Discovering and Tracking Events From News, Blogs and Microblogs on the Web

Sen XU
Department of Geography, the Pennsylvania State University, PA, U.S.A.

Abstract. Using three data sources, news, blogs, and microblogs, this study proposes a framework for discovering and tracking events embedded in free form online text. Existing methods for text mining are discussed for the three sources. Because three sources have different perspective, event analysis, region-topic model and rare keywords are proposed respectively. In order to integrate three data sources, extraction of temporal and spatial information is carried out and used as link to relate information across different sources.

Keywords. Text mining, geocoding, space-time analysis, event analysis.

1. Introduction

As the Web has become an inseparable part of modern life, it has become the platform for large scale information exchange. From an information analyst’s perspective, the Web is also the largest data repository. Every day, news providers broadcast events happening all over the world (mostly in forms of text, some in forms of photo or video) on the Web; Millions of internet users share their opinions and experience by posting to blogs or microblogs (e.g., twitter). The abundance of information has become more overwhelming than ever, which also brought up research opportunities to analyze and discover patterns within the large scale data sets. Although these online text data are free-form text, they also carry spatial and temporal footprints. Challenges are how to extract and analyze the spatial and temporal information embedded in the large scale data set. In this study we look into three data sources on the Web, news, blogs and microblogs and search for a solution to integrate the three sources for discovering and tracking events and event patterns.

Events in this study are referring to activities that are of interest to the researcher which have spatial and temporal footprints. In political science, event analysis has been one of the main research focuses and is defined as “categorical data showing who did what to whom (and when) derived from news reports” [1]. Methods developed in detecting political events are adopted and integrated with other feature extraction techniques to adapt to the more flexible text form in online text. The three data sources have different focus and text properties, text mining methods such as topic modeling as well as rule-based information extraction were used to process for different purposes. The final goal is to integrate the three sources and develop a workflow for analyst to discover unusual events as well as event patterns with regard to both spatial and temporal attributes.
2. Background

News articles have been used for event analysis [1, 2]. TABARI (Textual Analysis By Augmented Replacement Instructions), a software developed using dictionary-based event detection, can achieve close to human coder’s performance for political events using a predefined taxonomy (CAMEO). However the analysis is mostly focusing on the event type itself, lacking a thorough examination of the spatial patterns or temporal co-occurrences.

Blogs have attracted research interest for opinion mining, for example, to discover key topics from product reviews [3]. Recently a number of interesting researches were conducted on blogosphere. Meme-tracking [4] examined popular phrases (e.g., “lipstick on a pig”) and discovered a “heart-beat” pattern where the frequency of popular phrases oscillates between blogs and mainstream news. Other related research topics include sentiment analysis [5] and predicting movie success [6].

Microblogs (e.g., twitter), have provided a way for everybody to conveniently broadcast to the entire Web. As a short status updater, twitter user often posts information about events happening around the user, which could provide information where news reporter cannot reach on time. On May 1st, 2011, a twitter user with ID “ReallyVirtual” unknowingly reported the Osama Bin Laden raid (see http://tweetlibrary.com/damon/osamaraidlivetweets). It is not hard to imagine that if twitter data can be filtered and interpreted effectively, it can be helpful for various purposes. Monitoring influenza [7] is a successful use case.

The three growing data sources have attracted researchers from various disciplines, and have initiated an interesting research field. However, the analysis of existing studies usually focuses on one or two specific topic. Without a theoretical framework that covers the three sources, it is difficult to expand individual methodologies from existing studies for an integrated analysis. By examining the data, it is not difficult to see that spatial and temporal information are two fundamental properties that are expressed universally across all sources. How to extracting space and time and integrating them into event analysis across the three sources is a demanding challenge.

3. Research Goal

The overall system for achieving event analysis using web data is depicted in Figure 1. This study aims at the data collection and processing parts. Given characteristics of event information available from the three web data sources, the goal for this study is to find out how to design an event-ontology-based method that can better serve spatio-temporal event analysis than current available entity extraction methods. A set of tools for extracting events information from news RSS feeds, blogs and microblogs should be designed, implemented and evaluated. Using a case study (on Yemen Terrorism events versus social economical events), how the event-ontology based data collection method support the analytical process of spatio-temporal event analysis should also be evaluated.
4. Methods

Each data source requires tailored data processing. In data collection, a RSS feeds crawler is built for automatic fetching news data every day. A RSS feed contains <Title>, <Description>, <pubDate> and <link> to the full article. While temporal information is directly provided in pubDate, spatial information, actors (“who” in the news), and event types still needs to be extracted from the text <Title> and <Description>. Methods in TABARI are adopted for automatic code actors and event types using a comprehensively developed dictionary [1, 2]. The dictionary consists of verbs and actors, where verbs are associated with event types and actors are nouns of organization or people’s names. As a case study for event analysis, news feeds on Yemen are collected and currently I am working on extending event type and dictionary based the CAMEO codebook [2]. This case study aims at finding temporal patterns [9] for terrorism events and social-economical events.

Blogosphere has been indexed by Spinn3r, which provides access to the content and comments of A-list blog posts. The size of indexed blogs is over 40 million (reported by Spinn3r, May 2011). Blogs provides a different data perspective than news: news aims to provide an unbiased report on current events; blogs aims at providing opinions, describing experience, knowledge, personal emotions, etc. Nonetheless temporal and spatial information can also be found in blogs. Region-topic models [8] can be used to reveal word association with regions; considering daily updated blogs as a corpus, region-topic can be used to monitor the topic changes associated with a particular region by time.

Microblogs refer to a short (e.g., in twitter, less than 140 characters) form of blogs, which has characteristics of being prompt and concise. As microbloggers is a very large use group (twitter has over 200 million users as of March, 2011), discovering
tweets of interest can be challenges. Twitter API allows for query searches, and it also offers the registered location of tweet authors and time stamp of tweets. Developing a rare keywords list (e.g., "helicopter", "earthquake", "gunshot") and spatially restrict search range can help analyst to discover uncommon events firsthand reported by microbloggers.

Geolocated blogs shares similar property with VGI (volunteered geographic information [10]). As pointed out by Goodchild, the credibility of VGI is worth evaluating. Geolocated blogs (especially microblogs) can be difficult to validate by the data itself. However, linking among different data sources can help validate if an event has happened or were falsely reported.

5. Future Work

The proposed workflow is implemented for Yemen social economic events detection and to be analyzed with Yemen terrorism events. Using collected Yemen event data as a case study, the next step is to evaluate the design of data collection using the proposed methods. Potential extension on different topics than terrorism, such as emergency response, nature disaster and epidemic outbreak, regarding suitability of analysis methods and data source, could also be evaluated.

References