Integrating Textual Data into Heterogeneous Data Ingestion Processing

Mathieu Roche[1,2] and Maguelonne Teisseire[1,3]

[1] UMR TETIS, Univ Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, Montpellier, 34398 Montpellier, France.
[2] CIRAD, Montpellier, France
[3] INRAE, Montpellier, France

Abstract. In this abstract, two methods for integrating textual data and textual features into ingestion processing are summarized. The first method involves integrating all features, including textual features, into dedicated frameworks, such as by using machine learning techniques. In the second method, text and textual features, such as keywords, are used to explain results returned by heterogeneous data mining. In this context, it is necessary to link data (e.g., databases, images, etc.) and/or obtained results with textual data (e.g., documents and keywords).

Keywords: Data mining · Text mining · Natural language processing · Data integration · Image analysis

1 Context

Big data is traditionally characterized in terms of three Vs, i.e., volume, variety and velocity. The SONGES[1] (Heterogeneous Data Science) project was focused on the variety criterion. The project addressed the following research question: how can textual data be exploited to process heterogeneous data (e.g., databases, images, and video)? This abstract presents a discussion of this issue, which was studied in the SONGES project and is now being studied in the MOOD[2] (Monitoring outbreak events for disease surveillance in a data science context) project.

In the following subsections, two methods are summarized that could be implemented to incorporate textual features into heterogeneous data ingestion tasks.

2 Integration of textual features with heterogeneous data

Heterogeneous data can be used to predict prices and stock market trends [17, 11], perform person identification [8], analyze food security [4], monitor health

[1] http://textmining.biz/Projects/Songes/

[2] https://mood-h2020.eu/

information [18], etc. Recently, different types of features, such as textual and visual content-related features, have been considered in fake news detection approaches [15, 1].

The aforementioned applications can be performed using ingestion processing by integrating different features, including textual features (see Figure 1). For instance, events from web news and user sentiment from social media can be used to improve stock market predictions [17]. In this paper, the authors propose a coupled matrix and tensor factorization scheme to implement heterogeneous information integration and multitask learning simultaneously. In [3], the extraction of different features from multisource heterogeneous data, including trading transaction data, comments from user discussion boards, and news events, is proposed.

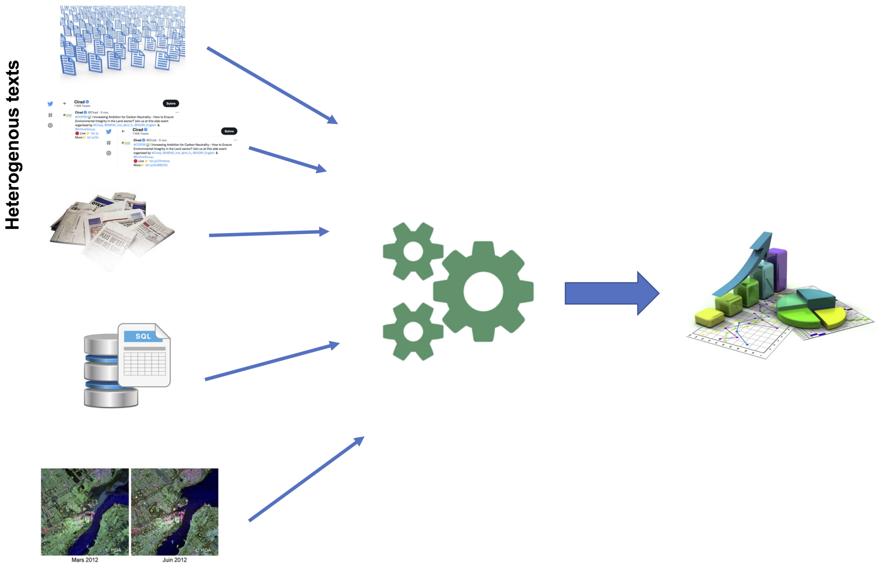

Fig. 1. Integrating textual data into a heterogeneous data ingestion pipeline.

Other approaches based on visual analytics [9] integrate textual data into heterogeneous data ingestion pipelines [5].

3 Using textual features to explain data and results

Methods based on machine and deep learning approaches using heterogeneous data produce good results [16, 4]. However, the results are challenging to explain [2]. In this context, mining textual data associated with heterogeneous data can be used to gain qualitative insights. For instance, many heterogeneous satellite images are currently available that require analysis [12]. Image-text matching based on spatial information [6] improves information retrieval and image annotation techniques [13]. Thus, a more global data context is provided that may be useful for experts involved in land-use planning [10].

In summary, mapping a text (e.g., news, tweets, and articles) using additional data (see (a) in Figure 2) can be useful in expert analysis and annotation tasks. Another challenge (see (b) in Figure 2) is to highlight textual features (e.g., keywords) to explain results obtained by mining heterogeneous data (satellite images, databases, etc.). For instance, data mining algorithms can predict food insecurity or health problems in a country over a given period, and text mining of media and/or social media data can highlight discriminative keywords to provide explanations [14, 7].

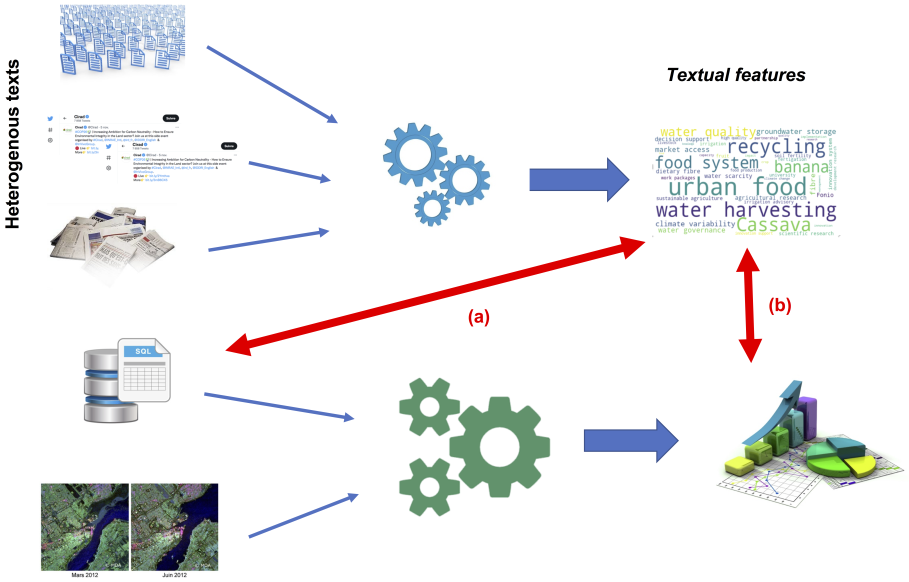

Fig. 2. Using textual features in a data ingestion pipeline.

Acknowledgements

This study was supported by the French National Research Agency under the Investments for the Future Program, Reference No. ANR-16-CONV-0004 (#DigitAg). This study was partially funded by EU Grant Number 874850 MOOD (Monitoring outbreak events for disease surveillance in a data science context) and the SONGES project (FEDER and Occitanie). The content of this document is the sole responsibility of the author and does not necessarily reflect the views of the European Commission.

References

1. Anoop, K., Gangan, M.P., P, D., Lajish, V.L.: Leveraging Heterogeneous Data for Fake News Detection, pp. 229–264. Springer International Publishing, Cham (2019)
2. Belle, V., Papantonis, I.: Principles and practice of explainable machine learning. Frontiers in Big Data **4**, 39 (2021). https://doi.org/10.3389/fdata.2021.688969, https://www.frontiersin.org/article/10.3389/fdata.2021.688969
3. Chai, L., Xu, H., Luo, Z., Li, S.: A multi-source heterogeneous data analytic method for future price fluctuation prediction. Neurocomputing **418**, 11–20 (2020). https://doi.org/https://doi.org/10.1016/j.neucom.2020.07.073
4. Deléglise, H., Bégué, A., Interdonato, R., d'Hôtel, E.M., Roche, M., Teisseire, M.: Linking heterogeneous data for food security prediction. In: Proc. of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD) Workshops, Springer. pp. 335–344 (2020)
5. Fadloun, S., Sallaberry, A., Mercier, A., Arsevska, E., Roche, M., Poncelet, P.: Epidvis: A visual web querying tool for animal epidemiology surveillance. Inf. Vis. **19**(1) (2020)
6. Fize, J., Roche, M., Teisseire, M.: Could spatial features help the matching of textual data? Intelligent Data Analysis **24**(5), 1043–1064 (2020). https://doi.org/10.3233/IDA-194749, https://doi.org/10.3233/IDA-194749
7. Hu, Y., Huang, H., Chen, A., Mao, X.L.: Weibo-COV: A large-scale COVID-19 social media dataset from Weibo. In: Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020. Association for Computational Linguistics, Online (Dec 2020). https://doi.org/10.18653/v1/2020.nlpcovid19-2.34, https://aclanthology.org/2020.nlpcovid19-2.34
8. Kampman, O., J. Barezi, E., Bertero, D., Fung, P.: Investigating audio, video, and text fusion methods for end-to-end automatic personality prediction. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 606–611. Association for Computational Linguistics, Melbourne, Australia (Jul 2018). https://doi.org/10.18653/v1/P18-2096, https://aclanthology.org/P18-2096
9. Keim, D., Andrienko, G., Fekete, J.D., Görg, C., Kohlhammer, J., Melançon, G.: Visual Analytics: Definition, Process and Challenges. In: Kerren, A., Stasko, J.T., Fekete, J.D., North, C. (eds.) Information Visualization - Human-Centered Issues and Perspectives, pp. 154–175. No. 4950 in LNCS, Springer (2008), https://hal-lirmm.ccsd.cnrs.fr/lirmm-00272779, state-of-the-Art Survey

10. Kergosien, E., Alatrista-Salas, H., Gaio, M., Güttler, F., Roche, M., Teisseire, M.: When textual information becomes spatial information compatible with satellite images. In: Proc. of the International Conference on Knowledge Discovery and Information Retrieval (KDIR). pp. 301–306 (2015)
11. Khan, W., Malik, U., Ghazanfar, M.A., Azam, M.A., Alyoubi, K.H., Alfakeeh, A.S.: Predicting stock market trends using machine learning algorithms via public sentiment and political situation analysis. Soft Comput. **24**(15), 11019–11043 (2020). https://doi.org/10.1007/s00500-019-04347-y, https://doi.org/10.1007/s00500-019-04347-y
12. Liu, P.: A survey of remote-sensing big data. Frontiers in Environmental Science **3**, 45 (2015). https://doi.org/10.3389/fenvs.2015.00045, https://www.frontiersin.org/article/10.3389/fenvs.2015.00045
13. Neptune, N., Mothe, J.: Automatic annotation of change detection images. Sensors **21**(4) (2021). https://doi.org/10.3390/s21041110, https://www.mdpi.com/1424-8220/21/4/1110
14. Roche, M.: Covid-19 and media datasets: Period- and location-specific textual data mining. Data in Brief **33**, 106356 (2020). https://doi.org/https://doi.org/10.1016/j.dib.2020.106356, https://www.sciencedirect.com/science/article/pii/S235234092031249X
15. Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H.: Fake news detection on social media: A data mining perspective. SIGKDD Explor. Newsl. **19**(1), 2236 (sep 2017). https://doi.org/10.1145/3137597.3137600, https://doi.org/10.1145/3137597.3137600
16. Wang, P.Y., Chen, C.T., Su, J.W., Wang, T.Y., Huang, S.H.: Deep learning model for house price prediction using heterogeneous data analysis along with joint self-attention mechanism. IEEE Access **9**, 55244–55259 (2021). https://doi.org/10.1109/ACCESS.2021.3071306
17. Zhang, X., Zhang, Y., Wang, S., Yao, Y., Fang, B., Yu, P.S.: Improving stock market prediction via heterogeneous information fusion. Knowledge-Based Systems **143**, 236–247 (2018). https://doi.org/https://doi.org/10.1016/j.knosys.2017.12.025
18. Zhou, X., Yang, F., Feng, Y., Li, Q., Tang, F., Hu, S., Lin, Z., Zhang, L.: A spatial-temporal method to detect global influenza epidemics using heterogeneous data collected from the internet. IEEE/ACM Transactions on Computational Biology and Bioinformatics **15**(3), 802–812 (2018). https://doi.org/10.1109/TCBB.2017.2690631