Data Analytics para Variedade de Dados

Data Analytics for Data Variety

Tiago Emanuel Senra da Cruz, Universidade do Minho, Portugal, a66785@alunos.uminho.pt
Jorge Oliveira e Sá, Centro Algoritmi, Universidade do Minho, Portugal, jos@dsi.uminho.pt

Resumo

A internet fez com que os gestores das organizações tivessem acesso a grandes quantidades de dados e esses dados são apresentados em diferentes formatos, em concreto, estruturados, semiestruturados e não estruturados. Esta variedade de dados é essencialmente proveniente das redes sociais, mas não só, também são provenientes da Internet of Things. Verifica-se para os dados estruturados que existem técnicas validadas, estudadas e maduras, mas para os outros tipos de dados, ou seja, semiestruturados e não estruturados tal já não se verifica. Neste poster, é apresentado um conjunto de técnicas de análise de dados para os dados semiestruturados e não estruturados, utilizando como principal bibliografia conferências de investigação na área de análise de dados.

Palavras-chave: Análise de dados, Variedade de dados, Tipo de dados

Abstract

Through the Internet, the organizations managers had access to massive amounts of data and these data are presented in different formats, namely, structured, semi-structured and unstructured. These variety of data is essentially generated from social networks, but not only, they also are generated from the Internet of Things, from machines, sensors, among others. While the structured data has techniques well studied, mature and validated, otherwise the other types of techniques, semi-structured and unstructured, this is no longer true. In this poster, a set of data analysis techniques is presented for the semi-structured and unstructured data by using as main bibliography data analytics conferences.

Keywords: *Data Analytics, Data Variety, Data Types.*

1. DESCRIÇÃO DO TRABALHO (PÓSTER)

Com o aumento da utilização da internet por parte das pessoas e organizações, a quantidade de informação cresceu exponencialmente. O universo digital apresenta uma enorme diversidade de dados gerados através das redes sociais, onde os utilizadores geram conteúdos diversificados como imagens, vídeos, textos, sites, entre outros (Fan e Gordon, 2014), mas não só, também a Internet of Things (Uckelmann et al., 2011), tornou possível que máquinas pudessem comunicar entre si automaticamente, utilizando sistemas de endereçamento exclusivos, e desta forma, consigam trabalhar em conjunto para atingir um fim comum. Esta diversidade de dados é acompanhada por uma variedade de tipo dados, estruturados, semiestruturados e não estruturados, tanto gerados por

ISSN 2183-489X
DOI http://dx.doi.org/10.18803/capsi.v17.00-00

pessoas como máquinas e consiste numa característica a ter em consideração, pois a sua análise traz valor para as organizações (Russom, 2011).

Os gestores para poderem tirar proveito destes tipos de dados, nomeadamente dos dados estruturados, semiestruturados e não estruturados, terão que utilizar diferentes técnicas de análise de dados capazes de retirar informações valiosas para os ajudar na tomada de decisão. As técnicas de análise de dados do tipo estruturado estão maduras e validadas pela comunidade científica, mas tal não se verifica para os dados do tipo semiestruturado e não estruturado.

Com o trabalho realizado pretende-se identificar técnicas de análise de dados para os tipos de dados semiestruturados e não estruturados com o objetivo de perceber qual o valor que se pode obter através da sua análise.

Nas seguintes tabelas são ilustradas um conjunto de técnicas de análise de dados retirados das conferências: Conference of Knowledge Discovery and Data Mining dos anos 2014, 2015 e 2016, Conference of Web Search and Data Mining do ano 2015 e por último, International Journal of Big Data Intelligence do ano 2014 a 2017; e estão agrupadas pelo tipo de dados (semiestruturado e não estruturado).

Para Dados Semiestruturados

Referência	Técnica
(Dong, et al., 2014)	HTML trees (DOM)
(Li et al.2016), (Zhou, Liu, & Buttler, 2015)	Link prediction
(Bi, et al., 2015)	probabilistic Three-way Entity Model (TEM)
(Blanco, Ottaviano, & Meij, 2015)	Entity linking
(Caballero Barajas & Akella, 2015)	Naive Bayes classifier
(Oliveira, Barbar, & Soares, 2016)	multilayer perceptron, (MLP)
(Makrynioti, et al., 2017)	entity recognition
(Makrynioti, et al., 2017)	sentiment analysis

Tabela 1 Técnicas para Dados Semiestruturados

Para Dados Não Estruturados

Referência	Técnica
(Sudhof, Goméz Emilsson, Maas, & Potts, 2014)	Conditional random Fields (CRFs) as a modeling technique
(Dong, et al., 2014)	Natural Language Processing
(Chen & Liu, 2014)	topic modeling
(Kurashima, Iwata, Takaya, & Sawada, 2014)	Probabilistic Latent Semantic Visualization
(Geerdink, 2015)	Data discovery
(Alshareef, Bakar, Hamdan, Abdullah, & Alweshah, 2015)	Association Rule Learning
(Makrynioti, et al., 2017)	entity recognition
(Makrynioti, et al., 2017)	sentiment analysis

Tabela 2 Técnicas para dados Não Estruturados

De forma a perceber qual o valor que esta variedade de dados pode fornecer, irá ser realizado uma experimentação com o objetivo de utilizar diferentes técnicas de análise de dados, nomeadamente o reconhecimento de imagens e análise de textos através do processamento de linguagem natural. A experimentação irá recorrer à ferramenta denominada de Watson Analytics da IBM Bluemix.

Através de técnicas de reconhecimento de imagens irá ser possível obter um conjunto de descritores sobre a imagem e o tipo de taxonomia presente. Estes descritores possibilitam catalogar um conjunto de imagens com características semelhantes.

A técnica de análise de textos através do processamento de linguagem natural possibilita processar textos não estruturados e obter um conjunto de descritores, como entidades, conceitos gerais, palavras-chave, categorias, relações e análises de sentimento.

2. CONCLUSÕES

O trabalho realizado até ao momento possibilita perceber que a variedade de dados é um aspeto importante a ter em consideração no âmbito das organizações, pois as organizações podem estar a não aproveitar informações valiosas e que facilitem o processo de tomada de decisão.

Retirar valor de dados semiestruturados e não estruturados, obriga à perceção de técnicas analíticas existentes para essa variedade de tipos de dados, para tal foi realizada uma pesquisa em conferências científicas para identificar e perceber quais as técnicas utilizadas pela comunidade científica.

Os próximos passos consistirão em realizar várias experimentações utilizando diferentes técnicas de análise de dados do tipo semiestruturado e não estruturado, de forma a perceber o valor que a análise deste tipo de dados pode trazer a uma organização.

REFERÊNCIAS

Fan, W., & Gordon, M. D. (2014). The power of social media analytics. *Communications of the ACM*, *57*(6), 74-81.

Russom, P. (2011). *Big data analytics*. TDWI best practices report, 1-35.

Uckelmann, D., Harrison, M., & Michahelles, F. (2011). An architectural approach towards the future internet of things. In *Architecting the internet of things* (pp. 1-24). Springer Berlin Heidelberg.

APÊNDICE TRABALHOS INVESTIGADOS

Alshareef, A. M., Bakar, A. A., Hamdan, A. R., Abdullah, S. M., & Alweshah, M. (2015). A case-based reasoning approach for pattern detection in Malaysia rainfall data. *International Journal of Big Data Intelligence,* 285-302.

Anagnostopoulos, C., & Triantafillou, P. (2014). *Scaling out big data missing value imputations: pythia vs. godzill*a. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 651-660.

Bi, B., Ma, H., Hsu, B. J., Chu, W., Wang, K., & Cho, J. (2015). *Learning to recommend related entities to search users*. In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, pp. 139-148.

Blanco, R., Ottaviano, G., & Meij, E. (2015). *Fast and space-efficient entity linking for queries*. In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, pp. 179-188.

Caballero Barajas, K. L., & Akella, R. (2015). *Dynamically Modeling Patient's Health State from Electronic Medical Records: A Time Series Approach*. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 69-78.

Chen, N., Hoi, S. C., Li, S., & Xiao, X. (2015). *SimApp: A framework for detecting similar mobile applications by online kernel learning*. In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, pp. 305-314.

Chen, Z., & Liu, B. (2014). *Mining topics in documents: standing on the shoulders of big data*. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 1116-1125.

Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., & Zhang, W. (2014). *Knowledge vault: A web-scale approach to probabilistic knowledge fusion*. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 601-610.

Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., & Zhang, W. (2014). *Knowledge vault: A web-scale approach to probabilistic knowledge fusion*. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 601-610.

Geerdink, B. (2015). A reference architecture for big data solutions-introducing a model to perform predictive analytics using big data technology. *International Journal of Big Data Intelligence*, 236-249.

Kurashima, T., Iwata, T., Takaya, N., & Sawada, H. (2014). *Probabilistic latent network visualization: inferring and embedding diffusion networks*.In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 1236-1245.

Li, L., Yao, Y., Tang, J., Fan, W., & Tong, H. (2016). QUINT: On Query-Specific Optimal Networks.

Liu, J., Aggarwal, C., & Han, J. (2015). *On integrating network and community discovery*. In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, pp. 117-126.

Makrynioti, N., Grivas, A., Sardianos, C., Tsirakis, N., Varlamis, I., Vassalos, V., & Tsantilas, P. (2017). PaloPro: a platform for knowledge extraction from big social data and the news. *International Journal of Big Data Intelligence*, 3-22.

Oliveira, T. P., Barbar, J. S., & Soares, A. S. (2016). Computer network traffic prediction: a comparison between traditional and deep learning neural networks. *International Journal of Big Data Intelligence*, 28-37.

Sudhof, M., Goméz Emilsson, A., Maas, A. L., & Potts, C. (2014). *Sentiment expression conditioned by affective transitions and social force*. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 1136-1145.

Tran, N. K., Ceroni, A., Kanhabua, N., & Niederée, C. (2015). *Back to the past: Supporting interpretations of forgotten stories by time-aware re-contextualization*. In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, 339-348.

Ulanova, L., Yan, T., Chen, H., Jiang, G., Keogh, E., & Zhang, K. (2015). *Efficient Long-Term Degradation Profiling in Time Series for Complex Physical Systems*. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2167-2176.

Wlodarczyk, T. W., & Hacker, T. J. (2014). Current trends in predictive analytics of big data. *International Journal of Big Data Intelligence*, 172-180.

Zhao, Z., Liu, J., & Cox, J. (2014). *Safe and efficient screening for sparse support vector machine*. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 542-551.

Zhou, Y., Liu, L., & Buttler, D. (2015). *Integrating vertex-centric clustering with edge-centric clustering for meta path graph analysis*. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1563-1572.