Distributed News Analytics Framework: Collecting News Feed Sources from social media

Gábor I. Nagy, Sándor Kazi, Győző Papp

Text mining in news articles and social media in the financial context became a vibrant research topic in the past years [1]. Distributed computing helps overcome the difficulties of processing vast amounts of various, heterognic textual data. The implementation and extension of an earlier version of such a system is discussed in this paper: a distributed adaptive news analytics framework that gathers, stores, curates and process vast amounts of textual data from conventional news feeds and social media. The main components of the system is shown on Figure 5.

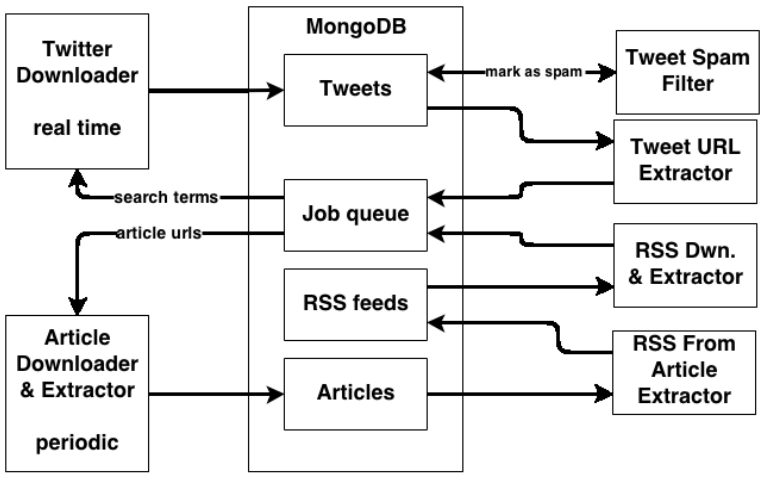

Figure 5: Distributed real-time content monitoring service - system components

The Twitter Downloader component uses Twitters Streaming API to listen to the conversation in social media and stores messages in MongoDB. The API is real-time: tweets are pushed to the data collection server as soon as they are published by Twitter. However one can only search for a limited number of terms (be it ticker symbol, hashtag or user) and can have updates on these terms only. There will be search terms that generate most of the content stored in our database, on the other hand there will be obsolete search terms with little or no messages as topics grow older and interest falls. These obsolete search terms can be replaced by popular terms and ticker symbols found in recent news articles, extracted from either URLs found in Twitter posts. Twitter content, and specifically URLs posted can be monitored for interesting new news RSS feeds, that are currently not used by the Article Downloader & Extractor.

This adaptive behaviour of search terms allows monitoring larger amounts of content than one could monitor with setting constant search terms in Twitter Stream API. However one should be careful with posted content, as Twitter is a prime target of content spam. [2] [3] The Twitter Spam Filter allows ranking user messages as being spam. The Spam Filter is discussed in details in [4]. The Article Downloader & Extractor downloads conventional news articles given by their URL. Extracts various features from these articles: links on the page, raw text extract of the news for text minign, links to rss feeds, or number of ads on a given page from the downloaded HTML. Content is stored in MongoDB. RSS Downloader & Extractor downloads feeds given by their URL and dispatches their links to the Job Queue for to the Article Downloader. The Tweet URL Extractor periodically extracts the URLs from the latest tweets and dispatches them as jobs for the Article Downloader. This module of the system was

described in an earlier work [5]. The module named RSS From Article Extractor copies the interesting new entries from extracted documents originating from Twitter, that containing RSS feed URLs. The purpose of the Job Queue is to be a centralized temporary storage for tasks and messages of the processes. Performance tests of the proposed system is the core focus in the paper together with implementation details.

Acknowledgements

This work is supported by the grant: FUTURICT, TÁMOP-4.2.2.C-11/1/KONV, Financial Systems subproject.

References

[1] J. Bollen, H. Mao, X. Zeng. Twitter mood predicts the stock market, *Journal of Computational Science*, Vol. 2, 2011, 1-8

[2] K. Thomas, C. Grier, D. Song, V. Paxson. Suspended accounts in retrospect: an analysis of twitter spam, *IMC '11 Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, 243-258

[3] A. H. Wang. Detecting Spam Bots in Online Social Networking Sites: A Machine Learning Approach *Data and Applications Security and Privacy XXIV Lecture Notes in Computer Science*, Volume 6166, 2010, 335-342.

[4] G. I. Nagy, S. Kazi. Filtering noise from stock related Twitter messages. *(CS)^2 - The 9th Conference of PhD Students in Computer Science*. Submitted.

[5] G. I. Nagy, S. Kazi. Distributed News Analysis Framework for Text Mining. *IEEE, CogInfoCom 2013 - Workshop on Future Internet Science and Engineering*, 2013.