Measuring the Interestingness of News Articles

Raymond K. Pon
University of California - Los Angeles, USA

Alfonso F. Cardenas
University of California - Los Angeles, USA

David J. Buttler
Lawrence Livermore National Laboratory, USA

INTRODUCTION

An explosive growth of online news has taken place. Users are inundated with thousands of news articles, only some of which are interesting. A system to filter out uninteresting articles would aid users that need to read and analyze many articles daily, such as financial analysts and government officials.

The most obvious approach for reducing the amount of information overload is to learn keywords of interest for a user (Carreira et al., 2004). Although filtering articles based on keywords removes many irrelevant articles, there are still many uninteresting articles that are highly relevant to keyword searches. A relevant article may not be interesting for various reasons, such as the article's age or if it discusses an event that the user has already read about in other articles.

Although it has been shown that collaborative filtering can aid in personalized recommendation systems (Wang et al., 2006), a large number of users is needed. In a limited user environment, such as a small group of analysts monitoring news events, collaborative filtering would be ineffective.

The definition of what makes an article interesting – or its "interestingness" – varies from user to user and is continually evolving, calling for adaptable user personalization. Furthermore, due to the nature of news, most articles are uninteresting since many are similar or report events outside the scope of an individual's concerns. There has been much work in news recommendation systems, but none have yet addressed the question of what makes an article interesting.

BACKGROUND

Working in a limited user environment, the only available information is the article's content and its metadata, disallowing the use of collaborative filtering for article recommendation. Some systems perform clustering or classification based on the article's content, computing such values as TF-IDF weights for tokens (Radev et al., 2003). Corso (2005) ranks articles and new sources based on several properties, such as mutual reinforcement and freshness, in an online method. However, Corso does not address the problem of personalized news filtering, but rather the identification of interesting articles for the general public. Macskassy and Provost (2001) measure the interestingness of an article as the correlation between the article's content and real-life events that occur after the article's publication. Using these indicators, they can predict future interesting articles. Unfortunately, these indicators are often domain specific and are difficult to collect for the online processing of articles.

The online recommendation of articles is closely related to the adaptive filtering task in TREC (Text Retrieval Conference), which is the online identification of articles that are most relevant to a set of topics. The task is different from identifying interesting articles for a user because an article that is relevant to a topic may not necessarily be interesting. However, relevancy to a set of topics of interest is often correlated to interestingness. The report by Robertson and Soboroff (2002) summarizes the results of the last run of the TREC filtering task. Methods explored in TREC11 include a Rocchio variant, a second-order perceptron, a SVM, a Winnow classifier, language modelling, probabilistic models of terms and relevancy, and the Okapi Basic Search System.

The recommendation of articles is a complex document classification problem. However, most classification methods have been used to bin documents into topics, which is a different problem from binning documents by their interestingness. Traditional classification has focused on whether or not an article is relevant to a topic of interest, such as the work done in TREC. Typical methods have included the Rocchio (1971) algorithm, language models (Peng et al., 2003), and latent Dirichlet allocation (Newman et al., 2006; Steyvers, 2006). Despite the research done in topic relevancy classification, it is insufficient for addressing the problem of interestingness. There are many reasons why an article is interesting besides being relevant to topics of interests. For example, an article that discusses content that a user has never seen may be interesting but would be undetectable using traditional IR techniques. For example, the events of the September 11 attacks had never been seen before but were clearly interesting. Furthermore, redundant yet relevant articles would not be interesting as they do not provide the user any new information. However, traditional IR techniques are still useful as a first step towards identifying interesting articles.

MAIN FOCUS

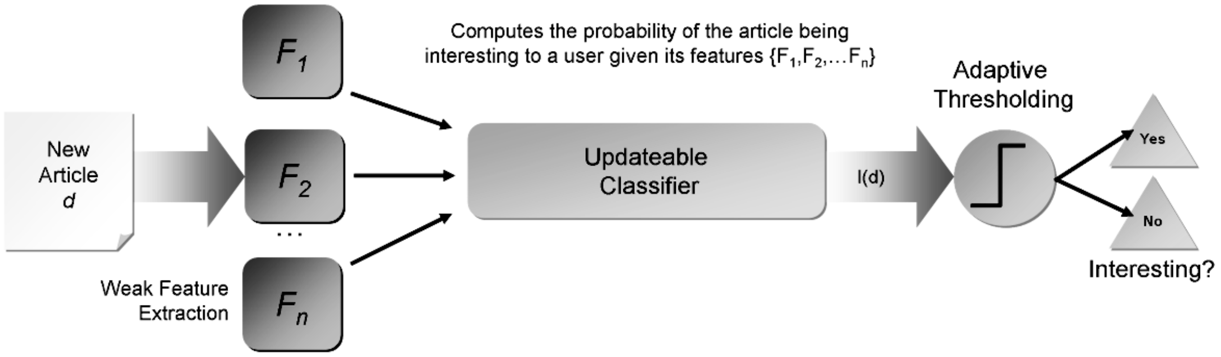
The problem of recommending articles to a specific user can be addressed by answering what makes an article interesting to the user. A possible classification pipeline is envisioned in Figure 1. Articles are processed in a streaming fashion, like the document processing done in the adaptive filter task in TREC. Articles are introduced to the system in chronological order of their publication date. The article classification pipeline consists of four phases. In the first phase, a set of feature extractors generate a set of feature scores for an article. Each feature extractor addresses an aspect of interestingness, such as topic relevancy. Then a classifier generates an overall classification score, which is then thresholded by an adaptive thresholder to generate a binary classification, indicating the interestingness of the article to the user. In the final phase, the user examines the article and provides his own binary classification of interestingness (i.e., label). This feedback is used to update the feature extractors, the classifier, and the thresholder. The process continues similarly for the next document in the pipeline.

Interestingness Issues

The "interestingness" of an article varies from user to user and is often complex and difficult to measure. Consequently, several issues arise:

1. There are a variety of reasons why an article is interesting. There is no single attribute of a document that definitively identifies interesting articles. As a result, using only traditional IR techniques for document classification is not sufficient (Pon et al, 2007).
2. Some interestingness features are often contradictory. For example, an interesting article should be relevant to a user's known interests but should yield new information. On the other hand, random events may be new and unique but may not necessarily be of interest to all users.

Figure 1. Article classification pipeline


Computes the probability of the article being interesting to a user given its features {F1,F2,...Fn}
F_1
New Article d
F_2
...
Weak Feature Extraction
F_n
Updateable Classifier
I(d)
Adaptive Thresholding
Yes
No
Interesting?