A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News

Kelwin Fernandes[1(✉)], Pedro Vinagre[2], and Paulo Cortez[2]

[1] INESC TEC Porto/Universidade Do Porto, Porto, Portugal
[2] ALGORITMI Research Centre, Universidade Do Minho, Braga, Portugal
kelwinfc@gmail.com

Abstract. Due to the Web expansion, the prediction of online news popularity is becoming a trendy research topic. In this paper, we propose a novel and proactive Intelligent Decision Support System (IDSS) that analyzes articles prior to their publication. Using a broad set of extracted features (e.g., keywords, digital media content, earlier popularity of news referenced in the article) the IDSS first predicts if an article will become popular. Then, it optimizes a subset of the articles features that can more easily be changed by authors, searching for an enhancement of the predicted popularity probability. Using a large and recently collected dataset, with 39,000 articles from the Mashable website, we performed a robust rolling windows evaluation of five state of the art models. The best result was provided by a Random Forest with a discrimination power of 73%. Moreover, several stochastic hill climbing local searches were explored. When optimizing 1000 articles, the best optimization method obtained a mean gain improvement of 15 percentage points in terms of the estimated popularity probability. These results attest the proposed IDSS as a valuable tool for online news authors.

Keywords: Popularity prediction · Online news · Text mining · Classification · Stochastic local search

1 Introduction

Decision Support Systems (DSS) were proposed in the mid-1960s and involve the use of Information Technology to support decision-making. Due to advances in this field (e.g., Data Mining, Metaheuristics), there has been a growing interest in the development of Intelligent DSS (IDSS), which adopt Artificial Intelligence techniques to decision support [1]. The concept of Adaptive Business Intelligence (ABI) is a particular IDSS that was proposed in 2006 [2]. ABI systems combine prediction and optimization, which are often treated separately by IDSS, in order to support decisions more efficiently. The goal is to first use data-driven models for predicting what is more likely to happen in the future, and then use modern optimization methods to search for the best possible solution given what can be currently known and predicted.

F. Pereira et al. (Eds.) EPIA 2015, LNAI 9273, pp. 535–546, 2015.
DOI: 10.1007/978-3-319-23485-4_53

Within the expansion of the Internet and Web 2.0, there has also been a growing interest in online news, which allow an easy and fast spread of information around the globe. Thus, predicting the popularity of online news is becoming a recent research trend (e.g., [3,4,5,6,7]). Popularity is often measured by considering the number of interactions in the Web and social networks (e.g., number of shares, likes and comments). Predicting such popularity is valuable for authors, content providers, advertisers and even activists/politicians (e.g., to understand or influence public opinion) [4]. According to Tatar et al. [8], there are two main popularity prediction approaches: those that use features only known after publication and those that do not use such features. The first approach is more common (e.g., [3,5,9,6,7]). Since the prediction task is easier, higher prediction accuracies are often achieved. The latter approach is more scarce and, while a lower prediction performance might be expected, the predictions are more useful, allowing (as performed in this work) to improve content prior to publication.

Using the second approach, Petrovic et al. [10] predicted the number of retweets using features related with the tweet content (e.g., number of hashtags, mentions, URLs, length, words) and social features related to the author (e.g., number of followers, friends, is the user verified). A total of 21 million tweets were retrieved during October 2010. Using a binary task to discriminate retweeted from not retweeted posts, a top F-1 score of 47% was achieved when both tweet content and social features were used. Similarly, Bandari et al. [4] focused on four types of features (news source, category of the article, subjectivity language used and names mentioned in the article) to predict the number of tweets that mention an article. The dataset was retrieved from Feedzilla and related with one week of data. Four classification methods were tested to predict three popularity classes (1 to 20 tweets, 20 to 100 tweets, more than 100; articles with no tweets were discarded) and results ranged from 77% to 84% accuracy, for Naïve Bayes and Bagging, respectively. Finally, Hensinger et al. [11] tested two prediction binary classification tasks: popular/unpopular and appealing/non appealing, when compared with other articles published in the same day. The data was related with ten English news outlets related with one year. Using text features (e.g., bag of words of the title and description, keywords) and other characteristics (e.g., date of publishing), combined with a Support Vector Machine (SVM), the authors obtained better results for the appealing task when compared with popular/unpopular task, achieving results ranging from 62% to 86% of accuracy for the former, and 51% to 62% for the latter.

In this paper, we propose a novel proactive IDSS that analyzes online news *prior* to their publication. Assuming an ABI approach, the popularity of a candidate article is first estimated using a prediction module and then an optimization module suggests changes in the article content and structure, in order to maximize its expected popularity. Within our knowledge, there are no previous works that have addressed such proactive ABI approach, combining prediction and optimization for improving the news content. The prediction module uses a large list of inputs that includes purely new features (when compared with the literature [4,11,10]): digital media content (e.g., images, video); earlier popular-