*IOSR Journal of Computer Engineering (IOSR-JCE)*
*e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 6, Ver. I (Nov. - Dec. 2016), PP 61-64*
*www.iosrjournals.org*

# Applications for Big Data in of Intelligent Distributed Processing

Dr Farheen Siddiqui

***Abstract:*** *Today, "Big Data" has posed new problem of over-information in many different areas. Such areas include health care (e.g., hospitals, bioinformatics), e-sciences (e.g., physics, chemistry, and geology), and social sciences (e.g., politics) [Bizer et al. 2011, Jung 2009]. Thus, as we have various types of hetrogeneous data from a number of available sources, it is becoming increasingly more difficult to efficiently process such Big Data. Distributed computing technologies (e.g., Hadoop, Hive and Pig) are strongly related to the "Big Data" issues [Hogarth and Soyer 2015, Jung 2012]. Current big data issues are efficient distributed data processing and management for example, information acquisition and stream processing, as well as data integration [Madden 2012]. Also, the big data involves heterogeneous information processing system architectures in various application areas. These information processing systems need to exploit relevant solutions to support a number of intelligent services (e.g., knowledge management and decision making). The aim of this paper is to discuss state of art infrastructure and solutions in areas of distributed computing in different application areas involving big data. This will give an opportunity to push further the discussion upon the potential of knowledge and semantic systems across many communities. This paper will also discuss and analyse "Big data" sources and what is more important to identify the areas where Big data can be applied and provide the knowledge that is not accessible for other types of analysis. Additionally, applications of Big data can be investigated either from static or dynamic perspective.*

## I. Introduction

In current era the advent of Social Media (Facebook and Twitter) and LOD are semantically rich global information source. These information sourses have potential to generate a significant added value in business operations and decisional processes. This value addition in business by information sources is firmly supported by deep research being currently conducted towards exploring such benefits in decision support systems eg brand recognition [Hoffman and Fodor 2010], competitive intelligence [Vuori 2011] or bench marketing [Bingham and Conner 2010]. Studies have been carried out regarding customer relationship management by identifying potential customers or improving and enriching the stored information about the client portfolio of the company at hand. Likewise, there are very few contributions to the literature addressing competition analysis based on public information; the existing ones (e.g. [He, Zha and Li 2013]) focus exclusively on data repositories of a certain class (e.g. Social Media), hence discarding its combination with related information sources of different characteristics. Furthermore, from a technical perspective the heterogeneity of the data coming from these sources comes along with non-standard, unreliable schemas that require a significant human effort to extract, format and assimilate knowledge. Indeed, the removal of noisy content (understood as the process of filtering out data due to their semantic irrelevance or lack of integrity) is mandatory before any knowledge inference stage. Another related issue inheres in how to merge these datasets with traditional business core systems such as relational database management servers or corporative repositories [Pham and Jung 2014].

## II. Big Data Application For Heterogeneous Databases

There exist many interesting big data analytical functionalities for heterogeneous databases. Particularly, two complementary use cases illustrate the potential of using the open data in the business domain. The first represents the creation of an existing and potential customer knowledge base, exploiting social and linked open data based on which any given organization might infer valuable information as a support for decision making. The second focuses on the classification of organizations and enterprises aiming at detecting potential competitors and/or allies via the analysis of the conceptual similarity between their participated projects Research in this area allow stepping further beyond the issues identified above by and many novel Client Relationship Management (CRM) system with extended analytical functionalities (information aggregation/fusion and knowledge discovery) applied over a semantically aggregated information database are developed . Technically speaking, these setup follows a semantic aggregation approach that allows retrieving, combining and analyzing information from emerging datasets (in particular, Social Media and LOD) with other corporate databases. This embodies an integral, universal platform that implements diverse BI functionalities which, without loss of generality, will be exemplified within this manuscript by 1) the retrieval of extended information through the customer database; and 2) the analysis of competitors/allies based on the cosine

similarity of published projects and initiatives participated by every client within the CRM database. Research work will show how semantic tools and Big Data technologies for information collection and aggregation can be hybridized so as to yield BI insights leveraging not only corporate datasets, but also the information contained in LOD and social platforms.

## III. Big Data And Semantic Web Functionalities

One of the best application of big data and semantic web is to understand big data processing scheme to in network eg public bus networks. The process has studied modeling and linking accessibility data by using ontological knowledge. Currently, the Linked Open Data (LOD) [LOD, 14] initiative presents a challenge in the area of Information Technologies (IT). It provides the mechanism for publishing, enriching and sharing data, information and knowledge on the Web, using semantic web technologies. It comprises the following principles: (a) using Universal Resource Identifiers for identifying all kinds of elements ("things"), (b) making these URIs accessible via the HTTP protocol and (c) providing a description for these things using the Resource Description Format (RDF) [Klyne, 12] along with (d) URI links to related information. This paradigm can be applied to heterogeneous distributed data sources, which are therefore integrated, and published as LOD. This way, a particular application is now able to use the information of public transport networks, which of course must be considered of public interest. Due to the large size and variety of these data, LOD provides both the mechanism to publish them and to do it in a flexible way, able to support the definition of new mobility services for citizens. A classic example could be route planning systems, which combine data from different public transit networks. The relevance of open data in the context of transport networks is significant, and this can be shown using both research papers and communications from several big organizations, for instance [Epsi, 12], [Hobbs, 14], [ODUK, 14]. Having this in mind, we intend to apply Information Technologies to the task of improving mobility services – specifically considering the case of citizens, both in usual or occasional trips, and trying to optimize their intended routes by using any available means of transport: both public transport and the rational sharing of private transport. Most interesting is an IT platform, called CoMobility [Cuesta, 13], to assist those citizens in the use of intermodal transport sharing, and integrating carpooling with public transport, as well as other private transport media. Besides, This paper also consider the accessibility of these media as a relevant aspect of public transit networks, as it makes possible the mobility of people with special needs. This paper want to emphasize that, in this article, the term "accessibility" is specifically referring to "accessibility for people with special needs" (that is, accessibility for blind people, or people using a wheelchair, who have permanent mobility issues; but also people pushing a baby carriage or carrying luggage, etc). Which data is required from transport networks, when users in transit make a request to a service to define a route? Is there any support information about the transport network already available? Trying to answer these questions, This paper have studied the already existing transport metamodels, specifically designed to be generic: Transmodel [Transmodel, 14] and IFOPT [IFOPT, 14]. Our intention was to match the actual data from public transport networks with such standards. Transmodel is a European Reference Data Model for Public Transport Information, which provides a model of public transport concepts, and data structures that can be useful to build information systems related to the different kinds of public transport. But Transmodel does not provide relevant information about accessibility. For this reason, This paper have also studied the related IFOPT metamodel, conceived itself as an extension of Transmodel. It defines a model (and also the identification principles) for the main fixed objects related to public access to Public Transport (e.g. stop points, stop areas, stations, connection links, entrances, etc.). It already includes specific structures to describe accessibility data about the equipment of vehicles, stop places and access areas. Therefore, This paper have defined a conceptual model by means a UML class diagram, supported by the IFOPT model as a basis, which includes new (additional) accessibility elements (classes and attributes) related to vehicle equipment. This model describes our reference ontology, designed with the purpose to define accessibility features in public transport, and specifically on buses. There are already several existing works dealing with ontologies for public transport in the literature. [Timpf, 02] proposes an ontology of wayfindings, from a traveller's perspective. His work is not based on any transport metamodel or standard, and it does not take into account any information about accessibility elements of public transport. The same happens with [Becker, 97], which describes an ontology for public transport which is actually based on a pre-existing, generic ontology for scheduling; it lacks a lot of detail about the transport domain, and does not consider accessibility issues. Other proposals define more evolved ontologies for public transport, either following or considering Transmodel: [Houda, 10], [Marçal de Oliveira, 13]. But again, like in the works mentioned above, they do not include any information about accessibility elements in public transport. To simplify the exposition, this article focuses on a specific issue in accessibility, namely how to publish the information about the equipment of public buses and their accessibility, as open data. However the intention is not to lose any generality: only the presentation is simplified – our models, ontology and applications still consider many relevant aspects of accessibility, besides equipment itself. The format of public data, within many existing open data initiatives, prevents non-experts from using them directly, and thus it requires

additional semantics, as provided by Linked Open Data [Heath, 11]. There are several proposals in this regard, within the specific context of transport networks. [Colpaert, 14] presents an immature proposal of a route planning system which uses Linked Open Data, as an initial idea of a doctoral thesis; but it never takes into account accessibility elements. [Pham and Jung, 14] presents a workflow for publishing and linking transport data on the web; moreover, they apply their linked transport data proposal in two different real-world datasets. But again, they do not take into account accessibility issues, either. In this article, our proposal focuses on modelling the accessibility and transport data for the public bus network. To do so,

## IV. Big data And Digital Content

In this area there exist many the software platform for discovering contents and stories in the movies. They claims that it is an important big data sources for digital cultural contents and understanding our society. The system automatically understands the movies by discovering social networks and measuring various social measurements. Discovering content and stories in movies is one of the most important concepts in research studies. To consider how to efficiently determine the relationships between characters, This paper focus on the appearance of each character during movie play and analyze the characters' relationships to build a CoCharNet, a social network, to determine characters' relationships in a movie. There are many innovative method which extracts and analyzes characters' relationships based on social network measurement and evaluation.

## V. Big Data And Parallel Processing

This section focuses on parallel data processing and parallel streaming systems for big data analytics. One of the key components of these systems is the task scheduler which plans and executes tasks spawned by the application on available CPU cores. The proposed task scheduler combined with the new memory allocator achieve up to speed up on a NUMA system and up to 10% speed up on an older SMP system with respect to the unoptimized versions of the scheduler and allocator. Parallel data processing and parallel streaming systems become quite popular. They are employed in various domains such as real-time signal processing, OLAP database systems, or high performance data extraction. One of the key components of these systems is the task scheduler which plans and executes tasks spawned by the application on available CPU cores. The multiprocessor systems and CPU architecture of the day become quite complex, which makes the task scheduling a challenging problem. In this paper, This paper propose a novel task scheduling strategy for parallel data stream systems, that reflects many technical issues of the current hardware. In addition, researchers have implemented a NUMA aware memory allocator that improves data locality in NUMA systems. The task scheduler combined with the new memory allocator achieve up to 3× speed up on a NUMA system and up to 10% speed up on an older SMP system with respect to the unoptimized versions of the scheduler and allocator. Many of the ideas implemented in parallel framework may be adopted for task scheduling in other domains that focus on different priorities or employ additional constraints.

## VI. Big Data And Distributed Regression System

Another major application area of big data is distributed regression problem. Distributed Regression System makes use of a discrete representation of the probability density functions. Neighbourhoods of similar datasets are detected by comparing their approximated pdfs. This information supports an ensemble-based approach, and the improvement of a second level unit, as it is the case in stacked generalization. When distributed data sources have different contexts the problem of Distributed Regression becomes severe. It is the underlying law of probability that constitutes the context of a source. A new Distributed Regression System is presented, which makes use of a discrete representation of the probability density functions (pdfs). Neighborhoods of similar datasets are detected by comparing their approximated pdfs. This information supports an ensemble-based approach, and the improvement of a second level unit, as it is the case in stacked generalization. Two synthetic and six real data sets are used to compare the proposed method with other state-of-the-art models. The obtained results are positive for most datasets.

## References

[1]. [Bizer et al. 2011] Bizer, C., Boncz, P., Brodie, M.L., Erling, O.: The Meaningful Use of Big Data: Four Perspectives - Four Challenges. SIGMOD Record, 40(4):56-60, 2011.
[2]. [Hogarth and Soyer 2015] Hogarth, R.M., Soyer, E.: Using Simulated Experience to Make Sense of Big Data. MIT Sloan Management Review, 56(2):49-54, 2015.
[3]. [Jung 2009] Jung, J.J.: Contextualized query sampling to discover semantic resource descriptions on the web. Information Processing & Management, 45(2):283–290, 2009.
[4]. [Jung 2012] Jung, J.J.: Evolutionary Approach for Semantic-based Query Sampling in Large-scale information Sources. Information Sciences, 182(1):30-39, 2012.
[5]. [Madden 2012] Madden, S.: From Databases to Big Data. IEEE Internet Computing, 16(3):4-6, 2012. 756 Jung J.J., Camacho D., Badica C.: Intelligent Distributed Processing ...

[6]. [Hoffman and Fodor 2010] Hoffman, D. L., Fodor, M.: "Can You Measure the ROI of Your Social Media Marketing?"; *MIT Sloan Management Review* 52, 1 (2010), 41-49.

[7]. [Vuori 2011] Vuori, V.: "Social Media Changing the Competitive Intelligence Process: Elicitation of Employees Competitive Knowledge"; Julkaisu-Tampere University of Technology (2011) 1001.

[8]. [He, Zha and Li 2013] He, W., Zha, S., Li, L.: "Social Media Competitive Analysis and Text Mining: A Case Study in the Pizza Industry"; *International Journal of Information Management* 33, 3 (2013) 464-472.

[9]. [Pham and Jung 2014] Pham, X. H., Jung, J. J.: "Recommendation System Based on Multilingual Entity Matching on Linked Open Data"; *Journal of Intelligent & Fuzzy Systems*, 27, 2(2014) 589-599.

[10]. [LOD, 14] Long, N.H., Jung, J.J.: "Privacy-aware Matching Online Social Identities for Multiple Social Networking Services," Cybernetics and Systems, 46, 1-2, 69-83, 2015

[11]. [Klyne, 12] Klyne, G., Carroll, J.: Resource Description Framework (RDF): Concepts and Abstract Syntax. W3C Recommendation. Available at: http://www.w3.org/RDF/

[12]. [Epsi, 12] e-Public Sector Information Platform (ePSI Platform): Open Transport Data Manifesto, September 2012. Available at: http://www.epsiplatform.eu/

[13]. [Hobbs, 14] Hobbs, A., Hanley, S.: Big and Open Data in Transport. POST notes POST PN 472, August 2014. Available at: http://www.parliament.uk/briefing-papers/post-pn-472/bigand-open-data-in-transport

[14]. [ODUK, 14] London Underground Open Data (ODUK), accessed Nov 2015: http://www.tfl.gov.uk/info-for/open-data-users/

[15]. [Cuesta, 13] Cuesta, C.E., Cáceres, P., Vela, B., Cavero, J.M.: CoMobility: A Mobile Platform for Transport Sharing, ERCIM News 2013(93), 2013.

[16]. [Timpf, 02] Timpf, S.: Ontologies of Wayfinding: a traveler's perspective. Networks and Spatial Economics, Kluwer Publishers 2(1): 9-33, 2002.

[17]. [Becker, 97] Becker, M., Smith, S. F.: An ontology for multi-modal transportation, planning and scheduling. Technical Report CMU-RI-TR-98-15, Robotics Institute, Carnegie Mellon University, Pittsburgh, USA.

[18]. [Houda, 10] Houda, M., Khemaja, M., Oliveira, K., Abed, M.: A public transportation ontology to support user travel planning in Proceedings of the Fourth International Conference on Research Challenges in Information Science, 2010.

[19]. [Marçal de Oliveira, 13] Marçal de Oliveira, K., Bacha, F., Mnasser, H., Abed, M.: Transportation ontology definition and application for the content personalization of user interfaces, Expert Systems with Applications, 40, 3145–3159, 2013.

[20]. [Heath, 11] Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space, Morgan & Claypool, 2011.

[21]. [Colpaert, 14] Colpaert, P.: Route Planning Using Linked Open Data, The Semantic Web: Trends and Challenges, Lecture Notes in Computer Science Volume 8465, 827-833, 2014.

[22]. [Pham and Jung, 14] Pham, X.H., Jung, J.J.: "Recommendation System Based on Multilingual Entity Matching on Linked Open Data," Journal of Intelligent & Fuzzy Systems, 27, 2, 589- 599, 2014