Sequential and Unsupervised Document Authorial Clustering Based on Hidden Markov Model

Khaled Aldebei[†*], Helia Farhood[†], Wenjing Jia[†], Priyadarsi Nanda[†] and Xiangjian He[†]
[†]Global Big Data Technology Centre, University of Technology Sydney, Australia.
[*]Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, Minjiang University, Fuzhou, Fujian, 350121, China.

***Abstract*—Document clustering groups documents of certain similar characteristics in one cluster. Document clustering has shown advantages on organization, retrieval, navigation and summarization of a huge amount of text documents on Internet. This paper presents a novel, unsupervised approach for clustering single-author documents into groups based on authorship. The key novelty is that we propose to extract contextual correlations to depict the writing style hidden among sentences of each document for clustering the documents. For this purpose, we build an Hidden Markov Model (HMM) for representing the relations of sequential sentences, and a two-level, unsupervised framework is constructed. Our proposed approach is evaluated on four benchmark datasets, widely used for document authorship analysis. A scientific paper is also used to demonstrate the performance of the approach on clustering short segments of a text into authorial components. Experimental results show that the proposed approach outperforms the state-of-the-art approaches.**

I. Introduction

Document clustering groups documents of certain similar characteristics in one cluster. It has received more and more attention recently due to the advantages of document clustering in the applications for organization, retrieval, navigation and summarization of a huge amount of text documents available on the Web [1]. According to literature, document clustering can be achieved based on different characteristics, where a certain characteristic is chosen to fulfil human needs in information retrieval and understanding. Many of the approaches in literature have considered topic-based document clustering, such as the works of [2], [3] and [4]. For authorship-based document clustering, there are very little work reported. In the work of [5], the authors presented an unsupervised approach for document clustering. However, since this approach deals with documents in Hebrew language only and requires the concordance between synonyms, it can only be applied to specific types of documents such as Bible books written in Hebrew. In the work of [5], the authors have considered another highly relevant authorship-based problem, named "multi-author document segmentation", where the sentences of a document written by multiple authors are segmented into components based on their authorship. The authors of [6] and [7] have investigated the limitation in the approach of [5] and presented a generic unsupervised method for multi-author document segmentation . In [8], the authors presented an unsupervised approach for the same problem, which utilized the difference of the posterior probabilities of a Naive-Bayesian Model in order to improve the performance of the approach. Another approach has been presented in [9] for multi-author document segmentation. In this approach, a simple HMM was constructed in order to segment the sentences of a multi-author document into authorial components. Recently, the authors of [10] proposed an unsupervised approach for clustering documents according to authorship. They employed a spectral clustering and random forest technique to cluster a group of documents. All of the aforementioned approaches, except for the approaches of [5] and [10], have been designed mainly for multi-author document segmentation. In this paper, we consider the problem of clustering a group of single-author documents. This problem is very similar to the problem of multi-author document segmentation, as the authors of [10] have also emphasized. The reason is because, in most cases of authorship segmentation, there is a very little amount of texts for which it can be affirmed that only a single author presents. Therefore, this amount of texts can be considered as a document written mainly by only one author.

Numerous approaches have been reported in literature for document clustering. Most of these approaches have simply applied general data clustering techniques, where a vector of features is firstly generated from a document and then some clustering algorithm is employed to cluster the resulted vectors of different documents into different components. In this paper, in order to cluster documents of various topics and various lengths, we propose an approach that is based on capturing writing styles of authors from each sentence, instead of each document.

Typically, document clustering is achieved by applying some classical clustering models, such as K-means [11], Gaussian Mixture Model (GMM) [12] and similarity measurements (e.g., cosine similarity, KL divergence, generalized I-divergence, etc.) [13], [14], [15]. The main assumption made with regard to these models is that the data are independently and identically distributed (iid), i.e., the correlation coefficients among the data are null. In this paper, instead of assuming that the data are iid, we propose a novel idea to utilize the sequence of data for clustering, i.e., the contextual information hidden among sentences are used in order to group documents based on their authorship.

The contributions of this paper are highlighted as follows:

1) We propose a new authorship-based approach for clustering documents into groups according to authorship by capturing writing styles of authors based on sentences, rather than documents. For this purpose, we propose a two-level model.
2) We propose to depict authors' writing styles by extracting the sequential correlations among sentences in order to cluster documents into groups of authorial components. For this purpose, we construct HMMs and propose this two-level, unsupervised framework.
3) A sentence-majority document clustering procedure is developed to cluster a group of documents into authorial components using sentence labels. It clusters a document into the author who has written most of the sentences of that document.

When tested on benchmark datasets, our approach has demonstrated superior performance over the state-of-the-arts. Our proposed approach is not restricted to any type of documents and it is effectively applicable even when the topics among documents are not distinguishable. When tested on a scientific paper to cluster short sections, our approach has also achieved very promising results, showing its independence on the length of a text.

The rest of the paper is organized as follows. Section II presents the framework of our proposed approach. This is followed by Sections III and IV describing the two levels of learning respectively. Experimental results are given in Section V. Finally, Section VI concludes the paper.

II. Framework of Our Proposed Approach

Precisely, the problem we are interested in can be formulated as follows. Given n documents written by l authors, $n \geq l$, it is assumed that each document is completely written by only one of the l authors. It is also assumed that there is no information about the documents and the authors available other than the number of authors, l. Our objective is to cluster the n documents into l authorial components.

In our work, we propose to address this problem by utilizing the sequential correlations hidden among sentences and develop an unsupervised, sequential approach for document clustering. A Hidden Markov Model (HMM) is constructed to model the relations between the authorships and the sentences of documents. The results obtained from the HMM are then used to cluster the n documents into l authorial components.

Our approach has two main levels, *First Level Learning* and *Second Level Learning*. Each level of learning goes through a series of steps, as follows.

- First Level Learning.
  - Estimating initial values of HMM parameters using chunks of sentences of documents.
  - Learning HMM parameters using an algorithm called $Baum - Welch$ algorithm.
  - Performing an initial sentence decoding process using an algorithm called $Viterbi$ algorithm.
- Second Level Learning
  - Creating a new training dataset of sentences using a procedure called "Consecutive-Sentence Dataset".
  - Re-estimating the initial values of HMM parameters using the newly created training dataset.
  - Estimating new learning HMM parameters using the $Baum - Welch$ algorithm.
  - Performing a final sentence decoding process using the $Viterbi$ algorithm.
  - Clustering n documents of l authors into l authorial components using a procedure called "Sentence-Majority Document Clustering".

In the following two sections, we give the details of our approach based on the above two levels of learning.

III. First Level Learning

In our proposed approach, we construct a HMM to depict the authorial writing styles hidden among sentences for authorial document clustering. In this section, we first review the concepts of a HMM. Then, we illustrate each step included in the first level learning.

A. Hidden Markov Model

HMM [16] is considered as a very efficient statistical method for characterizing the relations between the observed data arranged in series, called "observations", and the hidden variables, called "hidden states". Let us consider the V observations as $T = \{t_1, t_2, \cdots, t_V\}$ and the hidden states as $Q = \{q_1, q_2, \cdots, q_V\}$, where the q_i is the hidden state of the i^{th} observation (i.e., t_i). Each observation takes one value from a set of observation values $W = \{w_1, w_2, \cdots, w_m\}$ and each hidden state also takes one value from a set of state values $S = \{s_1, s_2, \cdots, s_l\}$. In this case, m and l represent the number of distinct observations and the number of distinct states in the model, respectively. The graphical structure of HMM is shown in Figure 1.

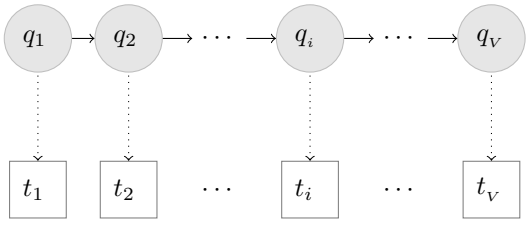

Fig. 1. A graphical structure of HMM with V hidden states $(q_1, \ldots, q_V)$ and V observations $(t_1, \ldots, t_V)$.

Formally, HMM is defined with three parameters as follows.

1) **A** is a set of transition probabilities where $\mathbf{A}_{jk}$ is the probability of making a transition from state j to state k, i.e., $\mathbf{A}_{jk} = p(q_i = s_k | q_{i-1} = s_j)$.
2) **B** is a set of emission probabilities which show the conditional probabilities of observations given certain

states. Each conditional probability is given by $\mathbf{B}_e(k) = p(t_i = w_k | q_i = s_e)$, where $w_k \in W, s_e \in S$.

3) $\boldsymbol{\pi}$ is the initial state probabilities, where $\pi(i) = p\ (q_1 = s_i)$.

For simplicity, we denote the three HMM parameters as $\boldsymbol{\lambda} = \{\mathbf{A}, \mathbf{B}, \boldsymbol{\pi}\}$.

The proposed approach is based on exploiting the sequential correlations hidden among sentences in order to capture the authorships of sentences in documents. Therefore, we employ the HMM in order to model the correlations, where the sentences represent observations and the authorships of sentences represent hidden states. In our model, we assume that there are V observations (i.e., sentences) with V distinct values (i.e., $W = \{w_1, w_2, \cdots, w_V\}$) and V hidden states with l distinct values (i.e., $S = \{1, 2, \cdots, l\}$). Note that the number of distinct values of hidden states is equal to the number of authorial components in which the n documents should be clustered. The objective of this model is to find the most probable sequence of authors, Q, for a given sequence of sentences, T, so we can cluster the n documents into l authorial components.

B. Estimating Initial HMM Parameters

Usually, in HMM, we learn the model by maximising the likelihood function of HMM in order to estimate the best values of HMM parameters (i.e., $\boldsymbol{\lambda}$) so that the probability of observations is maximised. In order to apply that, initial values of $\boldsymbol{\lambda}$ should be assigned to start learning the model. In our approach we estimate the initial values of HMM parameters, $\boldsymbol{\lambda}$, as follows.

We first randomly group all the n documents of l authors in one document. Assume that the resulted document contains V sentences. The document is then segmented into a group of consecutive segments, of which each contains v successive sentences from the document. The value of v is estimated according to the number of sentences in the document. We set 30 and 10 to v for a long document (containing more than or equal to 500 sentences) and a short document (containing fewer than 500 sentences), respectively. Each resulted segment is then vectorized using a feature set containing all words that have occurred three or more times in the document. Assume that the feature set is denoted by $R = \{word_1, word_2, \cdots, word_{R_1}\}$, where R_1 is the number of features (i.e., words) in the set. We employ a binary vector, of which each dimension on the vector represents whether an individual feature does or does not occur in the segment. With the resulted binary vectors, we cluster them into l components. Gaussian Mixture Models (GMMs) [17] are employed in order to cluster the segment feature vectors into l multivariate Gaussian densities. Based on the results of the clustering process of the segment vectors, each segment vector is given a label according to the Gaussian component that the vector is assigned to during the clustering process. The label, which can be considered as a state, of a vector takes one value from a set of l elements (i.e., $1, 2, \cdots, l$). Then, with the labels of segment vectors, which form a sequence of states, the transition probabilities (i.e., $\mathbf{A}$), the emission probabilities (i.e., $\mathbf{B}$) and the initial state probabilities (i.e., $\boldsymbol{\pi}$) are estimated as follows.

Given the sequence of segment vectors with its states (i.e., labels), the estimation of $\mathbf{A}$ is simply performed by finding the probability of moving each state toward the others in the segment sequence. The values of $\boldsymbol{\pi}$ are estimated by computing the fraction of each state in the segment sequence. The estimation of $\mathbf{B}$, which represents the emission probabilities of sentences, is done as follows. First, the conditional probability of each feature in the feature set R, given a label (i.e., $p(r|s)$ where r is a feature from the set R and s is a state from the set S), is estimated using the labeled segment sequence. Second, each sentence in the resulted grouped document is represented as a binary vector using the feature set R. Finally, the emission probability of each sentence (i.e,, observation) given a state is computed using the Bernoulli distribution [18] as shown in Eq. 1.

$$p(t_i|s) = \prod_{r=1}^{R_1} p(r|s)^{t_{ir}} (1 - p(r|s))^{1 - t_{ir}}, \tag{1}$$

where t_i, $i = \{1, 2, \cdots, V\}$, represents an observation, $s \in \{1, 2, \cdots, l\}$ represents a state, r represents a feature, t_{ir} represents the value of feature r in observation t_i, and R_1 is the number of features.

In the next subsection, the initial, estimated values of $\boldsymbol{\lambda}$ will be used for learning the HMM parameters to obtain the best estimations of $\boldsymbol{\lambda}$. The best estimations of $\boldsymbol{\lambda}$ will then be used to find the most probable sequence of authors for the given sequence of sentences.

C. Learning HMM Parameters and Initial Sentence Decoding

The initial values of $\boldsymbol{\lambda}$ are now used to start learning the HMM parameters by maximizing the likelihood function of HMM. The learning process is proceeded by using an algorithm, often called the Baum-Welch algorithm [19], which is basically a derived form of the Expectation-Maximization (EM) algorithm for HMM. The initial values of the HMM parameters are assigned first, then more accurate parameters are computed in each iteration until the algorithm converges.

The newly learned HMM parameters, $\boldsymbol{\lambda}$, are now used in order to find the best sequence of authors that represents the corresponding sequence of V sentences of the document, as shown in Eq. 2. This is done by employing the Viterbi algorithm [20], which can efficiently determine the most likely sequence of states for the given sequence of observations.

$$Q^* = \arg\max_{Q}\ p(Q|T). \tag{2}$$

IV. Second Level Learning

The results obtained from the first level learning (i.e., the sequence of authors for the given sequence of V sentences of the document), are used in the second level learning that aims to enhance the performance of the decoding process by

providing a more accurate training dataset. The second level learning, as shown in Section II, includes five steps discussed in the following three subsections.

A. Creating Consecutive-Sentence Dataset

In the first level learning, as shown in the previous section, the best sequence of authors for the corresponding sequence of sentences is determined based on the HMM, of which initial parameter values (i.e., $\boldsymbol{\lambda}$) are estimated based on clustering groups (i.e., segments) of consecutive sentences.

In order to enhance the performance of the HMM, we propose a procedure, called "Consecutive-Sentence Dataset", to produce a more powerful training dataset with more accurate labels. The training dataset can then be used to re-estimate the initial values of HMM for more accurate values used again to learn a more powerful HMM to enhance the sentence decoding process. The procedure is as follows.

Given the most probable sequence of authors for the sequence of sentences of the document, which is obtained in the decoding process of the first level learning, each sequence of minimum five successive sentences that have the same label is inserted into a new training dataset with that label.

Note that, the new training dataset contains a sequence of sentences, rather than a sequence of segments as in the first level training.

The new dataset is used to re-estimate the initial values of HMM (i.e., $\boldsymbol{\lambda} = \{\mathbf{A}, \mathbf{B}, \boldsymbol{\pi}\}$), as shown in the next subsection.

B. Re-Estimating and Learning HMM Parameters

The estimation process of initial values of $\mathbf{A}$ and $\boldsymbol{\pi}$ is the same with the estimation process done in the first level learning (see Section III-B). The only difference is that instead of having a sequence of segment vectors with its labels, now we have a sequence of sentence vectors with its labels.

In order to estimate the initial values of $\mathbf{B}$ using sentences, rather than segments, a new feature set is used to vectorize the sentences. The new feature set contains all words that have occurred in the document. Assume that the new feature set is denoted by $R' = \{word_1, word_2, \cdots, word_{R_2}\}$, where R_2 is the number of features in the set. The same binary representation, which is used for representing the feature set R, is also utilized for representing the feature set R'. After that, the same computations as those performed in estimating the initial values of $\mathbf{B}$ in the first level learning are applied to compute the initial values of $\mathbf{B}$. The only difference is that we replace the sequence of segments of R_1 features by the sequence of sentences of R_2 features.

The new estimated values of HMM parameters (i.e., $\boldsymbol{\lambda}$) are now used to learn the HMM again. The same learning process, which is used in the first level learning, is used to learn the HMM.

C. Final Sentence Decoding Process and Document Clustering

After learning the HMM parameters using the new initial values, we use these learned parameters to find the most probable sequence of authors for the corresponding sequence of V sentences of the document. The same decoding process as that explained in the first level learning is used to apply the decoding process on this level.

Now, the best sequence of authors for the sequence of sentences of the document is obtainable, i.e., the author of each sentence in the document can now be known. We propose a procedure, called "Sentence-Majority Document Clustering", to make use of the label of each sentence in the document and cluster the n documents into l authorial components. The procedure works as follows.

The document, which is created by grouping the n documents of l authors in first level learning, is divided back into original n separate documents. Then, we assign each document to the author who has most of the sentences of that document. If there is no majority for any author in a certain document, then we keep that document unclustered. Note that, in all of our experiments to be presented in the next section, no such case has been found.

V. Experiments

In this section, the performance of our approach is presented and compared against state-of-the-art approaches using four benchmark datasets widely used for authorship document analysis. We use these datasets because the author of each document is well recognized. Furthermore, a scientific document is employed in order to test our approach on clustering short segments (i.e., sections in this document) of a text.

A. Datasets

The first corpus consists of chapters of five biblical books written by five authors. The authors are Jeremiah (52 chapters), Ezekiel (48 chapters), Isaiah (35 chapters), Proverbs (31 chapters) and Job (39 chapters). The chapters are written in Hebrew language and related to two literatures. The chapters of the first three authors (i.e., Jeremiah, Ezekiel and Isaiah) are related to the prophetic literature and the chapters of the last two authors (i.e., Proverbs and Job) are related to the wisdom literature. This dataset offers an opportunity to evaluate our approach in non-English documents and in documents of the same literature.

The second corpus is the uncompleted novel $Sanditon$. The novel was launched by Jane Austen, who wrote 11 chapters, but interrupted by her death. Several years later, the novel had been completed by "an Other Lady", who had written other 19 chapters. She carefully emulated Austen's writing style and used her notes to complete the novel. A lot of studies have been done on $Sanditon$, but most of them have fallen out of style [10].

In the third corpus, we apply our approach on datasets containing articles of *New York Times*. The articles were written by four columnists and cover a variety of topics. The columnists are Gail Collins (274 articles), Paul Krugman (332 articles), Maureen Dowd (298 articles) and Thomas Friedman (279 articles). Using this corpus gives us a way to verify our

approach on clustering articles when their topics of authors are not differentiated.

The fourth corpus consists of 690 blogs written by the Nobel Prize-winning Gary Becker and the legal scholar Richard Posner. Becker has written 346 blogs and Posner has written 344 blogs. Through these blogs, both authors (i.e., Becker and Posner) presented their ideas and opinions on different topics. This corpus is considered as an important one because it covers a variety of topics and makes a process of clustering documents according to authorships, rather than topics, be more challenging.

Furthermore, in order to show that the proposed approach is able to cluster the short segments of a text, we test our approach on the sections of a very early draft of a scientific paper. The paper has been written by two Ph.D students (Students A and B). Each student has contributed to the paper by writing three sections. In order to apply our approach on this paper, we have deleted all the figures as well as all metadata (e.g., titles, author affiliations, references and citations).

B. Results

We evaluate the performance of our approach through a set of experiments on different groups of documents. For each experiment, we perform our approach for 10 trials and the average result over the 10 trials is presented. In each trial, the experimental result is measured by using purity [21].

Even that the approaches of [6], [7], [8] and [9] have addressed a slightly different problem of segmenting a multi-author document, and acknowledging the variation between the two problems, we compare the performance of our approach with these approaches. Furthermore, we compare the performance of the proposed approach with the approaches of [5] and [10], which exactly handle the same problem of ours.

1) Results on the Bible Books Dataset: In our first set of experiments, we use a dataset containing chapters of five biblical books written by five authors and related to two literatures. In order to evaluate our proposed approach using this dataset, we use all the chapters of any two authors in order to cluster them into two authorial components. The chapters of two authors are related to the same literature or different literatures.

Table I presents the purity results obtained by applying our approach on chapters of the same literature or different literatures. As shown in Table I, the purity results of our approach are compared against the approaches of [5], [7], [7]-SynonymSet, [8] and [9].

From the purity results presented in Table I, we can see that the results achieved using our proposed approach are very promising with a purity equal to 100.0% obtained on most experiments. We can also observe that the overall purities of the proposed approach are better than those obtained using the other five approaches. These also outperform the overall purity of 78.0% acquired by [6] when chapters of the same literature are clustered. The authors of [10] have performed their document clustering approach using the biblical chapters of only one pair of authors. In that approach, the chapters of Ezekiel and Jeremiah (i.e., 100 chapters) have been used in order to cluster them into two clusters (i.e., Ezekiel cluster and Jeremiah cluster). The approach achieves a result of 99.0%, i.e., 99 chapters are clustered into correct clusters and one chapter is clustered into a wrong cluster. Looking at Table I, one can notice that our approach is able to cluster all the 100 chapters of Ezekiel and Jeremiah into the correct clusters.

	Chapters	1	2	3	4	5	6
Different	Eze-Prov	76.6%	98.7%	90.8%	97.9%	98.8	100.0%
	Jer-Prov	72.7%	97.0%	75.0%	99.0%	99.5%	100.0%
	Jer-Job	87.3%	98.0%	93.1%	97.8%	98.5%	100.0%
	Isa-Job	82.2%	98.7%	89.1%	98.7%	99.4%	100.0%
	Eze-Job	85.9%	98.7%	95.0%	99.0%	99.4%	100.0%
	Isa-Prov	70.4%	95.0%	85.0%	97.9%	98.7%	98.5%
	Overall	***79.2%***	***97.7%***	***88.0%***	***98.4%***	***99.1%***	***99.8%***
Same	Jer-Eze	82.0%	96.6%	95.9%	97.0%	97.3%	100.0%
	Isa-Eze	78.9%	80.0%	88.0%	82.7%	83.2%	88.0%
	Job-Prov	84.5%	93.9%	82.0%	95.2%	98.2%	100.0%
	Isa-Jer	71.8%	66.7%	82.7%	71.0%	72.1%	81.6%
	Overall	***79.3%***	***84.3%***	***87.2%***	***86.5%***	***87.7%***	***92.4%***

TABLE I
PURITY RESULTS OF CLUSTERING BIBLICAL CHAPTERS OF DIFFERENT LITERATURES AND SAME LITERATURE USING THE APPROACHES OF 1- [5], 2- [7], 3- [7]-SYNONYMSET, 4- [8], 5- [9] AND 6- OUR APPROACH.

2) Results on Sanditon: For the second set of experiments, we apply our approach on the $Sanditon$ novel. The novel had been written by Jane Austen (11 chapters) and an unknown lady (19 chapters). Our proposed approach distinguishes Austen's chapters from Another Unknown Lady's chapters with 100.0% purity results. However, in [10], a purity result of only 93.8% is presented. Therefore, the proposed approach has produced a great result.

3) Results on New York Times Dataset: In these experiments, we use the *New York Times* articles of four columnists. We apply the proposed approach using articles of any pair of the four columnists in order to cluster them into two authorial components. This produces six sets of experiments. Table II shows the purity results of clustering the articles of any two columnists using the proposed approach in the six experiments. Table II show the excellent purity results of our approach. The proposed approach outperforms the ones shown in [8] and [9] in all six experiments.

Our approach further proves its superiority when compared against the results of 88.0% to 97.0% shown in [6] and [7], and the results of 90.0% to 98.8% with an average of 94.5% shown in [10].

4) Results on Becker-Posner Blogs Dataset: For this set of experiments, we utilize 690 blogs written by two authors, Gary

Articles	[8]	[9]	Our Approach
MD-PK	95.5%	96.3%	99.1%
MD-TF	93.3%	93.9%	97.1%
MD-GC	93.8%	93.9%	98.8%
TF-PK	95.6%	95.2%	98.2%
GC-PK	93.7%	94.1%	98.2%
GC-TF	96.1%	94.9%	98.0%
Overall	***94.7%***	***94.7%***	***98.2%***

TABLE II
PURITY RESULTS OF CLUSTERING ARTICLES OF ANY PAIR OF THE FOUR *New York Times* COLUMNISTS (GC = GAIL COLLINS, PK = PAUL KRUGMAN, TF = THOMAS FRIEDMAN, MD = MAUREEN DOWD) USING OUR APPROACH AND THE APPROACHES OF [8] AND [9].

Authors	Cluster 1	Cluster 2
Becker Blogs	**342**	4
Posner Blogs	3	**341**

TABLE III
RESULTS WHEN CLUSTERING THE 690 BLOGS WRITTEN BY GARY BECKER AND RICHARD POSNER INTO TWO CLUSTERS..

Becker (346 blogs) and Richard Posner (344 blogs). These blogs cover a variety of different topics, and some of these topics are discussed by both authors. Therefore, the topics are hard to be distinguished according to the authorship. Table III presents the results when clustering the 690 blogs written by Gary Becker and Richard Posner into two clusters.

As shown in Table III, our approach is able to correctly cluster 683 blogs from the 690 blogs, mislabelling only seven blogs.

Table IV presents the purity results of clustering Becker's blogs from Posner's blogs using the proposed approach. This table also shows the results obtained from the approaches in [6], [7], [8] and [9].

From Table IV, it is clear that the purity result of clustering the 690 blogs of Becker-Posner blogs using our approach is also great and surpasses those obtained from the other four approaches.

5) Results on a Scientific Paper: In order to show that our approach is workable on clustering short segments of texts, we employ a scientific paper written by two students (Students A and B). Each student has written three sections. We use our approach in order to cluster the six sections into two clusters (i.e., Cluster A and Cluster B) according to the authorship. Our approach clusters the sections of Student A from the sections of Student B with 100.0% purity result. Therefore, the result obtained from clustering sections of the scientific paper is perfect. As a comparison, the approach shown in [9] achieves only a result of 93.0% when tested on this same document.

Blogs	1	2	3	4	5
Becker-Posner Blogs	94.0%	94.9%	96.6%	96.7%	99.0%

TABLE IV
PURITY RESULTS OF CLUSTERING THE 690 BECKER-POSNER BLOGS INTO TWO CLUSTERS USING THE APPROACHES OF 1- [6], 2- [7], 3- [8], 4- [9] AND 5- OUR APPROACH.

Student A			
Section No.	No. of Sentences	No. of Correctly Clustered Sentences	Percentage of Correctly Clustered Sentences
1	11	11	100.0%
2	56	56	100.0%
3	64	60	93.8%
Student B			
Section No.	No. of Sentences	No. of Correctly Clustered Sentences	Percentage of Correctly Clustered Sentences
4	140	140	100.0%
5	34	24	82.4%
6	8	5	62.5%

TABLE V
THE NUMBER OF SENTENCES OF EACH SECTION OF STUDENTS A AND B WITH THE NUMBERS AND PERCENTAGES OF CORRECTLY CLUSTERED SENTENCES.

In our work, we use sentences of documents in order to group documents based on authorship, i.e., the labels of sentences are used in order to group the documents. In fact, it would be interesting to know the numbers and percentages of sentences that are correctly clustered in each section of the scientific paper. Tables V lists the numbers of sentences in each section of Students A and B. Furthermore, Tables V shows the numbers and percentages of correctly clustered sentences in each section of both students.

VI. CONCLUSION

In this paper, we presented an unsupervised approach for clustering documents into authorial components. We have proposed to utilize the contextual correlations hidden among sentences and developed a two-level learning procedure in order to group documents. The proposed approach has been evaluated using four benchmark datasets widely used in authorship document analysis. A scientific paper has also been tested to verify that the proposed approach can cluster short segments of a text (i.e., sections in a scientific paper) with promising results. The experiment results have shown that the proposed approach has produced excellent results on all of these datasets and have exceeded the state-of-the-art approaches.

ACKNOWLEDGEMENT

This project is partly supported by an Open Fund Project of Fujian Provincial Key Laboratory of Information Processing and Intelligent Control (Minjiang University) (No. MJUKF201701).

REFERENCES

[1] W. Xu, X. Liu, and Y. Gong, "Document clustering based on non-negative matrix factorization," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM, 2003, pp. 267–273.

[2] M. Franz, T. Ward, J. S. McCarley, and W.-J. Zhu, "Unsupervised and supervised clustering for topic tracking," in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2001, pp. 310–317.

[3] A. Tagarelli and G. Karypis, "A segment-based approach to clustering multi-topic documents," *Knowledge and information systems*, vol. 34, no. 3, pp. 563–595, 2013.

[4] C.-K. Yau, A. Porter, N. Newman, and A. Suominen, "Clustering scientific documents with topic modeling," *Scientometrics*, vol. 100, no. 3, pp. 767–786, 2014.

[5] M. Koppel, N. Akiva, I. Dershowitz, and N. Dershowitz, "Unsupervised decomposition of a document into authorial components," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011, pp. 1356–1364.

[6] N. Akiva and M. Koppel, "Identifying distinct components of a multi-author document," in *Intelligence and Security Informatics Conference (EISIC), 2012 European*. IEEE, 2012, pp. 205–209.

[7] ——, "A generic unsupervised method for decomposing multi-author documents," *Journal of the American Society for Information Science and Technology*, vol. 64, no. 11, pp. 2256–2264, 2013.

[8] K. Aldebei, X. He, and J. Yang, "Unsupervised decomposition of a multi-author document based on naive-bayesian model." in *ACL (2)*, 2015, pp. 501–505.

[9] K. Aldebei, X. He, W. Jia, and J. Yang, "Unsupervised multi-author document decomposition based on hidden markov model," in *The 54th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2016.

[10] A. Daks and A. Clark, "Unsupervised authorial clustering based on syntactic structure," *ACL 2016*, p. 114, 2016.

[11] P. Willett, "Document clustering using an inverted file approach," *Information Scientist*, vol. 2, no. 5, pp. 223–231, 1980.

[12] X. Liu, Y. Gong, W. Xu, and S. Zhu, "Document clustering with cluster refinement and model selection capabilities," in *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2002, pp. 191–198.

[13] A. Strehl, J. Ghosh, and R. Mooney, "Impact of similarity measures on web-page clustering," in *Workshop on artificial intelligence for web search (AAAI 2000)*, vol. 58, 2000, p. 64.

[14] H. Chim and X. Deng, "A new suffix tree similarity measure for document clustering," in *Proceedings of the 16th international conference on World Wide Web*. ACM, 2007, pp. 121–130.

[15] A. Huang, "Similarity measures for text document clustering," in *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*, 2008, pp. 49–56.

[16] S. R. Eddy, "Hidden markov models," *Current opinion in structural biology*, vol. 6, no. 3, pp. 361–365, 1996.

[17] C. M. Bishop, "Pattern recognition," *Machine Learning*, vol. 128, pp. 1–58, 2006.

[18] M. Evans, N. Hastings, and B. Peacock, "Statistical distributions," 2000.

[19] Z. Ghahramani, M. I. Jordan, and P. Smyth, "Factorial hidden markov models," *Machine learning*, vol. 29, no. 2-3, pp. 245–273, 1997.

[20] J. Hagenauer and P. Hoeher, "A viterbi algorithm with soft-decision outputs and its applications," in *Global Telecommunications Conference and Exhibition'Communications Technology for the 1990s and Beyond'(GLOBECOM), 1989. IEEE*. IEEE, 1989, pp. 1680–1686.

[21] E. Amigó, J. Gonzalo, J. Artiles, and F. Verdejo, "A comparison of extrinsic clustering evaluation metrics based on formal constraints," *Information retrieval*, vol. 12, no. 4, pp. 461–486, 2009.