Pattern Recognition Letters 69 (2016) 42–48

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Buzzword detection in the scientific scenario☆

Danielle Caled[a,*], Pedro Beyssac[a], Geraldo Xexéo[a,b], Geraldo Zimbrão[a,b]

[a] Programa de Engenharia de Sistemas e Computação, COPPE/UFRJ, Caixa Postal 68.511 – 21941–972 - Rio de Janeiro, RJ, Brazil
[b] Departamento de Ciência da Computação, IM/UFRJ, Caixa Postal 68.530 – 21941–590 – Rio de Janeiro, RJ, Brazil

ARTICLE INFO

Article history:
Received 25 May 2015
Available online 22 October 2015

Keywords:
Buzzword detection
Trend detection
Time-series analysis
DBLP database
Data mining

ABSTRACT

This paper addresses a relatively new concept: the buzzword. Buzzwords are fashionable words that continue gaining popularity until a tipping point is reached and then their popularity declines. Our goal in this study is to identify buzzwords through their frequency of occurrence over the years, using two clustering techniques: *k*-means and the self-organizing map (SOM). We also used the DBLP database to run experiments with data from published papers in an attempt to find terms that could be classified as buzzwords, in accordance with the defined meaning. Clusters generated by both k-means and SOM are very similar, indicating that it is very likely that buzzwords were correctly identified as such. We were able to identify terms such as "android" and "mapreduce", which were clearly buzzwords for 2012, as well as terms such as "pomdp", which was not an obvious buzzword. As a contribution, we highlight common characteristics identified for buzzwords and make comparisons between the two methods for finding buzzwords which were analyzed in this paper.

1. Introduction

All fields of research have topics which are the major focus of studies by their communities. Sometimes, such topics arouse interest gradually and eventually become the most discussed topic in a certain area, but at other times their appearance already indicates the timing of the main exploration of the subject. Knowing in advance which words will become buzzwords can help enterprises to make strategic decisions about which fields are promising and deserve more attention, thus dictating a possible pioneering position in certain areas of knowledge.

Buzzwords are new terms or phrases (neologisms) created in one language that acquire great popularity as fashionable words [22]. Informally, a buzzword is a word or phrase related to a specialized field or group at a particular time, or in a particular context, used mostly to impress lay persons. Through the use of these terms it is possible to identify the latest trends of what is happening around the world; that is, what is being most discussed by the population or the most interesting topics at that moment. Buzzword detection consists of important information, especially in the areas of marketing, business, politics, and intelligence [21,30]. Therefore, it is very useful to identify these words as early as possible.

The difference between buzzwords and most of the new terms in a language is the exponential growth of buzzwords. It is a difficult task to predict whether a new word used by a community is destined to become a common term in dictionaries or if it is heading towards a tipping point from which it will decline. Neuman et al. [22] cite the example of the term "Web 2.0", created in 2001 by Tim O'Reilly to describe a turning point for the Web. After a year and a half, the term had gained huge popularity, being quoted in Google more than 9.5 million times and, in 2009, the number of citations in Google had reached 422 million [22]. Currently, however, the term "Web 2.0" seems to have lost popularity.

The popularity of buzzwords comes from their use in media such as TV, magazines, newspapers, and social networks. However, there is usually a smaller group that uses these terms before they become popular with the masses. In other words, buzzwords emerge from a restricted community and gradually spread to other communities, to then become widely known among most people. By identifying this type of behavior, it is possible to find potential buzzwords.

The term "buzzword" has its historical use related to the language of the business and technology sector [20]. Some studies about it have been done, mainly in the blogosphere [7,19,21,22,30], or with a view to finding ways to model bursts of topics [4,12,24]. This is justified due to the fact that blogs are sources of information in which users can express their opinions and interests in real time and, thus, reflect the most current trends. It also provides an ideal place to study the

☆ This paper has been recommended for acceptance by Jie Zou.

* Corresponding author. Tel.: +55 21 97675 5397.
E-mail addresses: dcaled@gmail.com, dcaled@cos.ufrj.br (D. Caled), pedrobeyssac@cos.ufrj.br (P. Beyssac), xexeo@cos.ufrj.br (G. Xexéo), zimbrao@cos.ufrj.br (G. Zimbrão).

http://dx.doi.org/10.1016/j.patrec.2015.09.017
0167-8655/

dynamics of a language's environment. Studies on buzzword detection evaluate the possibility of a topic becoming popular by considering its temporal variation in the text of blogs, which allows researchers to observe the emergence of new topics and the concentration of interests over time [30]. Thus, a common approach in detecting buzzwords in blogs is to evaluate the growth rate for the citation of a topic in such communities.

An unexplored field in which it would be interesting to study buzzwords is the scientific scenario. Since it is possible to see trends in the development of innovations in this scenario, there is also a high propensity for the emergence of buzzwords. According to [3], buzzwords frequently appear in the titles of conference papers and in comments and questions addressed to conference speakers. This occurs mainly because of the strong relationship between innovations and buzzwords [3]. Moreover, in the academic context, a buzzword can represent the interests of a community in relation to a particular subject, and the frequency at which a particular term is used by the academic community in scientific publications should be accompanied over the years.

The identification of new buzzwords in the scientific field may indicate the rise of a new research or business area. To detect the emergence of these words, we can conduct technological forecasting studies, in which it would be possible to predict the impacts of a given innovation. Early buzzword detection is an important contribution to the decision-making process and market trend analysis.

The organization of the rest of this paper is as follows: Section 2 explains how we obtained and prepared the corpus for the experiments; Section 3 describes the clustering experiments; Section 4 analyzes and compares the results; and, finally, the conclusions of this work are presented in Section 5.

2. The corpus

2.1. DBLP

For the present study, the DBLP database [15] was used. The DBLP project is currently maintained by the Universität Trier, in Germany. This database consists of more than 2,947,000 documents of bibliographic information in the computer science area, including conference papers, journals, series, books, and even Master's and Doctorate degree theses.

2.2. Preprocessing

The preprocessing stage included data cleaning, data transcription to a database, and the selection of articles for use in this work.

Formatting tags were removed in the data cleaning stage. Also, non-English characters were replaced with the equivalent characters and, finally, symbols and reserved characters were removed.

We chose not to remove stopwords from the titles of publications. This decision was made based on the fact that a buzzword may be an expression; that is, it may include a set of words, and although one or more words used in the expression may be stopwords, the expression, as a whole, may be a buzzword. An example of this type of situation is the expression "big data", which is a buzzword. The words "big" and "data" by themselves, however, could be considered to be stopwords.

It was also decided not to implement stemming techniques to reduce the words to their radicals. If a stemmer were applied, some buzzwords that appear in the title of the articles could be overlooked due to the loss of information, and the results of the study would be compromised.

The articles used in this study were chosen based on the conference or journal in which they were published. Hence, we selected 100 conferences and journals with the highest number of publications in the DBLP database. Then, we searched the Microsoft Academic Search, website[1] in an attempt to find conferences and journals in the computer science field. In this website, the query results are presented along with an evaluation index in the research area. The index for each conference and journal is generated similarly to the *h*-index [10], which is calculated by the number of publications per author and the number of citations per publication. The evaluation of the Microsoft Academic Search only involves calculation of the publications and citations within a specific field and shows the impact of the conference or journal in the field.

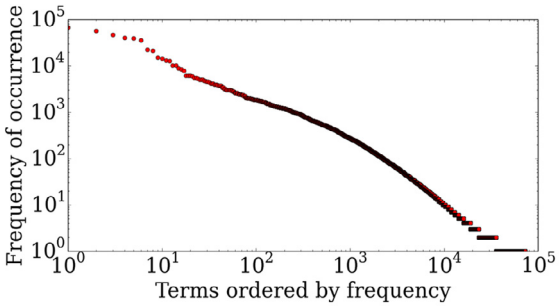

Fig. 1. Plot of word frequency in the DBLP base. The coordinates are in a log–log scale. In this distribution, few terms are extremely common, whilst many terms have a low frequency of occurrence.

Once the rating of major conferences and journals in the computer science field were obtained, a second selection was made in which only conferences and journals with an evaluation index greater than 100 in the research area were chosen. Thus, it was possible to reduce the large number of articles in the database and keep only those that were published in relevant journals or conferences in the computer science area, with a considerable number of publications from the DBLP database — a total of 29 conferences and journals were picked. Finally, we filtered the articles from the selected journals and conferences by picking only the publications between 1990 and 2012. Reducing this sample, the number of items used was restricted to 185,130 with a total of 72,884 distinct words in the titles.

2.3. Pruning

After the initial data analysis, we found that the presence of many terms occurring with a low frequency and few terms occurring with a high frequency could undermine the results. For example, the number of terms that appeared only once in the database was more than 37,000. Also, the most frequent term occurs 11,568 times more than the second most frequent term (e.g., the most frequent term occurs 68,520 times and the second most frequent term occurs 56,952 times).

From the analysis of a log-log rank/frequency plot of words in our corpus (Fig. 1), we identified behavior similar to the Zipf distribution [31]. The distribution of the words in the corpus follows Zipf's law, which states that the frequency of a ranked word is inversely proportional to its rank [17]. According to [8], it is in the transition zone between the words of high frequency and low frequency that the terms of higher semantic content of a given text are located.

Faria [5] recommends dividing the Zipfian curve into four different regions:

- Empty words: zone consisting of words with the highest frequency. Includes words that do not contain any important information — articles and prepositions are examples of these kinds of words;

[1] http://academic.research.microsoft.com/