Available online at www.sciencedirect.com
SCIENCE DIRECT®
Information Processing and Management 42 (2006) 1491–1512
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

A comparison of feature selection methods for an evolving RSS feed corpus

Rudy Prabowo *, Mike Thelwall [1]

School of Computing and Information Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1SB, UK

Received 16 March 2006; accepted 16 March 2006
Available online 16 May 2006

Abstract

Previous researchers have attempted to detect significant topics in news stories and blogs through the use of word frequency-based methods applied to RSS feeds. In this paper, the three statistical feature selection methods: χ^2, Mutual Information (*MI*) and Information Gain (*I*) are proposed as alternative approaches for ranking term significance in an evolving RSS feed corpus. The extent to which the three methods agree with each other on determining the degree of the significance of a term on a certain date is investigated as well as the assumption that larger values tend to indicate more significant terms. An experimental evaluation was carried out with 39 different levels of data reduction to evaluate the three methods for differing degrees of significance. The three methods showed a significant degree of disagreement for a number of terms assigned an extremely large value. Hence, the assumption that the larger a value, the higher the degree of the significance of a term should be treated cautiously. Moreover, *MI* and *I* show significant disagreement. This suggests that *MI* is different in the way it ranks significant terms, as *MI* does not take the absence of a term into account, although *I* does. *I*, however, has a higher degree of term reduction than *MI* and χ^2. This can result in loosing some significant terms. In summary, χ^2 seems to be the best method to determine term significance for RSS feeds, as χ^2 identifies both types of significant behavior. The χ^2 method, however, is far from perfect as an extremely high value can be assigned to relatively insignificant terms.

Keywords: Feature selection; Chi-square; Mutual information; Information gain

1. Introduction

Rich Site Syndication (RSS) is an XML format for publishing concise information updates. It is mainly used by news sites to publish summaries of their latest stories and by blogs for summaries of their latest postings. Both researchers and commercial companies have noticed that blogs and RSS feeds have the potential to be used for public-opinion gathering or marketing purposes, and hence there has been a drive to develop

* Corresponding author. Tel.: +44 1902 518584; fax: +44 1902 321478.
E-mail addresses: rudy.prabowo@wlv.ac.uk (R. Prabowo), m.thelwall@wlv.ac.uk (M. Thelwall).
[1] Tel.: +44 1902 321470; fax: +44 1902 321478.

0306-4573/$ - see front matter
doi:10.1016/j.ipm.2006.03.018

effective tools and techniques for the automatic analysis of RSS feeds (e.g., Gill, 2004; Pikas, 2005). Previous researchers have used word/noun/noun phrase time series analysis methods with simple frequency statistics to identify significant topics (Glance, Hurst, & Tomokiyo, 2004; Gruhl, Guha, Liben-Nowell, & Tomkins, 2004). However, there is an established body of research for identifying the most significant words in collections of documents, a task called 'feature selection'. It is therefore logical to assess whether any of the established statistical feature selection methods can help the identification of significant topics in RSS feeds. A first stage of this process is to compare these methods to assess which are the most suitable for this new data source.

In this paper, a 'term' is a noun or noun phrase and a 'feature' is a term that is judged to be significant within a collection of documents. Feature selection methods such as χ^2, Mutual Information (*MI*) and Information Gain (*I*), have been commonly used in different application domains. One example is automatic text classification (Yang & Pedersen, 1997): determining document categories based upon a set of significant terms representing the document features (Sebastiani, 2002). The role of feature selection in this context is in condensing documents by removing redundant words, in order to speed document classification without reducing classification quality. Feature selection that is too 'aggressive', in terms of removing too many words, will result in poor document classifications. A very different example is Topic Detection and Tracking (TDT) (Allan, Papka, & Lavrenko, 1998; Yang, Pierce, & Carbonell, 1998), which focuses on identifying a new event/topic, and tracking the previously identified event/topic with regard to new incoming stories. This may be achieved by identifying and clustering collections of related terms, but other methods are also used, such as Information Extraction (e.g., Luo, 2004). In comparison to document classification, TDT implicitly requires much more aggressive feature selection because its purpose is to identify significant events across documents rather than to capture the essence of individual documents. A relevant application of TDT is the automatic generation of overview timelines (Swan & Allan, 2000) through determining which terms are significant over a given time period.

Despite previous research into term selection methods, there is no clear indication of the superiority of any particular method for all types of data: each has its own strengths and weaknesses. Yang and Pedersen (1997), supported by Sebastiani's (2002) automatic text categorisation review article, suggest that *I* and χ^2 have similar performance in supporting document classification, with both being significantly better than *MI*. Of these, χ^2 seems to be gaining support in practice for classification, perhaps for its calculation efficiency (e.g., Ng, Goh, & Low, 1997) although selecting terms by using χ^2 has also been criticised for being unreliable when the cell value is less than 5 (Dunning, 1993). There is no strong evidence, however, to suggest which method is the most effective for the less studied TDT task of selecting significant terms from document collections, although Swan and Allan (2000) have adopted χ^2 for this purpose. It is not clear that methods which work best for document classification also work best for TDT, for example because for classification purposes it can be useful to eliminate terms based upon a high degree of association with remaining terms, which is not a consideration for TDT (Swan & Allan, 2000). Moreover, each method uses a probabilistic function based upon assumptions about the distribution of the data, such as independence, which are violated in practice to varying degrees (Cooper, 1995). Hence experiments are required with each new type of data source and for each new type of task to assess the strengths and weaknesses of the leading methods in practice.

In this paper the Mutual Information (*MI*), χ^2 and Information Gain (*I*) feature selection methods are evaluated for an evolving RSS feed corpus in order to decide which is the most suitable for identifying features that are significant across a number of documents within the collection (the TDT type of task). In particular, the suitability of the three methods for selecting significant features on a given date is assessed. Term Strength (TS) and Document Frequency (DF) Thresholding are significantly different from the other three methods, as these two methods only consider the document space, rather than individual parameters, such as category or date, and require a training corpus. TS, in particular, requires computationally expensive document clustering (Yang & Pedersen, 1997). For these reasons, they were rejected as inappropriate for this evolving data set.

In addition, the assumption that very large values indicate highly significant terms is investigated; and the extent to which the three methods agree with each other about the significance of a term on a given date. This is particularly relevant for TDT. An evaluation method that is inspired by the way in which conference papers are reviewed is also proposed and used. A paper may be accepted for publication if two or three