Mustafa Sofean, J Comput Eng Inf Technol 2018, Volume 7
DOI: 10.4172/2324-9307-C1-020

A Large-Scale Scientific Information Enrichment and Analysis

Mustafa Sofean
FIZ Karlsruhe, Germany

Patent documents are important intellectual resources of protecting interests of organizations, individuals, and companies, as well as they provide an extensive source of technical information. The management and analysis of large-scale patent data involves easy-to-use and adaptable platform for vastly improved efficiencies and better decision-making. The purpose of this presentation is to describe and summarize a scalable text mining framework built on top of big-data architecture and scientific workflow system for semantic annotation and analyzing large-scale patent information. This framework uses different scalable MapReduce-based services for adding semantic enrichment to texts of a single patent. Four scalable services will be presented: structuring detailed description texts of patent into several segments such as background, summary, and embodiments; structuring patent claims to identify the complete claim hierarchy such as independent and dependent claims; technical term extraction to extract the significant keywords and phrases; and extraction of chemical entities and map them with external resources. In addition, the framework uses Spark-based services for large-scale multi-document patent processing and analytics, such as natural language processing, text or metadata clustering, term or metadata counting, and topic modeling. This framework can be used by front-end users to search, process, analyze large-scale patent information in a visual workflow system, and produce results quickly. Two case studies will be presented: analyzing the texts of technology areas of inventions, and technology concepts extraction via applying text mining methods to discover a set of themes that are covering the topic of interest, as well as explore new themes such as emerging technologies. In addition, trend detection methodologies will be used to analyze, extract, and visualize the interesting trends. The execution performances of scalable services are tested with huge amount of patent documents.

Recent Publications

1. Mustafa Sofean. Automatic Segmentation of Big Data of Patent Texts. Big Data Analytics and Knowledge Discovery Conference, DaWaK, Lyon 2017.
2. Aras, Hidir, Rene Hackl-Sommer, Michael Schwantner and Mustafa Sofean. Applications and Challenges of Text Mining with Patents. IPaMin@KONVENS (2014), Hildesheim, Germany
3. Mustafa Sofean and Matthew Smith. Sentiment analysis on smoking in social networks. In Proceedings of the 14th World Congress on Medical and Health Informatics, Copenhagen, 2013.
4. Mustafa Sofean and Matthew Smith. A real-time architecture for detection of diseases using social networks: Design, implementation and evaluation. In Proceedings of the 23rd ACM Conference, HT'2012, Milwaukee, USA, 2012
5. Mustafa Sofean and Matthew Smith. A real-time disease surveillance architecture using social networks. In Proceedings of the European Medical Informatics Conference (MIE), Italy, 2012.
6. Mustafa Sofean, Kerstin Denecke, Avare Stewart, and Matthew Smith. Medical case-driven classification of microblogs: Characteristics and annotation. ACM SIGHIT, IHI. Miami, 2012

Biography

Mustafa Sofean holds a Master degree from University of Freiburg and a Doctorate from University of Hannover. His research interests include Text and Data Mining in Business Application and Machine Learning. He has published many research papers in international conferences. Currently, he is working at FIZ-Karlsruhe, and his current research interests focus on patent domain and include "Text and Data Mining Techniques, and Big Data Analytics".

Mustafa.Sofean@fiz-Karlsruhe.de