Expert Systems With Applications 99 (2018) 1–11

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Discovering sentiment sequence within email data through trajectory representation

Sisi Liu, Ickjai Lee*

Information Technology Academy College of Business, Law and Governance James Cook University, PO Box 6811, Cairns, QLD 4870, Australia

ARTICLE INFO

Article history:
Received 8 August 2017
Revised 7 December 2017
Accepted 18 January 2018

Keywords:
Sentiment analysis
Traclus
Trajectory clustering
Sentiment sequence

ABSTRACT

Traditional document-level sentiment analysis fails to consider sentiment sequence within documents. This research paper proposes a novel perspective of sequence-based document sentiment analysis for discovering sentiment sequence and clustering sentiments for Email data. The proposed scheme of approach applies a trajectory clustering algorithm to Email trajectories transformed from sentiment features generated from SentiWordNet lexicon for discovering sentiment sequence within topic and temporal pattern distributions on the basis of trajectory clusters and their representative trajectories. Experiments conducted on real Email data provide evidence on proving the feasibility of the proposed technique and justifying the indispensability of sentiment sequence within documents in the determination of sentiment polarity.

1. Introduction

Sentiment analysis is generally considered as a multitudinous problem composed of several subproblems such as aspect extraction and grouping, feature extraction, and sentiment classification (Liu, 2015). Sentiment analysis tasks are categorised into document-level, sentence-level and aspect-level classification tasks in terms of granularity (Liu, 2015). Document sentiment analysis is the most fundamental and crucial granularity among the three in a broader perspective as it extracts opinions or sentiments from an entire document (Tang, Qin, & Liu, 2015a). For years, document-level sentiment analysis has focused on the refinement and development of feature extraction and sentiment classification techniques (Bhatia, Ji, & Eisenstein, 2015; Li, Liu, Du, Zhang, & Zhao, 2015; Liu, Lee, & Cai, 2016; Moraes, Valiati, & Neto, 2013; Tang, 2015; Tang, Qin, & Liu, 2015b). For text mining problems involving feature identification or extraction process, sequence is a prominent concept applied to various term weighting schemes such as *n*-gram model and Conditional Random Field (CRF) (Bao, Shen, Liu, Liu, & Zhang, 2004; Garcíahernández, Martíneztrinidad, & Carrascoochoa, 2006; Mao & Lebanon, 2007; Matsumoto, Takamura, & Okumura, 2005).

As a crucial factor for correctly identifying sentiments of a document, sequence within documents is to be recognised among characters for feature orders. Mao and Lebanon (2007) introduce the concept of local sentiment for the first time using modified CRF for analysing sentiment flow from sentences within a document. As an increasingly appealing subject, deep learning techniques, such as word embedding that incorporates sequence in feature selection process, have been applied to document sentiment analysis studies (Tang et al., 2015a; 2015b). Nevertheless, to the best of our knowledge, no study has been conducted on discovering sentiment sequence within documents in sentiment clustering or classification process.

To better justify the feasibility of the proposed technique, Email data is to be used as a source for experiments. Due to its unique characteristics of diversified length, implicit and formal language, and topic-oriented communication, Email is a better option than other social media data, such as microblog and review data. In detail, the length of microblog and review data is tend to be consistent since they are restricted by character limitations. On the other hand, Email data could be rather short or long, depending on whether it is an original Email or response. As for the second unique characteristic, implicitness in the use of sentiment words hardens the process of assigning sentiment polarity based on individual feature. In terms of topic orientation, Email messages are topic oriented through “reply and forward communications” such as “re” and “fw”. In addition, previous studies on Email sentiment analysis indicate difficulties of applying traditional sentiment analysis techniques directly to Email messages due to its unstructured format and richness in noise (Blanzieri & Bryl, 2008; Bogawar & Bhoyar, 2012; Hangal & Lam, 2011). Specifically, traditional techniques for other social media data mainly focusing on enhanc-

* Corresponding author.
E-mail addresses: Sisi.Liu@my.jcu.edu.au (S. Liu), Ickjai.Lee@jcu.edu.au (I. Lee).

https://doi.org/10.1016/j.eswa.2018.01.026
0957-4174/

ing emoticon and irregular expression detection are inadequate. In addition, past studies suggest the appropriateness of conducting document sentiment analysis using Email data as its meta-information, such as subject and sender, contains necessary information relevant to entity and opinion holder (Mohammad & Yang, 2011; Shen, Brdiczka, & Liu, 2013). Therefore, it is necessary to explore a novel approach for discovering sentiment sequence within documents, as well as conducting sentiment analysis task on Email data.

Apart from introducing Email data as a source of sentiment analysis task, the other essential component of this research is to propose a sequence-based sentiment clustering technique for improving document-level sentiment analysis results. The essence of sequence identification within documents in sentiment clustering lies in the way of extracting feature words. Feature-based document sentiment classification extracts frequency or weighting of features in a document for analysis. For example, the following two review fragments both convey positive sentiments in a broader perspective; however, they are not identical. Conventional document sentiment classification rules generally treat features in a static way without considering the interaction among documents, whereas two documents classified as positive may express different sentiments based on the position of features within documents as shown in the given example (positive → positive → positive → negative for the first review while positive → positive → negative → positive for the second review). On the contrary, this novel sequence-based sentiment analysis introduces the concept of sequence within documents in sentiment analysis considering the chronological presence of features, which minimises the opportunity of clustering sentences conveying the same sentiments into different categories.

```
"Overall, I like this hotel.
The room is clean and service is good.
But the food in the hotel café is awful."

"I would stay here again.
The location more than made up for any
problems we had with the room.
The staff is excellent and very friendly."
```

As a result of incorporating spatial information of text into feature extraction process, trajectory clustering, a means of clustering algorithm particularly developed for spatial dataset, is used in comparison with other traditional sentiment classification algorithms. Associated with the unique characteristics of Email data discussed above, trajectory clustering algorithm is capable of handling instances with various attribute lengths and assigning instances with a set of sentiments instead of sole polarity.

This paper proposes an approach for clustering sentiments in accordance with sentiment sequence in a trajectory representation using trajectory clustering algorithm. Four major contributions of the study are described as follows:

- introducing a novel direction for solving sentiment analysis task based on sequence within documents using a trajectory representation for Email sentiment pattern recognition;
- proposing a technique for transforming features into a 2-dimensional trajectory representation;
- discovering sentiment sequence within documents in temporal categories and clustering sentiment polarity using trajectory clusters;
- visualising sentiment sequence aligning with original Email messages represented in sentiment features as well as Email messages in topic and temporal distribution.

2. Related work

2.1. Sentiment analysis with sequence

Studies relevant to sentiment sequence involve document sequence and temporal sentiment analysis. In the previous few decades, some techniques have been proposed and developed for studying document sequence. Most studies conducted on document sequence focus on linguistic comparison and grammatic relationship. For instance, Wei and Chang (2007) develop a technique for discovering evolution patterns in sequential documents based on temporal relationships; Bao et al. (2004) apply semantic sequence kin and word sequence kernel to document copy detection. Furthermore, Jindal and Liu (2006) propose an approach known as class sequential rule mining with the combination of machine learning techniques for identifying comparative sentences.

Apart from studies on document sequence, Matsumoto et al. (2005) propose a novel feature selection technique using syntactic relations for the extraction of word subsequences and Mao and Lebanon (2007) develop a revised CRF for the prediction of ordinal sequence in word sets. However, traditional studies share problems such as no temporal information involved and limitations in discovering sentiment sequence. Temporal sequence is considered as another form of sentiment sequence in previous studies. An increasing quantity of studies has been undertaken on temporal sentiment analysis in the past few years as incorporating temporal feature with sentiment analysis is progressively appealing to researchers. For example, Fukuhara, Nakagawa, and Nishida (2007) implement a coefficient model for displaying patterns and relationships among topic, timestamp and sentiment using graphs; Diakopoulos, Naaman, and Kivran-Swaine (2010) display a temporal trend of topic and keyword extracted from news data generated from social media using an automated visualisation tool called Vox Civitas.

However, a review of past studies indicates that sentiment sequence within documents has not yet been studied, as well as a rare usage of Email data as a source for sentiment analysis and linkage between temporal clustering with sentiment sequence identification. Therefore, a methodology regarding the discovery of sentiment sequence within documents is to be developed.

2.2. Trajectory clustering

Trajectory is the representation of movement of mobile objects. Yao (2003) states that “spatiality and temporality are two unique dimensions in geography”. As mentioned earlier, this research conducts sentiment analysis in a sequence-based perspective for discovering sentiment sequence within documents. To achieve this aim, traditional way of transforming documents into features represented by vectors is inadequate. Since sentiment variation within documents is denoted by the position of feature in combination with its sentiment value, trajectory space that models the movement of spatio-temporal datasets is an ideal option for representation. Therefore, trajectory clustering algorithm is utilised in the proposed framework for clustering document sentiments represented in trajectories. By transforming text features into spatio-temporal features, sentiment sequence detection from spatio-temporal represented documents is different from general sentiment classification task. Therefore, traditional sentiment analysis algorithms are infeasible to solve the problem as most adaptable classifiers, such as Support Vector Machines (SVM) and Naive Bayes, are only able to handle points rather than sequence.

Clustering is a process of assigning a set of randomly generated objects into groups based on a certain similarity measurement. Trajectory clustering is specifically developed for grouping moving objects, known as spatial-temporal data, and for discover-