Information Sciences xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Information Sciences

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

Extracting opinionated (sub)features from a stream of product reviews using accumulated novelty and internal re-organization☆

Q1 Max Zimmermann[a,1], Eirini Ntoutsi[b], Myra Spiliopoulou[c,*]

[a] Swedish Institute of Computer Science, Stockholm, Sweden
[b] Ludwig-Maximilians-University of Munich, Germany
[c] Faculty of Computer Science, Otto-von-Guericke University of Magdeburg, D-39106 Magdeburg, Germany

ARTICLE INFO

Article history:
Received 3 April 2014
Revised 6 June 2015
Accepted 28 June 2015
Available online xxx

Keywords:
Product feature extraction
Opinion mining
Stream classification
Stream clustering
Opinionated streams
Stream mining

ABSTRACT

Opinion stream mining extends conventional opinion mining by monitoring a stream of reviews and detecting changes in the attitude of people toward products. However, next to the opinions of people on concrete products, *product features*—on which people also bestow their opinions—are equally important: such features appear on all products of a given brand and can deliver clues to product vendors on what improvements should be done in the next version of a product. In this study, we propose an opinion stream mining framework that discovers implicit product features and assesses their polarity, while it also monitors features and their polarity as the stream evolves. An earlier version of this framework has been presented in Zimmermann et al. (2013). The extended framework encompasses an additional mechanism that merges clusters representing similar product features. We report on extensive experiments for both the original framework and the extended one, using two opinionated streams.

© 2015 Published by Elsevier Inc.

1. Introduction

Opinion mining technologies are used in e-commerce to assess and predict the popularity of products, but also to identify those product features that people consider important and therefore express their opinion about them [2,3]. Since opinions are usually unstructured texts, product *features* are implicit, expressed by one or more words, e.g., "lens", "weight of camera" or "weight of the camera lens". Opinion stream mining is a recent research area that focuses on detecting drifts and bursts in opinionated document streams [4–6]. In this study, we present our work on discovering product features from a stream of opinionated data and on deriving and monitoring the sentiment associated with these features.

Opinion mining on product features is useful to decision makers, because it delivers insights on which properties are considered important for a product category, independently of the concrete product, and how these properties are perceived by the consumers, i.e. in a positive or negative way. Opinion stream mining on product features captures changes in the polarity of a given feature, as well as the evolution of the features themselves. In our approach, we recognize emerging features, forget unpopular ones and derive their sentiment from the sentiment of the individual documents referring to them.

☆ This paper comprises an extended version of [1].

* Corresponding author. Tel.: +493916758965.
E-mail addresses: max.zimmermann@sics.se (M. Zimmermann), ntoutsi@dbs.ifi.lmu.de (E. Ntoutsi), myra@iti.cs.uni-magdeburg.de (M. Spiliopoulou).

[1] Work done while with the Otto-von-Guericke University Magdeburg.

http://dx.doi.org/10.1016/j.ins.2015.06.050
0020-0255/© 2015 Published by Elsevier Inc.

Please cite this article as: M. Zimmermann et al., Extracting opinionated (sub)features from a stream of product reviews using accumulated novelty and internal re-organization, Information Sciences (2015), http://dx.doi.org/10.1016/j.ins.2015.06.050

UNCORRECTED PROOF

Our original method [1], denoted as T-SentiStream hereafter, is a framework with components for unsupervised and supervised stream learning. The unsupervised learning part is responsible for product feature extraction. It encompasses (a) a definition of *polarized feature* and a stream clustering algorithm that identifies and adapts features, defined as centroids in a two-level hierarchy of clusters, (b) a notion of *document importance* and a mechanism that depicts important documents from the stream, so that product features are computed only on the basis of the terms in these documents, and (c) a forgetting mechanism that removes features that are no more present in the document stream, giving space to emerging ones. The supervised learning part assigns polarity labels to features by cluster specific classification. In this work, we extend T-SentiStream [1] by a mechanism that merges clusters to allow for possibly less compact but potentially longer lasting features that are represented in larger sets of documents; we term this extension T-SentiStream*.

From the earlier version [1], we have taken over the related work discussion (Section 2); we have expanded the related work though w.r.t. to our extension. Furthermore, we took over the first three parts of Section 3, where we introduce our definitions and present the earlier T-SentiStream before extending it in Section 3.4 with a cluster merging component for better adaptation to the drifting stream. The experimental Section 4 is completely rewritten: (i) we run extensive experiments on T-SentiStream and we compare it with the T-SentiStream* extension; (ii) for the evaluation, we use the criteria introduced in [1], and we introduce further criteria to quantify the behavior of the algorithms; (iii) we vary the values of several parameters and observe their impact on the performance of T-SentiStream and T-SentiStream*; (iv) next to the originally used dataset, we consider one further, larger dataset. The last concluding section is also rewritten.

2. Related work

Relevant to our work are studies on sentiment analysis over streams, on feature extraction from a stream of opinionated documents and on stream clustering.

2.1. Sentiment analysis over streams

Sentiment analysis aims at recognizing the sentiment associated with a review, usually either as positive or negative. A typical goal is to build a model that predicts the sentiment of unseen documents, given a set of documents with known sentiment. Pang and Lee [7] provide a comprehensive overview of the area, together with an evaluation of the performance of different machine learning methods—Naive Bayes, Maximum Entropy classification and support vector machines. Lately, there is a large amount of work on sentiment analysis in social media and especially in Twitter, see e.g., [8].

Even in a static setting, sentiment analysis is a difficult task due to e.g., ambiguity of the language, the usage of abbreviations and colloquial language, and due to unbalanced datasets. The problem becomes even more challenging in a dynamic setting, where new opinionated documents arrive over time, so that the models are subjected to drifts and shifts. Another challenge is the unavailability of labeled documents over the whole course of the stream.

One of the first approaches on sentiment analysis over a stream was proposed by Silva et al. [6]. Their method consists of a rule-based classifier extracted upon a small seed of labeled documents. The seed is gradually expanded with new relevant documents in order to deal with the changes in the underlying stream population. We also use a small initial seed for learning, however instead of a single generic classifier as in [6], we train feature-specific classifiers.

Bifet and Frank [4] investigate sentiment classification on a stream of tweets. They consider unbalanced classes with drifts and shifts in the class distribution, with the requirement of quick response under memory constraints. Closest to our approach is their follow-up framework [5] that consists of (i) a twitter filter to convert tweets into TF-IDF vectors, (ii) an adaptive frequent itemset miner that stores the frequency of the most frequent terms and (iii) a change detector that detects changes in the frequency distribution of the items. The framework monitors changes in the frequency of words. We also propose a framework for stream learning over opinionated documents, but our objective is to first identify the features and subfeatures of the products we study, and then to assess the polarity associated with them. Moreover, we allow for new features.

In [9] we propose an adaptive semi-supervised opinion stream classification algorithms that adapts itself in two ways: through forward adaptation, by incorporating into the training set new documents that convey enough information for the classification task, and through backward adaptation, by gradually eliminating outdated documents from the model. Replacing the supervised stream learning task with a semi-supervised one is reasonable, since the availability of fresh labeled documents in the stream should not be taken for granted. We address this issue in [10]. In this paper, our emphasis is on enhancing the unsupervised part of the sentiment-learning framework, though.

2.2. Feature extraction from reviews

Feature extraction and monitoring from an opinionated stream is a young subject. For feature extraction on a static set of reviews, Liu identifies four research subtopics [11], of which the identification of frequent nouns and of noun phrases are closest to our research.

Long et al. [12] extract core words for a product feature, compute their frequencies, estimate their distance to other words and use these distance values in order to acquire further words related to the specific feature. Zhu at al. [13] consider the frequency of terms that contain other terms. Mukherjee et al. [14] extract features and relationships among them: for feature extraction, they consider all nouns. In contrast, we suppress very frequent nouns with the help of TF-IDF weighting. Moghaddam and Ester

Please cite this article as: M. Zimmermann et al., Extracting opinionated (sub)features from a stream of product reviews using accumulated novelty and internal re-organization, Information Sciences (2015), http://dx.doi.org/10.1016/j.ins.2015.06.050

UNCORRECTED PROOF

Download English Version:

https://daneshyari.com/en/article/6857563

Download Persian Version:

https://daneshyari.com/article/6857563

Daneshyari.com