struggle to correctly attribute authorship when given more than a few hundred documents or more than a few authors. Our method has reasonable accuracy with 10,000 documents and several hundred authors. While it did not successfully scale further in our experiments, this approach is nonetheless much more effective than previous methods and is a clear demonstration that authorship attribution can be applied on realistic collections.

References

1. H. Baayen, H. V. Halteren, A. Neijt, and F. Tweedie. An experiment in authorship attribution. *6th JADT*, 2002.
2. R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman, May 1999.
3. R. Bekkerman, R. El-Yaniv, N. Tishby, and Y. Winter. Distributional word clusters vs. words for text categorization. *J. Mach. Learn. Res.*, 3:1183–1208, 2003.
4. D. Benedetto, E. Caglioti, and V. Loreto. Language trees and zipping. *The American Physical Society*, 88(4), 2002.
5. J. N. G. Binongo. Who wrote the 15th book of Oz? an application of multivariate statistics to authorship attribution. *Computational Linguistics*, 16(2):9–17, 2003.
6. S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. In A. Joshi and M. Palmer, editors, *Proc. 34th Annual Meeting of the Association for Computational Linguistics*, pages 310–318. Morgan Kaufmanns, 1996.
7. W. B. Croft and J. Lafferty. *Language Modeling for Information Retrieval*. Kluwer Academic Publishers, Norwell, MA, USA, 2003.
8. J. Diederich, J. Kindermann, E. Leopold, and G. Paass. Authorship attribution with support vector machines. *Applied Intelligence*, 19(1-2):109–123, 2003.
9. J. Goodman. Extended comment on language trees and zipping, 2002.
10. D. Harman. Overview of the second text retrieval conf. (TREC-2). *Information Processing & Management*, 31(3):271–289, 1995.
11. D. Heckerman, D. Geiger, and D. Chickering. Learning bayesian networks: the combination of knowledge and statistical data. *Machine Learning*, 20:197–243, 1995.
12. D. Hiemstra. Term-specific smoothing for the language modeling approach to information retrieval: the importance of a query term. In *Proc. 25th ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 35–41. ACM Press, 2002.
13. D. I. Holmes, M. Robertson, and R. Paez. Stephen Crane and the New York Tribune: A case study in traditional and non-traditional authorship attribution. *Computers and the Humanities*, 35(3):315–331, 2001.
14. D. L. Hoover. Statistical stylistics and authorship attribution: an empirical investigation. *Literary and Linguistic Computing*, 16:421–444, 2001.

15. P. Juola and H. Baayen. A controlled-corpus experiment in authorship identification by cross-entropy. *Literary and Linguistic Computing*, 2003.
16. A. Kaster, S. Siersdorfer, and G. Weikum. Combining text and linguistic doument representations for authorship attribution. In *SIGIR workshop: Stylistic Analysis of Text For Information Access*, August 2005.
17. D. V. Khmelev and F. Tweedie. Using markov chains for identification of writers. *Literary and Linguistic Computing*, 16(4):229–307, 2002.
18. M. Koppel and J. Schler. Exploiting stylistic idiosyncrasies for authorship attribution. In *Exploiting Stylistic Idiosyncrasies for Authorship Attribution. In IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, 2003.
19. M. Koppel and J. Schler. Authorship verification as a one-class classification problem. In *Proc. 21st Int. Conf. on Machine Learning*. ACM Press, 2004.
20. O. Kurland and L. Lee. Corpus structure, language models, and ad hoc information retrieval. In *Proc. 27th ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 194–201. ACM Press, 2004.
21. Y. S. Lai and C. H. Wu. Meaningful term extraction and discriminative term selection in text categorization via unknown-word methodology. *ACM Transactions on Asian Language Information Processing*, 1(1):34–64, 2002.
22. D. D. Lewis, Y. M. Yang, T. G. Rose, and F. Li. Rcv1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res.*, 5:361–397, 2004.
23. B. Scholkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond.* MIT Press, 2002.
24. K. Spark Jones, S. Walker, and S. E. Robertson. A probabilistic model of information retrieval: development and comparative experiments. *Inf. Process. Manage*, 36(6):779–840, 2000.
25. E. Stamatatos, N. Fakotakis, and G. Kokkinakis. Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, 35(2):193–214, 2001.
26. Y. M. Yang. A study on thresholding strategies for text categorization. In *Proc. 24th ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 137–145. ACM Press, 2001.
27. C. X. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Transaction on Information System*, 22(2):179–214, 2004.
28. Y. Zhao and J. Zobel. Effective authorship attribution using function word. In *Proc. 2nd AIRS Asian Information Retrieval Symposium*, pages 174–190. Springer, 2005.
29. Y. Zhao and J. Zobel. Search with style: authorship attribution in classic literature. In *Proc. 30th ACSC Thirtieth Australasian Computer Science Conference*, page to appear. ACM Press, 2007.
30. Y. Zhao, J. Zobel, and P. Vines. Using relative entropy for authorship attribution. In *Proc. 3rd AIRS Asian Information Retrieval Symposium*, pages 92–105. Springer, 2006.
31. J. Zobel and A. Moffat. Exploring the similarity space. *ACM SIGIR Forum*, 32(1):18–34, Spring 1998.
32. J. Zobel and A. Moffat. Inverted files for text search engines. *ACM Computing Surveys*, 38:1–56, 2006.