Expert Systems with Applications 41 (2014) 7805–7819

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Mining language variation using word using and collocation characteristics

Peng Tang, Tommy W.S. Chow [*]

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

ARTICLE INFO

Article history:
Available online 2 June 2014

Keywords:
Language variation
Text mining
Frequency Rank Ratio
Overall Intimacy

ABSTRACT

Two textual metrics "Frequency Rank" (FR) and "Intimacy" are proposed in this paper to measure the word using and collocation characteristics which are two important aspects of text style. The FR, derived from the local index numbers of terms in a sentences ordered by the global frequency of terms, provides single-term-level information. The Intimacy models relationship between a word and others, i.e. the closeness a term is to other terms in the same sentence. Two textual features "Frequency Rank Ratio (FRR)" and "Overall Intimacy (OI)" for capturing language variation are derived by employing the two proposed textual metrics. Using the derived features, language variation among documents can be visualized in a text space. Three corpora consisting of documents of diverse topics, genres, regions, and dates of writing are designed and collected to evaluate the proposed algorithms. Extensive simulations are conducted to verify the feasibility and performance of our implementation. Both theoretical analyses based on entropy and the simulations demonstrate the feasibility of our method. We also show the proposed algorithm can be used for visualizing the closeness of several western languages. Variation of modern English over time is also recognizable when using our analysis method. Finally, our method is compared to conventional text classification implementations. The comparative results indicate our method outperforms the others.

1. Introduction

Different kind of text documents covering diverse topics and genres are created with the development of Web 2.0 (Chang Yang & Hong Lee, 2005; Thelwall & Buckley, 2013). The increasing size and amount of data make text processing more challenging than before, which requires more effective and efficient approaches of organization and presentation of huge amount of texts. In documents retrieval, data items are often organized by ranking texts according to the users' queries, or given keywords (Jansen & Pooch, 2001; Manning, Raghavan, & Schütze, 2008). The text classification and categorization, from another angle, measure and classify documents according to the document topics such as science, business, sports, and genres such as narrative, argumentative or hybrid texts (Bing Xue & Hua Zhou, 2009).

Besides these classification and categorization methods, drift of the semantic units in documents, i.e. language variation, leads to variations in languages, and can be very important clues that differentiate documents. For example, this type of cues can be helpful in telling whether an article was composed by native English speakers, or the article is informal or formal. Here, word using and collocation expressing the vocabulary and word combinations that generate semantic units can be very important clues that differentiate documents. The language variation delivered by the two textual characteristics, word using and collocation, different from genres or topics, are apparently affected by genre and topic information of specified documents, as well as the writers.

Previous studies on analyzing topic and genre characteristics share similar mechanism with our study. The original unstructured texts cannot be used as training data in empirical Natural Language Processing. First, training text data are first processed to generate structured data that machine can understand. Then, useful textural features are extracted from the preprocessed data. Next, machine learning approaches, combined with the extracted features, output trained models to handle users' requirements. Generally, text preprocessing techniques treat text documents as a set of arbitrary tokens which have little meanings or structures. The Term Frequency (TF) and Term Frequency-Inverse Document Frequency (TF-IDF), are the most important and popular features in text analysis such as document classifications and clustering. TF-IDF is applied to most document models like Vector Space Models or Probabilistic Models (Chowdhury, 2010; Rajan, Ramalingam,

[*] Corresponding author. Tel.: +852 34427756; fax: +852 27887791.
E-mail addresses: ptang@ee.cityu.edu.hk (P. Tang), eetchow@cityu.edu.hk (T.W.S. Chow).

http://dx.doi.org/10.1016/j.eswa.2014.05.018
0957-4174/© 2014 Elsevier Ltd. All rights reserved.

Ganesan, Palanivel, & Palaniappan, 2009). A wide range of document analysis methods also rely on TF. Statistical Language Model (SLM) techniques, for example *N*-gram (Clement & Sharp, 2003), have been used in many Natural Language Processing (NLP) applications, e.g., machine translation, automatic text generation, and information retrieval (IR).

In extracting appropriate textual features, genre-based approaches and topic-based measures exhibit similar ideas, in which features are usually derived from simple term-level statistics of texts. Among most of literatures on document analysis, it can be found that the features, like TF, TF-IDF, stop words and length of sentences, are the most widely used. For instance, word and sentence lengths have been employed in classification of text genre and authorship detection (Brinegar, 1963; Morton, 1965). Syntactic and semantic features (Finn & Kushmerick, 2006) have been used in genre text classification tasks. The POS tagging (POS) techniques can tag the words with diverse categories, like pronoun, verb, article, adverbs, corresponding to a particular part of speech based on its neighboring words in a natural sentence. Many previous researches have been established with the assistance of POS tagging. *N*-gram measures cooperating with POS tagged texts are used to evaluate the influence of syntax structure on classification results (Clement & Sharp, 2003). For example, POS tagging texts are utilized to detect genre information in documents (Kessler, Numberg, & Schütze, 1997). It has been proved that the POS approach outperforms the conventional TF-IDF features (Finn & Kushmerick, 2006). The concept of "style markers" are proposed as a set of measurable patterns in Biber (1995) and be used in many applications on text analysis. Kessler further categorized different style markers into four generic cues: structural cues, character-level cues, lexical cues, and derivative cues (Kessler et al., 1997). It is usually hoped that the document analysis are able to handle unrestricted text with low computational cost. Terms of high-frequency are also considered in document analysis. Using the TF and TF-IDF of the most frequently used words of a corpus has been investigated (Stamatatos, Fakotakis, & Kokkinakis, 2000; van Halteren, Tweedie, & Baayen, 1996). These studies indicate that the high frequency words are reliable discriminators for text analysis task. Textual features with deep structure and sematic meanings are also studied. For example, hierarchical concept dictionaries are employed to recognize and classify topics of document collections (Gelbukh, Sidorov, & Guzmán-Arenas, 1999). Ontology-based knowledge base is also proved efficient for capture text characteristics (Kitamura & Mizoguchi, 2003; Shing Lee, Juan Chen, & Wei Jian, 2003). The proximity-based information between words is also utilized to yield extra features. For example, Petkova and Croft propose a document representation model using the proximity between occurrences of entities and terms (Petkova & Croft, 2007). Neuhaus and Bunke used edit distance based string kernel to extract textual structural features (Neuhaus & Bunke, 2006). Lv and Zhai raised a so-called Positional Language Model using the proximity information among words, then use this model to propagate the word count (Lv & Zhai, 2009). There is a belief that authors can be an important factor in affecting results of documents analysis. Some research work has been done to identify the authorship of documents collections. Style markers are used in an authorship-based classification task, and a 50% or above accuracy has been reported when a 10-author corpus are processed (Stamatatos et al., 2000). Character-level *n*-gram features for authorship detection are also raised to deal with both western and Chinese texts, which deliver an overall accuracy about 75% (Peng, Schuurmans, Wang, & Keselj, 2003).

Machine learning approaches output trained models, using the extracted features, to meet users' requirements. Existing machine learning techniques, such as Naive Bayesian (Bum Kim, Soo Han, Chang Rim, & Hyon Myaeng, 2006), *k*-nearest neighbor (*k*-NN) (Han, Karypis, & Kumar, 2001; Tan, 2005), neural networks (NN) (Li, Song, & Park, 2009; Ou & Murphey, 2007; Rajan et al., 2009), genetic algorithms (Song, Li, & Park, 2009), and support vector machine (SVM) (Joachims, 1999b), deliver reasonable results. Unsupervised feature discretization and feature selection algorithms for feature reduction in documents are implemented in Ferreira and Figueiredo (2012), Shang et al. (2007) and Ogura, Amano, and Kondo (2009). A supervised feature selection approach based on conditional mutual information is also explored in text clustering (Martínez Sotoca & Pla, 2010). Peng et al. also Naive Bayes classifier and *n*-gram language models to conduct text classification (Peng, Schuurmans, & Wang, 2004). Self-organizing Map (SOM) are also utilized in document clustering (Corrêa & Ludermir, 2008). Joachims also use SVM to improve the performance of document classification (Joachims, 1999a).

The above reviewed approaches are able to output promising results on topic-based and genre-based analysis. Dealing with language variation in documents, as mentioned previously, however, differs from the topic-based and genre-based document classification. It is clear that generic and topic-based textual features are not accurate in describing the language variation cues in documents. Previous studies on such characteristic extraction highly rely on TF with respects to term occurrence and co-occurrence in a corpus. Seretan used syntactic patterns to recognize word collocations in different languages. This approach takes advantages of the recent advance in natural language parsing tools aiming to construct deep syntactic structures of texts (Seretan, 2010). An approach for selecting multi-word collocation candidates based on the syntactical bound collocation bigrams and patterns is also proposed in Seretan, Nerima, and Wehrli (2003). A set of diverse extension patterns are also defined using *n*-gram POS-tagged patterns for selecting word collocations (Petrović, Šnajder, & Bašić, 2010).

The conventional TF and TF-IDF are important single-term features for text mining and information retrieval. High frequency words, usually referring to function words and stop words, can be effective style markers for extracting word using characteristics of document genres. But in most topic and content classification studies, the high frequency words are excluded from bags of words; otherwise they can overshadow meaningful words due to their huge proportion. In addition, certain meaningless words usually count most in the whole document, while the most meaningful words usually occupy a relatively small proportion (Manning & Schütze, 1999). The language variation can then be impacted by both the topics and genres of documents. Hence the improvements, which can make the best of high-frequency terms and at the same time to balance high and low frequency words, are thus needed to be designed. In this study, to balance the high and low frequency words and avoid the sparsity problem caused by the low frequency words, we introduce a textual metric named Frequency Rank (FR). FR can deliver single-term-level information of documents, as well as balance the high and low frequency words.

Most of existing implementations for obtaining higher textual features other than single-term features heavily rely on *n*-gram features with POS taggers can deliver reliable grammatical or syntactic information. *N*-gram models are built on a basis that one word depends only on its last $n-1$ words while is independent from the rest of the words. This assumption significantly alleviates the computational problem of calculating all tokens in long sequences (Wikipedia, 2011). However, in practical applications, the value of n is usually set to less than 4, otherwise the computational demanding would become unbearably high. A small value of *n* means the relations between distant words are missing. Hence it limits the ability of extract inter-term level statistics, which is a major drawback for *n*-grams aiming to capture inter-term-level features. In this paper, Intimacy is introduced to measure popularity discriminating compatibility of a word and its neighbors, and its