Social Impact - Identifying Quotes of Literary Works in Social Networks

Carlos Barata[12], Mónica Abreu[1], Pedro Torres[2], Jorge Teixeira[23], Tiago Guerreiro[1] and Francisco M. Couto[1]

[1] Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, Portugal
[2] SAPO Labs, Lisboa, Portugal
[3] LIACC, Universidade do Porto, Porto, Portugal

Abstract. A non-neglectable amount of information shared in social networks has quotes to literary works that, most of the times, is not linked to the original work or author. Also, there are erroneous quotes that do not fully match the original work, for example by including synonyms and slang words. Moreover, users sometimes associate their quotes to the wrong author, which creates misleading information. This paper presents *Social Impact* framework as an approach to identify quotes in social networks and match them to the original literary works from a particular author. This framework was applied to two case-studies: *O Mundo em Pessoa* and *Lusica*. In the first case-study, *Social Impact* evaluation achieved 98% for precision measure and 59% for recall, whereas in the latter case-study it obtained 100% for precision and 53% of recall.

Keywords: Information Retrieval, Information Extraction, Web Mining, Text Mining, Pattern Recognition

1 Introduction

Social networks emerged in last decade and changed the way we communicate, becoming essential tools in the human interaction. This happened possibly due to the fact that, at the distance of a click, lays the possibility to send and share content. As Kwak et al.[4] refers, this wide use of social networks provide a great interest of investigation in many areas like extraction and information analysis.

Most of the information shared in social networks such as Twitter and Facebook is in text format, and an interesting amount of such information (messages) contains quotes to literary works (e.g.: "Tudo vale a pena, quando a alma não é pequena - Fernando Pessoa"). Nevertheless, in a non-neglectable number of cases there is no reference to which text, book or literary work the quote is referred.

Due the fact that quotes may have incoherencies (e.g.: quote is different from the original text), the identification of the original text or author can be very challenging. These incoherencies have a higher presence on social networks (against, for instance, opinion articles on news) because of particular characteristics of the network, namely: short messages or reduced context. The use of

synonyms or typos are some of the most common causes for the lack of accuracy in the quotes published in social networks. One may think hash-tags can substantially reduce the complexity of this task, but unfortunately the usage of hash-tags on messages literary work is low as referred in [8] study. This study concluded that in Twitter, the ratio of hash-tags per tweet is between 4%(in Japanese language) to 25%(in German language) of the total tweets using them.

This paper presents *Social Impact* platform as an approach to this problem. The framework major goal is to identify quotes from literary work on social network messages, supported on *SocialBus*[4] and Apache Lucene systems. Evaluation was performed on two case-studies: *O Mundo em Pessoa* and *Lusica*.

2 Related work

Social Impact platform is generically supported on two different technological blocks: *SocialBus*, a social network crawling and analysis platform, and Lucene, a high-scalable infrastructure for indexing and querying documents.

SocialBus: Social networks such as Twitter and Facebook provide APIs that allow access to public messages, within certain limits, giving the possibility of analysing such content for a variety of purposes, including quotes detection. We propose to use *SocialBus* platform[7,2], a framework that collects and analyse data from Twitter and Facebook for a pre-defined set of users representative of the Portuguese community.

Lucene: is an open-source software[5] for text searching and indexing through a document indexation, coded in Java programming language and developed by Apache Software Foundation. According to Gospodnetic et al.[3] this framework works through the indexation of documents, information parsers and queries to consult and retrieve the indexed information. The result is a ranked list of documents ordered by relevance [6,5,1].

3 Social Impact Platform

Social Impact main objective is to find quotes in messages published in social networks and subsequently link them to their original literary work. This framework stores such data in a relation database and provides such data as RESTful APIs. More importantly, *Social Impact* architecture is abstract enough to be applied on different contexts and scenarios.

3.1 Architecture

The *Social Impact*'s structure is based on a Service-Oriented Architecture(SOA), broadly used in web applications, due to its standardisation approach. This architecture is represented in Figure 1 and it has three main layers, described below.

[4] `http://reaction.fe.up.pt/socialbus/`
[5] `http://lucene.apache.org/core/`

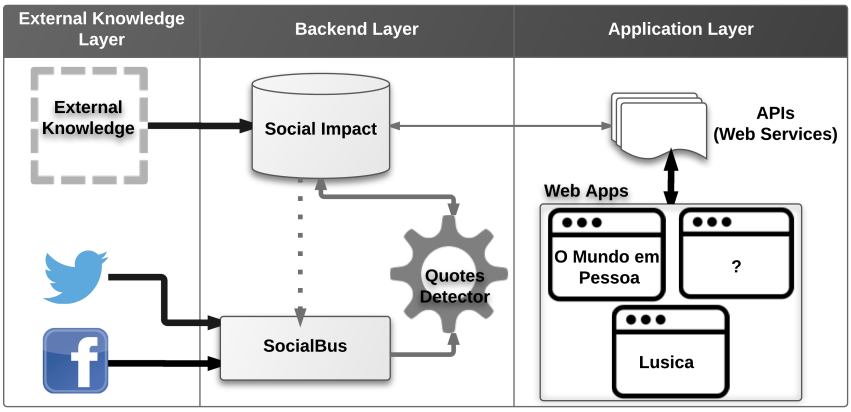

Fig. 1: *Social Impact* Global Architecture

External layer: represents information and knowledge external to *Social Impact* platform and that somehow is collected into the system. The leftmost block, External knowledge, represents data specific to each case-study, including literary work (e.g.: poems from Fernando Pessoa[6]) or domain specific keywords used to narrow the search over *SocialBus* collected data (e.g.: poems or musics authors). The remain two blocks represent Twitter[7] and Facebook[8] APIs to feed *Social Impact* with data from social networks.

Backend layer: is the core layer of *Social Impact*, and it is responsible for processing the messages coming from *SocialBus* and analyse them through the Quotes Detector, as well as store those messages and their subsequently generated meta-data on suitable a relational database (MySQL).

Application layer: represents the interface with the potential applications using *Social Impact* platform. This layer comprehends a set of RESTful APIs that provides information previously processed in the Backend layer to the web applications.

3.2 Quotes Detector

Figure 2 presents a detailed diagram of the Quotes Detector module, with two essential flows of information:

Pre-processing and indexing External Knowledge: represented in Figure 2 as "I" is imported only once and include, for instance, all the literary work from a particular author. Each of these documents (e.g.: a single poem) is submitted to Lucene engine, filtered through a stopwords filter and indexed.

Identification and indexing of Quotes (refer to "II" in Figure 2) module is listening to *SocialBus* and imports new data as new messages arrive to *Social-*

[6] Data obtained from "Arquivo Pessoa" available at `http://arquivopessoa.net`
[7] `https://dev.twitter.com/rest/public`
[8] `https://developers.facebook.com`

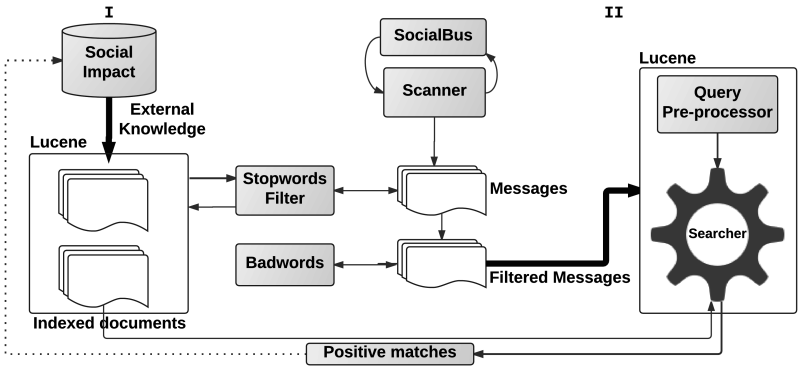

Fig. 2: Quotes detector workflow

Bus. Those messages are then filtered with a stopwords and a badwords (curse words) filter. Each filtered message is transformed into a lucene query syntax. The "Search" operation compares the indexed documents from *Social Impact* (External Knowledge) with each new message and retrieve, if the score if above a given threshold, the most relevant document (poem, music, etc.) as a positive match of a quote. Moreover, all tokens from the matched message are isolated and stored in the database.

4 Case studies

O Mundo em Pessoa[9] is a web based project that aims to depict the presence of Fernando Pessoa poems on social networks, based on quotes to his literary. This project is based on *Social Impact* platform and it covers Fernando Pessoa work and from all his heteronyms. The list of terms used to narrow the messages crawling (refer to Section 3.1) contains the names of all Fernando Pessoa heteronyms. This project is supported on a Web Application that displays the identified quotes from Fernando Pessoa organized by timeframes, going from one day to one month. For each quote, the user has the possibility to explore the number of social network users that publish that particular quote and access the original message, among other features.

Lusica[10] main purpose was study the lusophone music and its presence on the social networks, supported on *Social Impact* platform. There are two importante aspects that differentiate "Lusica" from "Mundo em Pessoa": (i) the domain is music instead of literary, and (ii) a large effort was put on the visualization of the information obtained from the quotes detection, through an interactive graph available online. "Lusica" external knowledge (refer to Figure 1) is based on the musics' and albums' titles from lusophone artists. Such in-

[9] `http://fernandopessoa.labs.sapo.pt/`
[10] `http://lusica.labs.sapo.pt/`

formation was obtained from LastFM APIs[11] (the list of lusophone artists) and from MusicBrainz service[12] (the albums and musics titles for each lusophone artist).

5 Results and Discussion

In this section will be displayed the results of the evaluation of *Social Impact* for both case studies. This evaluation aims to provide an overview of the system's performance.

Data collection: For evaluation purposes, we used a subset of data published between January 2014 and June 2014. Regarding *O Mundo em Pessoa*, from a set of 56.212 collected messages, only approximately 8% (4.720 messages) were identified as quotes (with Lucene score larger than 1,0). As expected, most of the collected messages are not classified as quotes, and this phenomenon can be explain by the fact that many of the collected messages are just references to Fernando Pessoa but are not actually a quote to his literary work. For *Lusica*, results are similar to *O Mundo em Pessoa*, with only less than 5% (7.628) of messages representing references to the songs' artists, in a set of approximately 420.000 messages.

Quotes Detector evaluation: Precision and recall metrics were calculated based on the following types of documents: True positive (TP): messages correctly classified as quotes; False positive (FP): messages incorrectly classified as quotes; True negative (TN): messages correctly classified as not quotes; False negative (FN): messages incorrectly classified as not quotes. Precision is measured as $P = TP/(TP+FP)$ while recall is $R = TP/(TP+FN)$. Regarding recall, our assumption is that *SocialBus* filtered messages correspond to all representative messages for the specific domain of the case-study. Evaluation was performed manually on a sample of 200 randomly chosen messages for each of the case-studies. Regarding "Mundo em Pessoa", the evaluation dataset was divided in 4 parts according to Lucene score and Twitter *versus* Facebook messages. Results for Twitter shown a precision of 19% and recall of 100% for low scores (between $0,5$ and $1,0$) and precision of 98% and recall of 100% for high scores (between $1,0$ and $2,0$). Concerning Facebook, precision value for low scores was 100% and recall 21% while for high scores was precision was 96% and recall was 100%. The average precision for "Mundo em Pessoa" was $P_{MundoemPessoa} = 98\%$, while recall was $P_{MundoemPessoa} = 59\%$. In respect to "Lusica", the same principle was followed, by selecting a sample of 200 messages and dividing them in two groups (Twitter messages with low and high Lucene score). Results shown a average precision value $P_{Lusica} = 100\%$, while recall was $P_{Lusica} = 53\%$.

Execution time: the performance of *Social Impact* platform was also evaluated, measuring the execution time of processing a single message in a desktop with an Intel(R) Xeon(R) CPU E5405 @ 2.00GHz processor and 3GB of RAM

[11] `http://www.last.fm/api`
[12] `https://musicbrainz.org/doc/MusicBrainz_Identifier`

memory. For *O Mundo em Pessoa* the results achieved an average time of execution of 0,01 seconds ($\pm 0,002$). Regarding *Lusica*, the result obtained was an average execution time of 0,02 seconds ($\pm 0,004$).

6 Conclusions

This paper presented an approach to find quotes of original literary work in shared messages on social networks. The proposed approach is supported on the *Social Impact* developed platform presented in this paper. This framework was applied to two distinct case studies: *O Mundo em Pessoa* and *Lusica*.

Evaluation shown that most of the collected messages from *SocialBus* are not classified as quotes (less than 8%) because they are just references to the author and do not contain any quotes. *Social Impact* evaluation achieved high precision values for both case-studies: $P_{MundoemPessoa} = 98\%$ and $P_{Lusica} = 100\%$.

Future work includes: (i) to automatically set the threshold for the Lucene score based on machine learning approaches; (ii) improve the domain specific list of keywords using automatic approaches; (iii) use user feedback to fill missing information about authors and song lyrics; and (iv) apply *Social Impact* platform on other non-literary corpora, such as plagiarism detection.

Acknowledgements

This work was partially supported by SAPO Labs and FCT through the project PEst-OE/EEI/UI0408/2013 (LaSIGE), and by the European Commission through the BiobankCloud project under the Seventh Framework Programme (grant #317871). The authors would like to thank to Bruno Tavares, Sara Ribas and Ana Gomes from SAPO Labs, João Martins, Tiago Aparício, Farah Mussa, Gabriel Marques and Rafael Oliveira from University of Lisbon and Arian Pasquali from Universidade of Porto for all their support, insights and feedback.

References

1. Baeza-Yates, R., Ribeiro-Neto, B., et al.: Modern information retrieval, vol. 463. ACM press New York (1999)
2. Boanjak, M., Oliveira, E., Martins, J., Mendes Rodrigues, E., Sarmento, L.: Twitterecho: a distributed focused crawler to support open research with twitter data. In: Proceedings of the 21st WWW. pp. 1233–1240. ACM (2012)
3. Hatcher, E., Gospodnetic, O.: Lucene in action. Manning Publications. (2004)
4. Kwak, H., Lee, C., Park, H., Moon, S.: What is twitter, a social network or a news media? In: Proceedings of the 19th WWW. pp. 591–600. ACM (2010)
5. Liu, B.: Web data mining. Springer (2007)
6. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to information retrieval, vol. 1. Cambridge university press Cambridge (2008)
7. Oliveira, E.J.S.L.: TwitterEcho: crawler focado distribuído para a Twittosfera portuguesa. Master's thesis, Faculdade de Engenharia da Universidade do Porto (2010)
8. Weerkamp, W., Carter, S., Tsagkias, M.: How people use twitter in different languages. Proceedings of the Web Science (2011)