Chapter 5
Analyzing Twitter User-Generated Content Changes

Luca Cagliero
Politecnico di Torino, Italy

Luigi Grimaudo
Politecnico di Torino, Italy

Alessandro Fiori
Institute for Cancer Research at Candiolo, Italy

ABSTRACT

User-generated content (UGC) coming from social networks and online communities continuously grows and changes. By analyzing relevant patterns from the UGC, analysts may discover peculiar user behaviors and interests which can be used to personalize Web-oriented applications. In the last several years, the use of dynamic mining techniques has captured the interest of the research community. They are focused on analyzing the temporal evolution of most significant correlations hidden in the analyzed data. However, keeping track of all temporal data correlations relevant for user behaviors, community interests, and topic trend analysts may become a challenging task due to the sparseness of the analyzed data.

This chapter presents a novel data mining system that performs dynamic itemset mining from both the content and the contextual features of the messages posted on Twitter. Dynamic itemsets represent the evolution of data correlations over time. The framework exploits a dynamic itemset mining algorithm, named HiGen Miner, to discover relevant temporal data correlations from a stream of tweet collections. In particular, it extracts compact patterns, namely the HiGens, that represent the evolution of the most relevant itemsets over consecutive time periods at different abstraction levels. A taxonomy is used to drive the mining process and prevent the discarding of knowledge that becomes infrequent in a certain time period.

Experiments, performed on real Twitter posts, show the effectiveness and the usability of the proposed system in supporting Twitter user behavior and topic trend analysis.

DOI: 10.4018/978-1-4666-4213-3.ch005
INTRODUCTION

In recent years, social networks and online communities have become a powerful source of knowledge. Social network users are used to publish and continuously update multimedia resources, posts, blogs, etc. Actions undertaken by Web users reflect their habits, personal interests, and professional skills. Hence, the analysis of the user-generated content coming from social networks has received an increasingly high attention in several application contexts. For instance, data mining techniques have already been applied to recommend personalized services and products based on social annotations (Wang et al., 2010; Shepitse et al., 2008; Xue et al., 2009), organize and make social knowledge accessible (Kas- neci et al., 2009), and perform email spamming based on social networks (Lam et al., 2007). In particular, data mining from UGC published on the popular Twitter microblogging Website has achieved promising results in the analysis of most notable user behaviors (Li, Guo, & Zhao, 2008; Mathioudakis & Koudas, 2010) and topic trends (Cheong & Lee, 2009).

Twitter user-generated content consists of a large collection of short textual messages (i.e., the tweets) posted by Web users and their contextual information (e.g., publication time and date). Since the Twitter user-generated content and contextual data continuously evolve over time, a relevant research issue is the application of data mining techniques to discover most significant pattern changes. Dynamic itemset mining (Agrawal & Psaila, 1995) entails discovering itemsets that (i) frequently occur in the analyzed data, and (ii) may change from one time period to another. The history of the main itemset quality indexes reflects the most relevant temporal data correlation changes. However, the sparseness of the analyzed data makes dynamic itemset mining from UGC a challenging task. In fact, potentially relevant itemsets discovered at a certain time period are likely to become infrequent (i.e., their support value becomes lower than a given threshold) in at least another one. Hence, the information associated with the discovered itemsets may be lost, unless lowering the support threshold and mining a huge amount of other (potentially redundant) itemsets.

This chapter presents the TwiChI (Twitter Change mIner) system that aims at supporting experts in the analysis of Twitter UGC changes targeted to user behavior and topic trend analysis. TwiChI exploits the Twitter Application Programming Interfaces (APIs) to retrieve both tweet textual contents and their contextual features (i.e., publication date, time, place). Data crawling is continuously executed using the Twitter Public stream endpoint to track the temporal evolution of the frequent itemsets occurring in the analyzed data. The retrieved data is analyzed by the proposed HiGen Miner algorithm (Cagliero, 2011), which discovers compact patterns, named the History Generalized Patterns (HiGens). HiGens represent the evolution of frequent itemsets across consecutive time periods. To avoid the discarding of rare but potentially relevant knowledge, itemsets that become infrequent in a certain time period with respect to the minimum support threshold are generalized at a higher level of abstraction by exploiting a taxonomy (i.e., a set of is-a hierarchies built on data items). A generalized version of a traditional itemset is an itemset that represents the same knowledge at a higher level of aggregation according to a given taxonomy (Agrawal & Sri-kant, 1995). Hence, the knowledge associated with itemsets that rarely occur at certain time periods is still maintained by replacing the low level itemset versions with their frequent generalizations with least abstraction level.

Consider, for instance, tweet messages and related contextual information (e.g., publication time, geographical location) retrieved in the period January and February 2012. The tweet collection may be partitioned into two distinct monthly time periods. Analyzing the two sub-collections, the TwiChI framework may discover the HiGens reported in Table 1. Suppose that