Information Fusion 23 (2015) 58–68

Contents lists available at ScienceDirect

Information Fusion

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

Finding a needle in the blogosphere: An information fusion approach for blog distillation search

José M. Chenlo [a,*], Javier Parapar [b], David E. Losada [a], José Santos [b]

[a] Centro de Investigación en Tecnoloxías da Información (CITIUS), University of Santiago de Compostela, Rúa de Jenaro de la Fuente Domínguez, 15782 Santiago de Compostela, Spain
[b] Information Retrieval Lab, Department of Computer Science, University of A Coruña, Campus de Elviña, 15071 A Coruña, Spain

ARTICLE INFO

Article history:
Available online 20 September 2014

Keywords:
Blog distillation
Particle swarm optimisation
Differential evolution
Line search
Information retrieval

ABSTRACT

In the blogosphere, different actors express their opinions about multiple topics. Users, companies or editors socially interact by commenting, recommending and linking blogs and posts. These social media contents are increasingly growing. As a matter of fact, the size of the blogosphere is estimated to double every six months. In this context, the problem of finding a topically relevant blog to subscribe to becomes a Big Data challenge. Moreover, combining multiple types of evidence is essential for this search task. In this paper we propose a group of textual and social-based signals, and apply different Information Fusion algorithms for a blog distillation search task.

Information fusion through the combination of the different types of evidence requires optimisation for appropriately weighting each source of evidence. To this end, we analyse well-established population-based search methods. Namely, global search (Particle Swarm Optimisation and Differential Evolution) and a local search method (Line Search) that has been effective in various Information Retrieval tasks. Moreover, we propose hybrid combinations between the global search and the local search method and compare all the alternatives following a standard methodology. Efficiency is an imperative here and, therefore, we focus not only on achieving high search effectiveness but also on designing efficient solutions.

1. Introduction

The blogosphere has become a key source of information to follow new trends and opinions about a wide variety of topics [1]. The raise and success of social networks (e.g., Facebook or Twitter) has motivated the evolution of blogs from informal discussion or informational sites to professional platforms specialised in specific domains (e.g., technology, fashion or finances). The proliferation of platforms to create and manage blogs, such as Blogger[1] or WordPress,[2] facilitates the development of immense communities of blogs on the Internet. This explosion of specialised information sources complicates the identification of new blogs related to people's interests.

Seeking information in the blogosphere is a Big Data problem. Because of the large number of publishers and followers, because of the huge volume of contents and topics, and because of the variety of signals needed to automatically estimate relevance. For instance, a blog search engine can be tracking more than 182 million blogs[3]; and the number of blogs doubles every six months [2].

Several studies have focused on search tools for the blogosphere [3,4]. In this area, different features and algorithms have been proposed and tested for detecting blogs potentially relevant to specific interests. For instance, some authors use the information provided by blog feeds[4] to estimate the relatedness between a blog and a particular topic [5]. Other authors go beyond these feeds and try to measure the relatedness between the topic and individual blog posts [6]. The use of a more heterogeneous set of features is also common. For instance, blog timestamps, link analysis or information provided from external resources (e.g., Wikipedia [5]). This array of experimental studies makes it difficult to understand what features are effective, and when and how they are best used. To shed light on these issues, in this paper we investigate different methods to automatically and effectively combine multiple sources of evidence to recommend blog posts based on a particular interest. We combine

* Corresponding author. Tel.: +34 881 816 396; fax: +34 881 816 405.
E-mail addresses: josemanuel.gonzalez@usc.es (J.M. Chenlo), javierparapar@udc.es (J. Parapar), david.losada@usc.es (D.E. Losada), santos@udc.es (J. Santos).

[1] https://www.blogger.com.

[2] https://wordpress.com/.

[3] http://smartdatacollective.com/matthewhurst/44748/farewell-blogpulse.

[4] A blog feed is a collection of the last entries – the blog's posts – from a particular blog.

http://dx.doi.org/10.1016/j.inffus.2014.09.001
1566-2535/© 2014 Elsevier B.V. All rights reserved.

evidence of relevance – at blog and post level – with more heterogeneous signals such as temporal or social features.

This combination of different types of evidence entails an optimisation task for setting the most appropriate weight to each source of evidence. To face this challenge, we have selected well-established population-based methods providing a global search (Particle Swarm Optimisation and Differential Evolution); and a local search method (Line Search) that has been effectively employed in Information Retrieval. The focus of the paper is not on providing a comparison of a wide range of search algorithms and variants of these. We have rather selected popular algorithms and standard implementations and, additionally, we implemented a hybrid version for integrating the advantages of global and local search. Besides search effectiveness, efficiency is an imperative in our application domain. Therefore, we will report on both effectiveness and efficiency results.

The main contributions of this paper are:

- We propose a set of relevance-based and social-based signals for estimating the relevance of blogs.
- We propose a fusion approach that combines these information signals for finding potentially relevant blogs to subscribe to.
- We evaluate different optimisation algorithms for weighting different features.
- We perform a thorough analysis of performance of the Information Fusion methods. Our evaluation considers both search effectiveness and efficiency.

The rest of the paper is organised as follows. Section 2 reviews some papers related to our research. In Section 3, we present our fusion methods and explain the approach to combine evidence, the search models utilised, and the features proposed to estimate the relevance of a blog feed. The experimentation design and results are reported in Section 4. The paper ends with Section 5, where we expose some conclusions.

2. Related work

2.1. Blog distillation task

The blog distillation task is defined as the process of searching for blogs with a recurring central interest, typically expressed as a textual query [3]. The task can be summarised as: *Find me a blog with a principle, recurring interest in T*. For a given target topic *T*, systems should suggest feeds that are devoted to *T* [3].

This problem can be seen as a particular instance of a resource selection problem in distributed Information Retrieval [7], in which the goal is to select the blogs that more likely contain documents relevant to the query *T* [1]. Many researchers utilised both the information provided by blog feeds and the information provided by the blog posts referenced by the feeds. For instance, Elsas et al. [5] combined relevance scores at feed level and at blog post level, and employed Language Models (LMs) to retrieve feeds related to a specific query. They also proposed a Query Expansion (QE) technique based on data extracted from the Wikipedia. This technique was effective and, in fact, the best performing systems in the TREC blog distillation task applied expansion [3,4].

In [8], Seo and Croft also followed the resource selection principle and considered different blog representations. Two main models were tested: (1) a blog represented as the concatenation of all its posts, and (2) a blog represented as a query-dependent cluster (containing only highly ranked blog posts for a given query). They found that the combination of both models was the most effective approach.

Other studies addressed the problem as an expert finding task [9]. Given a blog post and a query, the estimated relevance of the blog to the query can be seen as an indication of the interest of the post's author with respect to the query topic [1]. This approach was explored by Macdonald and Ounis [10], who defined a Voting Model [11] that represents the blogger (the blog writer) as a concatenation of all blog posts written by him. A query is run against all blog posts and every blog post retrieved that belongs to the profile is considered as a vote for the relevance of that blog. Combining evidence from feeds and posts has also been studied for other social media tasks, such as trend detection [12].

In this paper we explore an alternative approach to the blog distillation task. We consider blog distillation as an Information Fusion problem where multiple types of signals need to be combined. We study what signals are effective, and when and how they are best used to assess the recurring interest of a blog in a specific topic. We evaluate different fusion methods – with relevance-based, social-based, and temporal-based evidence – and compare their performance with that of baseline and state-of-the-art methods.

2.2. Combining evidence

A straightforward way to combine evidence in IR is linear interpolation [13]. However, optimising these models is a major bottleneck within the retrieval process. The traditional methodology in IR implies that parameter values are optimised with a training sample (for a certain evaluation measure); and, then, the trained parameter values are evaluated with a test sample. With one or two parameters, an exhaustive search for the best parameter setting – e.g., through line search – is expensive but affordable. But the computational cost of the optimisation process grows exponentially with the number of parameters. This problem has been previously addressed in the literature. Taylor et al. [14] studied retrieval models containing up to 375 parameters and agreed that the lack of efficient algorithms for parameter optimisation is one of the bottlenecks of current IR research.

The problem of combining multiple sources of evidence has been widely addressed in the literature of Information Retrieval (IR) and Machine Learning (ML). A well-known family of methods that naturally allows the combination of evidence is Learning To Rank (L2R). L2R [15] has received much attention lately as a result of the popularity of Web search engines. These methods learn how to combine predefined features for ranking by means of discriminative learning algorithms. L2R models can be grouped into three approaches. The *pointwise* approach assumes that we want to predict the relevance degree of each individual document. Regression-based algorithms and classification-based algorithms are examples of pointwise approaches [16,17]. The *pairwise* approach [15,18–21] takes a pair of documents and predicts their pairwise preference (e.g., +1 or -1). The *listwise* approach [22–26] takes the entire set of documents associated with a query and predicts the complete ranking. The listwise approach is the most natural method to rank documents in decreasing order of estimated relevance; and experimental results have showed that it has certain advantages over other algorithms [15,26]. In this paper, we follow a listwise approach and learn a ranking function from a certain set of signals. Given a training set of blogs and a vector of features, we directly optimise a global measure of performance – Mean Average Precision (MAP) –, which is computed over the ranked set of documents as a whole.

Evolutionary algorithms [27] have been applied to learn ranking functions. For instance, Particle Swarm Optimisation (PSO) was successfully applied in IR [28,29]. In [30], the authors experimented with a hybrid PSO-Line Search model and proposed a cooperative Line Search-Particle Swarm Optimisation (CLS-PSO)