Filtering Structures for Microblogging Content

Ryadh Dahimene, CNAM, Paris, France
Cedric du Mouza, CNAM, Paris, France

ABSTRACT

In the last years, microblogging systems have encountered a large success. After 7 years of existence Twitter claims more than 271 million active accounts leading to 500 million tweets per day. Microblogging systems rely on the all-or-nothing paradigm: a user receives all the posts from an account s/he follows. A consequence for a user is the risk of flooding, i.e., the number of posts received from all the accounts s/he follows implies a time-consuming scan of her/his feed to find relevant updates that match his interests. To avoid user flooding and to significantly diminish the number of posts to be delivered by the system, the authors propose to introduce filtering on top of microblogging systems. Driven by the shape of the data, the authors designed different filtering structures and compared them analytically as well as experimentally on a Twitter dataset which consists of more than 2.1 million users, 15.7 million tweets and 148.5 million publisher-follower relationships.

Keywords: Filtering, Indexing, Microblogging, Scalability, Social Networks

INTRODUCTION

Microblogging systems have become a major trend as well as an important communication vector. In less than seven years, Twitter\(^1\) has grown in a spectacular manner to reach more than 500 million users in August, 2013\(^2\) from which 54% are active users\(^3\) (posting at least a tweet per month). Other similar systems like Sina Weibo\(^4\), Identi.ca\(^5\) or Plurk\(^6\), to quote the largest, also exhibit dramatic growth. In such systems, the length of a published piece of news (called by post in the following) is limited to 140 characters which corresponds on average to 14.7 terms (Foster et al., 2011), so clearly greater than 4-5 terms adverts as reported in (König, Church, & Markov, 2009) but smaller than RSS items (Hmedeh et al., 2011), blogs (Ma & Zhang, 2007) or Web pages which have a size of 450 to 500 terms excluding tags (Levering & Cutler, 2006). The rationale for using such small messages is that these services were designed to be accessed with traditional cell phones through Short-Text Messaging protocol\(^7\).

In microblogging, a user, represented by his account, follows other accounts to be notified whenever they publish some information. Conversely, s/he becomes a publisher for the

DOI: 10.4018/ijiit.2015010103
accounts that follow her/him, which results in the existence of a large social graph. One of the main differences with other online social networks is that in microblogging, the graph linking the users can be seen as an interest graph, since users follow each other if they have interest in each other posts. Those networks work on an all-or-nothing fashion, i.e. if a user A follows a user B, A will receive all of B’s posts.

Microblogging is also characterized by the heterogeneous nature of users. In Twitter for instance, there exist some high update frequency accounts (newspapers, tech-blogs and journalists) and others that publish less than one post a week. Moreover, there exist very popular accounts (e.g. news channel accounts like @FoxNews8 or famous accounts like @BarackObama9 with more than 45 million followers), and others with 0 or 1 follower.

For various reasons (security, advertisement, control policy …) these systems rely on a centralized architecture. Each post published is received by the central system that forwards it to all the followers of the publishing account. The central system is also responsible of the search feature (Busch et al., 2012). Since the most active accounts are generally the ones with the highest number of followers, the system must face a tremendous amount of posts to forward. For instance Twitter, which claimed in 2011 more than 200 million tweets a day, had to deliver daily over 350 billion tweets. This traffic overload (especially for high-followed accounts) represents from an architectural point of view a scalability bottleneck.

On the follower’s point of view, the amount of posts received from the accounts s/he follows, between 30 and 200 depending on the system considered (Kwak, Lee, Park, & Moon, 2010) loses the reader in the middle of long feeds of posts. This results in poor data readability and potentially loss of valuable information.

In a UK-based questionnaire over 587 participant, Bontcheva, Gorrell, & Wessels (2013) report that 33.9% of users perceive that they receive too many posts and 70.4% have found the task of locating the interesting/relevant posts amongst the others difficult. Also, 66.3% of the interviewed users have felt at some time that they can’t keep up with the amount of received tweets. A direct consequence of this phenomenon is the high dynamicity of the graph: to avoid flooding, users who follow active accounts tend to unsubscribe because they can’t manage the continuous flow of posts as studied in (Kwak, Chun, & Moon, 2011; Kivran-Swaine, Govindan, & Naaman, 2011).

In order to improve the user experience and reduce the network load, we have chosen to introduce filtering in microblogging systems. The main underlying idea is that a user A follows another user B for some topics, and consequently s/he wants to receive only a subset of B’s posts that matches his interest. Such a structure must efficiently retrieve for an incoming post all followers of a publishing account whose filter is satisfied by the post.

While designing the filtering structures, we took a particular consideration about some specific aspects of microblogging systems which we can summarize as:

- **Short messages**: The size restriction (generally 140 characters) means that we handle short documents, the average length of a tweet is 14.7 terms (Foster et al., 2011);
- **Account heterogeneity**: Microblogging studies have revealed account heterogeneity in term of both update frequency and number of followers;
- **Graph evolution**: As observed in Twitter (Kwak et al. 2011; Kivran-Swaine et al. 2011), users follow and unsubscribe often to other accounts. The filtering structure must consequently handle the graph dynamicity in order to handle this phenomenon;
- **Centralized system**: The social graph is stored by the microblogging system. This system receives all posts and forwards them to followers according to the graph it stores. That means the whole task is supported by the centralized system. Therefore we must reduce the filtering process time by trying to manage the matching in central memory.
Facilitating Decision Making and Maintenance for Power Systems Operators through the Use of Agents and Distributed Embedded Systems
www.igi-global.com/article/facilitating-decision-making-maintenance-power/46960?camid=4v1a

A Heuristic Method for Learning Path Sequencing for Intelligent Tutoring System (ITS) in E-learning
www.igi-global.com/chapter/heuristic-method-learning-path-sequencing/71973?camid=4v1a