Available online at www.sciencedirect.com
ScienceDirect
Journal of Informetrics 1 (2007) 131–144
Journal of INFORMETRICS
An International Journal
www.elsevier.com/locate/joi

Generating overview timelines for major events in an RSS corpus

Rudy Prabowo [a,*], M. Thelwall [a], Mikhail Alexandrov [b]

[a] *School of Computing and Information Technology, University of Wolverhampton, Wulfruna Street, WV11SB Wolverhampton, UK*
[b] *Autonomous University of Barcelona, Barcelona, Spain*

Received 31 August 2006; received in revised form 19 October 2006; accepted 23 October 2006

Abstract

Really simple syndication (RSS) is becoming a ubiquitous technology for notifying users of new content in frequently updated web sites, such as blogs and news portals. This paper describes a feature-based, local clustering approach for generating overview timelines for major events, such as the tsunami tragedy, from a general-purpose corpus of RSS feeds. In order to identify significant events, we automatically (1) selected a set of significant terms for each day; (2) built a set of (term–co-term) pairs and (3) clustered the pairs in an attempt to group contextually related terms. The clusters were assessed by 10 people, finding that the average percentage apparently representing significant events was 68.6%. Using these clusters, we generated overview timelines for three major events: the tsunami tragedy, the US election and bird flu. The results indicate that our approach is effective in identifying predominantly genuine events, but can only produce partial timelines.

Keywords: Feature selection; Clustering; Overview timeline

1. Introduction

The task of identifying significant events from real time news feed data is a standard one in data mining and event detection and tracking (Allan, Papka, & Lavrenko, 1998b; Yang, Pierce, & Carbonell, 1998). The Internet now hosts a range of readily accessible information formats that are new candidates for event detection, and these may come to replace or supplement traditional types, or may give rise to new event detection applications. Really simple syndication (RSS) is one such technology and has already become a widely used standard: it allows blogs and news sources to post-timely information to subscribers, for example, hourly or daily summaries of the most recent updates. RSS feeds have great potential to be used for public-opinion gathering (Glance, Hurst, & Tomokiyo, 2004; Gruhl, Guha, Liben-Nowell, & Tomkins, 2004), mainly because of the large numbers of blog authors maintaining sites with RSS feeds, although bloggers are not typical citizens (Adar, Zhang, Adamic, & Lukose, 2004; Lin & Halavais, 2004) and have a wide variety of motives (Herring, Scheidt, Bonus, & Wright, 2004). In addition, the concise RSS formats allow relatively low-bandwidth data gathering, even for a large number of different sources. When a major (or world) event, such as the Asian tsunami (26/12/2004), occurs, RSS feeds could therefore be used to generate an overview timeline of the event.

Our contributions are to develop an automatic method to achieve the following using RSS data.

* Corresponding author. Tel.: +44 1902 518584; fax: +44 1902 321478.
E-mail addresses: Rudy.Prabowo@wlv.ac.uk (R. Prabowo), Mike.Thelwall@wlv.ac.uk (M. Thelwall), dyner1950@mail.ru (M. Alexandrov).

1751-1577/$ – see front matter
doi:10.1016/j.joi.2006.10.002

(1) Find daily sets of significant terms (either nouns or noun phrases) which maybe associated with important events, i.e. the most discussed happenings (Section 3).
(2) Use the significant terms to build a set of (term–co-term) pairs and cluster the pairs. The clusters are our candidates for the day's significant events (Section 4).
(3) Generate overview timelines for major events by sorting the clusters by date (Section 5).

In this paper, we are primarily interested in the precision of the clusters in (2). More specifically, we assess the extent to which human judges agree that the automatically generated clusters genuinely describe a single event. A human-based evaluation was important to discover whether the results could be understood by potential end users, i.e. human interpreters.

To illustrate the timeline generation, three major events, 'tsunami tragedy', 'US election' and 'bird flu spreading', were selected as our case studies. Tables 5–7 show the generated timelines for each major event. Each timeline refers to one particular major event along with many related, subsequent events.

2. Related work

This section reviews existing work in the area of (1) term selection, (2) topic and event detection and tracking (TDT) and (3) timeline generation.

2.1. Term selection

Given a set of terms (e.g. words, word stems, nouns and noun phrases) in a document collection, selecting the most significant terms is the first step. This stage is common to a range of specific tasks, including information retrieval (IR), automatic text classification and time series analysis. The selected terms represent document features in the form of a term vector: a list of the most significant terms from the document and their frequencies in the document.

Either *tf·idf* (Salton & McGill, 1986) or *lnu* weighting (Singhal, Buckley, & Mitra, 1996) can be applied to assign each term a value which estimates its significance. These formulae take into account both local and global term frequency. In IR, the assigned value is then used as a starting point to: (1) compute the similarity between the documents available in the corpus and a user query and (2) rank search results in order of relevance to a user query. In an ideal scenario, each term should be assigned a degree of significance, such that an IR system can achieve a high precision level at 100% recall (Baeza-Yates & Ribeiro-Netto, 1999; Belew, 2000). It is not suitable for our event detection task, however, because important events may be identified through single highly significant terms (Swan & Allan, 2000).

The three formulae that have previously been used to identify significant events in blog or RSS corpora are variants of *tf* and *tf·idf* (Gruhl et al., 2004; Glance et al., 2004; Thelwall, Prabowo, & Fairclough, 2006). The formulae do not, however, depend on the full document space, but on a fixed time period as a time window of observations, and are used to quantify the 'burstiness' of a term within the fixed, short time period, i.e. the degree of importance of terms within the time period. The result changes if another time window is used, for example, 1 week earlier or later. While this feature is useful to keep track of the burstiness of terms for different time windows, it is less suitable for the initial identification of significant events. For this, the degree of significance of a term over a long period of time is required, e.g. 1 year.

The commonly used formulae for identifying significant terms in the area of automatic text classification are: χ^2, Mutual Information (MI) and Information Gain (I) (Sebastiani, 2002). Swan and Allan (2000) use a χ^2-based method to determine the degree of significance of terms on given dates. Nevertheless, it is not yet clear whether this is the best method for all types of data, and in the context of RSS, it is also worth harnessing the power of the Information Gain method (Prabowo & Thelwall, 2006).

2.2. TDT

In the context of the TDT task, an event is defined to be something that happens at a specific time and place, whereas a topic is defined more widely as a seminal event or activity, along with all directly related events and activities (Allan, Carbonell, Doddington, Yamron, Yang, 1998a). The term 'story'; is often used to describe the natural unit of text in which the information arrives, such as a single newswire report. The topic detection and tracking tasks focus on