Expert Systems with Applications 40 (2013) 4301–4314

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Recognition of word collocation habits using frequency rank ratio and inter-term intimacy

Peng Tang, Tommy W.S. Chow *

Department of Electronic Engineering, City University of Hong Kong, Hong Kong

ARTICLE INFO

Keywords:
Text visualization
Text classification
Frequency rank ratio
Intimacy

ABSTRACT

An effective algorithm for extracting two useful features from text documents for analyzing word collocation habits, "*Frequency Rank Ratio*" (FRR) and "*Intimacy*", is proposed. *FRR* is derived from a ranking index of a word according to its word frequency. *Intimacy*, computed by a compact language model called *Influence Language Model* (ILM), measures how close a word is to others within the same sentence. Using the proposed features, a visualization framework is developed for word collocation analysis. To evaluate our proposed framework, two corpora are designed and collected from the real-life data covering diverse topics and genres. Extensive simulations are conducted to illustrate the feasibility and effectiveness of our visualization framework. Our results demonstrate that the proposed features and algorithm are able to conduct reliable text analysis efficiently.

1. Introduction

The World Wide Web has created huge amount of information in the form of text documents, making text processing increasingly challenging (Sun & Loh, 2009). In documents recognition and classification, articles are usually measured and classified according to their topics such as politics, sports, science and technology, and genres such as leisure or narrative or argumentative essays. A third type of text analysis can exist. For example, different writers can write in different style when they describe the same thing, and experienced English readers usually have no difficulty in telling whether an essay was written by native English speakers or the non-native ones. Here, the cues to differentiate the native and English as a Secondary Language (ESL) speakers are mainly the word collocation habits which express the vocabulary using and combinations of words that form lexical units, which differ from the genre or topic of texts.

The two textual habits are obviously affected by topics and genres of documents, as well as the writers. Consequently, approaches that handle topic-based or genre-based features in documents are of great help to our research on text analysis. A lot of researches have been done to analyze the topics and generic information of the past two decades. Most text analysis approaches treat language as a sequence of arbitrary symbols which do not have deep structure or meanings. The Term Frequency (TF), as well as Term Frequency-Inverse Document Frequency (TF-IDF), is one of the most important features in topic-based text analysis. TF-IDF is applied to most document models like Vector Space Models or Probabilistic Models (Jones, 1972; WenZhang, Yoshida, & Tang, 2011). A wide range of document analysis methods also rely on TF. These methods combined with machine learning techniques, such as Naive Bayesian (Fan, Zheng, Wang, Cai, & Liu, 2001), *k*-nearest neighbor (Han, Karypis, & Kumar, 2001) and support vector machine (SVM) (Joachims, 1999b; Joachims, 2001; Tong & Koller, 2002; Wang & Chiang, 2011), neural networks (Manevitz & Yousef, 2007), deliver acceptable results of text analysis. To extract more effective features, Ferreira and Figueiredo proposed an approach combined unsupervised feature discretization and feature selection techniques for textual feature reduction and selection (Ferreira & Figueiredo, 2012). A supervised feature selection approach combined conditional mutual information is also used in text clustering (Martínez Sotoca & Pla, 2010). Large-scale hierarchical dictionaries of concepts are utilized to identify the topics of the given documents and classify them to pre-set classes (Gelbukh, Sidorov, & Guzmán-Arenas, 1999). The Wikipedia database, thanks to its size, press coverage and copyright-free, is widely used to perform classification and discover and the topics of documents (Schonhofen, 2006). *N*-gram measures, which are mixed with Part-of-Speech (POS) tagging, function word lists, and naive Bayes classifiers, are used to investigate the influence of syntax structure on classification results (Clement & Sharp, 2003). Statistical Language Model (SLM) techniques have been used in many Natural Language Processing applications such as speech recognition, machine translation and information retrieval. A Bayesian and content-based text classification approaches are introduced for phishing web page detection (Zhang, Liu, Chow, & Liu, 2011). Peng et al. have presented a text classification approach, which involves Naive Bayes based classifier with statistical *n*-gram language models

* Corresponding author. Tel.: +852 27887756; fax: +852 27887791.
E-mail addresses: ptang@ee.cityu.edu.hk (P. Tang), eetchow@cityu.edu.hk (T.W.S. Chow).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.01.003

(Peng, Schuurmans, & Wang, 2004). Self-organizing Map (SOM) are also employed text clustering tasks (Corrêa & Ludermir, 2008; Liu, Wang, & Wu, 2008). Joachims has proposed an SVM based algorithm to improve the performance of text classification. It is suggested that the combination of SLM and SVM is capable of handling large datasets (Joachims, 1999a).

In selecting appropriate features of a document, genre-based approaches exhibit similar ideas with the topic-based or content-based measures, in which features are usually constructed of based words (or tokens) and phrases (or terms). Among most of the document research literatures, it can be noticed that the features, like word (or term) frequency and length of sentences, are the most widely used. For instance, word length and sentence length features have been used to test the genre classes and authorship (Brinegar, 1963; Brainerd, 1974; Morton, 1965). Tweedie has pointed out that the richness of vocabulary highly depends on text length and is very unstable (Tweedie & Baayen, 1998). Syntactic and semantic methods (Finn & Kushmerick, 2006; Luo, Chen, & Xiong, 2011), and advanced mathematical tools (Feldman, Marin, Ostendorf, & Gupta, 2009) have been introduced into genre text classification. The POS tagging marks up the words with different types, like nouns, verbs, adjectives, adverbs, corresponding to a particular part of speech based on its adjacent words in a sentence or paragraph. Many works have been established with the assistance of POS tagging. For example, Kessler et al. used POS tagging features to detect text genre (Kessler, Numberg, & Schütze, 1997). Finn also pointed out that the POS approach outperforms the traditional Bag of Words technique (Finn & Kushmerick, 2006). Feldman et al. extracted text genre features with POS histograms and Principal Component Analysis (PCA), then classified texts according to their genres using the Quadratic Discriminant Analysis (QDA) and Naive Bayes algorithms (Feldman et al., 2009). Biber (1995) defined "style markers", regarded as a formal definition of style of texts, as a set of measurable patterns. As a result, the computational cost for text genre classification can be significantly reduced (Biber, 1995). Kessler identified four generic cues: structural cues (e.g., POS tagging), lexical cues (e.g., Mr. and Ms.), character-level cues (punctuation cues and other separators, delimiters), and derivative cues (e.g., average sentence length), using existing text processing methods (Kessler et al., 1997). For modern language systems, it is generally believed that the system should be able to handle unrestricted or unpreprocessed text with low computational cost. Kessler's generic cues (Kessler et al., 1997) can result in a relatively low misclassification rate at the expense of computational cost. Despite POS being a widely used in genre text classification, it still suffers from certain shortcomings. For example, POS tagging has been found to be time-consuming (Lee & Myaeng, 2002; Oliva, Květoň, & Ondruška, 2003; Stamatatos, Fakotakis, & Kokkinakis, 2000) that may consume about 80% of the computational time required on extracting features. As a result, complex SLM approaches are rarely used in practice. In order to yield more effective style markers, approaches in a different way, e.g., high-frequency terms, are considered. For example, using the occurrence frequency of the most widely used words from a training corpus as style markers has also been studied (Burrows, 1987; Stamatatos et al., 2000; van Halteren, Tweedie, & Baayen, 1996). These studies suggest that the most frequently used words in written English are reliable discriminators for text genre classifications.

It is generally believed that writers can be a determining factor for genre/text classification. There are also research work focusing on identifying the authorship of given documents. Style markers are utilized to dealing with unrestricted text for an authorship-based classification, and a 50% or above accuracy has been reported when a 10-author corpus are processed (Stamatatos et al., 2000; Stamatatos, Fakotakis, & Kokkinakis, 1999). Another character-level *n*-gram authorship attribution approach is also raised to deal with both western and Chinese texts, and the overall accuracy is about 75% (Peng, Schuurmans, Wang, & Keselj, 2003).

There is another type of language models widely used for retrieving information. Instead of using term frequency itself, they use the proximity-based information between words to extract extra features of documents. For example, Petkova and Croft propose a document representation model based on the proximity between occurrences of entities and terms (Petkova & Croft, 2007). Lv and Zhai propagate the word count using a so-called Positional Language Model to obtain a virtual propagated word count and applied to other language models (Lv & Zhai, 2009). Different from the above method, our proposed method models a given text as a lexicon of weighted word pairs. In this paper, the weight of word pair, calculated by using proximity-based kernels in many applications, refers to the closeness between the two terms of the word pairs.

Generally speaking, a word-collocation extraction procedure highly relies on ranking which is usually computed using frequency information with respects to term occurrence and co-occurrence in a corpus. Seretan proposed syntax-driven criteria and syntactic patterns to identify word collocations in different languages. This method takes advantages of the recent advance in parsing tools in order to construct deep syntactic structures of texts (Seretan, 2010). A method for extracting multi-word collocation candidates based on the syntactical bound collocation bigrams and patterns is also described in Seretan, Nerima, and Wehrli (2003). A range of different extension patterns are defined using n-gram's POS-tagged patterns for extracting collocations tasks (Petrović, Šnajder, & Bašić, 2010).

The above-mentioned approaches, including topic-based and genre-based approaches, are able to deliver promising results, dealing with word collocation in documents, however, differs from classification documents using topic and generic cues. It is clear that genre-based and topic-based approaches are usually not accurate in describing word collocation in documents. In this paper, we mainly focus on recognizing these two textual characteristics. An important assumption is made for analyzing text documents in this paper. We assume that (1) words that construct complete documents, i.e., the word-using, and (2) relationship between words, i.e., the word-collocation, tell diverse documents apart. A compact and effective algorithm to extract features for analyzing the two characteristics from a text document is proposed. The two extracted features are called *Frequency Rank Ratio (FRR)*, and *Intimacy*. *FRR* is a conditional-defined feature based on a ranking index derived from the word frequency. *Intimacy*, in this paper, is proposed for modeling words proximity in a given document. We derive a compact language model to compute *Intimacy* which is stored in a series of lexicons. The lexicons functions like a ruler, i.e., a baseline. The original texts are firstly split into small slices of a pre-set length. Then, features extracted from given texts are compared to the baseline. All text slices are represented in the form of a two-dimensional vector. Consequently, the text slices can be visualized in a 2D space with the assistance of the vectors. Our obtained visualization results are promising as detailed in later sections of this paper. From our extensive experimental results, we suggest that the size of the training corpus is not necessarily large. Also, the classification results do not change much when different training corpora are used. Therefore, a small training corpus can always be used to conduct the text characteristics analysis. As a result, the time complexity of our proposed algorithm is far lower compared to other approaches. The reduction of computational complexity is caused by the exclusion of style markers on syntactic and semantic information.

The rest of this paper is organized as follows. In Section 2, we briefly overview our proposed framework derived for performing

text characteristics analysis. In Section 3, techniques for text preprocessing are discussed. Our main algorithms for textual feature calculation, as well as a concise language model, are detailed in Section 4. The textual features involved in our framework are addressed in Section 5. Extensive visualization and classification simulations are conducted, and corresponding results are shown in Section 6. We discuss the simulation results and conclude our proposed method in Section 7.

2. Overview of our framework

2.1. Framework of our approach

In the rest of this paper, we will show our approach that using only the two features can output promising results. Our framework is language-independent, which means semantic structures are unknown, the grammatical and semantic rules are not considered. The proposed approach works with original text, i.e., unrestricted text or text before preprocessed. There is no strict distinction among the terms, *document*, *essay* and *passage*, unless otherwise stated. In this paper, we also assume that the three words, *term*, *token* and *word*, are of similar meaning.

The proposed approach contains the following components.

1. A text preprocessor that splits the original text into sentences, each of which contains tokenized strings.
2. A document slicer. Here, a slice of length *n* refers to a small section of a given document with continuous *n* sentences.
3. A "Baseline" generator that outputs a series of lexicons as baseline for text feature extraction.
4. A text feature extraction approach based on our proposed algorithms: *FRR* and *Intimacy*.
5. A visualizer that simply projects the extracted text features into a 2D space.
6. A series of classifiers to evaluate the performance of our proposed approaches.

2.2. How the proposed framework works

Fig. 1 illustrates an overview of our framework. The framework includes a "training" section, which is to build lexicons that store the top frequent terms and two-word combinations. The goal of our framework is to calculate two different types of textual features, *FRR* representing the vocabulary using, and *Intimacy* which concerns with the word combination. First, a training dataset is preprocessed and then a set of lexicons are created using our proposed feature extraction algorithm and formed a "baseline". The baseline is "created once, used everywhere" in our framework. The input documents are first preprocessed into tokenized strings, and then cut into text slices. The baseline is utilized to compute the textual features of the text slices combined with our proposed algorithms. These features form vectors that represent slices of documents. After the features are obtained, we analyze the effectiveness by projecting the vectors into 2D space and measuring the results using classifiers. Our proposed algorithms are used in both the baseline preparation and textual feature extraction procedures. The details are shown in the following sections.

3. Text preprocessing

Previous literatures in natural language processing have pointed out that the distributions of term frequency obey the power-law. All the words form a scale-free complex network (Kr, Mukherjee, Mitra, Basu, & Banik, 2008; Motter, de Moura, Lai, & Dasgupta, 2002) in which most vocabularies are infrequent, while a few of tokens occupy a large proportion of the whole texts. The power-law distribution of words is illustrated in Fig. 2.

The power-law property is essential in this paper. Several preprocessing tricks are designed using this property. The power-law property is also employed to deliver a useful textual feature in the following section.

Common high-frequency words (typically 30–50 most frequent words) can be sensitive style markers for genre text classification (Burrows, 1987 and Riloff, 1995); this idea is adopted in this paper. In our developed algorithm, stop words and punctuations, as well as other words are all preserved for feature extraction. Note that all abbreviations, plural or singular forms of nouns, tenses of verbs, stop words, etc. are preserved in our algorithm. In other words, the "stem" procedure, which is usually employed in natural language processing (NLP) area for text preprocessing, are not adopted in our algorithm. This is because the "stemming" reveals original forms of tokens from their given forms in the original texts. The given forms, however, in most cases, reflect the word-collocation habits.

Merging low frequently used words to a unified term significantly reduces the computational cost, for it substantially reduces the numbers of distinct word. According to Section 4.1, there are huge amount of words with very low occurrence frequency, for example, the word "DIMENSIONALITY" occurs only once or twice in thousands of different news articles. These words, in fact, work only like placeholders. They can be regarded as the same token as people often ignore when reading. In this paper, the least 1–5%

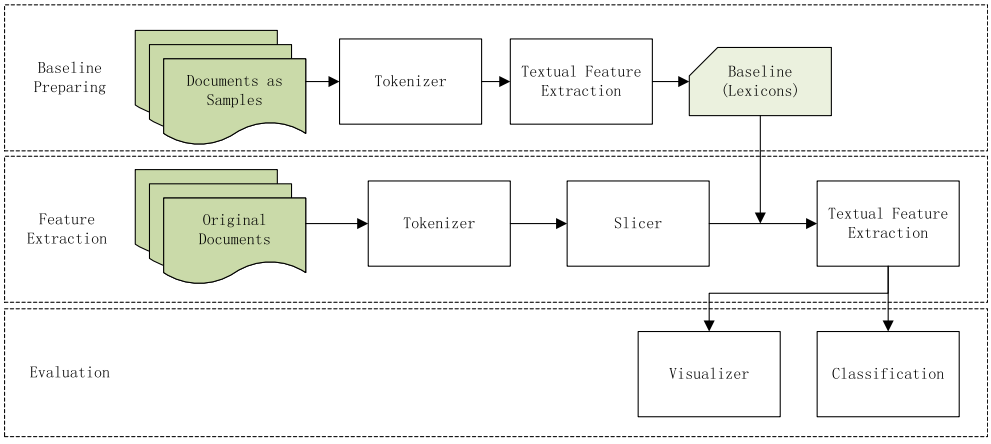


Fig. 1. Visualization framework.

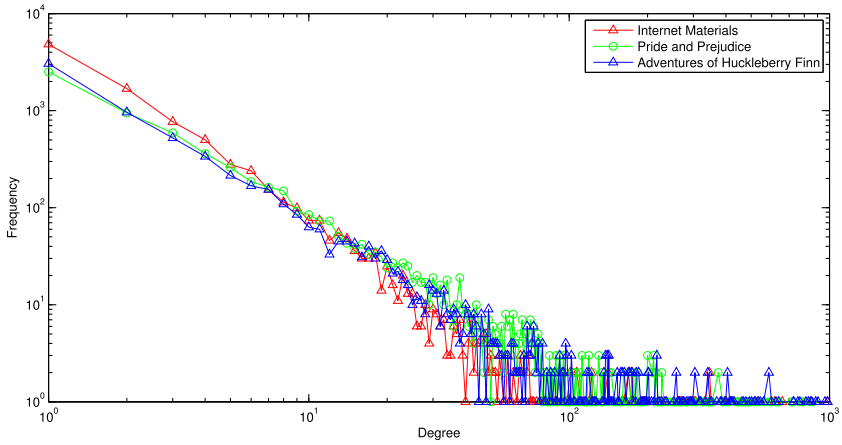


Fig. 2. The power-law distribution of words.

frequent words are merged in the preprocessing step. Numerical symbols, e.g., 1, 2, 3, I, II, III, IV, are merged to a special term called "numerical term" due to their similar usage.

To sum up, our proposed approach works with original unrestricted text, which means that original text without being preprocessed can deliver acceptable results. In order to further enhance the efficiency of the extraction procedures, the following preprocessing techniques are employed:

- All letters are converted to lower case.
- A virtual term, called "header", is attached to the start points of each sentence.
- Numerical symbols (e.g., 1, 2, 3, I, II, III, IV, "(a)", "(b)", etc.) are all replaced with a special term called "numerical term".
- Punctuations are treated as words.
- Terms with very low frequencies are merged to a unified term called "unknown term".

4. Features for recognizing word collocation cues

In this section, two newly introduced features, namely, *Frequency Rank* (*FR*) which describes the overall usage of words, and *Intimacy* which represents a complex connections between all words of a given document, are given. Here, *Intimacy* is computed for extracting the inter-term information of a given document. The mechanism of *Intimacy* is detailed in Section 4.3. The two features will form vectors that denote different document slices.

4.1. Frequency rank

Frequency of words is an important feature for document classification, clustering and retrieval. High frequent words, usually stop words, can be effective style markers. In most topic-based and content-based document classification work, meaningless words are considered as stop words and excluded from bags of words, because they can overshadow other meaningful words, because of their huge proportion. In the power-law distribution of word frequency as illustrated before, it is obvious that certain meaningless words such as THE/TO/OF that count most in the whole document, while the most meaningful words usually occupy a relatively small proportion. For example, "THE" and "TO" appears in a corpus 100,000 times and 20,000 times respectively, while the word "ALGORITHM" appears only 5 times. Genre-based text analysis, in another way, treats common words as useful elements, because of their constant existence and proportion in the whole text. For example, previous research works (Kr et al., 2008; Motter et al., 2002) prove that the top frequent words in English corpus do not change significantly, irrespective of the types and topics of the articles. According to our statistics, only the 3 words "THE", "TO", and "OF" constitute over 10% of the total number of words in our training corpus.

To make full use of high-frequency words and take a balance between the high and low frequent words, we introduce *Frequency Rank* (*FR*). The effect of an *FR*, i.e., the "action scope", can vary from a sentence to a paragraph, or even a whole chapter. In this paper, the "action scope" of *FR* is confined to each single sentence. *FR* of a word in a sentence is defined as the index of each word that indicates the presence of the corresponding words in the sorted word list by their global term frequency. For example, each word in the following sentence "This is a book." is attached with a corresponding bracketed global term frequency numbers (Attention that the stop symbol is considered as a word).

This (2300) is (5786) a (7777) book (23). (2710)
By ascending sorting each word using the global frequency, we get
book (1) This (2). (3) is (4) a (5)
Thus, the *FR* of each word is
This (2) is (4) a (5) book (1). (3)

The *FR* is used to measure relative "popularity" that the words appear in the corpus. From the comparison between a word networks and a scale-free network in Section 3, we can conclude that *FR* of a word is roughly in proportion to the log of its frequency if the "action scope" of each word is set to its whole document. *FR* of each word is always positive integer, which is preferred in computing procedures.

Let a n-word sentence is expressed in a form $S = \{w_1, w_2, \ldots, w_n\}$. A Normalized *FR* (NFR) is defined as:

$$\mathrm{NFR}(w_i, S) = \frac{\mathrm{FR}(w_i)}{\max_{w_j \in S}(\mathrm{FR}(w_j))}, \tag{1}$$

where w_i and w_j are an arbitrary words in a given sentence S. From (1) it can be observed that NFR $(w_i,S) \in [0,1]$.

4.2. The influence language model

Many document models describe a document by a vector:

$$x = (x_1, x_2, \ldots, x_v), \tag{2}$$

where $x_i = 0$ or 1 indicates the absence or presence, or the word frequency of the ith term. There are also models using Bayes, Markov

theory or other statistically based theories. Each type of model has certain advantages and drawbacks. For instance, vector based models define everything explicitly and strictly, making a range of existing mathematical tools directly applicable. For example, COSIN distance is a widely used method for measuring similarity. PCA is another widely used dimensionality reduction method for extracting useful features. Nevertheless, vector based models barely rely on structural information of sentences, like order of words, word combinations and so forth. For all these methods, only words are preserved, but structurally information has completely been overlooked. For instance, "This is a book" has an identical expression to "Is this a book" when only the occurrence of words is considered. In addition, the computing cost concerning with matrix is usually high when using vector based models. Statistically based approaches make full use of relations between words, implying that the structure and semantic contents of a document in some way are being taken into account. But the computational cost is too high for efficient application, especially when one needs to handle a huge number of large documents. Thus, n-gram methodologies are introduced to reduce the computational cost. Apart from these problems, data sparseness is a major problem that these methods are unable to solve.

An n-gram model is a statistical model for calculating the probability of the next item in a sequence. N-gram models are widely applied in NLP and other areas. It calculates $P(x_i|x_{i-(n-1)},\ldots,x_{i-1})$ of x_i based on $x_{i-(n-1)},\ldots,x_{i-1}$. N-gram models are built on a basis that one word depends only on its last "$n-1$" words. This assumption significantly alleviates the computational problem (Wikipedia, 2011). In practical applications, the value of n is often less than 4, or the computational cost would become unbearably high. This appears to be a major drawback for a statistic model aiming to extract useful features.

Prior to detailing the feature *Intimacy*, a compact language model inspired by the concept of cumulative probability, is first introduced. Our proposed language model, namely *Influence Language Model* (ILM), is mainly based on a simple idea that there exists a certain relationship between a word and its neighbors, and the neighbors of its neighbors. Some previously mentioned language models also use similar idea for propagating weights of words (Petkova & Croft, 2007; Lv & Zhai, 2009) and estimate similarity between terms (Petkova & Croft, 2007; Lv & Zhai, 2009). Different from the above method, our approach models a given text as a lexicon of weighted word pairs. Here, the weight of word pair refers to the closeness between the two terms of the word pairs. The weight is calculated using a so-called influence factor which is also applied as proximity-based kernels in many applications (Petkova & Croft, 2007; Lv & Zhai, 2009).

The interaction within a pair of terms depends on the distance between each other. The closer the neighbors, the larger the influence they have; conversely, the longer the distance between a pair of words, the weaker interaction force they have. We use to denote the influence of two words w_i and w_j in a sentence S. The influence $p(w_i,w_j,S)$ varies with the influence function $d(i-j)$ as shown in Fig. 3. In this study, the decrease rate is assumed to be Gaussian, because Gaussian function ensures every term of a long sentence receives a non-zero "interaction", whereas polynomial and linear functions will deliver zero "interaction" at a finite distance. A group of simulations are also conducted to illustrate their performance in later sections.

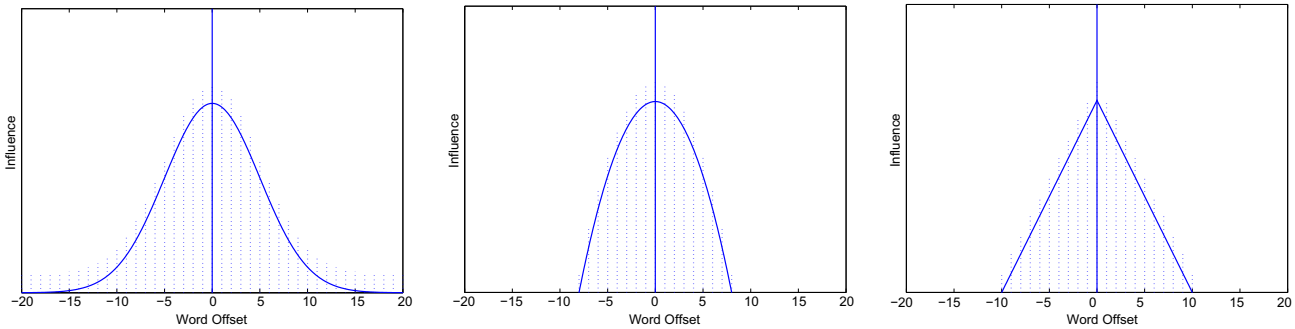
The interaction decreases with defined functions, such as Gaussian in Fig. 3(a), or simply linear shown in Fig. 3(c). These curves are often served as "proximity-based kernel" functions in many proximity-base language models (Petkova & Croft, 2007; Lv & Zhai, 2009). It is also interesting to notice that these "interaction" can be accumulative. For instance, given a sentence.

> *"a (1) bigram (2) is (3) a (4) sequence (5) of (6) two (7) items (8) from (9) a (10) given (11) sequence (12).".*

In the above sentence, each numbers is the corresponding index of the word. We assume that a word affects its direct neighbor with an influence factor 1.0, the next one with 0.9, and the further next one with 0.8, and so on. For example, the influence between the 4th term "a" and the 5th term "sequence" is 1.0, while the interaction between 10th term "a" and 12th terms "sequence" is 0.9. As a result, the overall influential effect of the term pair "a" and "sequence" is 1.9. It must be noted that the above approach is not strictly probabilistically defined. But using the above method, the word combinational information can be effectively collected.

4.3. Intimacy

FR based features provide token-level information for further processing. Yet the token-level information only considers the whole texts as "bags of words", in which most inter-term features are dropped. Thus, it is our attempt in this study to derive a new type of features, which aim to include higher level features, such as, the sentence-level or fulltext-level features. It must be noted that many approaches have been proposed to extract higher textual features, e.g., n-gram with POS tagged text can extract grammatical or syntactic features. There are also other relatively complicated methodologies, such as maximum entropy based model (Ratnaparkhi et al., 1996) and Bayesian based features (Goldwater et al., 2007). *Intimacy* is introduced to determine popularity discriminating compatibility of a word and its environment, i.e., its neighbors, and the neighbors of its neighbors, etc. When proficient readers are to classify if a given document is written by a professional English or naive English group, different readers



(a) Gaussian Shaped Influence factor (b) Polynomial Shaped Influence factor (c) Linear Shaped Influence factor

Fig. 3. Influence function $d(i-j)$ between word pairs. ±1 is supposed to have the most influence to the central words. The influence decreases along the distance between word pairs. Negative numbers represent a word in front of the "main" word.

may not often come up with total agreement in all cases. But simple errors such as "Thank (me) very much", "a plastic hand-making (white) toy" can be spotted without difficulty. Apparently, POS tagging based approach cannot detect these discrepancies because both "me" and "you" are pronouns.

In order to elaborate the feature *Intimacy*, we introduce the cumulative influence C between two words. First, we will elaborate *Intimacy* using our proposed language model. For a given sentence S_c, C can be yielded by (3) and (4). In (3), + means to get C between 2 particular words w and w' when w' appears after w; in (4), − means w' occurs ahead of w.

$$C_+(w,w') = \sum_{c=1}^{S} \sum_{\substack{a<b \\ w_a.w_b \in S_c}} d(a-b). \tag{3}$$

$$C_-(w,w') = \sum_{c=1}^{S} \sum_{\substack{a>b \\ w_a.w_b \in S_c}} d(a-b). \tag{4}$$

Here, $d(a-b)$ is a function representing how the influence between the two words decays in a sentence as the above-mentioned, where a,b are the index of the two words w and w' in sentence s_c, $w_a = w$,$w_b = w'$. Note that C is not statistically defined so that C can be much larger than 1. In the above definition, C can collect relationship between word pairs more effectively than n-gram based SLMs. Because C considers not only the direct neighborhood of a word, it takes the occurrence of all the words appeared in a sentence into account. Note that C for all words can be calculated and stored in another lexicon for further use. After C is calculated, I can be calculated by (3) directly.

Bayes' theorem is widely used methods in document classification and other classification area. For events A and B, provided that $P(B) \neq 0$.

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}. \tag{5}$$

Bayes' theorem predicts posterior probability, the uncertainty of a probability after observing the modeled system. In NLP area, Bayes' theorem often cooperates with other SLM methods such as n-gram. In Section 4, an essential feature using the concept of Bayes' theorem and our compact language model is described.

Similar to Bayes' theorem, we can express

$$I = \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{C(w_i,w_j)p(w_j)}{p(w_i)} d(i-j), \tag{6}$$

where $P(w_i|w_j)$ in the Bayes' theorem is replaced by $C(w_i, w_j)$. The local C_+ and C_- is calculated, which means only words in specified action scope, such as in a sentence scope and in a passage scope, is considered. They are subsequently compared with the previous lexicon, and the difference will be sorted. The smaller the differences, the more intimate the words within the specified area are. Here the training corpus is considered sophisticated and regarded as the "ruler". We must explain that all the above described features are not statistically defined. Strictly speaking, *Intimacy*, inspired by the Bayes' theorem, is empirical defined. It is worth noting that the computational complexity for calculating *FR* is $O(n\log n)$ when quick sort algorithm is utilized. The time complexity for intimacies is $O(n^2)$, where n is the word count of the sentence. The value of n is usually set to 3 to 20 which is far less than the total numbers of words used in a given document. This explains why our proposed feature extraction algorithm is fast.

We use the concept "Entropy" to prove that by using our IML, more uncertainty of sentences can be captured compared with other n-gram based approaches.

Generally, let $S = \{s_1,s_2,s_3,\ldots,s_n\}$ be a set of N sentences in a corpus. A section of sentence of length L where word are counted. We define the vocabulary $V = \{v_1,v_2,\ldots,v_{|V|}\}$ as the set of all unique words that occur in the corpus. We represent an arbitrary sentence t_i as a sequence of words $\{w_1,w_2,\ldots,w_n\}$ where n is the length of the sentence t_i. Different collocations starting with the word w_i examined in a sentence is written as $\{(w_i,w_{i+1}), (wi,w_{i+2}),\ldots, (w_i,w_{i+L})\}$, where L is the length of the collocation window.

Let $N(v_i,v_j)$ be the number of collocations with the first word v_i and the second word v_j in the collocation pair (v_i,v_j). The probability that the word v_j occurs as the second word given v_i as the first word is

$$p(v_j|v_i) = \frac{N(v_i,v_j)}{\sum_{j=1}^{j=|V_i|} N(v_i,v_j)},$$

where $|V_i|$ is the number of unique collocations examined with V_i as the first word. By the definition of conditional entropy, the uncertainty in the occurrence of a word from the vocabulary in the second position of a collocation, given the first word, can be expressed as

$$H(v_i,x) = -\sum_{j=1}^{j=|V_j|} p(v_j|v_i)\ln p(v_j|v_i),$$

where $|V_j|$ is the number of unique words from the vocabulary in the collocations with v_i as the first word.

Using our ILM, the $N(v,v')$ is replaced with $C(v,v') = \sum_{w_i,w_j \in s_k} d(i-j)$, where $v = w_i$, $v' = w'$ and $s_k \in S$. Because $0 < d(i-j) \leqslant 1$, $N(v,v') \geqslant C(v,v')$. Note that

$$f(x) = -x\ln x$$

This is a monotonic decreasing function. Consequently, after we replace $N(v,v')$ by $C(v,v')$, $H(v_i,x)$ will increase.

When the length of sentence section L is set to 2, only the word collocations that constructed by two adjacent words are considered. In this scenario, the IML is transformed to bigram model, i.e., n-grams for $n = 2$. When L is set to n, the IML is transformed to slice window model with a slice length n. It can be intuitively understood that using $L = |s|$, more distinct word-collocations will be scanned than the bigram model and slice window model. Therefore, we will obtain a larger value $H(v_i,x)$ when using our ILM model, which results in extracting more information from the text.

The sparsity of words in language models is always a problem, which often needs smooth techniques to solve the zero probability terms. In our proposed model, we use an effective preprocess technic described in Section 3 to minimize this problem caused. We merge the low frequent words to a special term which acts as a placeholder and merged to a unique symbol in the preprocessing steps. In other words, the low frequent words are regarded as the same term when Intimacy is calculated. Extreme individuals that still appear as zero-probability items in the trained lexicon will be abandoned directly. As a result, the sparsity of words can be solved. The presented results also show that classification accuracy can be improved by applying this approach.

5. Textual feature extraction

The above two features *FR* and *Intimacy* are capable to represent word collocation information of texts effectively. Acceptable visualization results can be achieved by projecting the above two features extracted from texts to a 2D space. To achieve better performance and more accurate results, two advanced feature *Frequency Rank Ratio* and *Conditional Intimacy* are calculated using *FR* and *Intimacy* as above defined. A "Baseline" is required in the feature calculation procedures. The construction of the Baseline is detailed in the followings.

5.1. "Baseline" generation

The baseline here refers to a set of Lexicons. The base line acts as rulers for textual feature extraction. Training corpus is first preprocessed. The size of training corpus is not necessarily large. This point will further discussed in the following section. The main purpose of this step is to get a sorted word rank list L_F, P_+ and P_- lexicon. It is assumed that there are n words in the training corpus and the average sentence length is l. The time complexity for getting rank list is $O(n)$ and $O(nl)$ for yielding P_+ and P_- dictionaries. The complete steps of baseline construction are shown in Algorithm 1.

Algorithm 1. The steps of baseline generation

Require: Original sample texts
Ensure: L_F,P_+,P_-
1. Preprocess the original sample texts to tokenized strings T.
2. Calculate the frequency f_i of a token $t_i \in T$
3. $L_F \Leftarrow (t_i, f_i)$, where $i = 1,2,3,\ldots,n$
4. Split T into a list of sentences S
5. For a sentence S_i in S, calculate $C_+(w,w')$ and $C_-(w,w')$, where $w, w' \in S_i$
6. $P_+ \Leftarrow (w, w', C_+)$, where $S_i \in S, w,w' \in S_i$
7. $P_- \Leftarrow (w, w', C_-)$, where $S_i \in S, w,w' \in S_i$

return L_F,P_+,P_-

5.2. Calculation of FRR

After obtaining the necessary lexicons, textual features can be extracted. All texts are firstly preprocessed and then split into slices with m sentences. The *FRR* and *Intimacy* are extracted from the slices. To avoid the problem of sparsity, only words with high word frequency are taken into account. In our proposed framework, a parameter "TOP", used to fetch the top x percentage of words from a vocabulary, is introduced to tune the actual number of terms when calculating the textual feature. The merging of these low frequent words in the text preprocessing step has an advantage of reducing the problem of sparsity. A word which has never appeared in lexicons is regarded with an occurrence of "1".

For a slice A with n sentences $S_1,S_2,\ldots,S_n$, the *FRR* of A given a lexicon L_F in a baseline is calculated in the following steps.

$$FRR(A|L_F) = \frac{1}{n}\sum_{i=1}^{n} \sum_{\substack{w\in S_i \\ w\in \text{top}\ (L_F)}} \text{NFR}(w, S_i)), \tag{7}$$

where the TOP $(\cdot, r)$ gives the top frequent words, r is proportion of the adopted words in the whole vocabulary. With the TOP operator, a term list can be automatically generated so that the manual specification of word list is avoided. This mechanism ensures our approaches a generalized method for all tokenized texts. The TOP $(\cdot, r)$ operator also acts as a parameter that controls the merging of the proportion $1 - r$ least frequent words.

Compared to *FR*, L_F in the specified baseline replaces the global term rank of the original texts that A belongs to when calculating *FRR* of A. The number of sentences in A is n, which is a constant. Therefore, (7) can be simplified as

$$FRR'(A|L_F) = \sum_{i=1}^{n} \sum_{\substack{w\in S_i \\ w\in\ \text{top}\ (L_F)}} \text{NFR}\ (w, S_i)). \tag{8}$$

The procedure of calculating *FRR* is shown in Algorithm 2.

Algorithm 2. The steps of calculating *FRR*

Require: Input text slice A, the baseline lexicon L_F
Ensure: $FRR(A|L_F)$
  Set the return value $R \Leftarrow 0$
  for all sentence $S_i \in A$ **do**
    for all term w S_i **do**
      if $w \in$ top (L_F) **then**
        $R \Leftarrow R$ + NFR (w,S_i)
      else
        Set the frequency of $w \Leftarrow 1$
        $R \Leftarrow R$ + NFR (w,S_i)
      end if
    end for
  end for
  return R

5.3. Calculation of overall intimacy

The Overall *Intimacy* $(\bar{I})$ is conditional-like defined, in which $\bar{I}$ is derived with constraints P_+ and P_-. Similar to *FRR* of A given L_F where L_F in the specified baseline replaces the global term rank of the original texts that A belongs to when calculating *FRR* of A, for a slice A with n sentences $S_1,S_2,\ldots,S_n$, only word pairs that exists in P_+ and P_- are taken into account when the Intimacy is calculated.

Observe that the number of word n in the training corpus is constant. Therefore, the $(\cdot)/n$ in (4) can be omitted, that is, $p(w) = f(w)/n$ can be simplified to $f(w)$. Consequently (3) can be rewritten as

$$\text{Intimacy}' = \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{C(w_i, w_j) f(w_j)}{f(w_i)} d(i-j). \tag{9}$$

The overall Intimacy $\bar{I} = \sum_{i=1}^{n} I(S_i)/n$ is used as one of the features for further text analysis. As all the slices are of the same length, the divisor in (9) can be eliminated as

$$I_{sum} = \sum_{i=1}^{n} I(S_i|P_+, P_-). \tag{10}$$

The procedure of calculating $\bar{I}$ is shown in Algorithm 3.

Algorithm 3. The steps of calculating overall intimacy

Require: Input text slice A, the baseline lexicon P_+, P_-
Ensure: $\bar{I}$
  Set the return value $R \Leftarrow 0$
  for all sentence $S_i \in A$ **do**
    for all word pair $w_i,w_j \in S_i$ **do**
      if $w_i,w_j \in P_+$ or P_- **then**
        $R \Leftarrow R + \frac{C(w_i,w_j)f(w_j)}{f(w_i)} d(i-j)$
      end if
    end for
  end for
  return R

The slices can be easily visualized with the assistance of the two extracted features. Our proposed two features, *FRR* and *Intimacy*, can construct higher dimensional vectors when they work together with other style markers. In this case, higher dimensional vectors can be processed using PCA to extract two principle components for visualization process.

6. Visualization and classification

To evaluate the performance of our proposed approaches in visualizing the different word collocation of diverse authors and genre, several groups of simulations are conducted with different corpus.

6.1. Corpus preparation

The popular corpora nowadays seldom concerns word collocation evaluation. For example, the *Brown* corpus is collected within wide range of topics, while the style is overlooked. The *Wall Street Journal* corpus considers less text styles except native news and reports. Take all things into account, our training and testing corpora are collected to evaluate our algorithm in different perspectives, two corpora are designed and collected.

6.1.1. Corpus 1

The first corpus is named "Corpus 1", the components of which are listed in Table 1. "Corpus 1" is constructed of a wide range of topics of documents with diverse word collocation habits. In this corpus, five novels are considered as one category (C-I). English diaries, mainly collected from a professional Chinese website for Chinese users for improving their English, exhibit significantly different characteristic from native English speakers, because the cultural backgrounds and language environment are totally different; these texts are thus considered as the second category (C-II). English news and reports obtained from several mainstream English media belong to the third category (C-III). Abstracts of undergraduate students Final year project (FYP) reports of an Asian University are considered as academic writings, and are classified as the fourth category (C-IV). Every category of corpora contains texts with diverse topics. For example, the news and reports include reportage covering politics, worlds, environmental problems, science and technologies, etc.; texts in C-IV include different research areas like electronic engineering, computer sciences, social sciences, etc. Newsgroup topics from *20 Newsgroups* are adopted as the fifth category (C-V).

6.1.2. Corpus 2

To analyze word collocation habits of documents of the same genre, another corpus, named Corpus "2" is collected. The components are listed in Table 2. To avoid the bias caused by changes of topics and text genres, the testing texts cover topics as diverse as breaking news, politics, sports, arts, culture, business, education, society, science, technologies, etc. And only recent items are collected in order to avoid possible bias due to the items of being out of date. Newsgroup topics selected from the corpus *20 Newsgroups* are introduced to make a comparison. All the texts are collected from well-acknowledged mainstream media on the internet.

Table 1
The components of our corpus.

Components	Component contents	Count
C-I	Novels (Pride and Prejudice, The Adventures of Huckleberry Finn, Gone with the Wind, Whtering Heights, The Call of the Wild)	5
C-II	English diaries by Chinese Learners from a Chinese website for Chinese users who want to improve their English	200
C-III	English news and reports available on CNN.com and Guardian.co.uk	400
C-IV	Abstracts of undergraduate students Final year project (FYP) reports of an Asian University	400
C-V	Newsgroup topics	400

Table 2
The components of Corpus 2.

Component Contents	Count
Native English Media (CNN, Guardian, NYTimes, Telegraph)	3000
English Media in CJK (Chinese, Japanese, and Korean) areas	3000
Newsgroup Topics	1000

In our simulations, the components of the corpora can be added or removed as needed.

6.2. Evaluation methods

Each vector formed by the calculated textual features represents its original text slices. To provide intuitive idea of the word collocation habits, the vectors of different categories of text slices are treated as data points and projected directly to a 2D space. The distribution of the data points reveals the cues of word collocation habits. We can directly observe the "Intimacy degree" of the data points by their distance from each other.

To quantitatively analyze the effectiveness of the features, three widely-used classification approaches, i.e., *k*-nn, naive Bayesian, and decision tree, are employed. We choose the classification accuracy using the textual features to measure the effectiveness. Higher classification accuracy indicates a more efficient group of features. The three classifiers are chosen because the decision tree is a more intuitive way, which is similar to human's decision procedure; *k*-nn and naive Bayesian are more computational oriented. Also, the three classifiers are widely used and tested by practice. The parameters and implementations of the three classifiers *k*-nn, naive Bayesian and decision tree can refer to Aha, Kibler, and Albert (1991), John and Langley (1995), and Quinlan (1993).

6.3. Simulation results of text of different contents

Corpus 1 is selected as the testing dataset. The novel *Pride and Prejudice* is taken as the training dataset to prepare the "baseline". All texts are split into slices, then, the two features are calculated using our proposed framework. All the corpora, except the novel components, are visualized in a 2D-space displayed in Fig. 4, in which each color point represents a slice of texts. Every category can be distinguished by the colors of the dots that they are

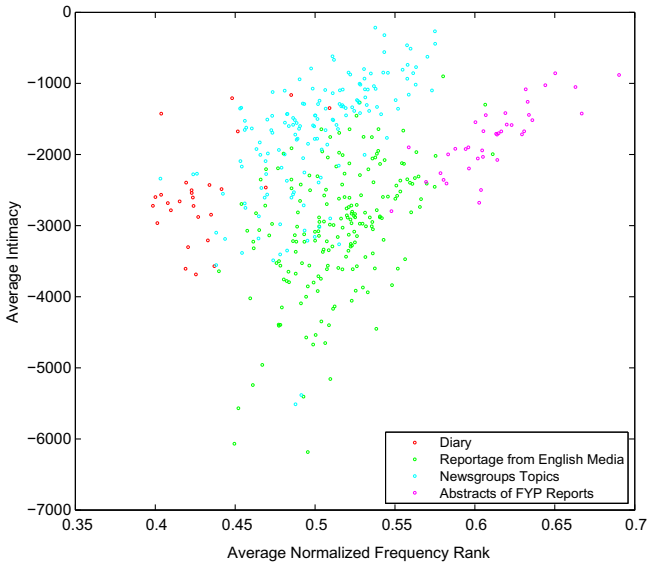


Fig. 4. Visualization the four components in Corpus 1.

distributed more or less in recognizable areas of the 2D space. It is observed from Fig. 4 that the feature *FRR* can effectively classify "Diary" and "Abstract of FYP Reports" from the other two types. As mentioned in Section 4.1, *FRR* models the spectrum of word usage. For the types of "Diary" and "Abstract of FYP Reports", it is clear that the selection of words is significantly different from the "Report from English Medias" and "Newsgroups Topics". But for newsgroup topics and English news and reports, the word-using cannot separate them from each other. The feature *Intimacy*, however, distributes the two categories into different areas in the 2D space.

All the five components in Corpus 1, including the novels, are visualized in Fig. 5. The black points represent the slices of novels. It can be seen that the five components of Corpus 1 are roughly scattered into their own domains. It is noted that there are black points scattered in the domains of other categories. This phenomenon is common in novels, in which the long body text consists of many types of word-using habits, and phrase-using habits as well as styles like narrative style, dialog style, and critical style, etc. After the long text is cut into slices, a distinct slice may be slightly different from other novel slices.

Another group of experiments are processed to illustrate the influence of text genre to the visualization result. Six different categories of texts, namely atheism, computer sciences on Windows OS, forsale, sports, space sciences and religions, are selected from the *20 Newsgroups* corpus. They are analyzed by our proposed algorithm; the results are visualized in Fig. 6(a). From the 2D visualization result, it is clear that the 6 categories of articles are largely overlapped. Despite them being in different categories of topics (appeared in different colors), all data points congregate together because they are all classified as a single category of newsgroup topics. But the distribution of each color is not uniform, implying the text characteristics of slices are not homogeneous. We can observe that slices about atheism and religion, appeared in red and pink, respectively, are totally overlapped, but they rarely appear in the region of computer and space sciences. This means that the writing styles between religion/atheism and computer/space sciences vary significantly. In Fig. 6(b), it shows the Holy Bible is located at a region close to the above six categories. From Fig. 6(b), it is clear that the "black crosses", representing slices of texts of the Bible, do not completely overlap with the 6 categories of newsgroup topics, because they are of different word collocation habits. But it is interesting to notice that the black cross, which represents the Holy Bible, are located in the close vicinity of atheism and religion, but are far from the regions that refers to sciences and engineering. This observation shows that the text characteristics information of texts is somehow affected by the topics.

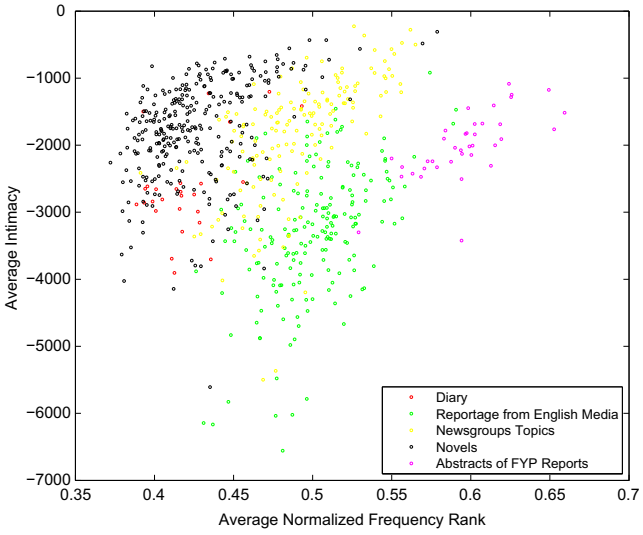
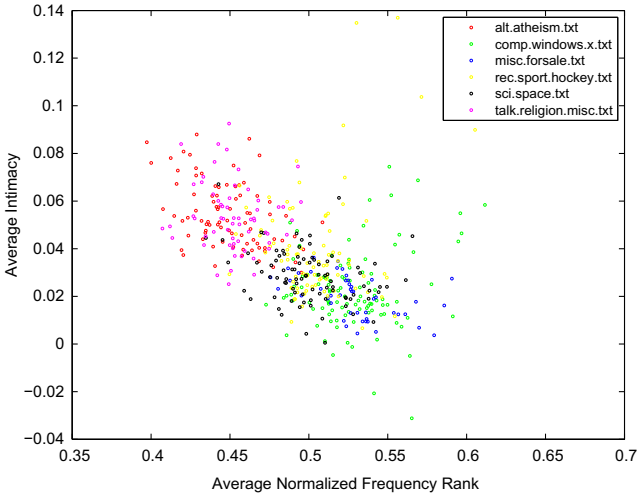


Fig. 5. Visualization results of Corpus 1.

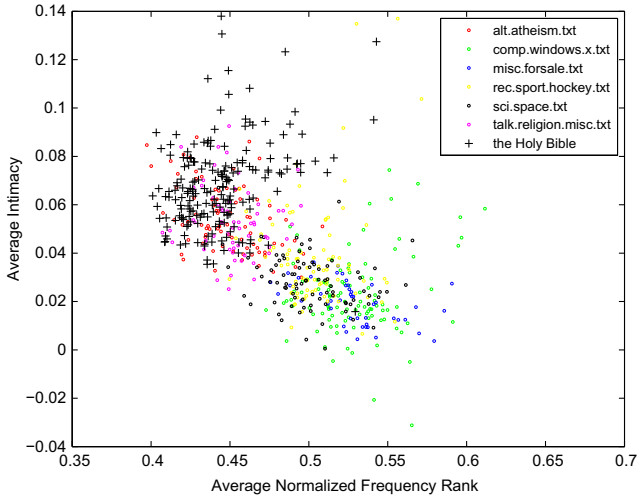
6.4. Visualization results of texts of the same genre and topic

Documents of the same genre can be in totally different word collocation habits. Corpus 2 are selected as the testing dataset. All categories of texts in Corpus 2 is visualized and shown in Fig. 7, in which baseline is still generated from the novel *Pride and Prejudice*.

It is observed that a portion of text slices from Native Media and CJK English Media categories are overlapped, which means the word collocation habits of the two categories of texts are already similar. However, an observable portion of the CJK English texts



(a) Visualization Result of Selected 6 Corpus from 20 Newsgroups Dataset. Basically, the six categories of articles are overlapped. But the distributions of points are still not homogeneous. Points concerning about atheism and religion slices are totally overlapped while rarely scattered in the domain of computer and space science, which means the word collocation habits vary with the change of the topics of newsgroups.



(b) Visualization Result of Selected 6 Corpora From 20 Newsgroups Dataset and the Holy Bible. The black cross representing the Holy Bible, are located in the close vicinity of atheism and religion, but are far from the regions that refer to sciences and engineering.

Fig. 6. Visualization result of Selected 6 Corpora from the *20 Newsgroups* dataset.

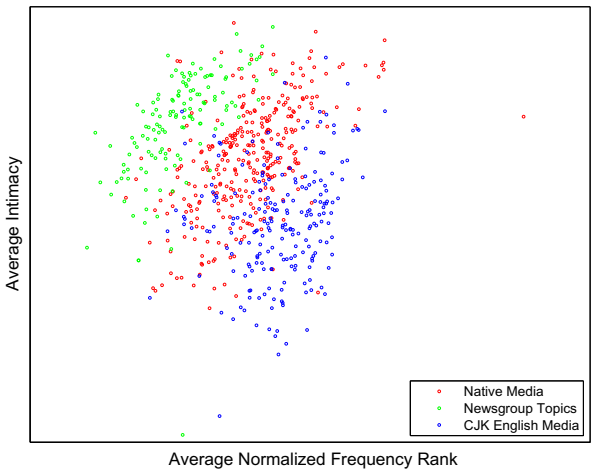


Fig. 7. Visualization Results of Corpus 2.

are located in different area from the native English Media. The text of Newsgroup Topics, written by native English speakers is visualized in Fig. 7. It can be seen that only a small number of points representing the Native Media and CJK English Media are overlapped with the Newsgroup topics. It is worth noting that, the points representing Native Media and Newsgroup Topics are closer to each other, compared to the Newsgroup Topics and CJK English Media. This is because most newsgroup topics are written by native English speakers, and they are expected to be more similar to the native English media than the CJK English media.

6.5. Alternative baselines

Our simulation results indicate that the alternative baseline, i.e., using different training dataset to generate the lexicon, does not have noticeable effect on the final visualization results. Fig. 8 shows a brief group of results using different training corpus, including different total length of texts and different styles. Detailed results can be found in Table 3. Here, the naive Bayesian Classifier implemented in Weka is utilized to calculate the accuracy. Corpus 1 without novel component, Corpus 1, Corpus 2, and Corpus 2 without newsgroup topics are used as the testing corpus of Test1, Test2, Test3 and Test4, respectively. From the results it is found that the overall distribution does not change substantially with different training corpus. As discussed in Section 4.1, a few frequent words (usually less than 30, such as "the", "a", "to"),

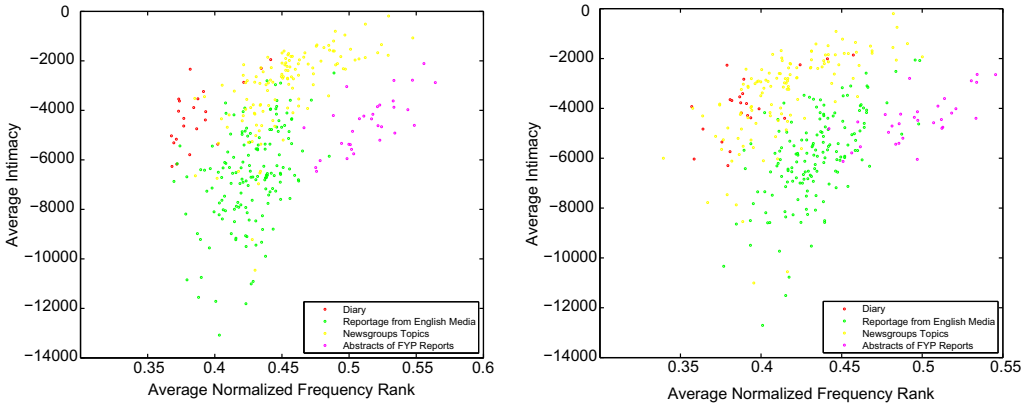
Table 3
Classification accuracy of the visualization results using different baseline.

Baseline	Test1	Test2	Test3	Test4
Pride and Prejudice	0.773	0.744	0.752	0.712
Gone with the Wind	0.780	0.747	0.759	0.720
Whtering Heights	0.764	0.741	0.739	0.698
FYP reports	0.755	0.731	0.756	0.707
Diaries	0.736	0.704	0.748	0.713
CNN (1000 items)	0.805	0.753	0.771	0.734
CNN (200 items)	0.792	0.748	0.769	0.731
CNN (100 items)	0.788	0.743	0.767	0.726

dominate the total amount of words. They have a significant effect on inter-term based feature. And it is worth noting that these words are relatively constant, i.e., they do not change substantially with the change of texts, either in topics or in styles. This is a useful characteristic, since it implies that a relatively small training corpus can be used to classify a large amount of testing data. This means the computational cost can be significantly reduced by using a relatively small baseline lexicon generated using small training corpus.

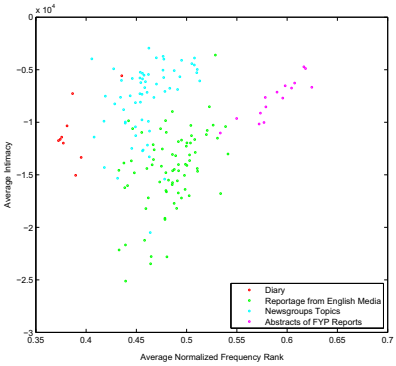
6.6. Classification using extracted features

Two parameters, namely *top words* (*TOP*) and *length of slices* (*LEN*), are provided by our visualization framework. To test the two parameters, classifiers are used for evaluating the data of extracted features using different pre-set parameters. Length of slices, namely the number of sentences in a slice, can be an effective factor for visualization. The comparative classification results using different values of LEN are displayed in Fig. 9. In this group of simulations, the TOP is set to 0.7. In this paper, cross-validation is performed to evaluate the classification accuracy. The number of folds for cross-validation is set to 10. It can be observed that when LEN increases, the numbers of slices get smaller, while the overall classification results appear to be more accurate. It is natural that with a smaller pre-set value of LEN, the slice points are more overlapped due to the fact that smaller slices may be of similar word collocation habits to slices in other categories. This phenomenon can be observed in Fig. 9. In this figure, Fig. 9(a) and (b) are the samples of real visualized distributions of text slices. Fig. 9(c) and (d) are the accuracy curves using Corpus 2 and Corpus 1, respectively. Compared to points in Fig. 9(a), a number of slice points in Fig. 9(b) are scattered in their neighbored categories. This feature can actually be used in two ways: if we want to extract the overall style of texts, LEN can be set larger; if we intend to capture the detailed genre information, the length can be set smaller so

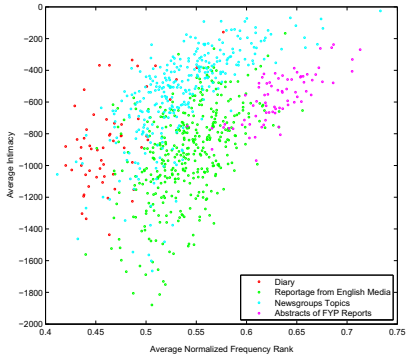


(a) Novel as baseline (b) Diary as baseline

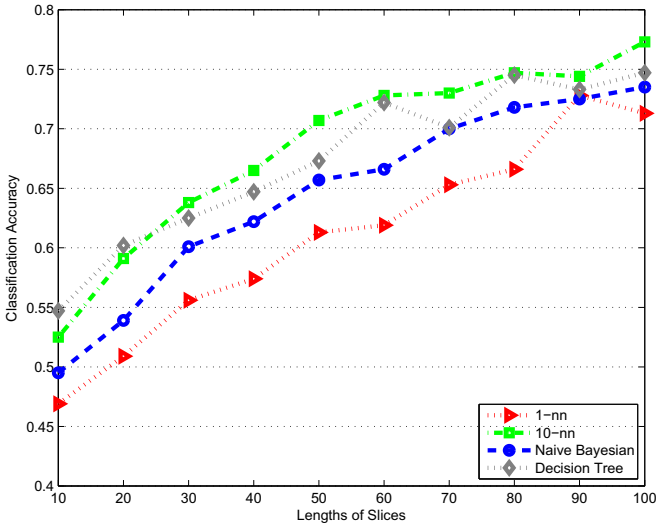
Fig. 8. Visualization results with alternative baselines.



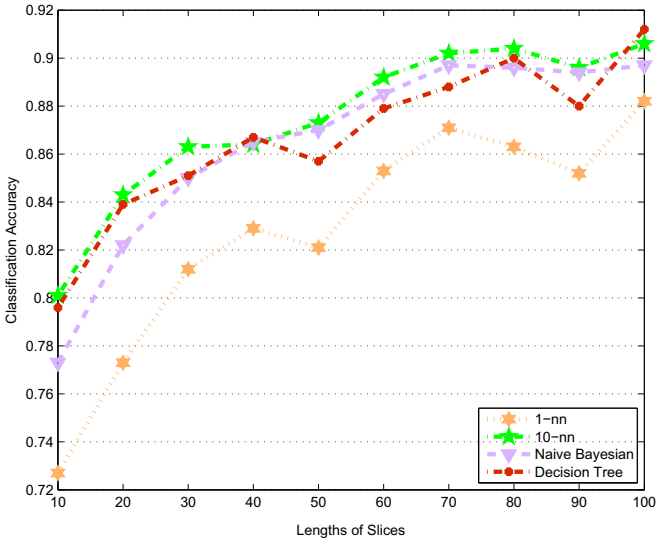
(a) LEN=100, Corpus 2 is utilized.



(b) LEN=20, Corpus 2 is utilized.



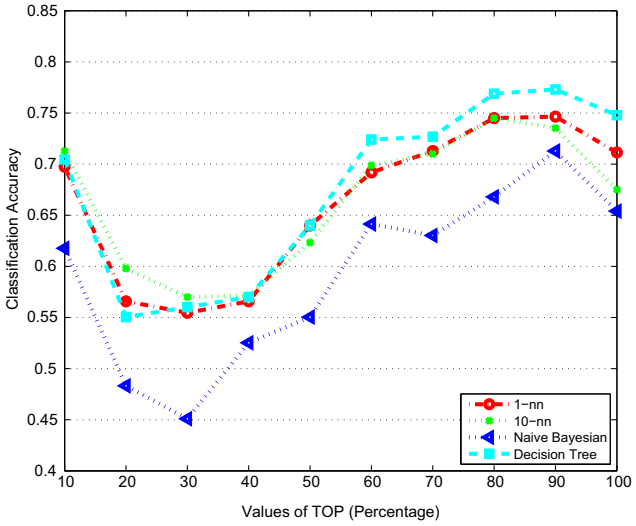
(c) Classification Accuracy with different values of LEN. Corpus 2 is utilized.



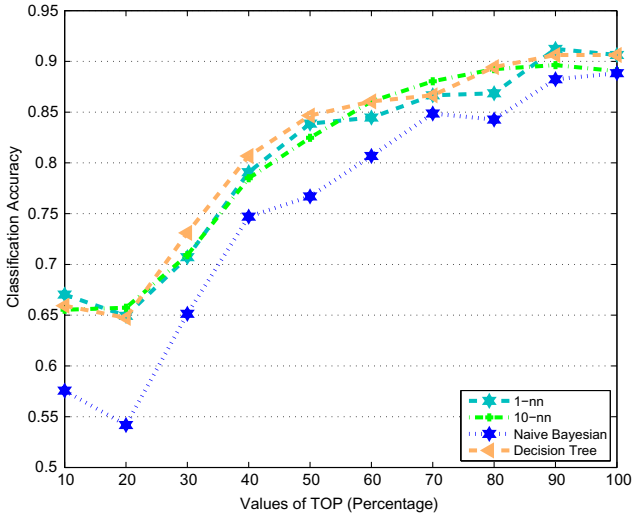
(d) Classification Accuracy with different values of LEN. Corpus 1 is utilized.

Fig. 9. Classification results using different LEN.

that detailed information in a text can be displayed. From Fig. 9(c) and (d), the overall accuracy using Corpus 1 is better than using Corpus 2. It is intuitively understood that it is an easier task to classify different word collocation categories of text of significantly different genres and topics than to separate slices of the same genre. From the accuracy curves it can also be observed that the



(a) Classification Accuracy using Corpus 2



(b) Classification Accuracy using Corpus 1

Fig. 10. Classification results using different TOP.

decision tree classifier delivers better classification results when LEN is set a small value.

The other parameter TOP, i.e., the proportion of the introduced words, also affects the classification results. Fig. 10 illustrates the classification results using different values of TOP. The value of LEN is set to 100. From the curves it can be observed that a "turning point" appears when TOP is set to 60% to 70%. When TOP is greater than 0.7, the accuracy is not significantly increasing when the value of TOP gets larger. Our experience shows that it is reliable using words that occupy the top 70% to 80% of the total number of words to calculate *FR* and merging words that occupy the last 10% of the total number of words to calculate *Intimacy*.

The feature *Intimacy* can also be calculated from *n*-gram language model. Another group of simulations are conducted to test the performances of *n*-gram and ILM. The visualization results are shown in Fig. 11, in which the horizontal axis is constantly the Feature *FRR* while the vertical axis is the Feature *Intimacy* which calculated by the corresponding language models. The red, green and blue curves represent bigram, linear ILM and Gaussian ILM, respectively. From Fig. 11, it can be observed that our ILM scheme outperforms the bigram models. Also, Gaussian ILM slightly outperforms the linear and polynomial ILMs.

6.7. Comparative analysis

6.7.1. FRR + intimacy v.s. style markers

Relatively few researches have been done for automated text characteristics analysis. To compare the effectiveness of other approaches to our proposed method, classification simulations with other popular sets of style markers relating to term-level and structural features are processed. Table 4 lists the selected style markers. All the listed style markers are English-dependent.

Style markers (Finn & Kushmerick, 2006; Kelih, Antić, Grzybek, & Stadlober, 2005; Kessler et al., 1997; Lim, Lee, & Kim, 2005) M1 ∼ M6 are considered as term level features while M7 M40 are considered as grammatical and semantic features. PCA is applied to extract the principle components from the extracted features of the above style markers. For our approach, the TOP and LEN are set to 0.7 and 100, respectively. The Gaussian ILM and Corpus 1 are chosen. The *k*-NN classifier is used to analyze the accuracy. The classification is displayed in Table 5. In Table 5, we can observe that our proposed approach can deliver better classification performances compared with other features. It is also worth noting that our proposed approach requires far less computational time than the grammatical and semantic based methods. It can also be observed that the more features we add, the higher accuracy we get. The computational cost, however, increases rapidly when more style markers are included, i.e., textual features, especially those grammatical and semantic features.

6.7.2. Use different language models to extract textual features

We have mentioned that the bigram language model is a special case of the ILM in the above. In this section, several different language models are compared with our proposed models. The ILM, trigram, bigram, bag of words (BOW), the Positional Language Models (PLMs) (Lv & Zhai, 2009) are employed for the text classification task. In this group of simulations, a 1-nn classifier is used for performing classification. The classification result is shown in Table 6. It can be observed that in most cases the Gaussian-ILM outperforms the other language models.

7. Discussions and conclusion

In this paper, extensive simulations are conducted to examine the performances of our text analysis method based on two proposed features: *FRR* and *Intimacy*. The *FR* provides single-term-level information when a baseline containing a sample vocabulary-using lexicon is given. By employing the same baseline, different *FRR* of single-term-level feature values can be calculated on the bases of FR. The *Intimacy*, a new concept modeling relationship between a word and others, is derived to capture the inter-term-level features of texts. Similar to *FRR*, different values of *Intimacy* represents different inter-term-level features. Promising visualization and classification results are delivered with the cooperation of the two extracted features. Compared to other style markers for different levels such as term and grammatical levels, *FR* and *Intimacy* are found to be more capable of representing useful word collocation characteristics.

It is worth noting that our proposed method can recognize the difference caused by the alternative word collocation habits among documents of the same genre information. The visualization results in general agree with humans perspective, that is the distance between texts from native English media and from native English newsgroup topics are closer than the distance from ESL English media and the newsgroup topics. It is noticed that certain text slices representing novels are found scattered into other categories of slices other than the novel category. This result makes sense because a complete novel is usually long enough to contain diverse word collocation scenes rather than a monotonous word-collocation type. All

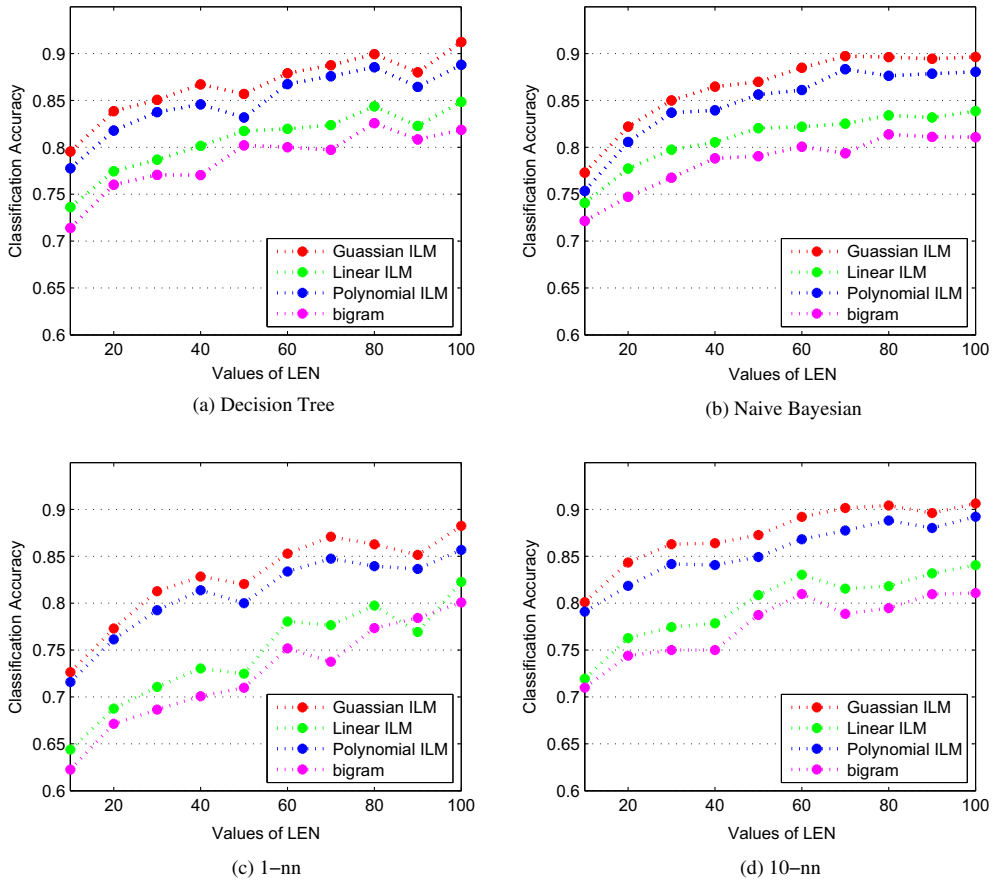


Fig. 11. Visualization results of *n*-gram language model and ILM.

Table 4
The selected style markers.

Index	Description
M1	Frequency of 30 most frequently used function words/total frequency of function words
M2	Frequency of 20 most frequently used punctuations/total frequency of punctuations
M3	Number of usual words/total number of words
M4	Number of distinct words/total number of words
M5	Number of imperative sentences/number of total sentences
M6	Number of question sentences/number of total sentences
M7 ∼ M23	Number of phrase/total number of phrases in a document for 17 phrases: NP, VP, AJP, AUXP, AVP, CONJP, SENT, IMPR, etc.
M24 ∼ M40	Average number of words per phrase for 17 phrases: NP, VP, AJP, AUXP, AVP, CONJP, SENT, IMPR, etc.

Table 5
Comparison of classification results: FRR and intimacy and other style markers.

Features	C-I	C-II	C-III	C-IV	C-V
FRR + *Intimacy*	0.821	0.691	0.734	0.853	0.776
M1 ∼ M4	0.691	0.638	0.600	0.764	0.745
M5 ∼ M6	0.312	0.525	0.553	0.711	0.727
M1 ∼ M6	0.729	0.677	0.659	0.770	0.732
M7 ∼ M23	0.675	0.689	0.681	0.764	0.743
M24 ∼ M40	0.742	0.687	0.711	0.860	0.781
FRR + *Intimacy* + M1 ∼ M4	0.844	0.751	0.769	0.862	0.801
FRR + *Intimacy* + M5 ∼ M6	0.841	0.698	0.757	0.854	0.795
FRR + *Intimacy* + M1 ∼ M6	0.848	0.764	0.772	0.870	0.809
FRR + *Intimacy* + M7 ∼ M23	0.857	0.776	0.780	0.871	0.813
FRR + *Intimacy* + M24 ∼ M40	0.864	0.791	0.787	0.882	0.819

these results prove the feasibility of our framework as well as our proposed ILM.

The decision tree classifier outputs have better accuracy on average among the referred implementations of classification methods, especially with small values of LEN and TOP. With the values of LEN and TOP increasing, the classification results using different classifiers are with little distinction. Another property of our framework is that the alternation of different baseline, regardless of the different size of the lexicons in baseline or the topics/genre of the original texts that is used for the generation of the lexicons, does not affect the final visualization or classification results. In other words, the visualization/classification results are relatively

Table 6
Comparison of classification results: ILM, trigram, bigram, bag of words (BOW), the positional language models (PLMs).

Language models	C-I	C-II	C-III	C-IV	C-V
Gaussian-ILM	0.821*	0.691	0.734	0.853*	0.776
Linear-ILM	0.817	0.689	0.730	0.837	0.757
Trigram	0.790	0.712*	0.731	0.817	0.781*
bigram	0.774	0.615	0.689	0.826	0.727
Gaussian-PLM	0.809	0.663	0.739*	0.825	0.764
Triangle-PLM	0.808	0.651	0.724	0.823	0.759
BOW	0.675	0.472	0.591	0.720	0.524

stable. By considering this property, we can use baseline that is generated from documents with smaller size. Therefore, fewer terms and word combinations will be contained in the baseline, which will result in reducing the computational cost.

At last, the computational complexity analysis indicates that our proposed method is not computationally demanding. In our investigations, we showed that our approach took less than 1 s to handle 1000 slices, each of which contains 50 sentences.[1] This indicates that handling large volume of documents will be computationally feasible.

References

Aha, D., Kibler, D., & Albert, M. (1991). Instance-based learning algorithms. *Machine learning, 6*(1), 37–66.

Biber, D. (1995). *Dimensions of register variation: A cross-linguistic comparison*. Cambridge Univ Pr.

Brainerd, B. (1974). *Weighing evidence in language and literature: A statistical approach* (Vol. 19). University of Toronto Press.

Brinegar, C. (1963). Mark twain and the quintus curtius snodgrass letters: A statistical test of authorship. *Journal of the American Statistical Association*, 85–96.

Burrows, J. (1987). Word-patterns and story-shapes: The statistical analysis of narrative style. *Literary and Linguistic Computing, 2*(2), 61–70.

Clement, R., & Sharp, D. (2003). Ngram and bayesian classification of documents for topic and authorship. *Literary and linguistic computing, 18*(4), 423–447.

Corrêa, R., & Ludermir, T. (2008). A quickly trainable hybrid som-based document organization system. *Neurocomputing, 71*(16), 3353–3359.

Fan, Y., Zheng, C., Wang, Q., Cai, Q., & Liu, J. (2001). Using naive bayes to coordinate the classification of web pages. *Journal of Software, 12*(9), 1386–1392.

Feldman, S., Marin, M., Ostendorf, M., & Gupta, M. (2009). Part-of-speech histograms for genre classification of text. In Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on, IEEE, 2009 (pp. 4781–4784).

Ferreira, A., & Figueiredo, M. (2012). An unsupervised approach to feature discretization and selection. *Pattern Recognition, 45*(9), 3048–3060.

Finn, A., & Kushmerick, N. (2006). Learning to classify documents according to genre. *Journal of the American Society for Information Science and Technology, 57*(11), 1506–1518.

Gelbukh, A., Sidorov, G., & Guzmán-Arenas, A. (1999). A method of describing document contents through topic selection. In *Proceedings of the string processing and information retrieval symposium & international workshop on groupware*. *SPIRE '99* (pp. 73). Washington, DC, USA: IEEE Computer Society. http://dl.acm.org/citation.cfm?id=519452.830776.

Goldwater, S., & Griffiths, T. (2007). A fully bayesian approach to unsupervised part-of-speech tagging. In *Annual meeting-association for computational linguistics* (Vol. 45, p. 744).

Han, E., Karypis, G., & Kumar, V. (2001). Text categorization using weight adjusted k-nearest neighbor classification. *Advances in Knowledge Discovery and Data Mining*, 53–65.

Joachims, T. (1999a). Making large-scale SVM learning practical. In B. Scholkopf, C. Burges, & A. Smola (Eds.), Advances in kernel methods - support vector learning (pp. 169–184). Cambridge, MA: MIT Press.

Joachims, T. (1999b). Transductive inference for text classification using support vector machines. In *Machine learning-international workshop then conference* (pp. 200–209). Morgan Kaufmann Publishers, Inc.

Joachims, T. (2001). A statistical learning learning model of text classification for support vector machines. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, pp. 128–136.

John, G. H., & Langley, P. (1995). Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*. *UAI'95* (pp. 338–345). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. http://dl.acm.org/citation.cfm?id=2074158.2074196.

Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation, 28*(1), 11–21.

Kelih, E., Antić, G., Grzybek, P., & Stadlober, E. (2005). *Classification of author and/or genre? the impact of word length*. *Classification–The Ubiquitous Challenge*. Heidelberg: Springer. 498-505.

Kessler, B., Numberg, G., & Schütze, H. (1997). Automatic detection of text genre. In *Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the European Chapter of the Association for Computational Linguistics, ACL '98, Association for Computational Linguistics, Stroudsburg, PA, USA* (pp. 32–38). doi:http://dx.doi.org/10.3115/976909.979622. URL `http://dx.doi.org/10.3115/976909.979622`

Kr, P., Mukherjee, A., Mitra, P., Basu, A., & Banik, A. (2008). A comparative study of the properties of emotional and non-emotional words in the wordnet: A complex network approach. In *Natural language processing*.

Lee, Y., & Myaeng, S. (2002). Text genre classification with genre-revealing and subject-revealing features. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, pp. 145–150.

Lim, C., Lee, K., & Kim, G. (2005). Multiple sets of features for automatic genre classification of web documents. *Information processing & management, 41*(5), 1263–1276.

Liu, Y., Wang, X., & Wu, C. (2008). Consom: A conceptional self-organizing map model for text clustering. *Neurocomputing, 71*(4), 857–862.

Luo, Q., Chen, E., & Xiong, H. (2011). A semantic term weighting scheme for text categorization. *Expert Systems With Applications, 38*, 12708–12716. http://dx.doi.org/10.1016/j.eswa.2011.04.058.

Lv, Y., & Zhai, C. (2009). Positional language models for information retrieval. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval* (pp. 299–306). ACM.

Manevitz, L., & Yousef, M. (2007). One-class document classification via neural networks. *Neurocomputing, 70*(7), 1466–1481.

Martínez Sotoca, J., & Pla, F. (2010). Supervised feature selection by clustering using conditional mutual information-based distances. *Pattern Recognition, 43*(6), 2068–2081.

Morton, A. (1965). The authorship of greek prose. *Journal of the Royal Statistical Society, Series A (General), 128*(2), 169–233.

Motter, A., de Moura, A., Lai, Y., & Dasgupta, P. (2002). Topology of the conceptual network of language. *Physical Review E, 65*(6), 065102.

Oliva, K., Květoň, P., & Ondruška, R. (2003). The computational complexity of rule-based part-of-speech tagging. In *Text, Speech and Dialogue* (pp. 82–89). Springer.

Peng, F., Schuurmans, D., & Wang, S. (2004). Augmenting naive bayes classifiers with statistical language models. *Information Retrieval, 7*(3), 317–345.

Peng, F., Schuurmans, D., Wang, S., & Keselj, V. (2003). Language independent authorship attribution using character level language models. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics*. *EACL '03* (Vol. 1, pp. 267–274). Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.3115/1067807.1067843. http://dx.doi.org/10.3115/1067807.1067843.

Petkova, D., & Croft, W. (2007). Proximity-based document representation for named entity retrieval. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management* (pp. 731–740). ACM.

Petrović, S., Šnajder, J., & Bašić, B. (2010). Extending lexical association measures for collocation extraction. *Computer Speech & Language, 24*(2), 383–394.

Quinlan, J. R. (1993). *C4.5: programs for machine learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Ratnaparkhi, A., et al. (1996). A maximum entropy model for part-of-speech tagging. In *Proceedings of the conference on empirical methods in natural language processing* (Vol. 1, pp. 133–142).

Riloff, E. (1995). Little words can make a big difference for text classification. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval. SIGIR '95* (pp. 130–136). New York, NY, USA: ACM. doi:http://doi.acm.org/10.1145/215206.215349.

Schonhofen, P. (2006). Identifying document topics using the wikipedia category network. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*. *WI '06* (pp. 456–462). Washington, DC, USA: IEEE Computer Society. doi:10.1109/WI.2006.92. http://dx.doi.org/10.1109/WI.2006.92.

Seretan, V. (2010). *Syntax-Based Collocation Extraction* (Vol. 44). Springer-Verlag New York Inc.

Seretan, V., Nerima, L., & Wehrli, E. et al. (2003). Extraction of multi-word collocations using syntactic bigram composition. In: Proceedings of the fourth international conference on recent advances in NLP (RANLP-2003) (pp. 424–431).

Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (1999). Automatic authorship attribution. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*. *EACL '99* (pp. 87–106). Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.3115/977035.977057. http://dx.doi.org/10.3115/977035.977057.

Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2000). Automatic text categorization in terms of genre and author. *Computational linguistics, 26*(4), 471–495.

Sun, A., & Loh, H. T. (2009). Imbalanced text classification: A term weighting approach. *Expert Systems With Applications, 36*, 690–701. http://dx.doi.org/10.1016/j.eswa.2007.10.042.

Tong, S., & Koller, D. (2002). Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research, 2*, 45–66.

Tweedie, F., & Baayen, R. (1998). How variable may a constant be? measures of lexical richness in perspective. *Computers and the Humanities, 32*(5), 323–352.

van Halteren, H., Tweedie, F., & Baayen, H. (1996). Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Computers and the Humanities, 28*(2), 87–106.

Wang, T.-Y., & Chiang, H.-M. (2011). Solving multi-label text categorization problem using support vector machine approach with membership function. *Neurocomputing, 74*(17), 3682–3689. http://dx.doi.org/10.1016/j.neucom.2011.07.001. http://dx.doi.org/10.1016/j.neucom.2011.07.001.

WenZhang Yoshida, T., & Tang, X. (2011). A comparative study of TF*IDF, LSI and multi-words for text classification. *Expert Systems With Applications, 38*, 2758–2765. http://dx.doi.org/10.1016/j.eswa.2010.08.066.

Wikipedia, N-gram – wikipedia, the free encyclopedia [Online; accessed 8-October-2011]. http://en.wikipedia.org/w/index.php?title=N-gram.

Zhang, H., Liu, G., Chow, T., & Liu, W. (2011). Textual and visual content-based anti-phishing: A bayesian approach. *Neural Networks, IEEE Transactions on* (99). 1-1.

[1] We use python to implement our framework.