Decision Support Systems for the Industry

Research in Computing Science

Research in Computing Science es una publicación trimestral, de circulación internacional, editada por el Centro de Investigación en Computación del IPN, para dar a conocer los avances de investigación científica y desarrollo tecnológico de la comunidad científica internacional. **Volumen 148, No. 4,** abril de 2019. *Certificado de Reserva de Derechos al Uso Exclusivo del Título* No.: 04-2005-121611550100-102, expedido por el Instituto Nacional de Derecho de Autor. *Certificado de Licitud de Título* No. 12897, *Certificado de licitud de Contenido* No. 10470, expedidos por la Comisión Calificadora de Publicaciones y Revistas Ilustradas. El contenido de los artículos es responsabilidad exclusiva de sus respectivos autores. Queda prohibida la reproducción total o parcial, por cualquier medio, sin el permiso expreso del editor, excepto para uso personal o de estudio haciendo cita explícita en la primera página de cada documento. Distribuida por el Centro de Investigación en Computación, Av. Juan de Dios Bátiz S/N, Esq. Av. Miguel Othón de Mendizábal, Col. Nueva Industrial Vallejo, C.P. 07738, México, D.F. Tel. 57 29 60 00, ext. 56571.

Editor responsable: *Grigori Sidorov, RFC SIGR651028L69*

Research in Computing Science is published by the Center for Computing Research of IPN. **Volume 148**, **No. 4**, April 2019.

Decision Support Systems for the Industry

Cuauhtémoc Sánchez-Ramírez
Giner Alor-Hernández
Jorge Luis García-Alcaraz (eds.)

Instituto Politécnico Nacional, Centro de Investigación en Computación
México 2019

ISSN: 1870-4069

Instituto Politécnico Nacional (IPN)
Centro de Investigación en Computación (CIC)
Av. Juan de Dios Bátiz s/n esq. M. Othón de Mendizábal
Unidad Profesional "Adolfo López Mateos", Zacatenco
07738, México D.F., México

http://www.rcs.cic.ipn.mx
http://www.ipn.mx
http://www.cic.ipn.mx

Indexed in LATINDEX, DBLP and Periodica

Electronic edition

ISSN 1870-4069

Editorial

The development of models and methodologies to improve the industrial processes is the principal objective of Decision Support Systems (DSS), for this reason in this volume eight papers are presented that were carefully selected of 11 submissions about the use of different techniques for designing and developing Decision Support System (DSS) in industrial contexts. The papers were evaluated by an editorial board integrated for reviewers with international prestige in the area. The papers were selected by considering the originality, scientific contribution to the field and technical quality of the papers. The articles were considered in the following industrial contexts: (1) Operational risk management in a retail company; (2) Effect of knowledge transfer and SC complexity on human performance and flexibility; (3) An Urban Supply Chain Distribution Model; (4) Impact of Managers and Human Resources on Supply Chain Performance; (5) Design of a Language for IoT Service Composition; (6) Automated Fault Detection and Diagnostics for Aluminum Threads Using Statistical Computer Vision; (7) Inspection System with Neural Network and Vision Techniques for the Manufacture Industry; (8) Towards a proposal of personalized medial decision support systems: analysis of gene expression levels of diabetes mellitus, inflammation and oxidative stress in Alzheimer´s disease.

The volume also contains two regular papers on speech recognition and security in data warehousing.

The editors would like to express their gratitude to the reviewers who kindly contributed to the evaluation of papers at all stages of the editing process. They equally thank the Editor-in-Chief, Prof. Grigori Sidorov, for the opportunity offered to edit this special issue and for providing his valuable comments to improve the selection of research works. Guest editors are grateful to the National Technological of Mexico for supporting this work and the National Council of Science and Technology (CONACYT) as part of the project named Thematic Network in Industrial Process Optimization.

Cuauhtémoc Sánchez-Ramírez (Instituto Tecnológico de Orizaba, Mexico)
Giner Alor-Hernández (Instituto Tecnológico de Orizaba, Mexico)
Jorge Luis García-Alcaraz (Universidad Autónoma de Ciudad Juárez, Mexico)
Guest editors
June 2019

ISSN 1870-4069

Table of Contents

Page

ISSN 1870-4069

An Urban Supply Chain Distribution Model

Cristian Giovanny Gómez-Marín[1], Julian Andres Zapata-Cortes[2*], Martín Dario Arango-Serna[1], Conrado Augusto Serna-Uran[3]

[1] Universidad Nacional de Colombia – Sede Medellín, Medellín, Colombia
[2] Institución Universitaria CEIPA, Sabaneta, Antioquia, Colombia
[3] Universidad de San Buenaventura – Seccional Medellín, Medellín, Colombia
crggomezma@unal.edu.co; julian.zapata@ceipa.edu.co; mdarango@unal.edu.co; conrado.serna@usbmed.edu.co

Abstract. Increased activities in urban areas related with goods transportation lead companies to look for new strategies in order to develop those process in a more efficient way, aiming to reduce costs and increase customer's satisfaction. This paper presents an urban supply chain framework and a Mixed Integer Linear Programing Model for its optimization. The model uses different goods distribution actors, including several suppliers, one consolidation facility and several customers. The proposed framework and optimization model allow to generate optimal routes and the assignment of the customers and suppliers in the distribution network.

Keywords: distribution strategies, mixed integer linear programing, multi-echelon distribution system, supply chain, urban goods distribution.

1 Introduction

In 2018 the people living in cities is around the 55% of the world's population and it is expected to increase to 68% by 2050 [1]. One of the most important issues to face in city planning and administration is the freight exchange to satisfy citizen's needs [2]. For that reasons, Urban Goods Distribution (UGD) is an important research field that impacts directly those key aspects, analyzing and proposing feasible solution that improve company competitiveness and people quality of life [3].

In order to propose solutions to organizing an efficient UGD, several elements must be considered, as for example the coordination and individual goals of the involved actors [4-6], the physical infrastructure, the economic vocation and the sustainability of the city, among others [7,8]. The cooperation and coordination among several actors is a key activity to generate proposals that improve the costs and customers service level in the UGD [10,11], both in the private and the public sectors [9].

According to Danielis et al. [12] and Tachizawa et al. [13] there is a need to research in urban goods distribution processes with the aim of understand how different configuration, actors and their behaviors, impact the network performance. Several

* Corresponding author.

initiatives can be found in the literature to improve city goods distribution [9,14], in which coordination and collaboration between actor are highlighted as one of the more attractive initiatives to implement in Urban Goods Distribution [8]. Some research that integrate the coordination of multiple stakeholders, representing the diverse objective and behaviors could be founded in Nathanail [15], who realized that the cooperation among the actors is essential to obtain goods results of the operative. Bean and Joubert [16] analyzed the coordination between several actors, in this case considering the mutual interaction and responses of the carrier and customers, also finding improvements in the distribution performance. Liu et al [17] formulated a coordination strategy using two vehicles that can complement their planed routes, producing a decrease in the travel distance and optimizing the problem using a hybrid ant colony heuristic. Gutierrez et al [18] use a memetic algorithm to tackle the uncertainties in the routing problem improving the service and travel times. Aragao et al. [19] evaluated two different types of cooperative strategies (the use of additional auxiliary vehicles and negotiation between vehicles) to face the high degree of randomness presented in the urban freight distribution variables in a dynamic vehicle routing.

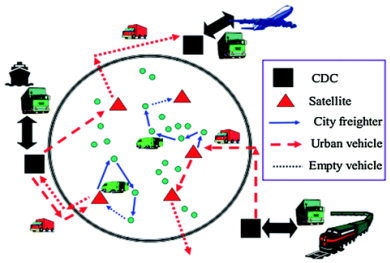
This paper presents a coordination model for the routing assignment optimization process in a two-echelon supply chain that consider suppliers, consolidation facilities and customers in the distribution process in the Hotel, restaurants and catering sector. In following section, the urban supply chain concept and some urban goods distribution strategies are provided by a literature review, followed by the urban supply chain model proposal. After that, the model is applied for several instances and the results are analyzed to finally present the conclusions and future research lines derived from the study.

2 Urban Supply Chain (USC) and Goods Distribution Strategies

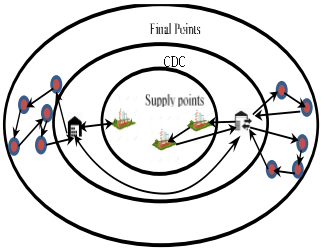
The aims of Urban Supply Chain (USC) are to minimize the total cost of the pick-up and delivery process or maximize the benefits for the main actors, in most of the cases, without considering the sustainability of the supply chain [20, 21], especially the social and environmental dimensions [11]. There are several goods distribution strategies in order to achieve those objectives, such as direct or multi-stop deliveries, or the use of one or several consolidation platforms in which products are consolidate to further deliveries to customers [8,22]. For the consolidation strategies it can be found several distribution networks, from which the most common are the Two tier [23, 24], the Multi-layer [25] and the Four-layer city distribution network [26, 27], which are depicted in figure 1.

The two-tier city logistic system is based on a City Distribution Center (CDC) located on the edge of the urban zone and a set of satellites platforms that receives the goods form the CDC and deliver those to the customers using two types of vehicles, one to move the freight form CDC to satellites and other one to make the final delivery [23,12-32]. In the Multi-layer scheme, a first layer is dedicated to the suppliers, a second layer is dedicated to the CDC (or Hub) to coordinate the inbound and outbound

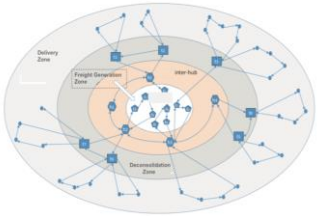
flow of freight, and a third layer focused on the customers or the final distribution points [3,25,33,34]; The four layers' distribution strategy divided the geographic operation area into consolidation and deconsolidation zones. The first layer focused on transportation between supply points and consolidation centers (hubs), the second layer is dedicated to transport freight from hub to hub, the third layer is a deconsolidation zone at terminals or satellites facilities and the fourth layer is the delivery zone from terminals to final customers [26,27].



a. Two tier.



b. Multi-layer.



c. Four-layer.

Fig. 1. Most common city consolidation distribution system.

The strategies mentioned are used in the Urban Goods Distribution for different commercial and industrial sectors, where the retail sector is one of the most studied to perform and assess different strategies that improve the supply chain [35]. Other sector that can be analyzed under those distribution strategies is the Hotels, Restaurants and Catering distribution sector (Ho.Re.Ca) due to its similarities with the retailer sector, especially in the replenishment decision-making processes [35]. In the following section an integrative urban supply chain framework is proposed for the Ho.Re.Ca sector in which different distribution models can be analyzed.

3 Urban Supply Chain Model for the Urban Goods Distribution

In this section it is proposed an urban supply chain network for the Ho.Re.Ca sector, where multiple products should be delivered to the final consumption points [36] and three delivery alternatives are considered: a) direct deliveries from the producer to the final customer using an own fleet of vehicles, b) use of wholesalers (hub) for a specific geographic zone and d) use a combination of such alternatives. Normally, in this sector there are multiple suppliers and customers, logistics service providers and distribution

centers (Hubs). The proposed model has three levels in which three different types of pick-up and delivery routes can be made:

Pick-up route - r1: This route is just dedicated to pick-up the products at supplier locations for the transportation to the consolidation center (hub).

Pick-up and delivery routes - r2: In this type of route both pick-up and delivery process are performed, first the pick-up from the supplier and then the delivery services are accomplished without consolidation at the hub. The hub could be a node to deliver goods.

Delivery Routes r3: In this type of routes goods deliveries from the hub are made to customers. These deliveries can be made directly or in tour way.

In this structure all vehicles must leave and arrive from the distribution center and each supplier produce only one type of product. This routes are depicted in figure 2.

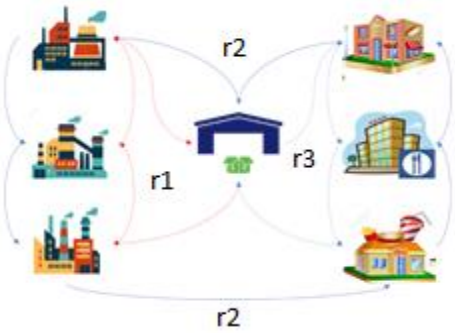


Fig. 2. Proposed Three level distribution network

4 Model Formulation

In order to define the mathematical model, a direct graph $G = (N, A)$ is considered, where N is the set of nodes $N = C \cup F \cup 0$ containing the subset $C \coloneqq \{j, \dots, j'\} \in \mathbb{N}$ that represent the customers, subset $F := \{i, \dots, i'\} \in \mathbb{N}$ representing the suppliers and node 0 representing the hub (only one hub is consider in this case). The set of arcs A is the link between the nodes. There are complete subgraphs consisting of suppliers $i, \dots, i'$ and the hub 0; and customers $j, \dots, j'$ and the hub 0. To ensure that there are direct transportation between suppliers and customers, G contain arcs $\{i, j\}$, $i \in F, j \in C$. The homogenous fleet of vehicles are indexed by $k \in K$ where $K \in \mathbb{N}$, and those start and finish their routes $R: \{r_1, r_2, r_3\}$ at the hub 0.

To ensure that the hub behaves as a consolidation center, the route r1 must be executed before r3 allowing that the products that arrive to the hub in route one, can be processed for the distribution in further transports in route r3.

The Objective Function that minimizes the transportation cost using mixed integer linear programming formulation (MILP) is presented in equation 1 and three groups of different constraints are considered (Common constraints for routing problems, constraints for vehicle flow conservation, constraints for capacity in each type of route R). The constraints are presented in annex A.

$$
\begin{aligned}
Min \sum_{k\in K}\left[\sum_{i\in F} C_{0,i}x_{0,i}^{k,r_1} + \sum_{i\in F}\sum_{j\in F, i\neq j} C_{i,j}x_{i,j}^{k,r_1} + \sum_{i\in F} C_{i,0}\, x_{i,0}^{k,r_1}\right] \\
+ \sum_{k=1}^{K}\left[\sum_{i=1}^{F} C_{0,i}x_{0,i}^{k,r_2} + \sum_{i}^{F}\sum_{j,i\neq j}^{F} C_{i,j}x_{i,j}^{k,r_2} + \sum_{i}^{F}\sum_{j}^{C} C_{i,j}x_{i,j}^{k,r_2}\right. \\
\left. + \sum_{i}^{C}\sum_{j,i\neq j}^{C} C_{i,j}x_{i,j}^{k,r_2} + \sum_{i}^{C} C_{i,0}\, x_{i,0}^{k,r_2}\right] \\
+ \sum_{k=1}^{K}\left[\sum_{j=1}^{C} C_{0,j}x_{0,j}^{k,r_3} + \sum_{i}^{C}\sum_{j,i\neq j}^{C} C_{i,j}x_{i,j}^{k,r_3} + \sum_{i=1}^{C} C_{i,0}\, x_{i,0}^{k,r_3}\right].
\end{aligned}
\tag{1}
$$

The model notation for parameters and variables is:

K={k} $Set\ of\ vehicles.$
C^k $Capacity\ of\ the\ vehicle\ k \in K.$
$d_{j,p}$ Demand form the customer j ∈ C of product p ∈ P.
$o_{i,p}$ Quantity of product p ∈ P supplied by supplier i ∈ F.
λ_i^k Time when the vehicle k ∈ K begins it service to the customer j ∈ C.
u_j $service\ time\ duration\ at\ customer\ j \in C.$
e_j, l_j Time window for the service at customer j ∈ C.
t_{ij}^k Travel time of vehicle k ∈ K between the nodes i, j ∈ N.
$q_{i,p}^{k,r}$ quantity of product p ∈ P transported in the vehicle k ∈ K before visit the node i ∈ {F∪0∪C} on the route r∈{R}.
$\rho_{j,p}^k$ Quantity of product p ∈ P in the vehicle k ∈ K that leave the hub and must be delivered at the nodes j ∈ C.
x_{ij}^{kr} Decision binary variable, 1 if the vehicle $k \in K$ uses the arc from i to j in the route $r \in R$, otherwise 0.

5 Discussion and Analysis of Results

To solve the problem presented in the former section in an exact way, the model was formulated on GAMS (General Algebraic Modeling System software) and solved using Mixed Integer Linear Programing (MILP) in the CPLEX Solver. For testing the model, five small and medium instances were randomly generated, considering less suppliers than customers like in real applications. The network configuration for the test instances and the cost obtained with the model are presented in Table 1, in all cases using a single hub, several suppliers and customers.

For the instances 1 and 2, with six and eight nodes respectively, the better routes to perform are route 2 (pick-up and delivery route). For the instance 3 with 10 nodes, the routes two and three are used for accomplishing the service to all the customers. The same occurs for instances 4 and 5 with 12 and 14 nodes respectively, in which routes two and three are generated as presented in Table 2. Figure 3 depicts the routes between suppliers, customers and the Hub for the instance 5 with 14 nodes at a Cartesian plane

that allows to obtain Euclidian distances and generate the routes construction according to the MILP optimization model.

Table 1. Instance configuration

Instance	Configuration (F – 0 – C)*	# nodes	Cost
1	2 – 1 – 3	6	244
2	2 – 1 – 5	8	318
3	3 – 1 – 6	10	470
4	3 – 1 – 8	12	493
5	3 – 1 – 10	14	494

* notation used in the model formulation (F: suppliers, 0: Hub, C: customers)

Table 2. Tour and routes for the instances.

Instance	Tour	Route	Distance
1	0 – F1 – F2 – C1 – C2 – C3 – 0	r2	244
2	0 – F1 – F2 – C1 – C5 – C4 – C3 – C5 - 0	r2	318
3	0 – F3 – F2 – F1 – C6 – C5 – C1 – 0.	r2	228
	0 – C4 – C3 – C2– 0	r3	242
4	0 – F1 – F2 – F3 – C6 – C5 – C1 – 0	r2	228
	0 – C2 – C3 – C4 – C7 – C8 – 0	r3	265
5	0 – F3 – F2 – F1 – C6 – C9 – C5 –C1 – 0	r2	228
	0 – C2 – C3 – C4 – C7 – C8 – C10 – 0	r3	266

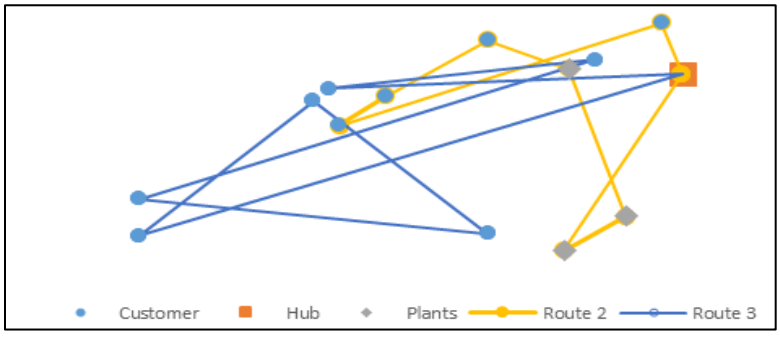


Fig. 3. Routes generated for the instance 5.

From table 2 it can be observed that the greater the number of nodes, the greater the total travel distance. Also, when there are few nodes just one route is activated. As the number of nodes increases, new routes are activated to divide the travel distance and equilibrate the loads.

When the instances include more than 15 nodes “the dimensionality curse” has their effect and the computation time and the computational memory use increases significantly, and the solver does not find a feasible solution in a reasonable

computational time, so for larger instances the problem must be solved using a more robust and fast computer which increase computational calculation cost, reason of why it is suggested the use of metaheuristic methods, such as genetic algorithms or Tabu search for techniques based on population and trajectory respectively [37].

6 Conclusions

Urban fright distribution includes several actors and transportation levels that must be consider in order to model and solve problems according to real conditions. Different strategies have been used in urban supply chain for the Urban Goods Distribution processes, in which is difficult to assign routes and vehicles to the facilities involved in the transport/distribution process. This article presented a Mixed Integer Linear Programming (MILP) model that allows configure the routes and the facility assignments to such routes for small and medium instances, that can be solved exactly for distribution network involving one hub and several suppliers and customers.

The results obtained for the model in the tested instances allows conclude about the ability of the model to obtain solution for a very common distribution problem in the Ho.Re.Ca sector. The behavior of the model is according to what was expected for the solutions of the tested instances. In this way, when solving small problems, a single route is required, but when bigger instances have to be solved, several routes and a more complex distribution plans are generated by the MILP model. However, when the proposed model has to solve instances with more than 14 nodes, the computational time and the memory requirements increase significantly, reason of why it is suggested the use of metaheuristics methods to solve those problems.

As future research lines, it is suggested to develop metaheuristics methods for solving the model in order to analyze complex distribution networks with more suppliers, hubs and customers. It is also interesting to incorporate more variables and conditions in the urban distribution model, such as traffic congestion and time windows. Another future research line is the use of multi-agent systems for asses the different behaviors and the response levels to dynamic changes in operational conditions, such as travel time, service time and dynamic demands.

References

1. United Nations Department of Economic and Social Affairs: World Urbanization Prospects: The 2018 Revision – Press release. Available in: https://population.un.org/wup/Publications/. Last Visited: September 1 of 2018 (2018)
2. Ros-McDonnell, L.R., De-la-Fuente-Aragón, M.V., Ros-McDonnell, D., Cardós, M.: Analysis of freight distribution flows in an urban functional area. Cities, Volume 79, pp 159–168 (2018)
3. Arango-Serna, M.D., Serna-Uran, C.A., Zapata-Cortes, J.A.: Multi-agent System Modeling for the Coordination of Processes of Distribution of Goods Using a Memetic Algorithm. In: García-Alcaraz J., Alor-Hernández G., Maldonado-Macías A., Sánchez-Ramírez C. (eds) New Perspectives on Applied Industrial Tools and Techniques. Management and Industrial Engineering. Springer, Cham (2018)

4. Arango, M.D., Zapata, J.A., Gutierrez, D.: Modeling The Inventory Routing Problem (IRP) With Multiple Depots with Genetic Algorithms. IEEE Latin American Transactions 13(12): 3959– 965 (2015)
5. Arango-Serna, M.D., Zapata-Corte,s J.A., Serna-Uran, C.A.: Collaborative Multiobjective Model for Urban Goods Distribution Optimization. In: García-Alcaraz J., Alor-Hernández G., Maldonado-Macías A., Sánchez-Ramírez C. (eds) New Perspectives on Applied Industrial Tools and Techniques. Management and Industrial Engineering. Springer, Cham (2018)
6. Arango, M.D., Zapata, J.A.: Multiobjective Model for The Simultaneous Optimization of Transportation Costs, Inventory Costs and Service Level in Goods Distribution. EEE Latin america Transactions 15(1), pp. 129–136 (2017)
7. Arango-Serna, M.D., Gómez-Marín, C.G., Serna-Urán, C.A.: Modelos logísticos aplicados a la distribución urbana de mercancías. Revista EIA 14(28), pp .57–76 (2017)
8. Zapata-Cortes, J.A.: Optimización de la distribución de mercancías utilizando un modelo genético multiobjetivo de inventario colaborativo de m proveedores con n clientes. Tesis Doctoral, Universidad Nacional de Colombia (2016)
9. Muñuzuri, J.: La logística urbana de mercancías: Soluciones, modelado y Evaluación. Doctor Ingeniero Industrial, Universidad de Sevilla, Sevilla. España (2003)
10. Febbraro, A., Di-Sacco, N., Saeednia, M.: An agent-based framework for cooperative planning of intermodal freight transport chains. Transportation Research Part C, 64, pp.72–85. Available at: https://ac-els-cdn-com.ezproxy.unal.edu.co/S0968090X15004301/1-s2.0-S0968090X15004301-main.pdf?_tid=spdf-1e85e57d-447d-4545-a6fa-d476dd10b5ef&acdnat=1518811138_c7cf8c8b75494658b8580d9737653630 [Accessed February 16, 2018] (2016)
11. Österle, I., Aditjandra, P.T., Vaghi, C., Grea, G., Zunder, T.H.: The role of a structured stakeholder consultation process within the establishment of a sustainable urban supply chain. Supply Chain Management: An International Journal 20(3), pp. 284–299. Available at: https://doi.org/10.1108/SCM-05-2014-0149 [Accessed May 15, 2018] (2015)
12. Danielis, R., Maggi, E., Rotaris, L.,.Valer, E.: Urban freight distribution: Urban supply chains and transportation Policies. In M. Ben-Akiva, H. Meersman, & E. Van de Voorde (eds). Freight Transport Modeling. Emerald Group Publishing Limited, pp. 377–403 (2013)
13. Tachizawa, E.M., Alvarez-Gil, M.J., Montes-Sancho, M.J.: How "smart cities" will change supply chain management. Supply Chain Managment: An International Journal 20(4): 237–248. Available at: https://doi.org/10.1108/SCM-03-2014-0108 [Accessed May 15, 2018] (2015)
14. Zenezini, G., De-Marco, A.: A review of methodologies to assess urban freight initiatives. IFAC-PapersOnLine, Volume 49, Issue 12, 2016, pp 1359–1364 (2018)
15. Nathanail, E.: A Multistakeholders Multicriteria Decision Support Platform for Assessing Urban Freight Transport Measures. In: Kabashkin, I., Yatskiv, I., and Prentkovskis, O. (eds.) Reliability and Statistics in Transportation and Communication. Lecture Notes in Networks and Systems, pp. 17–31. Springer International Publishing. Riga, Latvia (2018)
16. Bean, W.L., Joubert, J.W.: A systematic evaluation of freight carrier response to receiver reordering behaviour. Computers and Industrial Engineering 124, pp. 207–219 (2018)
17. Liu, C.S., Kuo, G., Huang, F.H.: Vehicle coordinated strategy for vehicle routing problem with fuzzy demands. Mathematical Problems in Engineering, pp. 1–10 (2016)
18. Gutierrez, A., Dieulle, L., Labadie, N., Velasco, N.: A multi-population algorithm to solve the VRP with stochastic service and travel times. Computers & Industrial Engineering 125, pp. 144–156 (2018)

19. Aragão, D.P., Galvão-Novaes, A., Mendes-Luna, M.M.: An agent-based approach to enable dynamic vehicle routing in milk-run OEM operations. In: XXVIII Congreso de Pesquisa e Ensino em Transportes. Curitiba (2016)
20. Jadhav, A., Orr, S., Malik, M.: The role of supply chain orientation in achieving supply chain sustainability. International Journal of Production Economics (2018)
21. Morais, D.O.C., Silvestre, B.S.: Advancing social sustainability in supply chain management: Lessons from multiple case studies in an emerging economy. Journal of Cleaner Production 199, pp. 222–235 (2018)
22. Rushton, A., Croucher, P., Baker, P.: The handbook of logistics and distribution management: Understanding the supply chain. 5th edition. Ed. Kogan Page Limited. ISBN 0749466278 (2014)
23. Crainic, T.G., Ricciardi, N., Storchi, G.: Models for evaluating and planning city logistics systems. Transportation Science 43(4), pp. 432–454 (2007)
24. Zhou, L., Baldacci, R., Vigo, D., Wang, X.: A Multi-Depot Two-Echelon Vehicle Routing Problem with Delivery Options Arising in the Last Mile Distribution. European Journal of Operational Research 265(2), pp. 765–778 (2018)
25. Lu, C.C., Ying, K.C., Chen, H.J.: Real-time relief distribution in the aftermath of disasters - A rolling horizon approach. Transportation Research Part E: Logistics and Transportation Review, 93, pp. 1–20. Available at: http://dx.doi.org/10.1016/j.tre.2016.05.002 (2016)
26. Serna-Uran, C.A.: Modelo multi-agente para problemas de recogida y entrega de mercancías con ventanas de tiempo usando un algoritmo memético con relajaciones difusas. Universidad Nacional de Colombia (2016)
27. Rieck, J., Ehrenberg, C., Zimmermann, J.: Many-to-many location-routing with inter-hub transport and multi-commodity pickup-and-delivery. European Journal of Operational Research 236(3), pp. 863–878. Available at: http://dx.doi.org/10.1016/j.ejor.2013.12.021 (2014)
28. Hemmelmayr, V.C., Cordeau, J.F., Crainic, T.G.: An adaptive large neighborhood search heuristic for Two-Echelon Vehicle Routing Problems arising in city logistics. Computers & Operations Research 39(12), pp. 3215–3228 (2012)
29. Amaral, R.R., Aghezzaf, E.H.: City Logistics and Traffic Management: Modelling the Inner and Outer Urban Transport Flows in a Two-tiered System. Transportation Research Procedia (2015)
30. Faccio, M., Gamberi, M.: New City Logistics Paradigm: From the "Last Mile" to the "Last 50 Miles" Sustainable Distribution. Sustainability (7)11, pp. 1–22 (2015)
31. Nguyen, V.P., Prins, C., Prodhon, C.: Solving the two-echelon location routing problem by a GRASP reinforced by a learning process and path relinking. European Journal of Operational Research, 216(1), pp.113–126. Available at: http://dx.doi.org/10.1016/j.ejor.2011.07.030 (2012)
32. Arango-Serna, M.D., Serna-Uran, C.A., Zapata-Cortes, J.A., Alvarez, A.F.: Vehicle routing to multiple warehouses using a memetic algorithm. Procedia - Social and Behavioral Sciences, 160(Cit), pp. 587–596. http://doi.org/10.1016/j.sbspro.2014.12.172 (2014)
33. Crainic, T.G., Montreuil, B.: Physical Internet Enabled Interconnected City Logistics. Available at: https://www.cirrelt.ca/DocumentsTravail/CIRRELT-2015-13.pdf [Accessed September 22, 2017] (2015)
34. Novaes, A.G.N., Bez, E.T. Burin, P.J., AragãoJr, D.P.: Dynamic milk-run OEM operations in over-congested traffic conditions. Computers and Industrial Engineering 88, pp. 326–340 Available at: http://dx.doi.org/10.1016/j.cie.2015.07.010 (2015)
35. Russo, F., Comi, A.: A model for simulating urban goods transport and logistics: The integrated choice of ho.re.ca. activity decision-making and final business consumers.

Procedia - Social and Behavioral Sciences, 80(Isttt), pp. 717–728. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1877042813010070 (2013)

36. Ponce-Cueto, E., Carrasco-Gallego, R., García-García, R.: Propuesta de una guía de selección del modelo de distribución en el sistema logístico del canal HORECA. Dirección y Organización, 000(37), pp. 67–75. Available at: http://www.revistadyo.com/index.php/dyo/article/view/40 (2009)
37. Arango, M.D., Zapata, J.A., Andres, C.: Metaheuristics for goods distribution. Proceedings of 2015 International Conference on Industrial Engineering and Systems Management (IESM), IEEE Publications. pp. 99 – 107. DOI. 10.1109/IESM.2015.7380143 (2015)

ISSN 1870-4069

BreastHealth: Technological Platform for the Prevention, Detection, Monitoring and Treatment of Breast Cancer

María del Pilar Salas-Zárate[1], Miriam Carlos-Mancilla[2], Emmanuel Lopez-Neri[2], Lorena Orellana Jimenez[3], Daniel Gonzalez Diaz[4]

[1] Tecnológico Nacional de México/I. T. Orizaba, Veracruz, Mexico
[2] Centro de Investigación, Innovación y Desarrollo Tecnológico (CIIDETEC-UVM), Universidad del Valle de México, Tlaquepaque, Jalisco, Mexico
[3] INFENEFNI S.A de C.V., Zapopan, Jalisco, Mexico
[4] Universidad Tecnológica del Centro de Veracruz, Cuitláhuac, Veracruz, Mexico
pilarsalasz12@gmail.com, miriam.carlos@uvmnet.edu, emmanuel.lopezne@uvmnet.edu, dralorenaorellana@hotmail.com, daniel.gonzalez@utcv.edu.mx

Abstract. Based on the information from the World Health Organization (WHO), breast cancer is one of the most common diseases that affect women and the largest worldwide disease causing of woman mortality. Currently both, patients and health personnel use the Internet to consult medical information; specifically, patients focus on seeking support for the care of their disease. However, the efforts made to support the care of patients with this disease are not enough. Hence, this paper presents a technological platform focused on the improvement of the processes of diagnosis, medical prescription, prevention, monitoring and treatment of breast cancer, which takes advantage of innovative technologies such as collaborative filtering, semantic Web, opinion mining, Big Data, and multi-device applications. In addition, a case study is presented, and an evaluation in terms of usefulness. Finally, from the analysis of the user interviews is shown that the Web application has a high degree of acceptance of patients with breast cancer.

Keywords: big data; breast cancer; opinion mining; medical recommendation.

1 Introduction

The World Health Organization (WHO) published in 2018 the latest estimates on the incidence and mortality of six different types of cancer in 186 countries. According to the WHO, lung and breast cancer are the leading types worldwide in terms of the number of new cases. Based on these results, it is estimated that cancer worldwide has increased to 18.1 million new cases, while 9.6 million people will lose in 2018 the battle against this disease [1].

The cancer diagnosis in an early stage is important for patient survival. Statistically, each year, 246,660 women are diagnosed with breast cancer, of which about 40,890 results in deaths. Depending on the stage in which the disease is detected, the survival

percentage may vary; if cancer is detected in a single breast, could reach 99 percent of survival rate; if it expands to the lymph nodes the survival rate decreases to 85%; and if disseminated in different parts of the body, the 5-year rate decreases to 26%. In the worst case, 5% of women are diagnosed for the first time when they are in the metastasis [2].

According to the WHO, there are three actions to improve the early stage diagnosis of cancer: 1) enhance awareness about the symptoms of cancer in the society and encourage them to seek medical attention when they detect it; 2) make investments in the equipment of health services and the training of medical personnel in order to make more accurate and timely diagnoses, and 3) ensure patients have a safe and effective available treatment neither costly nor personalized effort.

Currently, both patients and health personnel use the Internet to consult medical information without to be sure the information is accurate or true. Specifically, patients focus on seeking support for the care of their disease. However, the efforts made to support the care of patients with this disease so far are not enough.

This paper presents a technological platform focused on the improvement of the processes of diagnosis, medical prescription, prevention, monitoring and treatment of breast cancer.

Some of the benefits that this platform will provide are:

1) Guaranteed access to services and medical attention to people of any gender, personal condition, race or sexual condition, only Internet access is required.

2) Interactive exchange of opinions, an Internet interaction arises between patients, doctors or organizations through the use of social networks, forums, blogs, and open communities, among others.

3) Guaranteed and quality care.

4) Reliable medical information, the doctor supports the patient to access information reviewed/filtered by another doctor, delimiting the information available on the Web that is not relevant.

5) Patient training, the doctor offers suggestions about the patient's illness and then the patient becomes autonomous in the management of their daily health due to the process of searching for reliable information and the validation that follows with the doctor.

6) Effective time management, the patient saves time on trips for face-to-face consultations with the doctor, when doubts arise, he communicates with the doctor through his mobile device or computer.

7) Cost savings, personalized services are obtained at affordable prices. It is more expensive a private medical appointment than an online consultation.

8) Regardless of time and geographical location, the patient is guaranteed access to quality healthcare information and services.

9) Increase safety, in all stages of the medical process allowing better control in the organization of the patient information.

The rest of the paper is organized as follows. In section 2 related works discussion of the breast cancer research area are presented. In section 3 a platform architecture and main functions are described. Section 4 presents a medical specialist in oncology

ontology. Section 5 describes and present the results of a case study. In Section 6 a web application usefulness evaluation is obtained. Finally, some conclusions and concluding remarks ideas are given.

2 Related Work

This section describes some of the more representative research works reported in the literature related to the breast cancer subject. For example, in [3] a *mHealth* Peer To Peer application to connect Hispanic patients with cancer is developed. In [4] a ubiquitous m-health system based on the user-centric paradigm of Mobile Cloud Computing (MCC) and data mining techniques are described. The core of the client-side system is developed using an Android platform, and it is used for the collection of biological data from the breast; also, a data mining technique with the Naïve Bayes (NB) classifier to predict malignancy in breast tissue and storage of MCC data on the server side is described. In [5] a method of feeling analysis is proposed to understand the emotions and opinions of users of an online support group on breast cancer related to tamoxifen.

In [6] an approach to develop a mobile web application that focuses on breast cancer patients, the proposal allows analyzing the information related to specific dietary, physical and mental aspects according to the stage of their medical treatment is presented. The authors incorporate gamification and social networks to involve and motivate people to achieve their goals of adopting healthier eating habits while increasing physical activity to ensure a lifestyle change.

In [7] a web support system for clinical decision for oncologists and patients with breast cancer is proposed. This system comprises three different forecasting methodologies: the first one is the Nottingham Prognostic Index (NPI) used clinically; second is the Cox regression model and the third one is a Partial Artificial Neural Network with Automatic Relevance Determination (PLANN-ARD). The three models produce a different prognostic index that can be analyzed together to obtain a more accurate diagnosis of the patient.

In [8] the development of a Medical Decision Support System (MDSS) for a hospital case is proposed. The authors develop the services of the medical team for cancer in hospitals from Taiwan. An ontological basis using a knowledge engineering approach for breast cancer is presented. The results show that the system improves the internal administrative efficiency, medical care quality, and the hospitals decision making.

Although the works analyzed in this section provide innovative solutions to the breast cancer domain, it is important to mention that there are no reports that address the prevention, detection, monitoring, and treatment of breast cancer. In this sense, the present work proposes a technological platform for breast cancer that takes advantage of innovative technologies such as collaborative filtering, semantic Web, opinion mining, Big Data, and multi-device applications. In the following section, the proposed platform is described in detail.

3 Architecture

Figure 1 presents the general web platform architecture for the prevention, detection, monitoring, and treatment of breast cancer. The architecture is composed by four main layers, *Presentation*, *Web services API* (Application Programming Interface), *Business*, and *Data access*; each one with different modules that independently execute their tasks in order to enable the platform scalability.

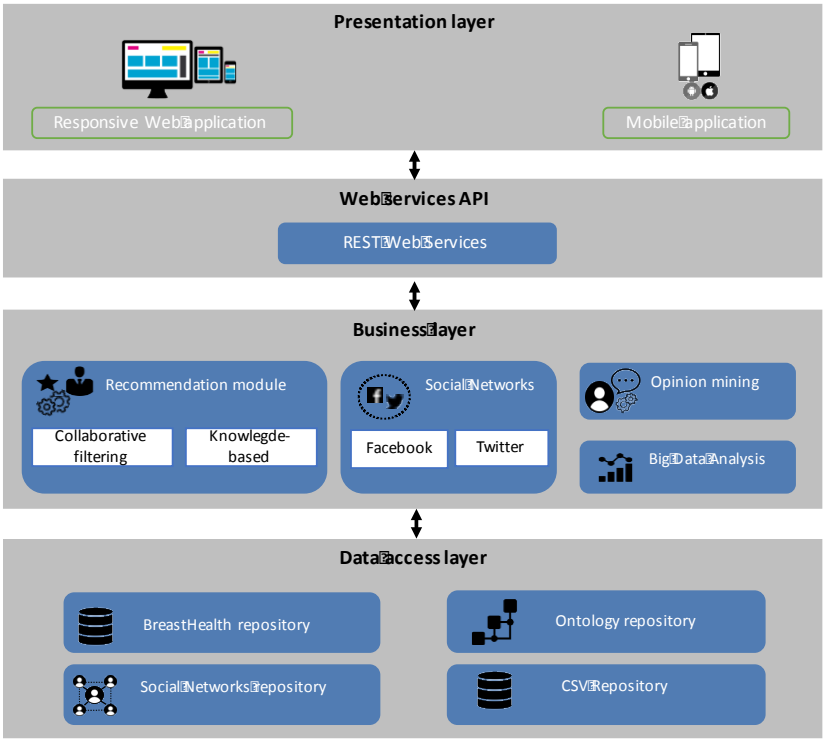


Fig. 1. Platform Architecture.

- **Presentation Layer**. The presentation layer enables the interaction between users and web platform, also known as a graphical user interface. The platform must be easy to use, understandable, and friendly to the user. This layer communicates with the platform through an API REST-based Web Services. Specifically, this layer covers the following applications:
  - ***Web Application***. This application enables the user access to the platform from any Web browser (Edge, Chrome, Mozilla Firefox, Safari, among others) thus enabling access from different operating systems such as Windows, Mac OS and Linux, as well as iOS, and Android. Also, a responsive User Interface approach is used, allowing the platform responds to different browsers and devices.

- ***Mobile app***. Consists of a mobile application for devices based on iOS and Android, being the most used versions in the market of smartphones and tablets.
- ***Web services API***. The Web Services API will provide a set of REST-based Web services grouped by a function that will allow exchanging data between applications, in this case, between the *presentation layer* (Web and mobile interface) and the *business layer* explained bellow.

- **Business Layer**. The business layer will contain the main logic of data processing within the Web application. It communicates with the presentation layer to obtain the user's inputs and present the resulting information, as well to the data access layer to carry out its operations. The main features or options that the user has in their interaction with the system are:
  - ***Recommendation module***. This module will generate recommendations based on two different approaches:
    - *Knowledge-Based*. The recommendations generated by this module are based on the knowledge we have about the items (oncologists and hospitals) that the user has valued (either implicitly or explicitly).
    - *Based on collaborative filtering*. The recommendations generated by this module will be based on similarities between the *active user* (the user to be recommended) and *the rest of the users* of the system. The items (hospitals and oncologists) will be recommended by those that have not been rated by the active user and that have been well evaluated by similar users.
  - ***Social networks Integration***. This module enables opinions collection from oncologists and hospitals from social networks like Facebook and Twitter.
  - ***Opinion mining***. Enables user opinions analysis regarding oncologists and hospitals from both the web platform and social networks like mentioned before. This analysis allows the generation of recommendations based on user opinions.
  - ***Big Data Analysis*.** This module considers the user data analysis available in the repositories of the platform. In addition, an analysis of large volumes of information (Big Data) will be carried out. User behavior patterns will be obtained to be useful for recommending treatments, diagnoses or follow-up the conditions not only based on similarity or previous clinical history experiences but according to the user's particular behavior patterns depending on the type of condition of the cancer. If the disease is in its terminal phase, its state of mind and other social factors acquired from medical social networks.

- **Data access layer.** Enables simplified the access to the stored data founded in this layer, which is comprised of the main logic of access and persistence of data within the Web application. This layer will have to support the storage of data, the recovery of information, and the concurrence of multiple users accessing the information. The stored information in this layer is:
  - ***Breast Health repository***. This repository will contain information related to recommendations, clinical history, and profiles of physicians and patients registered in the platform.
  - ***Social networks Repository*.** Contains opinions of registered oncologists and hospitals on the platform from social networks such as Facebook and Twitter.

- ***Ontology repository***. Used for storing and managing access to the information and knowledge based on an OWL (Web Ontology Language) format, such as, the ontology models, the oncology specialist data, and the breast cancer clinical presentation.
- ***CSV Repository***. This repository contains CSV files (simple text with commas separated values). These files allow representing the matrix scheme of the system in which the pertinent data are stored to make medical recommendations.

4 Medical Specialist Ontology

Since the purpose of this paper is to model the more representative oncology specialists, such as hematologist-oncologist, gynecologist-oncologist among others. We have designed an ontology of medical specialist in oncology. The proposed ontology is based on the web ontology language (OWL). The complete ontology is shown in Fig.2. The main class of the specialist is composed of *Medical Specialty*, *General medicine*, and *General practitioner*. The first one is subdivided into gynecologist, hematologist, oncologist, orthopedist, pathologist, and radiologist.

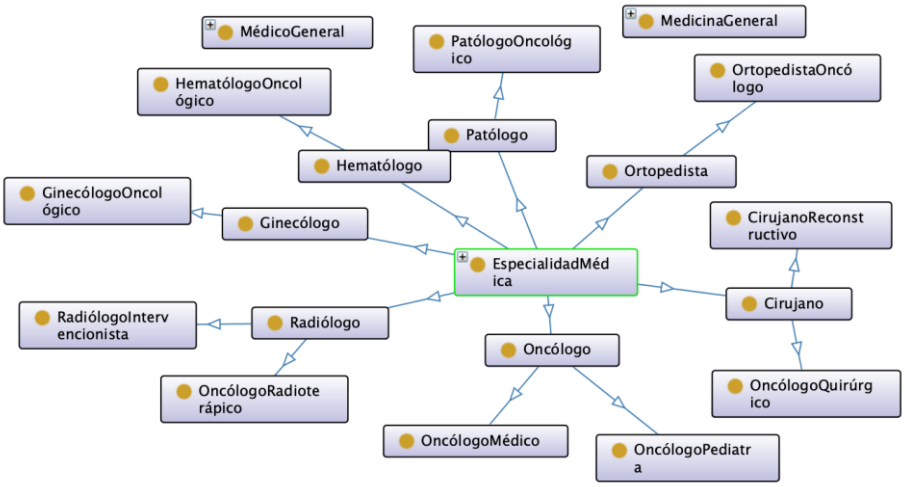


Fig. 2. Medical Specialist Ontology.

The Menthonlogy method presented in [9] was followed for the design of the ontology of medical specialists. Specifically, a series of activities were carried out, each of which focuses on one aspect of the conceptual model of knowledge: terms, taxonomy, relationships, axioms, and rules.

5 Case Study: Breast Cancer Patients' Recommendations

To this case study, the monitoring of the health status of breast cancer patients is based on the following conditions:

1. The user is able to supervise its activities, habits, symptoms, and vital signs for the purpose of learning and evaluate its state of health.
2. The user is looking for health's recommendations in order to stay as healthy as possible to reduce the possibilities incidence of breast cancer.
3. The user seeks to know the best oncologist experts for its kind of breast cancer according to its city, also the best options for hospitals in the breast cancer treatment.

A possible solution for any of these previous situations are solved using the BreastHealth platform. BreastHealth allows any user to register and consults symptoms, habits, and vital signs. The user will be able to visualize medical recommendations to improve and maintain its health. Furthermore, the platform allows to search hospitals based on its treatments and search medical specialists in breast cancer.

5.1 Medical Recommendations for Breast Cancer Patients

BreastHealth allows monitoring any habits, symptoms, and vital signs as a fundamental part of the platform to provide good recommendations. Every user is responsible to keep update its own information about any stride and setbacks during any treatment in order to provide a useful and personalized information; such information is useful to the system and the medical specialist who is giving recommendations.

Medical recommendations make easier for patients to keep control over diverse health aspects. The first recommendations to consider are those related to daily routines, since, through these it is possible to control the affectations that directly or indirectly impact on the vital signs of the users causing symptoms that are harmful to their health.

Fig.3 shows some recommendations provided by the system for user habits, which, if carried out, will reduce or maintain the negative impact they have on the users' health, particularly reflected through symptoms that can be registered by the users. To obtain the desired information, it is only necessary to enter the menu "Habits". As can be seen, the system provides habit recommendations by date.

5.2 Recommendations of Hospitals and Oncologists

The user has the possibility to search an oncologist by name, state, and specialty. Once the search has been made, the results will be displayed and the user will be able to visualize and consult every doctor resume and entire information on a detail screen. In this section, the system will deploy medical specialist recommendations according to the user preferences (see Fig. 4). With regards to hospitals, the system acts just like the medical specialist section and display hospitals recommendations according to the user preferences.

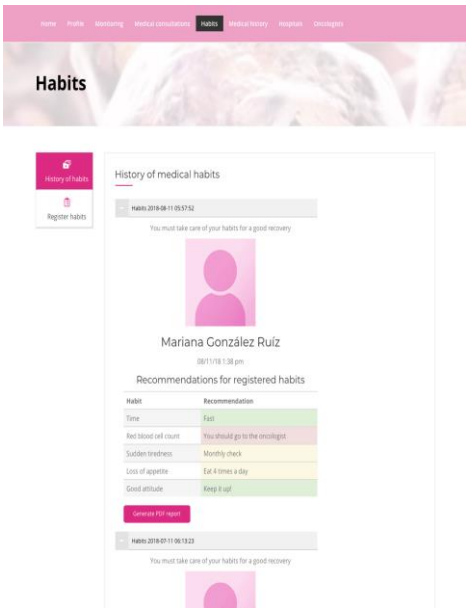


Fig. 3. Medical recommendations.

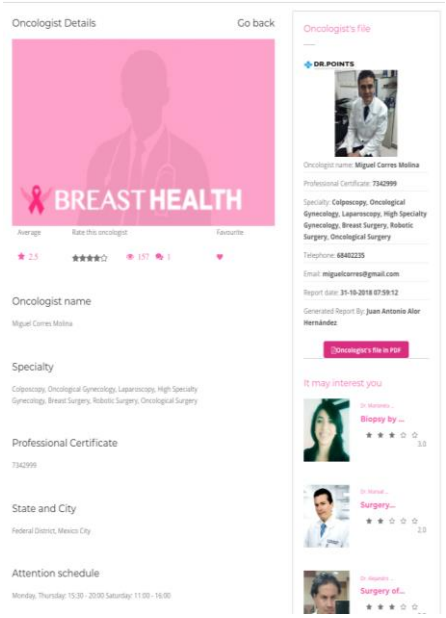


Fig. 4. Recommendation of oncologists

6 Results

The evaluation method is based on the satisfaction metrics of the user defined on SQuarRE ISO /IEC 25010: 2011 (International Standard Organization 2011). This standard ISO / IEC 25010: 2011 define two models of quality composed of general software characteristics; these features are composed of sub-general characteristics and specific attributes.

In the quality framework defined in ISO/ IEC 25010: 2011, user satisfaction is related to the grade of use of a system or product in a specific context. This model is composed of sub-characteristics such as usefulness, trust, satisfaction, and conform. In this proposal is only used the usefulness sub-characteristic to use a unified evaluation of the support system during the decision making. The usefulness is formally defined as the grade of client satisfaction in the goal fulfillment.

To evaluate the usefulness sub-characteristic, there are used four questions of the evaluation usability framework defined in ResQue [10] (Table 1). ResQue is a framework focused on the user experience with recommendation and decision making systems; which consist of constructions, metrics, and questions categorized by four dimensions (perceived system qualities, users' beliefs, subjective attitudes, and behavioral intentions.). Particularly, the ResQue usefulness is defined as a constructor and a metric named "user acceptance" dimension. This dimension allows grading two aspects: decision quality and the decision support through the four questions presented in Table 1.

Every answer of the questions must have a value 1 “Very disagree” to 5 “Strongly agree” in the rating scale of Likert according to ResQue. The questionnaire was applied to 20 people with breast cancer and they used the web application during a period of 10 days.

Table 1. Usefulness questions.

ID	Question
Q1	The recommended elements effectively helped to find the indicated hospitals and medical specialist (decision quality)
Q2	The recommended elements certainly influenced the selection of hospitals and medical specialist (decision quality)
Q3	The platform provides the necessary support to find the indicated hospitals and medical specialist for the breast cancer treatment (decision support)
Q4	The platform provides the necessary support to select the indicated hospitals and medical specialist for the breast cancer treatment (decision support)

According to Fig. 5., we are able to obtain the next conclusions: between 16 and 20 participants consider the web system as helpful, they were able to find recommendations for hospitals and medical specialist (Q1). Between 15 to 20 participants consider the recommended elements really affected during the selection of hospitals and cancer specialist (Q2). Between 17 to 20 participants were agreed in the platform provides the necessary support to find hospitals and cancer specialist according to different cancer medical treatments (Q3). Finally, between 16 to 20 participants consider the platform as useful to find locations of hospitals and cancer specialist for the cancer treatment (Q4).

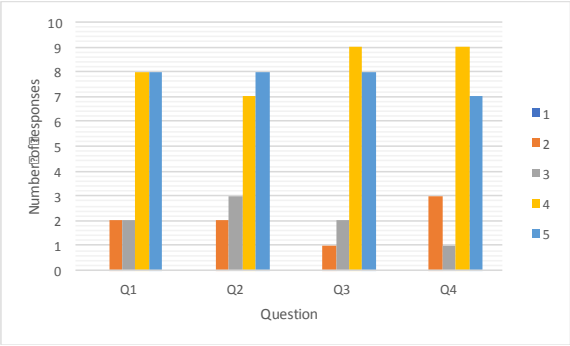


Fig. 5. Evaluation results.

7 Conclusions

A platform to support medical decision making to prevent, diagnose, treatment and monitoring of breast cancer is presented. This tool takes advantage of new technologies and innovations such as collaborative filtering, semantic web, opinion mining, big data, and multi-device applications. A case study was presented and it proves that Breast-

Health is easy to use, and the generated information is useful to the users such as patients and medical specialist enrollees in the platform who gives services to cancer patients.

In this work, an ontology that identifies medical specialists in Oncology is implemented. This ontology allows identifying the adequate medical specialist for monitoring and evaluation of every patient. The platform also provides medical recommendation related to daily routines, since, through these it is possible to control the affectations that directly or indirectly impact on the vital signs of the users causing symptoms that are harmful to their health

The web application was evaluated through usefulness terms. The obtained results show a high grade of acceptance between the main users of this proposal, breast cancer patients; the web application will be improved in a near future to bring a better treatment to users from specialists.

Acknowledgments. The authors are grateful to Mexico's National Council of Science and Technology (CONACYT) for supporting this work.

References

1. OMS: Las 10 principales causas de defunción.
2. Cancer.Net: Cáncer de mama: Estadísticas | Cancer.Net.
3. Banas, J.R., Victorson, D., Gutierrez, S., Cordero, E., Guitleman, J., Haas, N.: Developing a Peer-to-Peer mHealth Application to Connect Hispanic Cancer Patients. J. Cancer Educ. 32, pp. 158–165 (2017)
4. Gatuha, G., Jiang, T.: Android Based Naive Bayes Probabilistic Detection Model for Breast Cancer and Mobile Cloud Computing: Design and Implementation. Int. J. Eng. Res. Africa 21, pp. 197–208 (2015)
5. Cabling, M.L., Turner, J.W., Hurtado-de-Mendoza, A., Zhang, Y., Jiang, X., Drago, F., Sheppard, V.B.: Sentiment Analysis of an Online Breast Cancer Support Group: Communicating about Tamoxifen. Health Commun. 33, pp. 1158–1165 (2018)
6. Economou, D., Dwek, M., Roberston, C., Elliott, B., Kounenis, T., Azimi, T., Ramezanian, M., Bell, N.: PhytoCloud: A Gamified Mobile Web Application to Modulate Diet and Physical Activity of Women with Breast Cancer. In: 2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS). pp. 684–689. IEEE (2017)
7. Fernandes, A.S., Alves, P., Jarman, I.H., Etchells, T.A., Fonseca, J.M., Lisboa, P.J.G.: A Clinical Decision Support System for Breast Cancer Patients. Presented at the (2010)
8. Liao, S.H., Kan, S.L., Lu, S.: The Implementing of an Ontology-Based Medical Decision Support System on Breast Cancer. In: International Conference on Artificial Intelligence and Software Engineering (AISE). pp. 220–235 (2014)
9. Fernández-López, M., Gómez-Pérez, A., Juristo, N.: METHONTOLOGY: From Ontological Art Towards Ontological Engineering. Proc. Ontol. Eng. AAAI-97 Spring Symp. Ser. AAAI-97 Spring Symp. Ser. 24-26 March 1997. Stanford Univ. EEUU (1997)
10. Pu, P., Chen, L., Hu, R.: A user-centric evaluation framework for recommender systems. In: Proceedings of the fifth ACM conference on Recommender systems - RecSys '11. p. 157. ACM Press, New York, New York, USA (2011)

ISSN 1870-4069

Automated Fault Detection and Diagnostics for Aluminum Threads Using Statistical Computer Vision

Luis Alberto Arróniz Alcántara[1], Carlos Juárez Toledo[2], Irma Martínez Carrillo[2]

[1] Bocar Group, Lerma, Mexico
[2] Professional Academy of UAEMex Tianguistengo, Mexico
larroniz03@gmail.com, cjuarezt@uaemex.mx

Abstract. The present work describes the use of statistical computer vision to detect presence-absence of aluminum threads in automotive parts, the vision system is based on a Keyence IV 500 camera and its statistical software. A real case of detection of an industrial aluminum thread was used to demonstrate the effectiveness of the non-invasive machine developed. The results show that using three sequenced tools: brightness adjustment, position reference and area calculation, the repeatability improves by 38%. The study verified the usefulness of the statistical computer vision for fault detection and diagnostics in aluminum threads, a standard statistical analysis of the results presented in the study demonstrates that parts with threads have a punctual performance and, the parts without threads or even without the hole have a statistical normal behavior. The developed method has the ability to automate the correct segregation of good parts with enhanced accuracy avoiding damage to the part, normal in conventional manual methods. The used camera and its brightness compensation demonstrated that environmental light has no effect to the results.

Keywords: aluminum threads, fault detection, statistical computer vision, thread inspection, vision inspection.

1 Introduction

1.1 Manufacture of Threads in Aluminum

Automotive manufacture commonly uses screws to make an assembly between components, the housings of those screws are known as threads. Several normativity exists to define the basic shape of a thread and its manufacturing tolerances.

The basic shape of internal metric threads is described by ISO-68-1:1998 [1]; for aluminum there are two different manufacturing processes that fulfill this. For the first one a previous hole with a diameter equal to the minimum diameter of the thread is made and then a cutter removes material and generates the thread, in the second one a tool deforms the aluminum applying force in the previous hole superior to the elastic limit of the aluminum, the material flows through the male profile and becoming a female laminated thread. In aluminum it´s common to use laminated threads for its grater resistance to deformation.

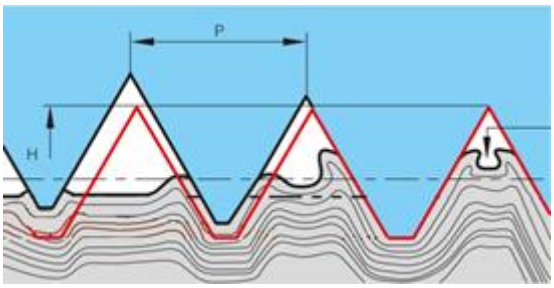


Fig. 1. Representation of laminated threads.

Figure 1, shows a representation of a laminated thread, the red line denotes the basic shape of the thread, the blue zone represents the male tool, the gray zone is aluminum being laminated. Figure 2 shows the real laminated.

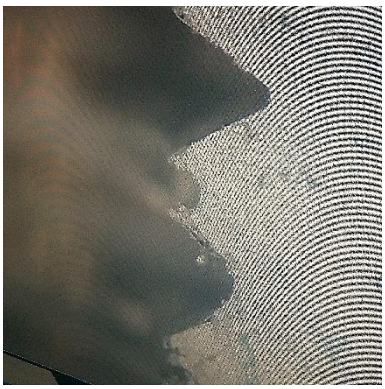


Fig. 2. Laminated thread shape viewed through microscope.

Commonly, laminated threads manufacturing process doesn't have any failure, however if the tool breaks, there is a possibility that the thread doesn't reach the desired depth or not being machined. This failure is what is needed to be detected.

1.2 Traditional Methods for Thread Inspection

There are several accepted methods for inspecting threads, however for mass production, go/no-go gauge is the most used. The MSA Manual [2] describes a Gage as a device used to obtain measurement; frequently used to refer to equipment used on production floor; including go and no-go devices. In other words, is an attribute characteristic tool easily used by operators that mechanically compares the interferences between the manufactured part and the device itself.

In the case of threads, gages use pitch diameter to do the inspection, first with the minimum material condition and after with maximum material condition. In the first case, the go gage must be inserted through the whole thread, then the length is measured with a Vernier calibrator, if both comparisons are under tolerance then the no-go gage

is tried to be inserted but the part must not accept the gage. This inspection method has a cycle time between 12 to 15 seconds for M5 to M12 threads.

The method has three principal disadvantages, cycle time, the probable damage caused in the manufactured part, and the use life of the gage. To solve this problem, the gage is used only for setup analysis and tracking every few hours, however the operator must do a visual inspection to the part, this inspection is called "presence-absence inspection".

1.3 Discrepant Parts Detection Systems

The industry is always pushing for suppliers to achieve high volume production with perfect quality rate at low cost. The only way to obtain this objective is to get more efficient process by reducing machine time cycles, and errors in parts and subassemblies.

Presence-Absence detection process have received considerable attention by the Automotive industry, because missing characteristics have associated elevated production and rework cost. Automotive Industry Action Group (AIAG) in FMEA Manual [3] foresees several types of detection systems that can be resumed in the table 1.

Table 1. Detection FMEA resumed table.

Criteria	Range	Detection probability
Postprocessing failure detection by visual aids	8	Remote
Postprocessing failure detection by operator using variable or attribute gages	7-8	Low
Postprocessing failure detection by automated controls that detects discrepant parts and locks to prevent posterior postprocessing	3-4	High

According to [3] automated validation machines are placed at the process to ensure that the parts are compliant with the customer's specifications. This validation becomes more important when the customer is Tear 1 or 2, because it doesn't use some of the missing characteristic, and the whole assembly goes to the OEM (Original Part Manufacturer), the cost of the claim increases, and the confidence of the whole productive chain is lost. Those characteristics are called PTC's (Pass Through Characteristics).

Automotive industry normally uses a go/no-go gage to inspect threads, if this verification is made in every part, it could be possible that some of the threads were damage. The gage itself could be affected in its own dimensions, and the part and process could be compromised. The use of gages for 100% inspection requires an operator dedicated to this verification, the usual time for the job is 12 seconds per characteristic to be verify.

The use of non-invasive media to detect presence-absence of threads has many advantages (see Table 2), however when a Repeatability-Reproducibility study is

performed, the non-invasive media has a low performance, therefore this kind of inspection methodologies are avoided.

Table 2. Comparison between gages and non-invasive media.

Criteria	Use of gage go No-go	Use of non-invasive media
Cycle Time	12 seconds minimum	0.5 seconds maximum
Operators	Dedicated	Shared
Damages	Highly possible	Non-Possible
Repeatability	High	Low
Reproducibility	High	Low

No invasive methods could be used to support the diagnosis of production, assuring the characteristic without damaging the part. The use of statistical methods to determine when a machine is trustworthy enough to detect not conforming product and guaranty the compliance with the customer regulations. The main porpoise of this paper is to present a visual machine that detects the presence-absence of CNC laminated threads with a high performance in repeatability and reproducibility, demonstrating the viability of the method.

2 Related Work

The main works for the detection of faults are enlisted below.

In 1982 Baumann [4] documented the first production implementation of the General Motors "Consight Vision System", foundry successfully sorted up to six different castings at up to 1,400 an hour from a belt conveyor us-ing three industrial robots in a harsh manufacturing environment.

Vedang [5] fully describe at least ten documented works dated between 1988 and 2014, he himself presented in 2015 a Comparative study of machine vision-based methods for fault detection in an automated assembly machine [6], that compares three different methods, Gaussian Mixed Models and Blob Analysis, optical flow method and running average method.

Gonzalez [7] describes in 2013 a vision-based system presence-absence inspection of automotive subassemblies, using artificial vision for identifying the correct position of holes, threads and welds points using a laser pro-filometer and stroboscopic light.

Urrea [8] in 2004 uses the Hough transform for detecting lines in a low-level vision system, he uses this algorithm in a toy mobile.

Cortes [9] presented in 2010 a computer vision system for quality control in production, in his paper he describes the use of Matlab for digital processing for RGB and CMYK decomposition.

In 2013 Londoño [10] generated an introduction to artificial vision through laboratory guides using Matlab.

Vedang [5] in 2014, presented the Effect of illumination techniques on machine vision inspection for automated assembly machines.

Braggins [11] provides the actual illumination features for machine vision.

There are many processes that uses vision systems to detect faults or follow tracks, in example, in 2007 Min-Goo Kang [12] presented an automatic weld seam tracking method using laser vision, in 2017 Sun Huaiyuan [13] used a vision based machine to assure the correct package of pharmaceutical drugs, in 2016 Ji Yeon Lee [14] describes a method to detect faults in the package of electronic chips.

The actual work presents a machine that detects the presence-absence of CNC formed threads, based in Keyence IV cameras and Keyence contrast detectors. The next section describes the camera and all the involved configurations.

3 Machine for PTC´s Inspection

3.1 Machine Design

The main parts of the machine are shown in Fig. 3, it consists of: one nest where the part is clamped and inspected, six cameras that verify the presence-absence of several product specification characteristics and a pneumatic cylinder that marks the conditions piece.

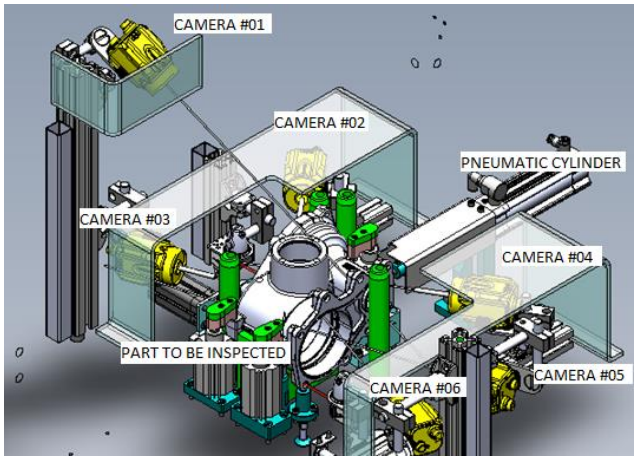


Fig. 3. Automated Vision Base PTC detection machine.

The machine uses a Siemens PLC to control all the functions, the human machine interface (HMI) helps to communicate failures and to accept reestablishment commands, opaque polycarbonate was installed to reduce environmental lighting to minimum, the inspection machine rate is 170 parts per hour. Siemens PLC and HMI are programed according the process, Keyence IV cameras were setup to fulfill three primary requirements, best focus, sufficient bright and the best contrast between an ok and nok part.

During the automatic cycle cameras are operated simultaneously, each one sends an ok or nok signal to the PLC and if the trial of all sensor is ok, led lights are turned on

at a visual aid (see fig 4) in this case the part is unclamped otherwise, the correspondent led blinks showing which characteristic is not accepted, the sequence is interrupted until the supervisor recognize the fault.

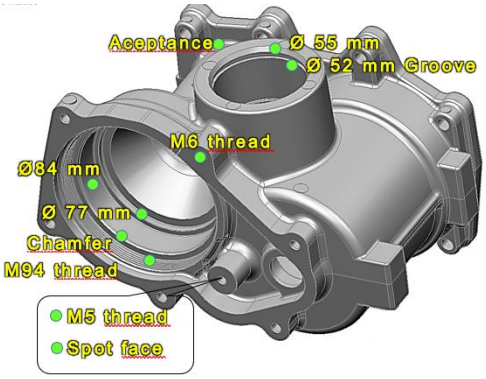


Fig. 4. Visual Aid.

The system design includes steps such as selection, location and adjustment of cameras, environmental light controls, load-unload nest design, clamping method, PLC-HMI program. The design accomplishes the next norms NOM-004-STPS-1999, ISO 12100, ISO TR 14121-2, ISO 1420, ISO 4413, IEC 60204-1, IEC 61140 and IEC 62061.

A critical task to implement the project is the calibration of the cameras, the present work will detail the configuration of the cameras to inspect a roll tap M5 thread in aluminum automotive part with a spot face in fig 5.

Fig. 5. M5 thread with spot face.

3.2 Camera Configuration for Thread Detection

Cameras have been used to detect the presence of threads, though, it has been proven that the variation in the illumination and the contrast of the image can cause measurement errors.

Keyence IV cameras have an autofocus system however, it could be set in different positions between 0mm and 500mm, brightness adjustment could be set between 1 (dark) and 120 (bright) and it has three different imaging modes:

- ***Normal*** less noise.
- ***HDR*** special adjustment for metals.
- ***High gain*** short exposure time but low quality of the image.

The parameters for M5 thread capture image were set as follows: Bright 71 with an exposure time of 0.36 ms and HDR mode is selected, focus selected was 110mm, and a magnification area of 2x with edge emphasis turned on. This increments the outline recognition. The image of figure 6 was set as master image and a compensation of brightness was set.

Three hundred images were captured to be used as samples of the production line, those images were compared with the master image in Keyence IV configured with an area tool. Figures 7 a) and 7 b) show the results using the compensation of brightness.

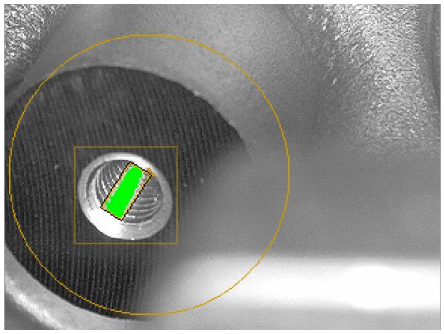


Fig. 6. Common configuration for thread detection.

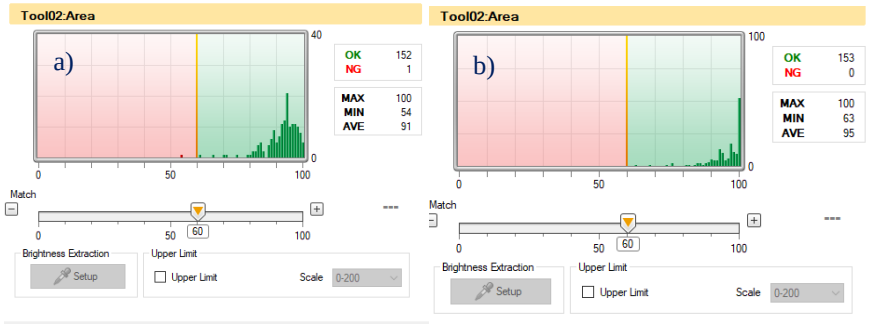


Fig. 7. a) Statistical results without brightness compensation, and b) Statistical result with bright compensation.

Two tools were used to configure the camera:

-*Position tool*, it looks for edges in the part that are located always at the same distance to the area to be analyzed, it is important to locate edges in axis x and y, fig 8 shows the configuration of the position tool. Green patterns are the edges to be validated, yellow patterns are erased from the images to avoid noise. As seen, an arc and a circle are validated, the result is that all the next tools must be referred to the position tool.

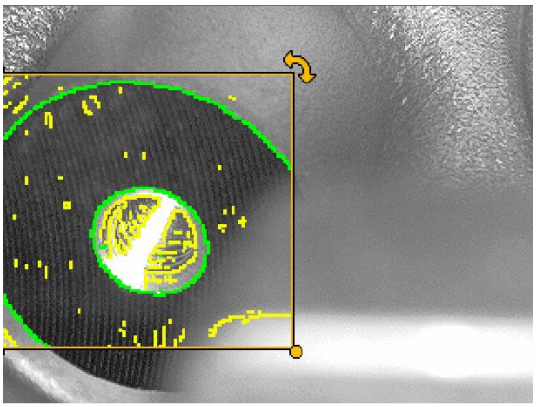


Fig. 8. Position tool adjustment.

-Area tool, the area tool is used to detect high bright in the selected area, the tool was configured to detect thread pitches avoiding the center bright. Blue square in fig 9 is the bright compensation for the analysis, green pattern are the thread pitches, a mask in the center allows the camera to ignore the high bright area that is coincident with holes without threads.

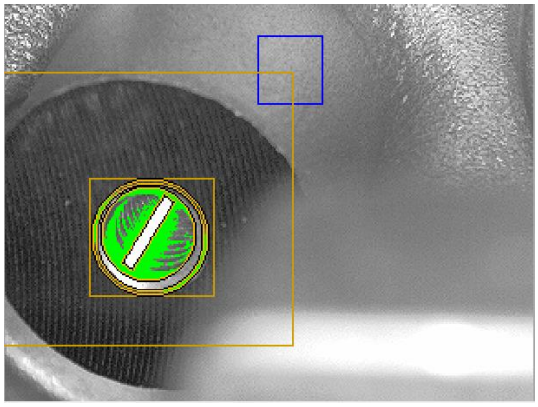


Fig. 9. Thread detection area tool configuration.

The inspection was made in less than a second, as the machine has 6 cameras and several sensors installed, the total cycle time is 20 seconds including load and unload.

4 Results and Discussion

The performance of the adjustment was tested on 300 samples out of which the 185 data sets were for normal operation sequence of the machine (CNC formed thread M5), 75 were for parts that only include a hole nor a thread, and 40 were for parts that has no hole. To evaluate the performances, the results obtained were analyzed with the

statistical tool included in the Keyence IV sensor simulation software. Fig. 10 shows the behavior, Green data are *good* parts, red data are for *not good* parts, as seen there are 2 sections in red that correspond to no presence of thread and no presence of hole.

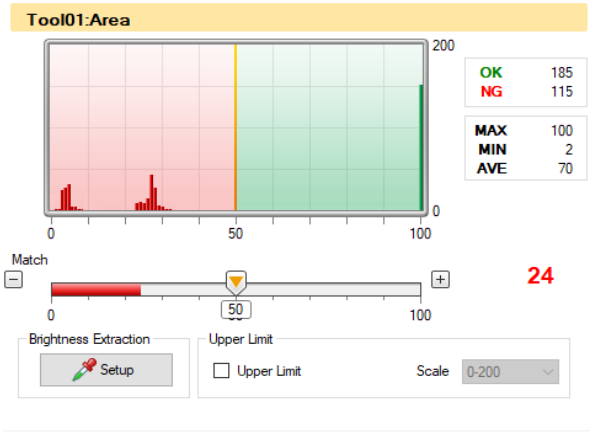


Fig. 10. Equipment performance using statistical results.

Even when the results of good parts have no statistical normal behavior, the difference between the good parts and the not good parts is 78 units of 100, that significantly contributed to increase the confidence on the detection of the absence of threads, Hence the led included in the visual aid helps the operator and the supervisor to protect the customer.

5 Conclusions

The effectiveness of the developed non-invasive machine using the statistical computer vision is presented, the article shows that a correct configuration of the camera improves the performance of the machine. Thread parametrization requires special training for process engineers, however operators have a confident tool to segregate ok and nok parts. It is concluded that presence-absence of threads with automated visual controls will assure the protection of the customer and complies the demand of AIAG.

The time difference between, 20 seconds inspecting 15 characteristics, with 12 seconds only for verify the M5 thread with the use of a gage demonstrated that automated fault detection machines are faster and effective to isolate *good* parts.

Plans for future work include parametrization in different internal threads between M5 to M94 using the actual work as a base, the parameters would be tested on their ability to detect faults.

References

1. ISO: ISO-68-1:1998. ISO general purpose screw Threads-Basic profile.

2. Automotive Industry Action Group, MSA MANUAL REV 04, AIAG (2010)
3. Automotive Industry Action Group, FMEA MANUAL REV 04, AIAG (2010)
4. Baumann, R., Wilmshurst, D.: Vision System Sorts Castings at General Motors Canada. Emerald, pp. 145–150 (1982)
5. Vedang, C., Surgenor, F.: Effect of Illumination Techniques on Machine Vision Inspection for Automaed Assembly Machines. In: Proceedings of The Canadian Society for Mechanical Engineering International Congress, pp. 1–6. Toronto, Ontario, Canada (2014)
6. Chauhan, V., Surgenor, B.: A comparative Study of Machine Vision Based Methods for Fault Detection in an Automated Assembly Machine. In: 43rd Proceedings of the North American Manufacturing Research, pp. 416–428 (2015)
7. González, A., Ramírez, A., Padilla, J.A., Morales, R.: Sistema basado en visión para inspección del tipo ausencia/presencia de subensambles automotrices. Científica 17(1), pp. 29–37 (2013)
8. Urrea, J. P., Ospina, E.: Implementación de la Transformada de Hough para la Detección de Líneas para un Sistema de Visión de Bajo Nivel. Scientia Et Technica X(24), pp. 79–84 (2004)
9. Cortés, J.A., Medina, F.A., Mendoza, J.A.: Computer Vision System for Quality Control in Production, Scientia Et Technica XVI(45), pp. 130–134 (2010)
10. Londoño, V., Marín, J., Arango, E.: Introducción a la Visión Artificial Mediante Prácticas de Laboratorio Diseñadas en Matlab. Tecno Lógicas, pp. 591–603 (2013)
11. Braggins, D.: Illumination for Machine Vision. Sensor Review, 20(1), pp. 20–23 (2000)
12. Kang, M.: Laser vision system for automatic seam tracking of stainless steel pipe welding machine. In: International Conference on Control, Automation and Systems, pp. 1045–1051. Seoul, South Korea (2007)
13. Huaiyuan, S.: The Detection System for Pharmaceutical Bottle-packaging Constructed by Machine Vision Technology. In: Third International Conference on Intelligent System Design and Engineering Applications, pp. 1423–1425. Hong Kong, China (2013)
14. Lee, J. Y.: Development of vision system for defect inspection of electric parts in the tape and reel package. In: 16th International Conference on Control, Automation and Systems (ICCAS), pp. 437–439. Gyeongju, South Korea (2016)

ISSN 1870-4069

Design of a Language for IoT Service Composition

Isaac Machorro-Cano[1], Giner Alor-Hernández[1], José Oscar Olmedo-Aguirre[2], Lisbeth Rodríguez-Mazahua[1], Mónica Guadalupe Segura-Ozuna[3]

[1] Tecnológico Nacional de México / I. T. Orizaba, Orizaba, Veracruz, Mexico
[2] CINVESTAV-IPN, Electrical Engineering, Mexico City, Mexico
[3] Universidad del Papaloapan (UNPA), Tuxtepec, Oaxaca, Mexico
imachorro@gmail.com, galor@itorizaba.edu.mx, lrodriguez@itorizaba.edu.mx,oolmedo@cinvestav.mx, msegura@unpa.edu.mx

Abstract. In the Internet of Things (IoT) ecosystem, multiple smart devices communicate among them and with people. Similarly, they are primarily characterized by remarkable detection and processing capabilities. On the other hand, a service composition (SC) task involves performing the orchestration or choreography of services. SC is frequently studied in the context of Web services (WS), where a series of standards have been developed and used in real-world implementations to support SC. Unfortunately, these standards are inadequate in the IoT paradigm, since IoT devices are based on data/events and the resources are restricted. This research work proposes the design of a language for IoT SC, while simultaneously discussing important literature on SC, business process, IoT service orchestration, and IoT service choreography. Finally, as a proof-of-concept, we present a case study of IoT service composition in the healthcare domain for patients with overweight or obesity.

Keywords: business process, IoT, orchestration, service composition.

1 Introduction

The IoT is the important evolution of the Internet where heterogeneous devices and machines are interconnected among them and with people. Recently, the microcontrollers used to communicate over the Internet have gained popularity, thus giving rise to a wide variety of smart and networked devices, such as digitally enhanced objects, motion detectors, health surveillance devices, electric meters, and even street lights. All these devices are chiefly characterized by their detection, processing, and network connection capabilities [1].

In order to support the real-time communication of smart objects through the Internet, some web protocols are being implemented in real time. These protocols are compatible with smart objects, Web open source standards such as the devices profile for web services and Constraint Application Protocols (CoAPs). In parallel, the services offered by smart objects are accessed directly on the Web and interact with a large

number of conventional web services to form a new generation of ubiquitous applications [2]; however, there seems to be a problem regarding the service composition (SC) of smart objects.

SC is a basic principle of service-oriented computing (SOC) [3] where different services are combined together in order to satisfy complex user requirements. Two important aspects of SC are service orchestration and service choreography [4]. SC is frequently studied in the context of web services and business processes, where a series of standards are developed and used in real-world implementations to support it. However, the current characteristics of IoT systems (e.g. devices are based on da-ta/events), as well as issues such as resource-restriction, make some of the techniques developed for Web SC inadequate when applied in the IoT, thereby implying that new SC mechanisms should be developed by taking into account the new requirements of IoT systems.

Real-time support for web-based protocols has paved the way for the arrival of new IoT applications. Moreover, SC seeks to reuse several services from existing components by joining them in a creative way; the idea is that when applied to the IoT context, streamlines the development of IoT applications. Likewise, SC applied in the IoT allows combining services from multiple smart objects to satisfy complex user needs across a wide range of application domains, and it is used to create innovative applications in a more efficient way [5].

This research work proposes an IoT-based language for service composition that considers both service orchestration and service choreography mechanisms. The remainder of the paper is structured as follows: Section 2 discusses relevant literature on SC applied in the IoT, service orchestration and choreography and business pro-cesses (BP). On the other hand, Section 3 proposes the design of our language for IoT SC, whereas Section 4 introduces a case study of SC for the IoT in the healthcare domain, specifically for patients with overweight or obesity. Finally, section 5 pre-sents the research conclusions and suggestions for future work.

2 Related Works

In their work, Yang and Li [4] proposed a strategy for selecting and aggregating sensory data to address the issue of IoT information service composition. The strategy considers the modeling and evaluation of the quality of service (QoS) in the IoT. Moreover, the authors implemented a binary genetic algorithm to identify the best SC solutions. On the other hand, Dar et al. [6] discussed the design and implementation of ROA, a generic architecture model with tools for integrating end-to-end systems and IoT-based business processes. Additionally, Ren et al. [7] presented a service selection model that highlights the global synergy effect based on collaboration requirements. The validity and advantages of the model and the algorithm were tested through the smart car manufacturing simulation experiment in the cloud.

Rapti et al. [8] proposed a decentralized service composition model for pervasive IoT environments that relies on artificial potential fields (APFs), while Dijkman et al. [9] proposed a framework for developing business models in IoT applications. The

framework was created from a literature survey on existing business model frameworks. Then, the authors adapted these frameworks based on surveys to 11 companies that develop IoT applications. Conversely, Vidyasankar [10] proposed a transaction model and a correction criterion for executions of SC in the IoT. The proposal defines relaxed atomicity and isolation properties for transactions in a flexible manner and can thus be adapted for multiple IoT applications. Additionally, Ju et al. [11] presented a generic business model for IoT services that took as its basis a literature analysis and interviews with eight experts working for IoT companies. Similarly, the authors identified the key components of an IoT business model: key partners, key resources, key activities, and value propositions.

Baker et al. [12] developed a multi-cloud IoT SC algorithm called E2C2, which seeks to create a conscious energy composition plan by looking for and integrating the least possible number of IoT services. To test E2C2's performance, the authors evaluated it against four SC algorithms in multiple cloud environments (i.e. All clouds, Base cloud, Smart cloud, and COM2). Meanwhile, Bergesio et al. [13] proposed an object-oriented model capable of orchestrating services and using the services in a stand-alone system to help users personalize smart spaces.

Furthermore, the model provides an automatic adaptation of a "personalization" when the environment is modified. On the other hand, Wen et al. [14] proposed a fog orchestrator to facilitate the centraliza-tion of a group of resources, map applications to specific requests, offer an automated workflow to physical resources, generate workload execution with control of runtime QoS, and create efficient directives over time to manipulate objects. In addition, Macker and Taylor [15] proposed the Network Edge Workflow Tool (Newt) for two use cases. In the first case, Newt was used to implement a causal workflow in a disaster response scenario, whereas in the second case, it was used to orchestrate William Shakespeare's Hamlet by distributing the actors across an emulated wireless environment and having them exchange information.

In their work, Duhart et al. [16] presented the Environment Monitoring and Management Agent (EMMA) framework, which relies on a set of elements for designing distributed architectures for receptive environments. The authors used the resource-oriented architecture (ROA). In addition, Chen and Englund [17] discussed a choreography platform for Internet-oriented services, which performs the choreography of heterogeneous services by means of an automatic synthesis of choreography dia-grams. The approach was particularly proposed for Cooperative Intelligent Transport Systems (C-ITS), where vehicles, infrastructure, and cloud services are interconnected and cooperate to achieve efficient transport solutions. Finally, Pahl et al. [18] pre-sented an architectural pattern with its underlying principles. The pattern combines IoT edge orchestration with a provenance mechanism, which relies on the chain of blocks for trusted orchestration administration (TOM) in the cloud.

As can be observed, the majority of the works do not draw upon traditional Web services or wearables for IoT service composition. The following section introduces our language for IoT service composition and thoroughly discusses its design and its basic elements that enable service orchestration.

3 Design of an IoT-based Language for Service Composition

Service composition is performed through either service orchestration or service choreography. Service orchestration is a centralized process for organizing interac-tions among the services of an activity or business process; however, orchestrators involved in a same service orchestration task rarely know each other. Similarly, note that service orchestration was designed to orchestrate both conventional web services described in WSDL (Web Services Description Language) and REST services (Rep-resentational State Transfer) that currently use the different providers of wearable devices.

In this paper, we present the design of a language for IoT service composition, which has the necessary elements to perform the orchestration of services. The lan-guage's service orchestration elements are explained below:

- **orchestrationIoT**: The beginning and the end of the services orchestration in the IoT.
- **definitionRoles**: Label to define business processes and their roles.
- **varorc**: Data/states to be used within the business processes.
- **colaborations**: Label for conversations in the business process.
- **exceptions**: Label for handling exceptions.
- **faultRecovery**: Label for error recovery.
- **eventCtrl**: Label for concurrent events.
- **busproflow**: Implement the business process flow.
- **invdoe**: Method for invocating data or events from the smart device.
- **recdoe**: Method for receiving data or events from the smart device.
- **repdoe**: Method for responding to data or events from the smart device.
- **deviceData**: Label used to obtain the description of the smart device (i.e. series, brand, model, size, weight, and status - active or inactive).

4 Case Study: Services Composition in Healthcare for Patients with Overweight or Obesity

This section presents a case study to represent the services' orchestration within the IoT-based service composition language. Namely, the case study addresses the composition of healthcare services for overweight/obese patients. The scenario is as follows:

- An adult patient with overweight/obesity needs to monitor and coordinate the services provided by the wearable provider and those from a smart scale. Likewise, the patient needs external services that would help them achieve their goal of controlling or losing weight. Note that all the service's data are visualized by the patient in real time.

Fig. 1 represents the scenario explained above. As can be observed, the patient's medical parameters (i.e. burned calories, physical activity, heart rate, weight, BMI) are collected by a wearable device and a smart scale, which are synchronized using a smartphone.

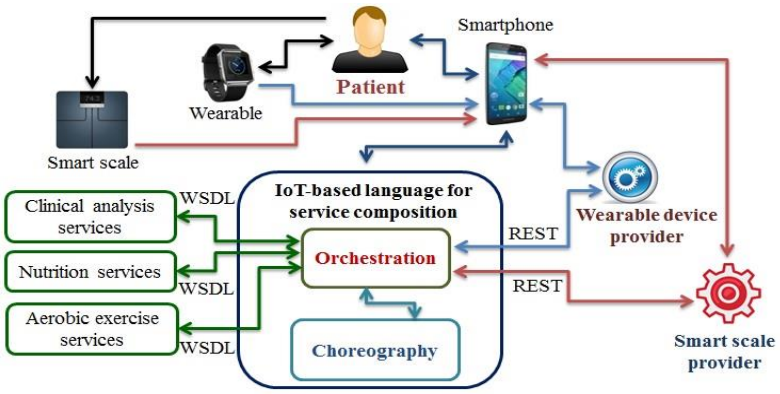


Fig. 1. Service orchestration for a patient with overweight or obesity.

The patient uses an application to visualize the services' composition. To perform the orchestration of the services (i.e. burned calories, physical activity, heart rate, weight, BMI), the SC mechanism relies on REST to request the medical parameters to both the wearable device provider and the smart scale provider. The SC mechanism does this to coordinate these services and establish an order of invocation, so that they can be visualized dynamically and in real time by the patient. Then, the external Web services (clinical analysis, nutrition services, and aerobic exercises) are orchestrationed. To this end, the SC mechanism relies on the WSDL to request information on each external Web service (e.g. availability, price, time, location). Then, using this information, the patient can select the most reliable clinical analysis laboratory, the best nutrition plan, and the most convenient type of aerobic exercise, according to their preferences and criteria.

The UML activity diagram of the previous scenario is depicted below in Fig. 2. Also, the figure shows how we mapped the labels for the services' orchestration to the language for IoT service composition.

- The <orchestrationIoT> tag is generated in the initial state of the diagram.
- The <definitionRoles> tag is generated according to the streets represented in the diagram (Language for IoT service composition, Web services, and REST-based Web service).
- The <varoc> tag is generated when the patient defines the amount of weight they want to lose.
- The <collaborations> tag is generated by identifying the smart devices and web services used to achieve weight loss or weight control.
- The <exceptions> tag is generated when exceptions occur (e.g. maintenance of the REST-based web service from the wearable provider).
- The <faultRecovery> tag is generated to handle exceptions or errors (e.g. incomplete data on the availability of web services).
- The <eventCtrl> tag is generated when the wearable device and the smart scale are concurrently working. This tag is represented in the diagram by the horizontal blue lines.

– The <invdoe>, <recdoe>, and <repdoe> tags are generated during the process of requesting the patient's medical parameters to the wearable device provider.
– The <deviceData> tag is generated during the process of receiving the patient's medical data.

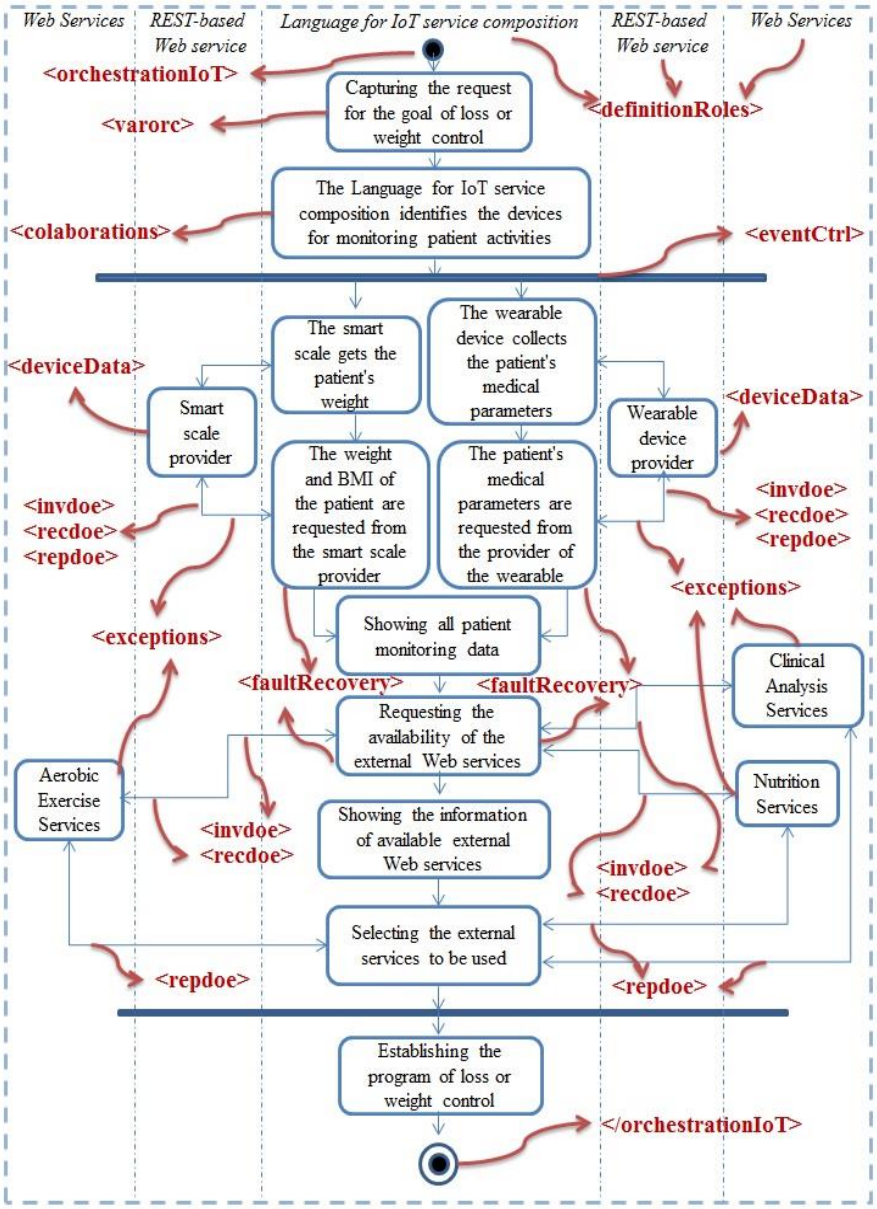


Fig. 2. UML-based activity diagram and the mapping process of the labels for the services orchestration

The language for IoT service composition is able to represent other scenarios, not only in the medical domain, but also in other contexts, such as home automation and the industrial sector.

5 Conclusions and Future Work

The increasing number of smart devices that nowadays communicate over the In-ternet has led to multiple heterogeneous devices connected in a network. Consequent-ly, there is a need for specific mechanisms to achieve the composition of all the ser-vices offered by these devices, and thus exploit the potential of the IoT. In this work, we propose the design of a language for IoT service composition. As a proof of con-cept, we present a case study for the orchestration of medical services for patients with overweight or obesity. The language requires all the involved smart devices to be intercommunicated through the Internet, and all the device providers to provide all the requested data.

As future work, we will seek to work on other case studies and thus expand the applicability of the language. Furthermore, we will intend to design the service chore-ography task and formalize the language to perform the composition of services in the IoT paradigm.

Acknowledgments. This work was supported by Tecnológico Nacional de México (TecNM) and sponsored by both the Mexico's National Council of Science and Tech-nology (CONACYT) and the Secretariat of Public Education (SEP) through the PRODEP program (Programa para el Desarrollo Profesional Docente).

References

1. Luthra, S., Garg, D., Mangla, S.K., Berwal, Y.P.S.: Analizing challenges to Internet of Things (IoT) adoption and diffusion: An Indian context. Procedia Computer Sci-ence 15, pp. 733–739, doi: 10.1016/j.procs.2017.12.094 (2018)
2. Tokognon C.A., Gao B., Tian G.Y., Yan Y.: Structural Health Monitoring Frame-work Based on Internet of Things: A Survey. IEEE Internet of Things Journal 4, pp. 619–635, doi: 10.1109/JIOT.2017.2664072 (2017)
3. Pisching, M.A., Junquiera, F., Santos Filho, D.J., Miyagi, E.: Service Composition in the Cloud-Based Manufacturing Focused on the Industry 4.0. International Fed-eration for Information Processing 450, pp. 65–72 (2015)
4. Yang, Z., Li, D.: IoT information service composition driven by user requirement. IEEE 17th International Conference on Computational Science and Engineering. Pp. 509–513, doi: 10.1109/CSE.2014.280 (2014)
5. Han, S.N., Khan, I., Lee, G.M., Crespi, N., Glitho, R.H.: Service composition for IP smart object using realtime Web protocols: Concept and research challenges. Computer Standards & Interfaces 43, pp. 79–90 (2016)
6. Dar, K., Taherkordi, A., Baraki, H., Eliassen, F., Geihs, K.: A resource oriented integration architecture for the Internet of Things: A business process perspective. Pervasive and Mobile Computing 20, pp. 145–159 (2015)

7. Ren, M., Ren, L., Jain, H.: Manufacturing service composition model based on synergy effect: Asocial network analysis approach. Applied Soft Computing 70, pp. 288–300, doi: https://doi.org/10.1016/j.asoc.2018.05.039 (2018)
8. Rapti, E., Karageorgos, A., Gerogiannis, V.C.: Decentralised Service Composition using Potential Fields in Internet of Things Applications. Procedia Computer Science 52, pp. 700– 06, doi: 10.1016/j.procs.2015.05.079 (2015)
9. Dijkman, R.M., Sprenkels, B., Peeters, T., Janssen, A.: Business models for the Internet of Things. International Journal of Information Management 35, pp. 672–678, doi: http://dx.doi.org/10.1016/j.ijinfomgt.2015.07.008 (2015)
10. Vidyasankar, K.: A Transaction Model for Executions of Compositions of Internet of Things Services. Procedia Computer Science 83, pp. 195–202 (2016)
11. Ju, J., Kim, M., Ahn, J.H.: Prototyping Business Models for IoT Service. Procedia Computer Science 91, pp. 882–890, doi: 10.1016/j.procs.2016.07.106 (2016)
12. Baker, T., Asim, M., Tawfik, H., Aldawsari, B., Buyya, R.: An energy-aware service composition algorithm for multiple cloud-based IoT applications. Journal of Network and Computer Applications 89, pp. 96–108 (2017)
13. Bergesio, L, Bernardos, A.M., Casar, J.R.: An Object-Oriented Model for Object Orchestration in Smart Environments. Procedia Computer Science 109C, pp. 440–447, doi: https://doi.org/10.1016/j.procs.2017.05.415 (2017)
14. Wen, Z., Yang, R., Garraghan, P., Lin, T., Xu, J., Rovatsos, M.: Fog Orchestration for Internet of Things Services. IEEE INTERNET COMPUTING. Pp. 16–24 (2017)
15. Macker, J.P., Taylor, I.: Orchestration and analysis of decentralized workflows within heterogeneous networking infrastructures. Future Generation Computer Systems 75, pp. 388–401, doi: http://dx.doi.org/10.1016/j.future.2017.01.007 (2017)
16. Duhart, C., Sauvage, P., Bertelle, C.: A Resource Oriented Framework for Service Choreography over Wireless Sensor and Actor Networks. International Journal of Wireless Information Networks, pp. 173–186, doi: 10.1007/s10776-016-0316-1 (2016)
17. Chen, l., Englund, C.: Choreographing services for smart cities: smart traffic demonstration. IEEE 85th Vehicular Technology Conference (VTC Spring), pp. 1–5, doi: 10.1109/VTCSpring.2017.8108625 (2017)
18. Pahl, C., Ioini, N.E., Helmer, S., Lee, B.: An Architecture Pattern for Trusted Orchestration in IoT Edge Clouds. 2018 Third International Conference on Fog and Mobile Edge Computing (FMEC), pp.1–8 (2018)

ISSN 1870-4069

Impact of Managers and Human Resources on Supply Chain Performance

José Roberto Mendoza-Fong[1], Jorge Luis García-Alcaraz[2], Liliana Avelar-Sosa[2], José Roberto Díaz-Reza [2]

[1] Universidad Autónoma de Ciudad Juárez, Department of Electrics and Computing, Chihuahua, Mexico
[2] Universidad Autónoma de Ciudad Juárez, Department of Industrial Engineering and Manufacturing, Chihuahua, Mexico
al164438@alumnos.uacj.mx,jorge.garcia@uacj.mx, liliana.avelar@uacj.mx, al164440@alumnos.uacj.mx

Abstract. Current competitive, complex, and uncertain markets push companies toward increasing active collaboration on the part of all human resources (HR) involved in the supply chain (SC), because increasing employee participation and SC collaboration among partners increase SC performance, competitiveness, and, consequently, financial and social success. In this article, we propose a structural equation model to measure the impact of human resources (independent latent variables) on SC efficiency (dependent latent variable). As data gathering instrument, we designed a survey that is responded in a Likert scale and then administered it to 284 participants, including company managers, SC managers, and operators in the Mexican manufacturing sector. For measure the dependence among variables, a structural equation model (SEM) integrates four latent variables: Role of Managers, Learning Environment, Employee Competencies, and Supply Chain Performance. The model is evaluated using Partial Least Squares (PLS) integrated in WarpPLS 6.0 software. Our findings revealed a positive interrelation among the four latent variables, yet in terms of magnitude, the Role of Managers reported the largest effect on the SC Learning Environment.

Keywords: human resources, structural equations model, supply chain, managers, learning focus.

1 Introduction

Supply chain (SC) has become an essential business tool to survive in current competitive markets [1], fueled by rapidly changing customer interests and loyalty. To subsist in these challenging environments, companies must act wisely in their SCs by improving the inventory's management, production and delivery times [2]. In such cases, it is important to coordinate all SC members to work as a unified front and move toward common goals, such as meeting customer demands and improving the efficiency of the procurement, production, and distribution processes, among others.

That collaboration strategies allow companies to take better advantage of human resources (HR), including their abilities, skills, and knowledge, and to understand better both suppliers and customers' concerns in an attempt to integrate, coordinate, and improve the production process and the information flow due to the contribution of all SC members [3]. However, ensuring a collaborative environment to reach common goals is not easy, as it means forging relationships, making adjustments and alignments, communicating effectively, making joint decisions, and sharing information and knowledge, among others [3].

While many studies have emphasized on the role of managers as leaders [4], or the impact of operators, because of their skills and knowledge, on SC performance [5], our research is the first one to quantitatively define the impact of these two human factors on SC performance within a learning environment. In other words, this paper aims at quantitatively defining how the role of managers, employee competencies, an appropriate learning environment, and SC performance are interrelated.

1.1 Role of Managers and a Learning Environment in the SC

The Company directors today, know best the company's strategic objectives, and they are responsible for aligning every SC activity with such objectives [6]. In addition, SC performance have reported that managers' perceptions regarding their environment directly influence their attitudes and commitment to the organization and their subordinates. Furthermore, the abilities and skills of company directors and employees represent a competitive advantage to improve SC performance. For this reason, HR managers must align every employee competency and organizational value with the SC, without neglecting policies, practices, and systems affecting the attitudes, behavior, and performance of SC members [7]. Another important role of managers is to encourage employee skills and creativity by promoting the generation of new ideas applicable to products, services, and work methods for continuous improvement. In fact, the success of organizations and SCs depends on the continuous improvement of employee capabilities and skills, developed when promoting empowerment, participation, and collaboration [8]. Considering our discussion on the role of company managers in the creation of an appropriate learning environment in the SC, we propose the first working hypothesis as follows:

H_1: The Role of Managers has a positive direct effect on the Learning Environment in the supply chain.

1.2 Employee Competencies

Employee capabilities are a competitive advantage in rapidly changing markets. Everything in an organization mirrors the abilities and skills of staff, from the production process to the product itself, including the company's organizational structure, brand(s), marketing strategies, management processes, customer service, and even the supply chain. All this is an open window to what employees can and actually do [7].

To reach high standards and fulfill customer needs, the SC must be collaboratively managed at all levels strategic, tactical, and operational. The appropriate flow of information and inputs is vital from the procurement stage [9] since a product's added value is generated from the beginning. It is also important to train multidisciplinary workers, so their knowledge and skills gained in different domains contribute to their efficacy and to a continuous SC improvement [5]. However, note that managers are responsible for planning such training programs with an appropriate focus that best suits the company needs. Considering thus the role of managers in the development of employee competencies, we propose the second working hypothesis as follows:

H_2: The Role of Managers has a positive direct effect on Employee Competencies in the SC.

It is widely recognized that the information flow and knowledge are key to boosting SC in terms of improving demand forecasts and streamlining the flow of goods and inputs. [2]. Yet, the big advantages of knowledge and information are linked to employee capabilities. Employees must be skilled, must make effective use of data, and ought to share and communicate information that is useful to the organization as a whole, not only to a single department. A lack of such employee competencies may be a major cause of business failure [10], which is why managers must provide an appropriate work environment, focused on learning opportunities for all. These claims allow us to propose the third working hypothesis below:

H_3: A Learning Environment has a positive direct impact on Employee Competencies in the supply chain.

1.3 Supply Chain Performance

Efficient managers make the company profitable to shareholders. Therefore, they must be the first ones to generate new ideas that can be translated to economic benefits [11], and they are responsible for classifying and sharing all necessary information along the SC. Similarly, managers' performance depends on their qualifications when managing the SC, communicating, promoting changes, and measuring the company's progress, especially in economic terms. For all these reasons, the fourth working hypothesis of this study can read as follows:

H_4: The Role of Managers has a positive direct impact on Supply Chain Performance.

HR training and an appropriate learning environment also contribute to a high-performance SC. That is, efficiently trained employees who refine their skills, become more commitment, and improve productivity are a safe source of competitiveness, since they add value to both products and the SC [12]. Nevertheless, learning in companies must not be focused only on generating added-value, but also on developing effective interaction skills, which promote a pleasant work environment in which employees feel proud to work. As an advantage of teamwork and employee motivation, companies manage to improve designs, processes, and even distribution systems [13]. In conclusion, we believe that if training increases HR commitment, flexibility, and quality, companies obtain significant economic benefits. For this reason, we propose the fifth working hypothesis below:

H_5: A Learning Environment has a positive direct effect on Supply Chain Performance.

SC performance results from the interaction among various elements. HR are perhaps the fundamental ones because employee abilities and skills applied to the SC and the production process bring significant economic benefits thanks to the added-value that is generated. Companies investing in education, training, and opportunities for skills development always have a competitive advantage and are capable of solving any complex problem [14]. That said, to find a relationship between HR competencies and SC performance, we propose the sixth and last working hypothesis of this research as follows:

H_6: Employee Competencies have a positive direct effect on Supply Chain Performance.

Fig. 1 below depicts the six research hypotheses, in which arrows directly connect one latent variable with another.

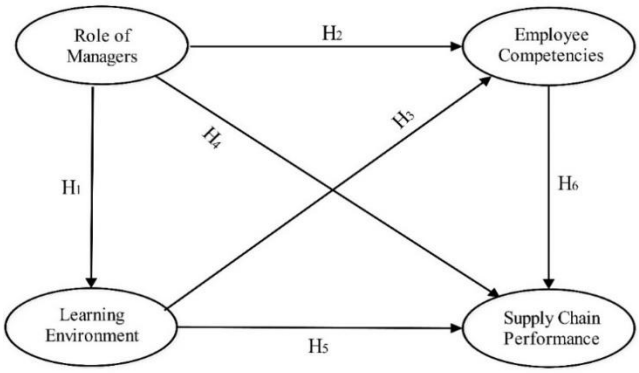


Fig. 1. Research hypotheses.

2 Methodology

This part of the paper describes in detail the methodology followed to conduct our research. The section is divided into six subsections.

2.1 Step 1. Survey Design

To design the survey, we conducted a review of the literature for the four latent variables to be studied: Role of Managers, Learning Environment, Employee Competencies, and Supply Chain Performance. This review of the literature also served to validate the items of latent variables (rational validation). Then, we constructed the first version of the questionnaire, which was composed of two sections. The former aimed at gathering sociodemographic information of participants, whereas the latter assessed latent variables throughout their corresponding items previously identified in the literature review. Finally, we tested the survey accuracy and reliability thanks to a

panel of subject matter experts (SME validation), composed of academics and SC managers. Table 1 shows the final list of items included in the survey.

Table 1. Latent variable.

Role of Managers	**Learning Environment**
Show commitment and support in all SC activities [6, 9]. Identify market fluctuations and rapidly provide the necessary resources for correct SC functioning [7, 9]. Plan and monitor the implementation of SC plans [2, 9]. Their actions are congruent with the company and with SC values [15, 16]. Train workers, promote collaboration, and provide support [6, 16].	Appropriate training and supportive environment [7, 8]. Focus on experimentation, taking the initiative, and responsibility [8, 15]. Secure and psychologically safe work environment [8, 15]. Knowledge sharing [7, 16]. Knowledge transfer among HR in the SC [6, 16].
Employee competencies	**Supply Chain Performance**
Adequate knowledge of corresponding SC activities [17]. Effective communication skills [6, 9]. Trained and skilled in SC operations [10, 15]. Implement new SC projects [2, 9]. Improve SC efficiency and effectiveness amid changes [7, 17]. Invest in talents acquisition for the SC [7, 10].	Increasing profitability [6, 17]. Improved return on investments (ROI) [6, 14]. Increasing sales [6, 9]. Market expansion [7, 17]. Improved product development [7, 17]. Costs reduction [2, 9]. Improved company performance [6, 15].

2.2 Step 2. Statistical Validation of Data

Screened data were validated using seven indices. We computed the Cronbach's alpha and the composite reliability index to analyze the internal reliability of latent variables, setting 0.7 as the minimum acceptable value [18]. The Average Variance Extracted was estimated to analyze convergent validity, looking for values above 0.5. Also, we computed R-Squared (R^2) and Adjusted R-Squared as indicators of the parametric predictive validity of latent variables, whereas the Q-Squared (Q^2) was estimated as a measure of nonparametric predictive validity [19]. Finally, the Full collinearity Variance Inflation Factor (Full Collinearity VIF) was computed to measure internal collinearity of latent variables, only accepting values below five.

2.3 Step 3. The Structural Equation Model

To accept or reject the six hypotheses proposed in Fig. 1, we built a model using the Structural Equation Modelling (SEM) technique, with the aid of WarpPLS 6.0®. This piece of software has algorithms based on Partial Least Squares (PLS), widely recommended for small sample sizes and non-normal data [20]. Then, we computed six

indices to assess the resulting model: the Tenenhaus Index, Average R-Squared (ARS), Average Adjusted R-Squared (AARS), Average Path Coefficient (APC), Average Variance Inflation Factor (AVIF), and Average Full collinearity VIF (AFVIF).

The Tenenhaus Index, also known as goodness of fit (GoF) index, indicates the model's explanatory power [19], and acceptable values usually must be higher than 0.36. For APC, ARS, and AARS, we analyzed their corresponding P-values, setting 0.05 as the cutoff and testing the null hypotheses, in which APC, ARS, and AARS = 0, against the alternative hypotheses, in which APC, ARS and AARS $\neq 0$. About AVIF and AFVIF, values must be equal to or lower than 3.3.

Finally, we measured three types of effects in the SEM: direct, indirect, and total. In Fig. 1, direct effects be arrows directly connecting two latent variables, whereas indirect effects are represented by paths with two or more segments. Finally, total effects between two latent variables are the sum of direct and indirect effects. To test the statistical significance we use 95% confidence level, testing the null hypothesis: $\beta_i = 0$, versus the alternative hypothesis: $\beta_i \neq 0$.

3 Results

This section first presents the descriptive analysis of the sample and the latent variables. Then, we discuss results from the model evaluation, including its effects.

Table 2. Latent Variable Coefficients.

Coefficients	Role of Managers	Learning Environment	Employee Competencies	Supply Chain Performance
R-Squared		0.579	0.535	0.553
Adjusted R2		0.577	0.532	0.548
Q-Squared		0.581	0.535	0.553
Composite reliab.	0.942	0.935	0.954	0.947
Cronbach's alpha	0.923	0913	0.942	0.934
Avg. var. ext.	0.766	0.743	0.776	0.717
Full collin. VIF	2.948	2.844	2.233	2.181

3.1 Validation of Latent Variables

Table 2 shows results from the validation performed on latent variables, using indices described in the methodology section.

3.2 Structural Equation Model

Fig. 2 shows results from the model evaluation. Every segment indicates a relationship between two latent variables and it includes a beta (β) parameter, a P-value for the hypothesis testing, and an R2 value to indicate the percentage of explained variance of dependent latent variables. Following the model evaluation, using indices discussed in the methodology section, we obtained these results:

- Average path coefficient (APC) = 0.394, P<0.001
- Average R-squared (ARS) = 0.556, P<0.001
- Average adjusted R-squared (AARS) = 0.552, P<0.001
- Average block VIF (AVIF) =2.504, acceptable if <= 5, ideally <= 3.3
- Average full collinearity VIF (AFVIF) = 2.552, acceptable if <= 5, ideally <= 3.3
- Tenenhaus GoF (GoF) = 0.646, small >= 0.1, medium >= 0.25, large >= 0.36

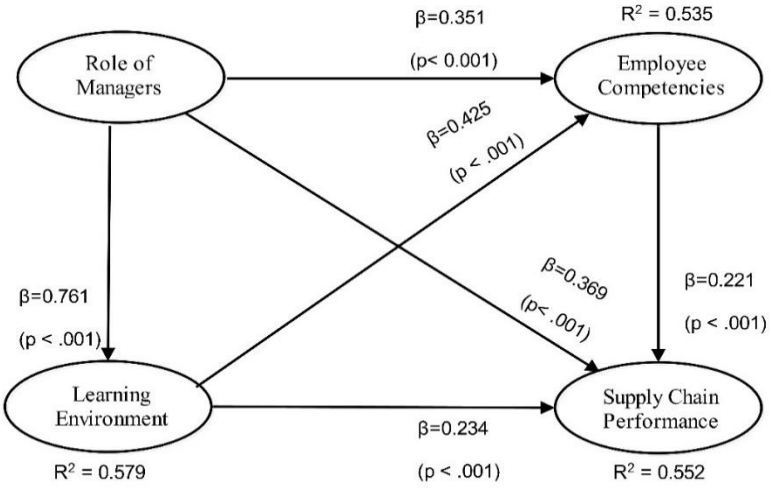


Fig. 2. Evaluated model.

3.3 Effects Analysis

Direct effects

Direct effects allowed us to validate hypotheses presented in Fig. 1 and analyzed in Fig. 2. From the model's evaluation, Table 3 shows results regarding our six research hypotheses.

Sum of indirect effects

Table 4 below shows the sum of indirect effects found in our model (see Figure 2).

Table 3. Hypotheses validation.

Hypothesis	Independent Variable	Dependent Variable	β	Effect Size	P-Value	Decision
H1	Role of Managers	Learning Environment	0.761	0.579	P< 0.001	Accepted
H2	Role of Managers	Employee Competencies	0.351	0.239	P= 0.014	Accepted
H3	Learning Environment	Employee Competencies	0.425	0.296	P< 0.001	Accepted
H4	Role of Managers	Supply Chain Performance	0.369.	0.257	P< 0.001	Accepted
H5	Learning Environment	Supply Chain Performance	0.234	0.156	P< 0.001	Accepted
H6	Employee Competencies	Supply Chain Performance	0.221	0.139	P< 0.001	Accepted

Table 4. Sum of indirect effects.

To	From	
	Role of Managers	**Learning Environment**
Employee Competencies	0.323 (P<0.001) ES = 0.220	
Supply Chain Performance	0.328 (P<0.001) ES = 0.228	0.094 (P<0.012) ES = 0.063

Table 5. Total effects.

To	From		
	Role of Managers	**Learning Environment**	**Employee Competencies**
Learning Environment	0.761 (P<0.001) ES = 0.579		
Employee Competencies	0.674 (P<0.001) ES = 0.459	0.425 (P<0.001) ES = 0.296	
Supply Chain Performance	0.697 (P<0.001) ES = 0.486	0.328 (P<0.001) ES = 0.219	0.221 (P<0.001) ES = 0.139

Total effects

Table 5 summarizes the total effects for every relationship in the model. Note that in three relationships, total effects equaled direct effects, meaning that no indirect effects

were found in those cases. For each one of the three remaining relationships, the table provides the sum of indirect effects.

4 Conclusions and Industrial Implications

We conducted this research in the manufacturing industry of Chihuahua, Mexico, but our conclusions, especially about the role of managers in supply chain performance, can extend beyond this territory and touch the whole Mexican manufacturing sector.

First, as [4] argues, SC performance is built upon trained managers that guarantee the well-being of the system. Our findings support this claim as we proved that the Role of Managers has a significant positive direct effect on *Supply Chain Performance*. Additionally, as [5] affirms, it is important to ensure an appropriate *Learning Environment* to support the development of Employee Competencies, which, in turn, have a positive effect on SC efficiency.

Our findings also revealed the strong impact of the *Role of Managers* on an appropriate *Learning Environment* and *Employee Competencies*. Managers must have outstanding qualifications to recognize the current and future needs of the companies and, then, train their employees based on such needs [11]. In fact, in this research, the direct effect of the *Role of Managers* on *Employee Competencies* is like the indirect effect, occurring thanks to the Learning Environment. In other words, appropriate opportunities for learning, granted by managers through effective training programs, have the potential to increase and sharpen employee abilities and skills.

Finally, although we may have expected a higher effect from *Employee Competencies* on *Supply Chain Performance*, this research still proves the importance of having skilled employees for the correct functioning of the SC. In fact, even though the effect of the first latent variable on the second one was relatively low (0.221), we found that Employee Competencies are of vital importance, since they serve as a mediating variable between *Supply Chain Performance* and other variables.

In conclusion, our findings stand out for the value of this research, which quantitatively validated the importance of Managers, Employee Competencies, and a Learning Environment for Supply Chain Performance. Likewise, we demonstrated that collaboration between HR and managers reflects on the success of companies and the correct functioning of SCs.

References

1. Mendoza-Fong, J.R., et al.: The Impact of Supplier's Administrative Attributes on Production Process and Marketing Benefits. In: Ethics and Sustainability in Global Supply Chain Management. IGI Global: Hershey, PA, USA, pp. 73–91 (2017)
2. Ramanathan, U.: Aligning supply chain collaboration using Analytic Hierarchy Process. Omega 41(2), pp. 431–440 (2013)
3. Montoya-Torres, J.R., Ortiz-Vargas, D.A.: Collaboration and information sharing in dyadic supply chains: A literature review over the period 2000–2012. Estudios Gerenciales 30(133), pp. 343–354 (2014)

4. Derwik, P., Hellström, D., Karlsson, S.: Manager competences in logistics and supply chain practice. Journal of Business Research 69(11), pp. 4820–4825 (2016)
5. Badea, A., et al.: Competency Training in Collaborative Supply Chain Using KSA Model. Procedia - Social and Behavioral Sciences 191, pp. 500–505 (2015)
6. Sánchez, A.A., Marín, G.S., Morales, A.M.: The mediating effect of strategic human resource practices on knowledge management and firm performance. Revista Europea de Dirección y Economía de la Empresa 24(3), pp. 138–148 (2015)
7. Aryanto, R., Fontana, A., Afiff, A.Z.: Strategic Human Resource Management, Innovation Capability and Performance: An Empirical Study in Indonesia Software Industry. Procedia - Social and Behavioral Sciences 211, pp. 874–879 (2015)
8. Kim, H.J., et al.: Is all support equal? The moderating effects of supervisor, coworker, and organizational support on the link between emotional labor and job performance. BRQ Business Research Quarterly (2017)
9. Alfalla-Luque, R., Medina-Lopez, C., Schrage, H.: A study of supply chain integration in the aeronautics sector. Production Planning & Control 24(8-9):769–784 (2013)
10. Wichitchanya, W., Durongwatana, S.: Human Resource Management and Organizational Innovation. The Business Review Cambridge 20, pp. 221–227 (2012)
11. Wilson, K., Barbat, V.: The supply chain manager as political-entrepreneur? Industrial Marketing Management 49, pp. 67–79 (2015)
12. Teixeira, A.A., et al.: Green training and green supply chain management: evidence from Brazilian firms. Journal of Cleaner Production 116, pp. 170–176 (2016)
13. McCrie, R.: 4 - Training and Development for High Performance, in Security Operations Management (Third Edition). Butterworth-Heinemann, Boston, pp. 113–143 (2016)
14. Hussein, N., et al.: Learning Organization Culture, Organizational Performance and Organizational Innovativeness in a Public Institution of Higher Education in Malaysia: A Preliminary Study. Procedia Economics and Finance 37, pp. 512–519 (2016)
15. Paşaoğlu, D.: Analysis of the Relationship Between Human Resources Management Practices and Organizational Commitment from a Strategic Perspective: Findings from the Banking Industry. Procedia - Social and Behavioral Sciences 207, pp. 315–324 (2015)
16. Zhang, X., Zhou, J.: Empowering leadership, uncertainty avoidance, trust, and employee creativity: Interaction effects and a mediating mechanism. Organizational Behavior and Human Decision Processes 124(2):150–164 (2014)
17. Qin, R., Nembhard, D.A., Barnes Ii, W.L.: Workforce flexibility in operations management. Surveys in Operations Research and Management Science 20(1):19–33 (2015)
18. Kottner, J., Streiner, D.L.: Internal consistency and Cronbach's α: A comment on Beeckman et al. (2010). International Journal of Nursing Studies 47(7):926–928 (2010)
19. Boon Sin, A., et al.: Structural equation modelling on knowledge creation in Six Sigma DMAIC project and its impact on organizational performance. International Journal of Production Economics 168, pp. 105–117 (2015)
20. Richter, N.F., et al.: European management research using partial least squares structural equation modeling (PLS-SEM). European Management Journal 34(6):589–597 (2016)

ISSN 1870-4069

Operational Risk Management in a Retail Company

Carlos Andres Pastrana-Jaramillo, Juan Carlos Osorio-Gómez

Universidad del Valle, Escuela de Ingeniería Industrial, Cali, Colombia
carlos.pastrana@correounivalle.edu.co,
juan.osorio@correounivalle.edu.co

Abstract. The risk management is one of the main activities inside modern management of supply chains. One of the main risks is operational risk, those risks are inherent to the daily activities of the company, and perhaps the effects of operational risks do not have the magnitude of the disruptive risks, but if they are not considered and managed, can to affecting significantly business results. A proposal is then presented to identify, prioritize and manage the operational risks present in the distribution process of a company in the retail sector in Colombia. Once the priority of the risks has been defined, the company must take mitigation or elimination actions on them.

Keywords: operational risk, risk management in supply chains, fuzzy QFD, risk prioritization.

1 Introduction

Nowadays companies try to mitigate different risks, addressing those in what their experience have a greater impact, however, companies can pay attention to risks that are not the priority at the time, causing the company to make decisions that have some impact but don't have the results over the most significant risks; this is where it is required to have arguments and the correct information that allows analyzing and identifying the defined processes, evaluate it and finally generate action plans for the mitigation and continuous monitoring of the evidenced risks.

The management of these risks accomplished an improvement in the process, where good practices become necessary and additionally they become daily, common and frequent in the operation, which requires constant monitoring that generates a continuous cycle where processes are guaranteed with minimum impact risks. In this way, this document highlights the importance that there is in controlling them and maintaining all the necessary conditions, on the actors of the supply chain to prevent negative impacts on the profits of the company.

In their work Sangwan and Liangro [7], risks are defined as an uncertain situation where an event can negatively affect the functioning of the organization, and has the probability to happen and may affect the performance of the company or process in the short or long term. Operational risks have an impact or relationship with the processes, equipment or environment. This is how different authors have addressed this issue of

risk management, and which involves different actors along the supply chain and has become a frequent topic of study that is increasing as say Fahimnia et al. [2].

In the handbook, Manotas, Osorio Gomez, & Rivera [5] define risk management in four stages: Risk identification, risk assessment and prioritization, risk management and risk monitoring. These phases are consider in this paper.

The authors Wee et al. [9] explain that the first step of risk management is to identify the sources or actors of the risk, in this way is also mentions in the article Giannakis and Louis [3] whom are agree that this is a fundamental step in the risk management process. To have an appreciation of the existing risks, one can first list the faults that can cause adverse results and then for each failure define the sources that can affect or influence the organization in Tummala and Schoenherr [8]. In addition, Manotas et al. [5] summarizes the most common tools among which are distinguished interviews, questionnaires, panels of experts or Delphi method and checklists mainly. Once the risks have been identified, it is necessary to rate this risk in order to generate strategies that mitigate their impact or even eliminate them. Lavastre et al. [4] proposed, this stage of risk management seeks to determine the severity of the risks, measuring the effect through the processes with the probability that the risks become for real and the potential scope of the impact.

The importance of the risk prioritization is that it show to the company which risks should be accepted and which one can be ignored due to their level of impact; the authors Giannakis and Louis [3] also emphasize that risks consider a wide range of criteria such as the probability of occurrence of the event, the level of risk and especially its impact. In this sense, the prioritization of risks must be based on the objectives set by the company, defined in a strategic way, seeking to be the first to be addressed and mitigate the negative impacts on the core of the company.

Understanding that this aspect of prioritization and evaluation provides the basis for establishing actions that seek to eliminate, reduce or simply ignore the impacts of previously identified risks. This criteria of impact definition when obtained from the experts uses scales such as (No impact, minimum impact, medium impact, high impact) as well as for the probability of occurrence is used (Improbable, moderate, probable, very likely) in the article of Giannakis and Louis [3], these qualitative data lead to look for tools that allow to analyze them. Some of the most commonly used tools according to Manotas et al. [5] are multi-criteria tools such as AHP and ANP and simulation. Additionally, Osorio-Gomez et al. [6] propose to prioritize risks using diffuse QFD, a tool that will be considered in this article.

2 Methodology

In Figure 1 the methodological proposal is presented. For the identification it is necessary to define the situations that can be considered risks in the operation and once this has been defined, a questionnaire is designed to effectively validate that they correspond to the risks of the process. Additionally, it is important to select a team that has all knowledge of the process to be evaluated, since they are who must define the pertinence of considering or not the identified risks.

From a linguistic scale defined in Table 1, to apply the designed questionnaire and decide if the failures evidenced in the distribution process within the organization correspond to operational risks or not; if it is considered a risk, both its probability of occurrence and its magnitude of impact must be defined, using the scale in Table 1.

Fig. 1. Methodology for the management of operational risk in a retail company.

Table 1. Linguistic scale for the risk identification and fuzzy equivalence for FQFD.

Linguistic Scale	**Very low (VL)**	**Low (L)**	**Medium (M)**	**High (H)**	**Very high (VH)**
Numerical equivalence	1	2	3	4	5
Triangular fuzzy number	(0,1,2)	(2,3,4)	(4,5,6)	(6,7,8)	(8,9,10)

Data should be consolidated so that it can be translated in proportion and quantitative questionnaire data, which will be the basis for related matrix Impact – Probability. These are obtained from define the percentage of the increased risk applicability by the corresponding values in the quantitative scale of the weighted averages of the scores made in both probability of occurrence and in impact according to Equation 1 (weighted average of the magnitude of risk i) and Equation 2 (weighted average probability of risk i):

$$\overline{X}_i = \frac{\sum_{j=1}^{n}(B_{i,j} \times M_{i,j})}{n} \quad ; \quad \forall\, i, \tag{1}$$

$$\overline{Y}_i = \frac{\sum_{j=1}^{n}(B_{i,j} \times P_{i,j})}{n} \quad ; \quad \forall\, i\,, \tag{2}$$

where $\overline{X}_i$ is weighted average of the magnitude of risk i,
$\overline{Y}_i$ is weighted average probability of risk i,
$B_{i,j}$ is expert's criterion j if i is applicable as risk (1,0),
$M_{i,j}$ is expert's qualification j on the impact of risk i,
$P_{i,j}$ is expert's qualification j on the probability of risk i.

Based on the impact matrix, it can be defined by a range of colors, those risks ranging from its lower impact and lower probability to a critical case of risk with a high impact

on the operation and in turn with high possibilities occurrence. With this result, we proceed to apply the FQFD for the risks located in the critical zone.

For prioritization through FQFD, the steps presented in Osorio-Gomez et al. [6] must be followed as they show. These steps will be developed in the following section.

Finally, from the previous ranking the company can define the strategies to be able to mitigate or eliminate the risks and in this way improve the analyzed process. It is important to highlight that the implementation of actions may include strategies associated with transferring risk, eliminating it, reducing it or applying strategies focused over the person or associated machine as showed Lavastre et al. in their article [4].

3 Results

Company operates in the retail sector in Colombia, where it has achieved a leading position in the home constructions. It seeks to satisfy the customer through multiple points of contact and sales channels that ensure the Omni-channel business model (Stores, Internet and Telephone). Its activity is focused on developing and providing solutions to the customer's remodeling and construction projects, in addition to satisfying their projects, offering good service.

The company currently has the distribution process through a third party, belonging to the corporate group of the owner organization. This company is responsible for managing deliveries to customers through contracted vehicles that meet the company's security and policy requirements.

According to the needs of the organization, a questionnaire was designed, which was applied in different stores of the region. This allows establish the initial risks that were considered in each one of the warehouses or stores, to finally elaborate the questionnaire that was applied to the defined experts.

According to the article of Avelar-Sosa et al. [1]; in a questionnaire, at least 7 of the respondents must agree with the points to be evaluated. For this reason, eleven people are selected representing the 4 branches of the Valle del Cauca region, experts in the logistics field. From the questionnaire, these people are asked to make the pertinent qualifications in order to consolidate the answers, and determine the viability of the previously selected risks, according to the observation of the process, and then obtain the weighted on the probability and the impact of the risk. These will allow building the probability and impact matrix of the preliminary risks as shown in Figure 2 based on Equation 1 and Equation 2.

According to the matrix and the managerial decision by the case study company, the risks found in the critical areas demarcated with red, listed in Table 2, are considered for the analysis, based on the FQFD methodology. The decision-making group; made up of the logistics coordinator, the manager, the dispatch coordinator, the product logistics coordinator and the operations manager; from red risk have to rate it's based on fuzzy logic.

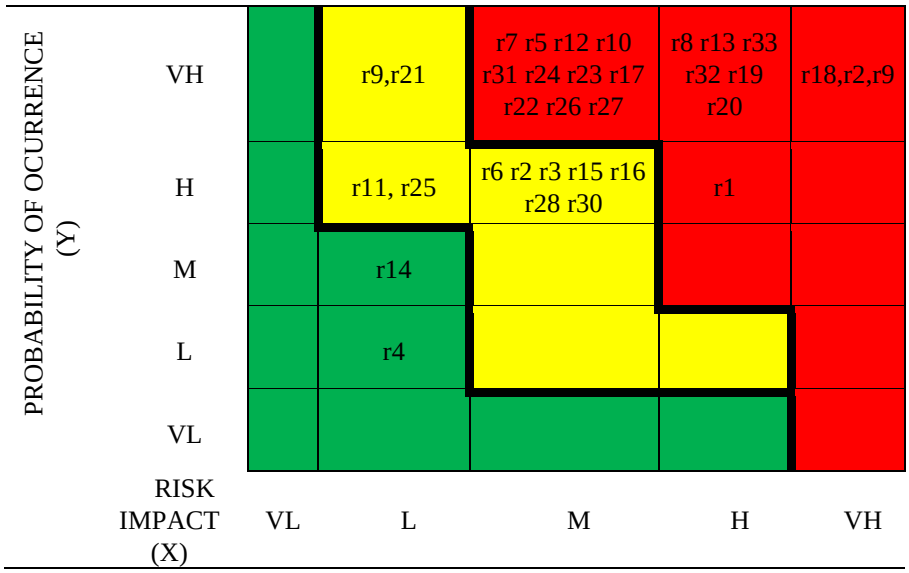


Fig. 2. Matrix of probability and impact of preliminary risks case study.

Table 2. Risks defined in the matrix to apply the FQFD.

RISK DESCRIPTION	ID	RISK DESCRIPTION	ID
Do not perform the sweep of enlistments in the defined times.	r1	Difference between physical and virtual inventory (Sale without existence)	r20
Do not have the necessary and obligatory courses.	r5	Failure to comply with agreed delivery.	r22
Pick a wrong SKU.	r7	Deliver an NP to the customer's home.	r23
Picking with color or size difference (batch).	r8	Do not place complete delivery seal on the bill.	r24
Enlisting wrong amounts of a SKU	r10	Generate delivery record before validating based on the Enrollment Sheet.	r26
Do not record the NP once enlisted and left in the distribution area.	r12	Do not check the quantities listed at the time of boarding.	r27
Failures in Saps at the time of generating enlistment sheets.	r13	Without space for storage of the enlistments.	r29
Do not label the orders with the talker.	r17	Mechanical failures of the vehicles while they	r31

RISK DESCRIPTION	ID	RISK DESCRIPTION	ID
		are in function of delivering NP.	
Orders stored in warehouse with more than 5 days of enlistment	r18	Do not count on the number of vehicles sufficient to deliver the NPs.	r32
No availability of forklift or Macalister for the preparation of the merchandise.	r19	Technological failures in Saps at the moment of generating delivery records.	r33

3.1 Phase 1 and 2. Identify the Internal Variables "What's" and Determine Their Weight

This phase is determined by the wishes of the decision-making group regarding the process that is being evaluated, which were recorded in Table 3, together with the relative importance assigned by the decision-making group.

Table 3. Internal Variables and their relative importance.

		Weight of WHAT'S		
W1	Deliver timely	7,2	8,2	9,2
W2	Deliver reliably	6,8	7,8	8,8
W3	Planned operation in times and effectiveness	7,6	8,6	9,6
W4	Efficiency in operational costs	5,2	6,2	7,2
W5	Have the correct layout	5,2	6,2	7,2
W6	Focused attention to the customer.	6,8	7,8	8,8

3.2 Phase 3. Identify the Strategic Objectives or "How's"

To determine how, the indicators that manage the analyzed process were established, since they were defined and focused on the fulfillment of the company's objectives or strategic guidelines. Therefore, they are listed in Table 4.

Table 4. Strategic objectives of the analyzed process. HOW.

	Strategic objectives or "How's
H1	Delivery on time
H2	Complaints and claims
H3	Impact NPS

	Strategic objectives or "How's
H4	Reprogramming of shipments
H5	Deliveries of transferred sales
H6	PIS Withdrawal in Store
H7	% Non-existent sales

3.3 Phase 4 and 5. Determine the Correlation Between the "What's and How's" and Define the Weight of the How's

In this phase the decision-making team qualifies the relationship that each of the WHAT has with respect to the HOW, for example for member E1 the relationship between timely delivery and timely deliveries has a HIGH (H) relationship, on the other hand for the same expert the relation that has the efficiency in operational costs with respect to deliveries on time is LOW (L). This sequence is followed for the rest of the experts and correlations and the weight of the HOW's is calculated as shown in Table 5.

Table 5. Weight of the How's for the case study.

	Strategic objectives or "How's	Weight of How´s		
H1	Delivery on time	44	58	74
H2	Complaints and claims	40	53	69
H3	Impact NPS	46	60	77
H4	Reprogramming of shipments	20	30	43
H5	Deliveries of transferred sales	38	51	67
H6	PIS Withdrawal in Store	41	55	70
H7	% Non-existent sales	24	35	48

According to Osorio-Gomez et al. [6], these diffuse triangular numbers correspond to the average of the multiplication between the weights of the "WHAT's and the assessment given for the relationship between each WHAT's and the corresponding strategic objective.

3.4 Phase 6 and 7. Determine the Impact of Risk on the Strategic Objectives "How's" and Establish the Priority of the Risks

The risks that are considered critical; selected from the red quadrants of the matrix were valued according to their relationship between each of them and the strategic objectives defined for the dispatch process, finally obtaining the order of priority shown in figure 3 where it is observed that risks r20 and r22 are the most critical for the process that was being considered. Additionally, it can be observed that of 20 risks that were evaluated, 11 were ranked higher and ranked among those with a criticality level between High and Very High, and each of these risks must be addressed and intervened in order to mitigate their risk, the impact or eliminate it.

3.5 Strategies or Actions to Mitigate Operational Risks

For the management group of the company this work was very useful since it could associate the FQFD methodology to its internal improvement processes, known as "Closed Cycle" and in this way make decisions based on the risks between the rating interval High and Very High; but additionally it was decided to group the risks by their common causes or common monitoring indicator, with this grouping the Cause-Effect tool is used, to finish with the establishment of actions and follow-up the indicators; which will show the improvement in the process of dismissals; established in table 6.

Then we proceeded to establish a defined order to intervene the risks, this order was established according to the specific needs of the organization and decided by the management of the branch. This order does not have any interference on the initial approach of establishing actions on the risks that affect the dispatch process, because in the end all the risks that after prioritizing through the FQFD have been intervened between the High and Very High levels.

RANKING	ID	RISKS	DIFFUSE RATING			DEFUSED RATING
1	VH	VH	287,3	439,9	419,7	**396,6914**
2	r20	Difference between physical and virtual inventory (Sale without existence)	282,3	432,9	413,6	**390,4122**
3	r22	Failure to comply with agreed delivery.	261,7	403,6	388,9	**364,4297**
4	r1	Do not perform the sweep of enlistments in the defined times.	255,2	395,7	382,8	**357,3352**
5	r32	Do not have the enough vehicles to deliver the NPs.	244,3	381,4	369,5	**344,1417**
6	r7	Enlist a wrong SKU.	243,5	379,2	368,8	**342,6766**
7	r23	Wrong deliver an NP to the customer's home.	241,8	378,1	365,0	**340,7562**
8	r10	Enlisting wrong amounts of a SKU	238,6	373,2	361,6	**336,6484**
9	r27	Do not check the quantities listed at the time of boarding.	231,8	364,4	353,8	**328,6145**
10	r8	Loading with lots difference.	231,5	362,8	352,5	**327,4328**
11	r26	Generate delivery record before validating based on the enlistment sheet.	221,5	350,0	340,8	**315,5825**
12	r31	Mechanical failures of the vehicles while they are in function of delivering notes ordered.	216,9	343,5	336,7	**310,1406**
13	H	H	215,5	342,2	335,7	**308,8743**
14	r33	Technological failures in Saps at the moment of generating delivery records.	198,5	317,0	316,0	**287,1364**
15	r13	Failures in Saps at the time of generating enlistment sheets.	191,2	309,1	307,6	**279,2343**
16	r19	No availability of forklift or nacalister for the preparation of the merchandise.	187,8	303,4	303,5	**274,5116**
17	r12	Do not record the NP once enlisted and left in the dispatch area.	182,1	296,9	297,9	**268,4278**
18	r29	Insufficient space for storage of the enlistments.	164,8	273,9	277,6	**247,5672**
19	r17	Do not label the orders with the talker.	154,8	259,8	265,9	**235,0701**
20	r18	Orders stored in warehouse with more than 5 days of enlistment	149,4	251,5	258,2	**227,6480**
21	M	M	143,6	244,4	251,8	**221,0571**
22	r5	Not having the approved forklift courses or heights.	139,6	238,0	247,4	**215,7333**
23	r24	Do not place complete delivery seal on bill.	110,8	199,7	212,5	**180,6690**
24	L	L	71,8	146,6	167,9	**133,2400**
25	VL	VL	0,0	48,9	83,9	**45,4229**

Fig. 3. Prioritization of risks in the case of study.

4 Conclusions

The identification of risks is very important, but this has no relevance if it is not included in your personal selection expert on the process, which validate and approve that such risks effectively impact the performance of the company.

It can be specified that prioritization is one of the most important steps since it is the crucial point where actions are directed or more focused strategies can be generated; about those risks that generate the greatest impact and are likely to affect the strategic objectives set by the company and finally be able to control, eliminate or mitigate them.

Table 6. Grouping of risks between High and Very High.

Ranking	Classification by group	Risk	KPI associate
7	2	Enlisting wrong amounts of a SKU	Mistakes in enlisting the products, 36 new features that represent 4% of sales made by deliveries.
9		Loading with lots difference.	
5		Enlist a wrong SKU.	
8		Do not check the quantities listed at the time of boarding.	
11	4	Do not have the enough vehicles to deliver the NPs.	Availability of vehicles in 97% to deliver orders
4		Mechanical failures of the vehicles while they are in function of delivering notes ordered.	
1	1	Difference between physical and virtual inventory (Sale without existence)	Noncompliance in the promise of delivery to the client. The indicator of delivery on time is 93.61%
6		Wrong deliver an NP to the customer's home.	
2		Failure to comply with agreed delivery.	
10	3	Do not perform the sweep of enlistments in the defined times.	Indicator notes lists vs. generated notes is 92% for the delivery of the merchandise to the customer
3		Generate delivery record before validating based on the enlistment Sheet.	

Through the implementation of the diffuse quality function deployment methodology or FQFD, it was possible to establish the priority of the risks in terms of their impacts on the strategic objectives of the company, this methodological scheme can be

applied throughout any process business. In this way the organization manages to have a clear picture of what are the critical risks associated with its processes.

Finally, the quantification of the impact of each risk on the financial scheme of an organization, that is, translating the occurrence of each risk and its impact to economic or financial terms, remains a study opportunity.

References

1. Avelar-Sosa, L., García-Alcaraz, J.L., Castrellón-Torres, J.P.: The Effects of Some Risk Factors in the Supply Chains Performance: A Case of Study. J Appl Res Technol 12:958–968. doi: 10.1016/S1665-6423(14)70602-9 (2014)
2. Fahimnia, B., Tang, C.S., Davarzani, H., Sarkis, J.: Quantitative Models for Managing Supply Chain Risks: A Review. Eur J Oper Res 247:1–15. doi: 10.1016/j.ejor.2015.04.034 (2015)
3. Giannakis, M., Louis, M.: A multi-agent based framework for supply chain risk management. J Purch Supply Manag 17:23–31. doi: 10.1016/j.pursup.2010.05.001 (2011)
4. Lavastre, O., Gunasekaran, A., Spalanzani, A.: Supply chain risk management in French companies. Decis Support Syst 52:828–838. doi: 10.1016/j.dss.2011.11.017 (2012)
5. Manotas, D.F., Osorio, J.C., Rivera, L.: Operational Risk Management in Third Party Logistics (3PL). In: Alor-Hernández G, Sánchez-Ramírez C, García-Alcaraz JL (eds) Handbook of Research on Managerial Strategies for Achieving Optimal Performance in Industrial Processes (2016)
6. Osorio-Gomez, J.C., Manotas-Duque, D.F., Rivera, .L, Canales, I. Operational risk prioritization in supply chain with 3PL using Fuzzy-QFD. In: New perspectives on applied industrial tools and techniques, management an industrial engineering. pp 91–109 (2018)
7. Sangwan, T., Liangro, J.: Risk Identification for Outbound Road Freight Transportation Service (2015)
8. Tummala, R., Schoenherr, T.: Assessing and managing risks using the Supply Chain Risk Management Process (SCRMP). Supply Chain Manag An Int J 16:474–483. doi: 10.1108/13598541111171165 (2011)
9. Wee, H.M., Blos, M.F., Yang, W.: Risk Management in Logistics. In: Handbook on Decision Making. Springer Berlin Heidelberg, pp 285–305 (2012)

ISSN 1870-4069

The Role of Employees' Performance and External Knowledge Transfer on the Supply Chain Flexibility

José Roberto Díaz-Reza; Jorge Luis García-Alcaraz; Liliana Avelar-Sosa, José Roberto Mendoza-Fong

Universidad Autónoma de Ciudad Juárez, Chihuahua, Mexico
al164440@alumnos.uacj.mx

Abstract. In this article structural equation models are reported, which relate four latent variables associated with employee performance, knowledge transfer, and supply chain flexibility, which incorporate 17 observed variables. In addition, the latent variables are related through 6 hypotheses that were tested with data from 269 questionnaires applied to the maquiladora industry in Ciudad Juarez, Mexico. Moreover, this model was executed in WarpPLS 6.0 software using the partial least squares technique to analyze the direct, indirect, and total effects. Additionally, the results show that the external knowledge transfer is crucial within the supply chains, since it explains 44.6% of the complexity, 19.5% of the employees' performance, as well as 10.6% of the supply chain flexibility.

Keywords: supply chain, maquiladoras, structural equation modeling.

1 Introduction

During the last two decades from the twentieth century, maquiladora industries have had a great importance in the Mexican economy. The maquiladoras are export assembly and processing plants specialized in labor-intensive products, and since 1965, favorable economic regulations have been established with the United States [1]. Since then, the proximity to the US market and the relatively cheap workforce labor have made Mexico one of the most favored offshore destinations for US companies for a long period of time. In addition, these maquiladoras have established strategies to reduce costs and waste, as well as generally apply advanced production processes and implement new methodologies. Likewise, they are distinguished by importing all raw materials and exporting all finished products, and because of materials flow in their supply chains (SC), which is an area of opportunity for further research [2].

Nowadays, due to the increase of uncertainty and complexity about the SC environment, companies must improve their competitiveness by reducing delivery time and changing the production level, since the company's operating capacity depends on the efficient operation network of supply chains in the company. Slack [3] proposed the concept of supply chain flexibility and noticed that this is the ability of its members to respond in a timely manner, according to customers' needs. Also, flexibility could

reduce the low demand impact, as well as reduce the maintenance costs because of unsold items [4].

Currently, the SC is no longer limited to the physical distribution, the information flow or funds flow [5]. Knowledge transfer (KT) is also added to the supply chains and it is considered as a strategic resource that affects the entire competitive advantage of the SC [5], which is the process where intra- and inter-organizational factors exchange, receive, and are affected by the knowledge from other ones. Also, external knowledge is transferred through collaboration agreements between external aspects (for example, clients, suppliers, and research institutes), and companies [6].

For this reason, SC management (SCM) improves competitive capabilities and performance by integrating the internal functions of the company and associating them with suppliers and customers operations effectively [7]. In order to be successful in SCM, applications that aim to achieve the high supply chain performance, external integration with suppliers and customers is needed, as well as the integration between the internal functions of the company [8], which generates knowledge that must be managed, and that is considered a critical success factor (CSF) in the SCM.

Furthermore, this knowledge is generated from people who are essential for the company success [9], because employees with high commitment consider their organization worthwhile and they are proud to work at [10]. Therefore, they will share all their efforts into working well for the organization, they will do it with greater autonomy, they will develop basic competences more quickly and, in addition, they will tend to be more receptive to any task that is given, and in this way, the probability that the company achieves a better performance will increase. Hence, the implementation of successful management in a supply chain requires effective management from human resources and superior employees' performance [11].

2 Literature Review and Hypothesis

CSF are the few key areas where the favorable results are absolutely necessary for a particular manager to reach their objectives, and it is because these areas are critical, and there must be the adequate information to determine whether the events are working adequately [12]. As a consequence, it is fundamental to identify the CSFs to manage the SCs, since they represent a wide variety of strategies dedicated to improving operational efficiency and competitiveness. [13]. In the literature review, CSF are reported for the appropriate SC management, for instance Kumar, Singh [13], have identified a total of 13 CSFs to implement SCM, such as Senior management commitment, Development of reliable suppliers, Higher flexibility in production systems, among other. On the other hand, Avelar-Sosa, García-Alcaraz [14] identified and classified 77 CSFs into four categories; risk attributes, regional attributes, manufacturing and performance practices, these categories are divided into latent variables, which are related by structural equation models to measure the impact on the SC performance. Also, Özdemir, Simonetti [7] report 25 CSFs, which are divided into 3 latent variables.

As it can be observed, there is a lot of literature review related to CSFs in the SC, and even that relates them through SCM, however, the CSFs from the External knowledge transfer (EKT), Supply chain complexity (SCC), Employee performance (EP), and Supply chain flexibility (SCF) from the maquiladora industries in Ciudad Juarez, Mexico. In the present article, these variables are linked through a SCM, therefore, the CSFs on each of these variables are described through a causal mode, as well as the impact that they have on when the SCF is measured.

2.1 Hypothesis

External Knowledge Transfer

The current globalization and innovation trend in the business environment has brought many external and internal challenges for modern companies, such as the volatile and changing market, large organizations, different product choices, etc. [15]. In addition, this change also increases the complexity and, therefore, threatens companies' performance. For instance, one area that is seriously affected is the SC, which requires the ascending and descending relationships with suppliers and customers management in order to deliver high quality to the customer at a lower cost along with the supply chain as a whole [16]. Consequently, the following hypothesis can be proposed:

H_1: The EKT has a direct and positive effect on the SCC.

The knowledge transfer is a system created to address client's needs and expectations in a more technical way to avoid misunderstandings and errors that may result in inefficiencies [17]. For companies that operate in the manufacturing industry, knowledge can be seen as the cornerstone of the business to complete quality products with success, elimination of waste, and defects in a short period of time, as well as deliver the product as needed according to clients [18]. In addition, it is a complex process involving education, learning, knowledge communication and promotion to employees and leaders [19]. Also, the employees' performance is key for the success of any organization; in manufacturing, employees are still relevant in the production process, but most important are the initiators and drivers of changes as well as improvements in design, monitoring and evaluation, therefore, the following hypothesis can be presented:

H_2: The EKT has a direct and positive effect on EP.

Supply Chain Complexity

The SC complexity grows as the client requirements, the competitive environment, and the industry changes, as well as SC companies are part of strategic alliances, carry out mergers and acquisitions, subcontract functions to third factors, adopt new technologies and launch new products/services, and extend their operations to new geographies, time zones, and markets. Due to its complexity, SC networks are difficult to understand, describe, predict, and control, in order to reduce the level of uncertainty in these networks, it is necessary to understand the different roles of the members in the SC, their interactions and the transition models that they use to interact with each other. In

addition, employee commitment is a relevant tool to help each organization to strive to obtain a competitive advantage over others, since people are a factor that cannot be duplicated or imitated by competitors, and it is considered the most important asset if it is properly managed and performed [20]. Also, employees' performance is basically the results obtained and the achievements in work. Performance refers to maintaining plans while aiming some results [20].

H_3: The SCC has a direct and positive effect on EP.

Employees Performance

Work performance is the result of three factors that work together: skills, efforts, and the nature of working conditions [21]. In addition, an appropriate employees' performance in an organization has many implications, such as great motivation, outstanding ability, an acceptable climate, and organizational infrastructure, excellent leadership that can maintain the relationship and productivity as well as an adequate relationship with the staff [21].

Moreover, KT is related to continuous improvement to achieve a high level of productivity [19]. In addition, significant sources of new knowledge and innovations are suggestions for employees and partners established in the production network (for example, suppliers and customers), as well as the commitment of highly qualified people. Also, the knowledge transfer is not only essential for people and/or companies to seek for a better performance, but it has also been increasingly recognized as a moral challenge in organizations [22]. Therefore, the following hypothesis can be proposed:

H_4: The EKT has a direct and positive effect on EP.

Supply Chain Flexibility

A SC is definitely a complex system that integrates a large number and a variety of relationships, processes, and interactions between and within companies, dynamic processes and interactions where several levels of the system and a large amount of data is involved. Also, some companies in cyclical industries increasingly face a volatile demand and must adjust their production volume quickly without incurring significant costs [23], since there is a high probability that customers will suddenly increase, reduce, cancel or advance or regress their orders, factors in the supply chain must be more flexible in many ways [24]. Therefore, the following hypothesis can be established:

H_5: The SCC has a direct and positive effect on the SCF.

Moreover, SCs must be more responsive to the customers changing requirements, as well as offer an added value above the average, therefore, the manufacture flexibility and the SC is becoming one of the key objectives for manufacturers [23]. Also, the commitment to training and development of multidisciplinary workforce may improve the workers capabilities to manage different products as well as handle different operations and tools, while contributing to increase the ability of companies to move from the production of a product to another in a combination with other products, and

minimize transition penalties, which contribute to higher levels of product flexibility [25].

In this way, the following hypothesis is proposed:

H_6: EP has a direct and positive effect on the SCF

Figure 1 presents the hypotheses that related the variables.

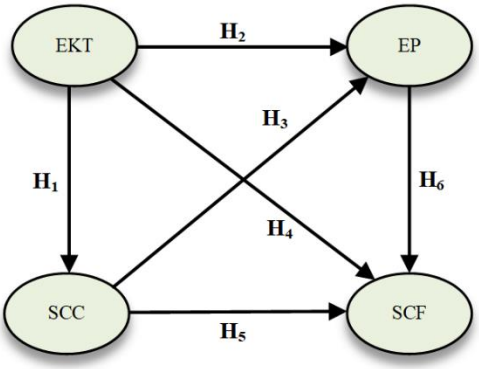


Fig. 1. Proposed model.

3 Methodology

3.1 Questionnaire Development

In order to carry out this research, the questionnaire by Blome, Schoenherr [26] was implemented, which other study variables were added, such as, for example, employees and SCs performance. In the first part, demographic data are requested while the second part consists of 33 items divided into seven different variables. In addition, in the current work, only four variables were used; EKT, SCF, SCC and EP.

Furthermore, the items for each of the latent variables are the following: EKT (Suppliers are able to share their experiences in new technology with researchers, there were frequent meetings with suppliers to develop new knowledge, where the purchaser-provider relationship is technical addressed to integrate the supplier into our new products and processes); SCF (Short-term adjusting of suppliers' order of goods and services, adjusting deliveries to customer changes, reducing manufacturing lead time, reducing development cycle times, adjusting manufacturing process capacity, increasing frequencies of new product introductions); SSC (The number of our direct suppliers is high, long-term plans of our procurement activities are hampered by high dynamism, our suppliers often do not supply on time or the desired quality), and EP (High employee morale, high employee productivity, fast troubleshooting, high usage of employees skills and abilities, internal customers' concept is widely understood).

Finally, in order to answer each of the questions, a Likert scale of five points was used, where 1 represents that the activity that is never performed whereas 5 indicates that it is always done.

3.2 Questionnaire Validity

The data registration and purification were performed in the SPSS 21® software, where the standard deviation of each questionnaire was calculated, as well as missing values are identified as extreme values, which are replaced by the median. In order to validate the latent variables analyzed in the model, several indexes are used, such as R^2 and Adj. R^2 to measure the predictive validity, the Compound Reliability Index and the Cronbach's alpha are used for internal reliability, the Average Variance Extracted (AVE) to measure the predictive validity, the average full collinearity VIF to measure the multicollinearity, finally, Q^2 is used to measure the nonparametric predictive validity.

3.3 Equation Structural Modeling

The proposed hypotheses in Fig. 1 are tested using the structural equation modeling (SEM) technique in WarpPLS 6.0® software. In addition, the efficiency indexes from the analyzed models are: average path coefficient (APC), average R^2 (ARS), average variance inflation factor (AVIF), and average full collinearity VIF (AFVIF), and Tenenhaus (GoF), all proposed by Kock [27].

Moreover, the relationships between variables are called effects, the direct effects are represented by arrows (each one represents a hypothesis), the indirect effects that occur between an independent variable on a dependent variable, through a mediating variable, and the total effects, which are the sum of the indirect effects plus the direct effects. In order to determine the significance of each effect, the P values associated with a β value are estimated for a hypothesis test with a level of significance of 0.95, that is, H0: $\beta = 0$; H1: $\beta \neq 0$, and the size effect (SE) is also reported for each dependent latent variable [28].

4 Results

4.1 Descriptive Analysis of the Sample

From the questionnaire application and after the database debugging, a total of 269 valid cases of individuals working in the companies were obtained, where, 53.15% (143) have between one to two years in the position, 17.47% (47) have up to 5 years, and, 29.3% (79) have over 5 years working within the industry. In the same way, the industrial sectors that participated are distributed as follows: automotive sector with 119, electronic sector with 42, machinery with 27, aeronautical with 25, medical with 15, also 10 questionnaires were from different sectors to those already mentioned. Finally, 22 participants left that question without an answer.

4.2 Questionnaire Statistic Validation

In Table 1, the values of the indexes for each latent variable used in the model are shown, where it can be observed that they are achieved, and the analysis is proceeded.

Table 1. Variables validation.

	SCF	*EKT*	*EP*	*SCC*
R^2	0.415		0.453	0.446
Adj. R^2	0.409		0.449	0.444
Composite Reliability	0.911	0.901	0.934	0.876
Cronbach´s Alpha	0.877	0.835	0.911	0.830
Avg. Var. Extracted	0.671	0.752	0.738	0.541
Full Collin. VIF	1.687	1.968	2.013	2.087
Q^2	0.418		0.455	0.444

4.3 Structural Equation Modeling

The results from the efficiency indexes of the model are the following: APC = 0.357 and a value P <0.001, ARS = 0.438, P value <0.001, AARS = 0.434, and a P value <0.001, which shows that they are statistically significant. In addition, the values AVIF = 1.892 and AFVIF = 2.628 demonstrate that there is no collinearity problems, finally, according to the GoF value = 0.589 index, it is concluded that the model has enough explanatory power.

Direct Effects

In Fig. 2 the β and P values can be observed for each of the direct effects or proposed hypotheses in the model from Fig. 1, it can be seen that each of these hypotheses are statistically significant since the P value for each is under 0.05.

Indirect Effects

Table 2 presents the four indirect effects, which are integrated by two segments, where it is observed that all values are statistically significant according to the associated p value.

Total Effects

Table 3 portrays the total effects (sum of the direct and indirect effects), likewise, the P values are shown for each of them, and it is observed that all values are statistically significant. Also, the largest effect within this model is caused by the EKT towards the SCC with a value of 0.668

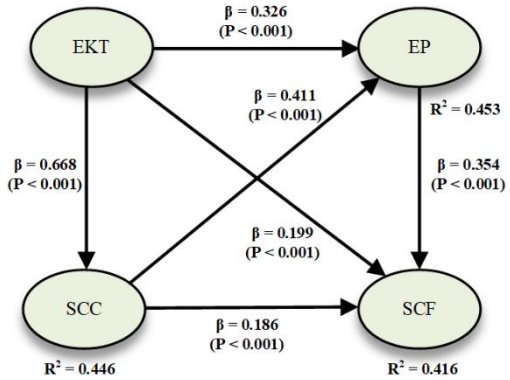


Fig. 2. Evaluated model.

Table 2. Indirect effects for 2 ways segments.

	Independent Variables	
Dependent Variables	EKT	SCC
SCF	0.240 P < 0.001 ES = 0.128	0.145 P < 0.001 ES = 0.079
EP	0.274 P < 0.001 ES = 0.164	

Table 3. Total effects.

	Independent Variables		
Dependent Variables	EKT	EP	SCC
SCF	0.536 P < 0.001 ES = 0.285	0.354 P < 0.001 ES = 0.209	0.332 P < 0.001 ES = 0.179
EP	0.600 P < 0.001 ES = 0.359		0.411 P < 0.001 ES = 0.258
SCC	0.668 P < 0.001 ES = 0.446		

5 Conclusions

According to direct effect results, the following can be concluded:

- It is observed that the largest effects are those from the EKT variable.
- The 0.453 from the EP is explained by 0.195 because of the EKT while 0.258 because of the SCC variable.
- Finally, the 0.416 from the SCF is explained in 0.106 by the EKT, in 0.209 by EP, and in 0. 101 by the SCC variable.

- As it was already mentioned, the external knowledge transfer is an essential for companies, since it is vital that the relationship between supplier-purchaser is very close, in order that the data is shared in real time, therefore overcome the changes in the demands.
- The employees' participation in the flexibility is relevant, since they are the ones who develop the activities to overcome the changes in the demand, as well as they are the ones that make them fit in the production processes to fulfill on it.
- It is important to have an adequate number of suppliers and avoid having more than required in order to avoid delays and low-quality products that may delay tasks.

References

1. Utar, H., Ruiz, L.B.T.: International competition and industrial evolution: Evidence from the impact of Chinese competition on Mexican maquiladoras. Journal of Development Economics 105:267–287 (2013)
2. García-Alcaraz, J.L. et al.: Structural equation modeling to identify the human resource value in the JIT implementation: case maquiladora sector. The International Journal of Advanced Manufacturing Technology 77(5):1483–1497 (2015)
3. Slack, N.: The flexibility of manufacturing systems. International Journal of Operations & Production Management 7(4):35–45 (1987)
4. Chan, H., et al.: Flexibility and adaptability in supply chains: a lesson learnt from a practitioner. Supply Chain Management: An International Journal 14(6):407–410 (2009)
5. Kang, P., Jiang, W.: The Evaluation Study on Knowledge Transfer Effect of Supply Chain Companies. in Advances in Education and Management. Berlin, Heidelberg: Springer Berlin Heidelberg (2011)
6. Ollila, S., Elmquist, M., Fredberg, T.: Exploring the field of open innovation. European Journal of Innovation Management 12(3):326–345 (2009)
7. Özdemir, A.İ., Simonetti, B., Jannelli, R.: Determining critical success factors related to the effect of supply chain integration and competition capabilities on business performance. Quality & Quantity 49(4):1621–1632 (2015)
8. Wook Kim, S.: Effects of supply chain management practices, integration and competition capability on performance. Supply Chain Management: An International Journal 11(3):241–248 (2006)
9. Wright, P.M., Dunford, B.B., Snell, S.A.: Human resources and the resource based view of the firm. Journal of management 27(6):701–721 (2001)
10. Marín García, J.A., Medina López, M.d.C., Alfalla Luque, R.: Is worker commitment necessary for achieving competitive advantage and customer satisfaction when companies use HRM and TQM practices? Universia Business Review 36:64–88 (2012)
11. Swart, W., Hall, C., Chen, H.: Human Performance in Supply Chain Management. Supply Chain Forum: An International Journal 13(2):10–20 (2012)
12. Bullen, C.V., Rockart, J.F.: A primer on critical success factors (1981)
13. Kumar, R., Singh, R.K., Shankar, R.: Critical success factors for implementation of supply chain management in Indian small and medium enterprises and their impact on performance. IIMB Management Review 27(2):92–104 (2015)

14. Avelar-Sosa, L., García-Alcaraz, J.L., Maldonado-Macías, A.A.: Models of Manufacturing Practices and Integrative Model, in Evaluation of Supply Chain Performance: A Manufacturing Industry Approach. In: L. Avelar-Sosa, J.L. García-Alcaraz, and A.A. Maldonado-Macías (eds.) Springer International Publishing: Cham., pp. 373–411 (2019)
15. Sun, C., et al.: Best Practice Sharing for Complexity Management in Supply Chains of the Semiconductor Industry. Procedia CIRP 41:538–543 (2016)
16. Christopher, M.: Logistics and Supply Chain Management (Financial Times Series), 4ta ed. Prentice Hall (2010)
17. Schomaker, M.S., Zaheer, S.: The role of language in knowledge transfer to geographically dispersed manufacturing operations. Journal of International Management 20(1):55–72 (2014)
18. Ooi, K.-B., et al.: TQM practices and knowledge sharing: An empirical study of Malaysia's manufacturing organizations. Asia Pacific Journal of Management 29(1):59–78 (2012)
19. Argote, L., Ingram, P.: Knowledge transfer: A basis for competitive advantage in firms. Organizational behavior and human decision processes 82(1):150–169 (2000)
20. Jagannathan, A.: Determinants of employee engagement and their impact on employee performance. International Journal of Productivity and Performance Management 63(3): 308–323 (2014)
21. Hassan, N., et al.: Critical Factors in Organizational Change and Employee Performance. in Proceedings of the Colloquium on Administrative Science and Technology. Singapore: Springer Singapore (2015)
22. Chih-Chien, W.: The influence of ethical and self-interest concerns on knowledge sharing intentions among managers: An empirical study. International Journal of Management 21(3): 370 (2004)
23. Seebacher, G., Winkler, H.. A Citation Analysis of the Research on Manufacturing and Supply Chain Flexibility. International Journal of Production Research 51(11):3415–3427 (2013)
24. Angkiriwang, R., Pujawan, I.N., Santosa, B.: Managing uncertainty through supply chain flexibility: reactive vs. proactive approaches. Production & Manufacturing Research 2(1): 50–70 (2014)
25. Chang, S.-C., et al.: Manufacturing flexibility and manufacturing proactiveness: empirical evidence from the motherboard industry. Industrial Management & Data Systems 105(8): 1115–1132 (2005)
26. Blome, C., Schoenherr, T., Eckstein, D.: The impact of knowledge transfer and complexity on supply chain flexibility: A knowledge-based view. International Journal of Production Economics 147: 307–316 (2014)
27. Kock, N.: WarpPLS 5.0 User Manual. Laredo, TX, USA: ScriptWarp Systems (2015)
28. Hayes, A.F., Preacher, K.J.: Quantifying and testing indirect effects in simple mediation models when the constituent paths are nonlinear. Multivariate Behav Res 45:34 (2010)

ISSN 1870-4069

Towards a Proposal of Personalized Medical Decision Support Systems: Analysis of Gene Expression Levels of Diabetes Mellitus, Inflammation and Oxidative Stress in Alzheimer's Disease

Sonia Lilia Mestizo Gutiérrez[1] , Nicandro Cruz Ramírez[2], Gonzalo Emiliano[2], Aranda Abreu[3]

[1] Facultad de Ciencias Químicas, Universidad Veracruzana, Xalapa, Veracruz, Mexico
[2] Centro de Investigación en Inteligencia Artificial, Universidad Veracruzana, Xalapa, Veracruz, Mexico
[3] Centro de Investigaciones Cerebrales, Universidad Veracruzana, Xalapa, Veracruz, Mexico
{smestizo,ncruz,garanda}@uv.mx

Abstract. The increased incidence of Alzheimer's disease (AD) and diabetes mellitus (DM) are emerging as major public health problems worldwide. Both sufferings share pathophysiological characteristics and have no cure. Inflammation of the central and peripheral nervous system has been shown to be the link between DM and AD. Oxidative stress is also associated with AD and DM. The increasing complexity of the problems and the continuous growth of information creates the need for the use of Decision Suport System (DSS) driven by the use of new technologies such as big data and machine learning. In this context, the objective of this work is to use decision trees and Bayesian networks as mechanisms of classification of AD gene expression levels, DM, inflammation and oxidative stress, MMSE (Mini-Mental State Examination) score and the number of neurofibrillary tangles to classify 31 individuals (9 healthy controls and 22 AD patients in three different stages of disease) that could be key in the development of AD. Our results allowed us to generate classification models of different states of AD severity, according to the MMSE and we found that the level of expression of the ADIPOQ gene could play an important role in the onset of AD. Our predictive model can contribute knowledge that could be incorporated into a personalized medical DSS in the future.

Keywords: medical decision support system, decision trees, Bayesian networks, Alzheimer disease, diabetes mellitus.

1 Introduction

Alzheimer's disease (AD) is the most common cause of dementia. It is a slowly advancing neurodegenerative disorder with cognitive impairment, progressive memory loss, and behavioral disorders. Despite major research efforts, there is still no cure, but new research is underway to determine the cause of the disease and detect changes in the

brain before the first symptoms appear. Worldwide, there are 50 million people with dementia [1]. The incidence of AD is increasing and constitutes a public health challenge in our society characterized by the increase of elderly people. There are two proteins involved in the development of AD: the beta amyloid protein (Aβ) that accumulates abnormally in the brain to form extracellular neuritic amyloid plaques and the tau protein that produces the formation of intracellular neurofibrillary tangles. Both alterations increase the levels of inflammation, oxidative stress and lead to the death of neurons. In the last 20 to 30 years, scientists have discussed which protein plays the most important role in the development of the disease. The certainty of the diagnosis of AD is approximately 85% and is only confirmed by post mortem examination. AD is multifactorial in nature and is considered a complex condition resulting from an interaction of environmental and genetic factors. The main risk factor is advanced age, however other potential risk factors have been found such as sex, diabetes mellitus, headaches, lifestyle, hypertension, obesity, dyslipidemia, metabolic syndrome, cerebrovascular disease, smoking, physical inactivity, depression and low levels of education [2]. There are reported findings from genetic studies that have pointed to APP metabolism, immune response, inflammation, lipid metabolism and intracellular trafficking/endocytosis that open the door for exploration of new pathways for genetic testing, prevention and treatment [3].

Diabetes Mellitus (DM) is a chronic disease characterized by a high concentration of glucose in the blood because the body does not produce insulin or does not use it properly. Globally, it is estimated that there are 425 million diabetics and it is estimated that by 2045 it will increase to 629 million [4] making it one of the major health challenges of this century. Several studies converge on the implication of inflammation as a key factor in the relationship of DM with AD [5, 6, 7].

The initial relationship between AD and DM was established in the Rotterdam study where it is revealed that Diabetes Mellitus type 2 (DM2) doubles the risk of patients to develop AD, while patients with Diabetes Mellitus type 1 (DM1) who receive insulin treatment quadruple the risk [8]. Several studies [9, 10, 11] propose the existence of a relationship between AD and DM and some authors have called it "type 3 diabetes" [12, 13, 14].

The existence of large volumes of biomedical data provides a great opportunity for better understanding, prediction and decision making of conditions. Microarrays are a powerful technique for the measurement of gene expression data that allow the comparison of the relative abundance of messenger RNA generated in different biological tests. The analysis of microarrays is a challenge due to its high dimensionality and complexity so machine learning techniques have been used with satisfactory results. Our work aims to use supervised learning techniques (decision trees and Bayesian networks) to classify gene expression levels of Alzheimer's disease, diabetes mellitus, inflammation and oxidative stress from a public database of 31 individuals, MMSE scores and number of NFT (neurofibrillary tangles) in order to contribute to a better understanding of AD and provide knowledge for the development of earlier and more accurate diagnosis, as well as the development of more appropriate treatments leading to future personalized treatments for incorporation into a personalized medical DSS.

In the next sections we present the state of the art, the methodology that we followed for building classifiers, the results and finally the conclusions and future work.

2 State of Art

Several works have used automatic learning techniques (neural networks, support vector machines, bagging, boosting, information gain, random forests, genetic algorithms) for the analysis of the levels of expression of AD [15, 16, 17]. Some work using Bayesian nets has also been carried out [18, 19].

Recently the use of machine learning for the classification of gene datasets has increased. In one study decision trees were used to classify a gene dataset of AD. Classification models were generated according to Mini-Mental State Examination (MMSE) scores to identify expression levels of different proteins that could determine the involvement of genes involved in various pathways of AD pathogenesis. The results showed that the MMSE score and relevance association score are the most significant attributes for gene classification. In the functional gene classification analysis, they reported that APOE, PSEN1, GRN, ACE, BCHE, PRNP, IL1A are strongly related to AD [20]. Machine learning techniques (decision trees, quantitative association and hierarchical cluster) have been used to identify potential genes for the prognosis of AD through the use of different biological sources (microarrays, PubMed, GO and PPI network). The results reported a set of significant genes (down/up) related to AD [21]. Park and colleagues formulated a new random forest-based method that allows the classification of gene-gene interaction of gene expression profiles. The proposed method was evaluated using AD data with remarkable accuracy, the result of gene-gene interaction could be used for the construction of a genetic network to explain underlying mechanisms of AD [22].

In a recent study, decision trees were used to report a genetic risk profile derived from a set of candidate genes related to cognitive performance selected a priori in order to explore the combined effect of these genes on cognitive impairment rates during the preclinical stage of AD. The results support the hypothesis that the combination of genes associated with cognitive performance makes it possible to identify groups with accelerated rates of cognitive impairment [23].

3 Methodology

The development of this study was divided into two main phases. During the first phase we made the selection of the microarrays database, the analysis of properties of the data and the preprocessing techniques were applied and in the second phase the genes of interest for our study were selected and the techniques of decision trees and Bayesian networks were applied to obtain the knowledge models that represent the patterns of behavior in the levels of genetic expression of Alzheimer's disease. Finally, we evaluated and interpreted the results obtained.

3.1 Database Selection

In this work we made use of the microarray database GDS810 obtained from the *National Center for Biotechnology* (NCBI) *Gene Expression Omnibus* (GEO) database ([HG-U133A] *Affymetrix Human Genome* U133A Array) [24]. Expression levels of 23,283 genes from 31 individuals were obtained from the CA1 region of the hippocampus and correspond to 9 control patients, 7 with incipient Alzheimer's disease, 8 with moderate Alzheimer's disease and 7 with severe Alzheimer's disease. The dataset includes MMSE scores and number of NFT [25].

3.2 Property Analysis and Preprocessing

For the analysis of the properties of the data we proceeded to explore, clean and adjust the data. We removed clones and pseudogenes from the database. Regarding the preprocessing of Affymetrix microarray data, RMA (Robust Multi-Array Average), GCRMA (GeneChip Robust Multi-Array Average), MAS5 (MicroArray Suite 5.0) and Expresso (Gautier, et al., 2004) were used in the normalization phase using the Affy R [26] Bioconductor package.

3.3 Gene Selection

Our interest focused on the expression values of genes related to AD: APP, APOE, BACE1, NCSTN, PSEN1, PSEN2, MAPT and INPP5D, MEF2C, HLA-DRB5/DRB1, NME8, ZCWPW1, PTK2B, CELF1, SORL1, FERMT2, SLC24A4, CASS4 [27], as well as DM genes: HLA-DQB1, TCF7L2, ACE, PPARG, HLA-DQA1, APOE, ADIPOQ and inflammation: TNF, IL6, IL1B, IL10, TLR4, IL1RN, LTA, IL1A, CD14, PTGS2, CRP reported by Genotator [28], and oxidative stress-related genes: ANXA6, ARAF, CBX7, DHX16, EBP, FGF13, HIF1A, TNIP1 or NAF1, NDUFS1, NFE2, POLD1, RAB15, SGK2, SMAD5, STAT5B, UBA7, WNT2B [29]. We added MMSE score and number of NFT.

3.4 Decision tree and Bayesian network

The performance of our classifiers is based on precision (number of correct classifications divided by the size of the test set), sensitivity (number of AD patients correctly identified) and specificity (correct identification of patients without AD).

For the analysis of gene expression levels using decision trees [30, 31] and Bayesian networks [32,33] the clones and pseudogenes were removed from the database. The data were analyzed using a WEKA software utilizing decision tree J48 classification algorithm and Bayesian network (Naive Bayes algorithm) with 10-fold cross-validation. We used a decision tree because they provide models that are easy to interpret and understand thanks to their ability to select and classify attributes according to their relevance [34].

In the generation of the Bayesian network we use the CAIM (Class-attribute Interdependence Maximization) [35] and MDL (Minimum Description Length) [36] methods provided by WEKA (Waikato Environment for Knowledge Analysis) [37, 38].

4 Results

In Fig.1, the decision tree obtained from the levels of genetic expression of AD, DM, inflammation, oxidative stress, MMSE score and the number of neurofibrillary tangles is presented with an accuracy of 87.09%, sensitivity of 90.90% and specificity of 77.77%.

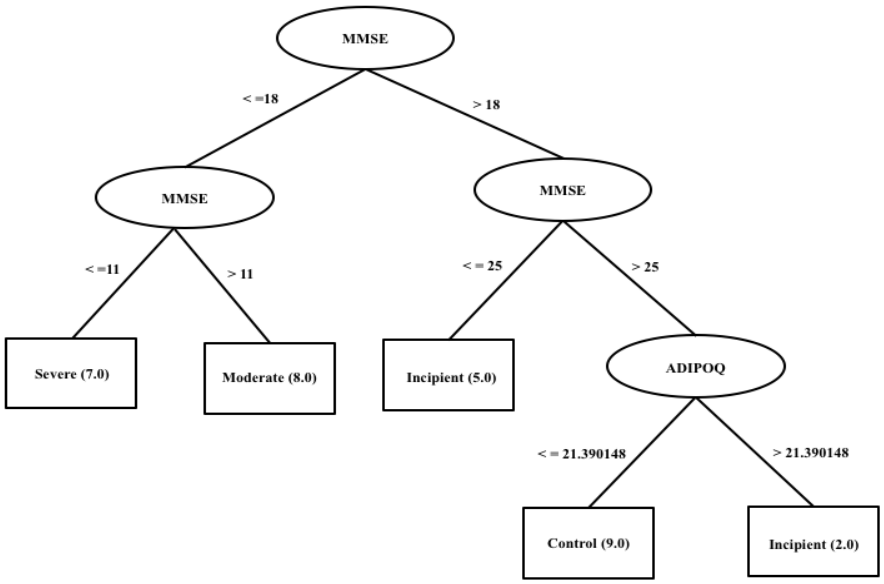


Fig. 1. Decision tree of the main genes related to AD, DM, inflammation, oxidative stress, MMSE and number of neurofibrillary tangles.

As we can see, the most informative variable is the MMSE. The J48 algorithm provides MMSE score cut-off values for each stage of the disease: normal > 25, incipient 19-25, moderate 18-12 and severe < 12, which are similar to those used in clinical practice to classify an individual's cognitive status. The importance of our model is that it allows us to identify individuals at an early stage of AD when the MMSE score is above 25 points and the level of expression of ADIPOQ (Adiponectin, C1Q and Collagen Domain Containing) is greater than 21.390148, an individual is classified as AD incipient. The ADIPOQ gene is only expressed in adipose tissue. Obesity has been reported to be a significant risk factor for the development of metabolic syndrome and other degenerative diseases. One study found that serum adiponectin level correlated positively ($r=0.683$, $P<0.001$) with MMSE score in patients with AD [39]. Our work corroborates the results obtained in this study, however it is more explanatory since it allows us to identify AD at an early stage. Another advantage is that our model is transparent and understandable for human experts who are not machine learning specialists.

The model generated by Naive Bayes with the discretization technique CAIM correctly classified 29 of the 31 samples: 7/7 of severe AD, 8/8 of moderate AD, 6/7 of incipient AD and 8/9 of healthy control. Model accuracy was 93.54%, sensitivity 95.45% and specificity 88.88%. In the model obtained by Naive Bayes with the MDL discretization technique, an accuracy of 90.32%, sensitivity of 90.90% and specificity of 88.88% were obtained. From our results, the best classification model was obtained using Naive Bayes with the CAIM discretization technique.

5 Conclusions

In the development of this work we evaluated classifiers with decision tree techniques and Bayesian networks of AD gene expression levels, DM, inflammation and oxidative stress, MMSE score and the number of neurofibrillary tangles. In the decision tree, the MMSE score was the most important attribute, however we found that the level of ADIPOQ expression can play a crucial role in distinguishing between a normal cognitive state and incipient EA when the MMSE score is considered normal. In summary, we successfully modeled different states of AD with accuracies of 87.09% (decision tree), 93.54% (Naive Bayes with CAIM) and 90.32% (Naive Bayes with MDL) and showed that the level of expression of ADIPOQ has potential to be considered in the early diagnosis of AD so our results could contribute with knowledge for a future implementation of a personalized medical DSS. The development of this work demonstrates that the use of machine learning techniques, provide favorable results for the early diagnosis of AD. The models obtained can become the knowledge base of a personalized medical DSS. A limitation of our is the sample size, so as future work is proposed the use of artificial instance generators to improve performance.

Acknowledgements. This work was supported by support for the reincorporation of former fellow PROMEP DSA/103.5/16/10415/EXB-552, 47856 project.

References

1. Alzheimer's Disease International: The state of the art of dementia research: New frontiers Homepage, https://www.alz.co.uk/research/WorldAlzheimerReport2018.pdf?2, last accessed 2018/09/15.
2. Kivipelto, M., Mangialasche, F., Ngandu, T.: Lifestyle interventions to prevent cognitive impairment, dementia and Alzheimer disease. Nature reviews. Neurology, doi: 10.1038/s41582-018-0070-3 (2018)
3. Reitz, C., Mayeux, R.: Alzheimer disease: epidemiology, diagnostic criteria, risk factors and biomarkers. Biochemical pharmacology 88(4):640–651 (2014)
4. International Diabetes Federation, Homepage, https://www.idf.org/, last accesed 2018/09/15.
5. Lue, L., Andrade, C., Sabbagh, M., Walker, D.: Is there inflammatory synergy in type II diabetes mellitus and Alzheimer's disease? J Alzheimers Dis. 1:1–9 (2012)
6. De Felice F., Ferreira S.: Inflammation, defective insulin signaling, and mitochondrial dysfunction as common molecular denominators connecting type 2 diabetes to Alzheimer´s disease. Diabetes 63:2262–72 (2014)

7. Jiang, C., Li, G., Huang, P., Liu, Z., Zhao, B.: The Gut Microbiota and Alzheimer's Disease. J Alzheimers Dis. 58(1):1–15 (2017)
8. Ott, A., Stolk, R., Hofman, A., van, H., Grobbee, D., Breteler, M.: Association of diabetes mellitus and dementia: The Rotterdam Study. Diabetologia 39:1392–1397 (1996)
9. Pasquier, F., Boulogne, A., Leys, D., Fontaine, P.: Diabetes mellitus and dementia. Diabetes & Metabolism 5(32):403–414 (2006)
10. Arab, L., Sadeghi, R., Walker, D., Lue, L., Sabbagh, M.: Consequences of Aberrant Insulin Regulation in the Brain: Can Treating Diabetes Be Effective for Alzheimer's Disease. Neuropharmacol 9(4):693–705 (2011)
11. Adeghate, E., Donáth, T., Adem, A.: Alzheimer disease and diabetes mellitus: do they have anything in common? Curr Alzheimer Res. 10(6) (2013)
12. Steen, E., Terry, B., Rivera, E., Cannon, J., Neely, T., Tavares, R., Xu, X., Wands, J., de la Monte, S.: Impaired insulin and insulin-like growth factor expression and signaling mechanisms in Alzheimer's disease—is this type 3 diabetes? J Alzheimers Dis. 7(1):63–80 (2005)
13. Kroner, Z.: The relationship between Alzheimer's disease and diabetes: Type 3 diabetes? Altern Med Rev. 14(4):373–9 (2009)
14. de la Monte, S.: Wands, J. Alzheimer's disease is type 3 diabetes-evidence reviewed. J Diabetes Sci Technol. 2(6):1101–13 (2008)
15. Walker, P., Smith, B., Liu, Q., Famili, A., Valdés, J., Liu, Z., Lach, B.: Data mining of gene expression changes in Alzheimer brain. Artif. Intell. Med. 31(2):137–154 (2004)
16. Scheubert, L., Lustrek, M., Schmidt, R., Repsilber, D., Fuellen, G.: Tissue-based Alzheimer gene expression markers comparison of multiple machine learning approaches and investigation of redundancy in small biomarker sets. BMC Bioinformatics 13(266) (2012)
17. Jain, M., Dua, P., Dua, S., Lukiw, W.J.: Data Adaptive Rule-based Classification System for Alzheimer Classification. J Comput Sci Syst Biol 6:291–297 (2013)
18. Armañanzas, R., Larrañaga, P., Bielza, C. Ensemble transcript interaction networks: a case study on Alzheimer's disease. Comput Methods Programs Biomed. 108(1):442–50 (2012)
19. Zhang, B., Gaiteri, C., Bodea, L.G., Wang, Z., McElwee J., Podtelezhnikov, Emilsson, V.: Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. Cell. 153(3):707–20 (2013)
20. Kumar, A., Singh, T.R.: A New Decision Tree to Solve the Puzzle of Alzheimer's Disease Pathogenesis Through Standard Diagnosis Scoring System. Interdisciplinary Sciences: Computational Life Sciences 9(1):107–115 (2017)
21. Martínez-Ballesteros, M., García-Heredia, J. M., Nepomuceno-Chamorro, I. A., Riquelme-Santos, J. C.: Machine learning techniques to discover genes with potential prognosis role in Alzheimer's disease using different biological sources. Information Fusion 36:114–129 (2017)
22. Park. C., Kim, J., Kim, J., Park, S.: Machine learning-based identification of genetic interactions from heterogeneous gene expression profiles. PLoS ONE 13(7):e0201056 (2018)
23. Porter, T., Villemagne, V. L., Savage, G., Milicic, L., Ying Lim, Y., Maruff, P., Laws, S.M.: Cognitive gene risk profile for the prediction of cognitive decline in presymptomatic Alzheimer's disease. Personalized Medicine in Psychiatry 7(8):14–20 (2018)
24. Gene Expression Omnibus, http://www. ncbi.nlm.nih.gov/geo, last accesed 2018/07/21
25. Blalock, E.M., Geddes, J.W., Chen, K.C., Porter, N.M., Markesbery, W.R., Landfield, P. W.: Incipient Alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses. Proc. Nat. Acad. Sci. U.S. A. 101:2173–2178 (2004)
26. Gautier, L., Cope, L., Bolstad, B.M., Irizarry, R.A. affy--analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 20:307–315 (2004)

27. Lambert, J., Ibrahim-Verbaas, C., Harold, D., Naj, A., Sims, R., Bellenguez, C., DeStefano, A., Amouyel, P.: Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nature Genetics 45:1452–8 (2013)
28. Wall, D., Pivovarov, R., Tong, M., Jung, J., Fusaro, V., DeLuca, T., Tonellato, P.: Genotator: a disease-agnostic tool for genetic annotation of disease. BMC medical genomics 3:50 (2010)
29. Walton, N., Shin, R., Tajinda, K., Heusner, C., Kogan, J., Miyake, S., Chen Q., Tamura, K., Matsumoto, M.: Adult Neurogenesis Transiently Generates Oxidative Stress. PLoS ONE 7(4) (2012)
30. Quinlan, J.: Induction of decision trees. Machine learning 1:81–106 (1986)
31. Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., Zhou, Z.H., Steinbach, M., Hand, D.J., Steinberg, D.: Top 10 algorithms in data mining. Knowl Inf Syst 14:1–37 (2008)
32. Friedman, N., Geiger D., Goldszmidt, M.: Bayesian networks classifiers. Machine Learning 29:131–163 (1997)
33. Needham, C.J., Bradford, J.R., Bulpitt, A.J., Westhead, D.R.: A primer on learning in Bayesian networks for computational biology. PLoS Comput Biol 3(8) (2007)
34. Geurts, P., Irrthum, A.L.: Wehenkel, Supervised learning with decision tree-based methods in computational and systems biology. Mol. Biosyst. 5(12):1593–1605 (2009)
35. Kurgan, L.A., Cios, K.J.: CAIM Discretization Algorithm. IEEE Transactions on knowledge and data engineering 16:145-153 (2004)
36. Fayyad, U., Irani, K.: Multi-interval discretization of continuous-valued attributes for classification learning. Thirteen. Int. Jt. Conf. Artificial Intell., vol. II, pp. 1022–1027 (1993)
37. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.: The WEKA Data Mining Software: an update. SIGKDD Explorations 11(1):10–18 (2009)
38. Waikato Environment for Knowledge Analysis (WEKA), http://www.cs.waikato.ac.nz/ml/weka/, last accesed 2018/07/10
39. Li, W., Tian, Y., Deng, Y.Y., Feng, X.L., Wang, Y., Feng, H., Hou, D.R.: Correlation between serum adiponectin level and cognitive function in patients with Alzheimer's disease. Nan Fang Yi Ke Da Xue Xue Bao 37(4):542–545 (2016)

ISSN 1870-4069

Improvement a Transcription Generated by an Automatic Speech Recognition System for Arabic Using a Collocation Extraction Approach

Heithem Amich, Mounir Zrigui

LATICE Laboratory, Research Department of Computer Science,
University of Monastir, Tunisia
heithem07@gmail.com, mounir.zrigui@fsm.rnu.tn

Abstract. The following study propose a novel heuristic to improve an automatic speech recognition system for Arabic language. Our heuristic relies on the collaboration of two approach: the first one ensures the extraction of collocations from a voluminous corpus then stores them in a database. It uses a combination of several classical measures to cover all aspects of a given corpus in order to exclude bigrams having a high probability of occurring together. The second one constructs a search space on the relations of semantic dependence of the output of a recognition system then, it applies phonetic filter so as to select the most probable hypothesis. To achieve this objective, different techniques are deployed, such as the word2vec or the language model RNNLM in addition to a phonetic pruning system. The obtained results showed that the proposed approach allowed improving the precision of the system.

Keywords: automatic speech recognition, multi-level improvement, collocation, semantic similarity, phonetic pruning.

1 Introduction

Automatic speech recognition has been growing interest in recent years. It aims to facilitate communication between people and system and allows to moving from an acoustic signal of speech to the transcription of the signal in a written version. Indeed, how does a transcription system work? From a recording, the system starts by calculating a transformation of the signal in acoustic parameters adapted to a recognition engine [1]. This latter makes use of acoustic and linguistic knowledge to produce the transcription [2]. The performances of the transcription systems are good when two critical elements are well mastered, the quality of the sound recording and the availability of recordings representative of the context of use. Although an ideal transcription system remains always nonexistent, several research efforts have recently been made to come up with robust systems [3]. Automatic speech processing still has a few defect. In fact, the main limitations that hinder the development of efficient systems are generally linked to the great deal of variability in speech. On this respect, we remind of the intra-speaker variability [4], due to the elocution (singing voice, shouting, whispering, hoarse, husky, under stress), inter speaker variability (male voice, female voice, or child voice) as well as the variability caused by the signal acquisition device (type of micro-

phone), or by the environment (noise, cross talk) [5]. Moreover, the degradation of performance is generally due to the lack of precise rules to formalize knowledge to different decoding levels (including, syntax, semantics, and pragmatics). On statistical methods with learning techniques from oral corpora where the correct transcription is known in advance. A statistical ASR is made up of several components following the acoustic and linguistic modeling of speech signal with a view to its recognition.

Many Techniques have been developed to improve each component of the system so as take account of or reduce the problems related to speech variability. Never the less, each technique has certain weaknesses. This leads us to develop an approach which takes account neither of the recognition modules adopted by an ASR, nor its search algorithms, or its smoothing techniques, which is the strong point of this approach. As a matter of fact, we considered the ASR as a black box device of any power of decision. Its role is limited to providing the transcription that will trigger our correction process. Finally, our approach in the only one responsible for correcting mis-recognized hypotheses and irrelevant word [6,7]. Also, if possible, it try to predict the next word that speaker probably will uttered. After a brief state of the art on the technique of improving transcriptions, we describe our first approach in section 3 and the precision improvement approach in section 4, we evoke the global steps of our idea. In section 5, we integrate the concept of colocation into our system. Finally, we discuss different evaluation results. In the last section, we discuss different evaluation results in section 6.

2 State of the Art

Improving the performance of ASR caught the attention of specialists in many languages. Many works were carried out to improve the competency of the various components of the system such as the linguistic and acoustic models and to significantly improve the decoding quality and the transcription quality a priori. In this framework, Lecouteux [8] presents a combinational method allowing to exploit a priori manual transcriptions and to integrate then directly into the heart of a SARP. This method allows to effectively guiding the recognition system with the help of auxiliary information. He also combined SRALs based on guided decoding [9]. With reference to previous research works, Benoıt Favre [10] proposed a fusion system between an original sentence containing an error and sentence of clarification. Thus, he proposed many alignments of levenshtein variants [11] and a reranker to select the best hypothesis. Antoine Laurent [12] came up with a method allowing to help the user in the step of correcting ASR outputs and to correctly transcribe proper names to facilitate the automatic indexing of transcribed reunions.

Fathi Bongares [13] studied the methods of combining transcription systems of large vocabulary speech. His study focuses a on the coupling of heterogeneous transcription systems with the aim of improving the transcription quality. Combining different transcription systems is based on the idea of exploiting the strengths of each system in order to obtain a final improved transcription. In order to overcome the essential problem of natural language processing that resides in the manipulation of large volumes of texts long Med Achraf presents a collocation extraction approach based on clustering technique. He used a combination of several classical measures which cover all aspects of

a given corpus in order to drew out the consecutive pairs (w_i,w_{i+1}) of a word commonly used from a voluminous corpus. Likewise, Christopher manning exposes a number of approaches to capturing collocations such as selection of collocation by frequency or the method based on the mean and variance of the distance in more than the t-test method and mutual information.

3 The Proposed Approach

In this section, we will present our system in details.

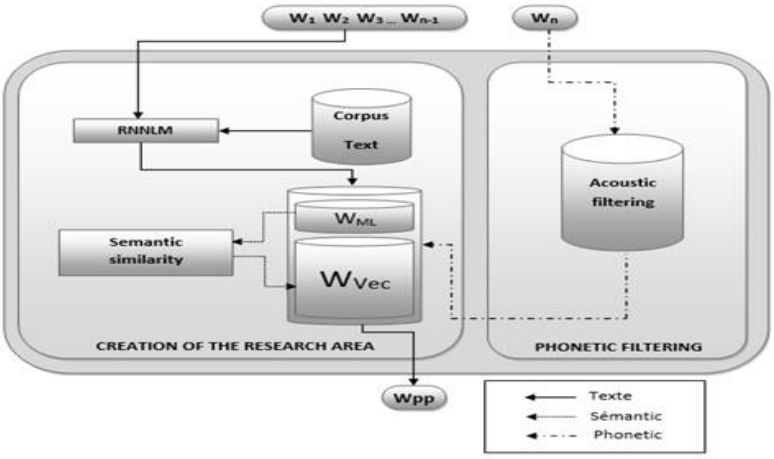


Fig. 1. The Verification and Correction System of Transcription (SyMAT).

The process of automatic correction of mis- spelt words from Arabic will be done in two main phases, as shown in figure 1. The steps of the left block scheme represent the first phase. It is particularly appropriate for extending the search space for the word to correct. The second stage is it at the right scheme. This phase is responsible for selecting the most likely word scheme.

3.1 Creation of Search Space

We expose to you the following case: the ASR has succeeded to transcribe the following word: w_0,w_1,…,w_{n-1} . By using our approach, we want to find the next word w_n badly recognized by the system ASR. The first step is to build a research space that may contain the word which we are seeking. This part is essential to develop the search space that will contain the words generated by the RNNLM language model and the semantic similarity.

Rnnlm. Let S=w_0,w_1,…,w_{n-1} be the context at a given instance our approach aims to estimate all of the most likely hypotheses w_n by using an RNNLM language model. This preliminary phase consists of passing the set of observations S to a language model in order to retrieve the set of the most likely words which could complete S. The

RNNLM model is based on the association of neural networks at word level. In what follows, we briefly remind of the mathematical strategies relevant to the model. Recently, deep neural networks have made a great success in the fields of image processing, acoustic modelling [13], language modelling [14,15], etc. Language models based on neural networks do better than standard back off n-gram models. Words are projected into low dimensional space similar words are grouped together. RNNLM could be a deep neural network LM due to its recurrent connection between input layer and hidden layer [16]. The network has an input layer x, a hidden layer S and an output layer y. We denote input to the network in time t as x (t) and output as $y(t)$. $S(t)$ refers to the state of the network (hidden layer). In put vector (x) is formed by concatenating vector $w(t)$ which represents current word. Output is made from neurons in context layer S at time $(t-1)$ [17]. The architecture of the neural network used to calculate conditional probabilities is organized in three layers [18]. The input layer reads a word $w(t-1)$ and a continuous $S(t-1)$. The hidden layer compresses the information of these two inputs and calculates a new representation S (t) for the input of the next propagation. The value is then passed on to the output layer, which provides the conditional probabilities P (w (t) $\mid w(t-1), S(t-1)$). RNNLM can be expressed as follows:

$$x(t) = w(t-1) + S(t-1), \tag{1}$$

$$S_j\ (t) = f\left(\sum_i\ U_i(t)U_{ij}\right), \tag{2}$$

$$y_k = g\left(\sum_i\ S_j\ (t)k_j\right), \tag{3}$$

where f (z) is a function of sigmoid activation:

$$f(z) = \frac{1}{1+e^{-z}}, \tag{4}$$

and g (z) is a softmax function:

$$g(z_m) = \frac{e^{zm}}{\sum K\ e^{zm}}. \tag{5}$$

Semantic Similarity. Identifying the similarity between words is an important TAL task regarding the domains where this technique could be useful, such as the search for information, automatic translation or even the automatic generation of text. The ability to correctly identify the semantic similarity between words is essential for our system. This is because of its contribution to the reconstruction of research space. The search for similarity is based on the word2vec techniques [19]. Word2vec is a neural network with two layers having as an input a text corpus and as an output a set of vectors representing the characteristics of the input word in this corpus. Word is then taken to measuring the cosines similarity where an angle of 0 degree expresses a total similarity, whereas an angle of 90 degrees expresses no similarity. The following table present a list of words associated with the word «July» rising word2vec, in order of proximity.

Table 1. A list of Words Associated with the Word "July= جويلية " using Word2vec

ASR	Cosine values
June (جوان)	0.9557317
April (افريل)	0.9386088
August (اوت)	0.9324805
March (مارس)	0.9314448
May (ماي)	0.9097166

Word2vec assigns a value equal to 0.6230781 to the word «France», so we deduce that France does not admit any semantic dependence with the word «July». The next step is to apply the text corpus learning and display the figure that shows the location of the words in a two dimensional space by a projection of the main component PCA, we notice that words with the same semantic meaning are adjacent. The figure below illustrates the locations of a set of words having the same semantic context.

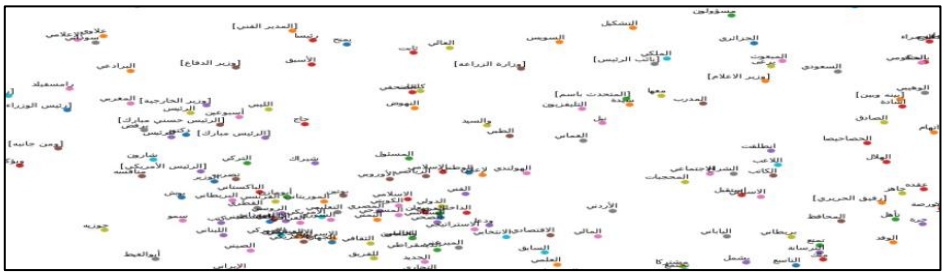


Fig. 2. The Distribution of Words According to the Cosine Value using PCA.

3.2 Selection of the Most Probable Word

Having collected a well-defined number of lexicons constituting the search space, we highlighted the techniques allowing filtering, classifying and finding the most appropriate hypothesis. We adopted two filtering methods: the syntactic filtering and the phonetic, filtering.

Phonetic Comparison. Having obtained a set of word W_{vec} +W_{ML}, we introduced another filtering mechanism operating at a phonetic level. This tool compares the frequency spectrum of the word W_n coming from an ASR and the frequency spectra of the word W_{vec} +W_{ML}. This method consists in aligning the signals of two words, then measuring the degree of similarity of two spectra. At the end of this phase, we estimate the word W_n having the most likely label and the highest degree of acoustic similarity. This example shows how to measure the similarities of signal. Whether they are correlated or not? The black and blue signals show the signals of two most likely words generated by search space. The third signal corresponds to the word signal generated by ASR. This figure shows that there is no phonetic similarity between the two candidates with the third signal. Just by looking at the time series, the signal seems not to

correspond to one of both models. A closer look reveals that the signals did different lengths and sample rates.

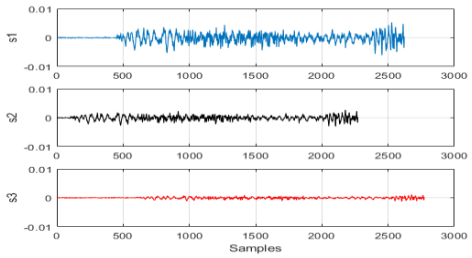


Fig3. Comparing the Similarity of Two Signals.

The Case of the First Word of the Sentence. Concerning the previous steps of our approach, we recalled the different phases of the automatic correction of transcriptions provided by an automatic speech recognition system. We elaborated architecture capable of sending back the next most likely hypothesis w_n after taking the n-1 hypotheses produced by an ASR as input: its worth mentioning that it is evident to find the words having indices between 2 and n given that there is data to manipulate. However, at the start of our procedure, we had w_0 data to activate our approach, so as to find the first word of the sentence. To overcome this limitation, we have partially changed our strategy. Indeed, we temporarily accepted the two most likely words generated by an ASR w_{11} and w_{12}. We remind that a speech recognition system uses these three pillars lexicon, the language model and the acoustic model to provide a text representing the transcription of a sound signal (the best one). It is also possible to retain several recognition hypotheses. The output world, then, be a list of best hypotheses N, a word graph or a confusion network. We limited ourselves to extracting the two most likely words among the retained N best hypotheses of an ASR of the first word of a sentence. This is simple due to the lack of data, which obliges us to accept w_{11} and w_{12}. However, the choice is not final. We have designed the method that reviews and verifies the first word of the sentence. The final result can accept w_{11} or rather w_{12} as well as a new lexicon retained by our approach based on a set of probabilities.

4 The Global Steps

In this section, we will present a detailed representation of our automatic correction system of the transcript provided from a speech recognition system. This procedure is carried out in 4 steps:

- The first step consists in extracting the two best hypotheses of first word of the sentence 1 from an ASR.
- Having acquired the two hypotheses W_{11} andW_{12}, we accept W_{11}. Then, we pass W_{21} to our search approach.

- It is essential to indicate the origin of the word. That is to say, if it is the result of the language model $W_{2\mathrm{M1}}$ or rather the result of word2vec $W_{2\mathrm{vec1}}$.

- Of the word comes from the language model, we pass W11 and $W_{2\mathrm{ML1}}$ to our approach in order to determine $W_{3\mathrm{ML1}}$ or $W_{3\mathrm{vec1}}$. Otherwise, shift back to by using an inverse language model choose either W_{11} or W_{12} or even another word proposed by the language model. This back shift is done only when the word, retrieved by our approach, comes from the tool word2vec. Needless to remind that we could also define a sort of in versed language model whose words were generated in a reverse order (from right to left):

$$P_{\mathrm{reversed}}\ \overrightarrow{(w)} \stackrel{\mathrm{def}}{=} \mathrm{P}(w_n)\ \mathrm{P}(w_{n-1}|w_n).\ \mathrm{P}(w_{n-2}|w_{n-1}w_n).$$
$$\mathrm{P}(w_{n-3}|w_{n-2}\ \mathrm{P}(w_{n-2}|w_{n-1}w_n).w_n).\ \ldots.\ \mathrm{P}(w_2|w_3w_4)\ \mathrm{P}(w_1|w_2w_3)$$

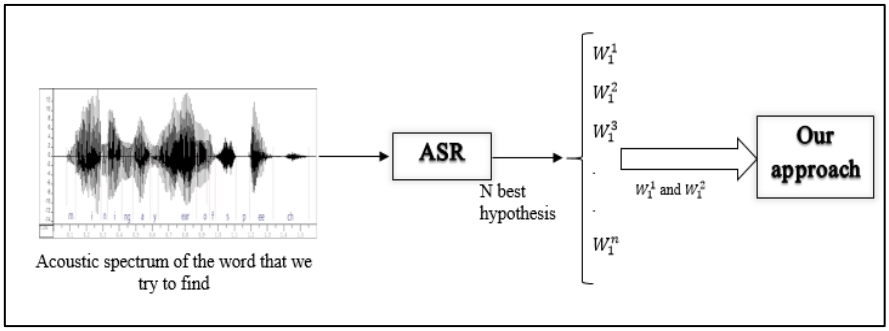


Fig. 4. First Phase of Our Approach.

Following each word generated by an ASR, it is susceptible to change the old word found by our approach during a back shift. The final choice is decided when we process the last word of the sentence, which can influence or substitute the previously executed hypotheses.

5 Improvement Precision

In order to increase the robustness and performance of our main system shown in Figure 1 and reduce its response time. We have added a new compartment called collocation. In this section, we will present in detail the process of extraction of collocations in the system as well as the integration steps of two approaches. Collocations refer to the most widespread pair of lexemes (l_i ,l_{i+1}) commonly used in the spontaneous Arabic language. They are necessarily consecutive whose existence of a lexeme li at position X_i in a corpus T certainly requires the presence of the lexeme l_{i+1} at the position X_{i+1}. A collocation is expression of two words that corresponding to some conventional method of saying things. There is considerable overlap between the concept of collocation and notion like term, technical term and terminological. Collocation are crucial for several

domain: natural language generation, computational lexicography and corpus linguistic research. It comprise:

- Proper names : الولايات المتحدة (United State)
- Verbal expression : (I saw the light) أبصر النور
- Terminologies : (Hello) السلام عليكم

5.1 Conventional Approaches for Extraction of Collocations

The t Test. If two words occur together many times, then we expect the two words to co-occur a lot just by chance. The t- test has been widely used for collocation discovery. It looks at the difference between the observant and expected means. If t is large enough the w1 and w2 are associated, we compute the t static:

$$t = \frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{N}}}, \tag{6}$$

where $\bar{X}$ is the sample mean, N is the sample size, μ is the mean distribution and S^2 is the sample variance [20].

Likelihood Ratio. is further method for hypothesis testing. In applying this test to collocation discovery, we have the ability to distinguish the occurrence of both common and rare phenomenon [20]. This method gives two hypotheses and test, which one is most probably the two hypotheses $H_1$and H_2 are:

- H_1: independence between w_1 and w_2: $p(w_2|w_1) = p(w_2|\neg w_1) = p$,
- H_2: dependence between w_1 and w_2: $p(w_2|w_1) = p_1 \neq p_2 = p(w_2|\neg w_1)$.

The likelihood ratio is:

$$\lambda = \frac{L(H_1)}{L(H_2)}, \tag{7}$$

where L is the likelihood function, assuming a binominal distribution L is given by:

$$L(p; n, r) = r^p (1 - r)^{n-p}, \tag{8}$$

where n is the number of trials, r the number of successes, and p is the probability of success.

Mutual Information. is a measure of how much one tell us about the other. It allows to compare the probability of observing w_1 and w_2 independently $p(w_1)\ p(w_2)$ mutual information is calculated by:

$$I(w_1|w_2) = \log_2 \frac{p(w_1|w_2)}{p(w_1)\ p(w_2)}. \tag{9}$$

If mutual information is large then w_1 and w_2 are related else, it is too low then w_1 and w_2 are independent [20].

5.2 The Steps of Extraction of Collocations

To extract all the most common collocations of the arable language, we have combined the three methods recently mentioned, called the t-test, the Likelihood ratio and the mutual information. Thus for each candidate of the collocation w_1 w_2, these thee measures will be used to calculate the dependency between w_1 and w_2. Then, we calculate the average value of the three measures for each bigram. We consider a collocation all bigrams corpus having a mean higher than a very high empirical threshold. The preliminary step consists to segmenting the corpus by identifying the basic units forming the corpus. This means identifying the separators used to isolate the morphemes. We also define a stop list to omit the words that cannot form a collocation as:

- The particles of coordination: (ثم، أم، أو، أما، إما).
- The interrogative particles : (أيّ, كيف, أين, متى).
- The particles of Appeal: (يا، أيا ،أيها هيا).

Once the bigrams have been identified, the next step is the calculation, for each bigram, we calculate the average of three measures mentioned previously. If the value found is greater than a threshold, then the bigrams is considered collocation and we add it to the list of collocations.

Notations used are summarized in the following:

- T: Corpus size.
- L_i : lexeme i, 1≤ i ≤ T .
- B_i : Bigram i.
- SL : Stop List.
- E_i : a real which designates the calculated average of each bigrams

```
1. //Bigrams extraction and Measures computation
2. For all lexemes l_i, 1 ≤ i ≤ T − 1 Do
3. B_j = { l_i, l_{i+1} / l_i ∉ SL ∧ l_{i+1} ∉ SL } End.
4. //calculate the average of each bigram
5. Ei = average( Mutual inf (Bi), t test (Bi), Likehood.ratio(Bi)
6. // add the bigram to all the collocations If (E_i > thrushold) {
7. Col = Col ∪ { w_i, w_{i+1} } }. End
```

Figure 6 illustrates some of the accumulated collocations in our database collocation [20].

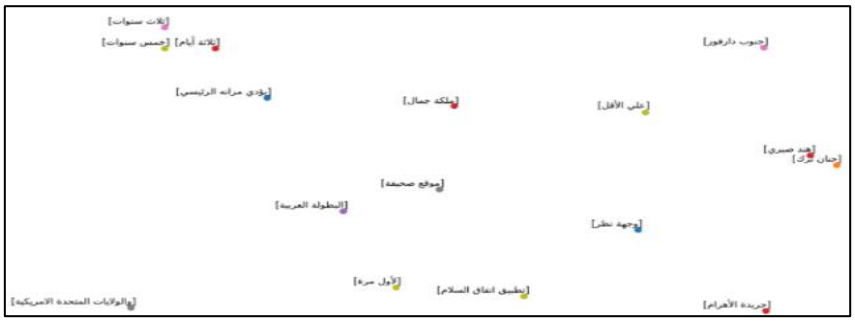


Fig. 6. Colocation Group in a Two-Dimensional Space.

5.3 The Integration of Collocations into the Main System

At this level, we have completed the construction of a collocation base. However, the obvious question is about the contribution of integrating the concept of collocations into our system?

Let S=w_0,w_1,…,w_{n-1} be the context at a given instance. S represents the words pronounced by the speaker. At this point, our system has completed the verification of the whole sequence in order word after word with success. Let w_n be the word that will be treated. If the word w_{n-1} does not belong to the stop list, our heuristic checks if the previous word w_{n-1} is part of one of the collocations previously collected. We recall that a colocation is composed of two lexemes (l_i,l_{i+1}). If w_{n-1} exists in the collocation base (w_{n-1}=l_i), then it is sufficient to apply the acoustic comparison of w_n with the second lexeme l_{i+1}.

That means that the steps for creating the search space provided by the RNNLM language model and word2vec be canceled, the general heuristic has two paths, if the last word processed by the system is part of the collocations, then we just perform the acoustic test. If this test is positive then this is the word to look for. If not we execute as usual our initial approach (SyMAT). The integration of the colocation approach into the SyMAT system is very beneficial to the level of confidentiality and accuracy of the final result.

Indeed, if the word belongs to the list of collocations stored, and the acoustic test established is positive. Doubtless, we are confident that this is the exact word uttered by the speaker.

6 Experimentation

To construct the language model, we have used an Arabic text corpus of 100M words collected from corpus available on the used. This same corpus served to the construction of the model based on label. As for the testing of our system, we recorded a caustic corpus of 40 hours. We set up our SyMAT system at the exit of two known SPAP namely Sphinx [21].

Table 2. Results of the system.

ASR	Precision	F-mesure
Sphinx	51,38	56,41
Sphinx + SyMAT	56,52	62,05
HTK	46,24	50,77
HTK + SyMAT	52,72	57,88

The obtained results show that the proposed approach effectively contributed to improving ASR. We may also note that our method is more efficient for the HTK system than for the Sphinx system. This is justified by:

- The high clean error rate of the HTK system as compared to the sphinx system [21].
- The acoustic models trained by Sphinx were much better than that of HTK.

Table 3. Samples of Collocation Candidates.

$w_1 w_2$	M.I	T.T.	L.R.	E_i
مليون دولار (Million dollars)	0.9999	0.9999	0.9999	0.9999
أشراط الساعة (Signs of the Hour)	0.9987	0.9750	0.8774	0.9503
الطبعة الاولى (First Edition)	0.6487	0.2548	0.3458	0.4164
اتفاق السلام (Peace Agreement)	0.7814	0.7895	0.8569	0.809

The obtained results show that if the sum of the three values exceeds a threshold equal to 0.8, then the bigram is considered collocation.

7 Conclusion

On this paper, we propose heuristics with the aim of improving the transcription generated by an ASR for Arabic. This method exploits semantic, phonetic levels and collocation's concept in order to evaluate the output of the ASR system and to propose the most likely word in case there is an error. To enforce this approach, we resorted to the techniques of word similarity, t test, mutual information, likelihood ration and to the RNNLM language model to establish a search space based on the history of a transcription $W_1...W_{n-1}$. After that, we carried out a phonetic pruning to choose the most probable word. We also resorted to the techniques of t test, mutual information and likelihood ration to extract collocations in order to increase the exactitude of final result. As a future work, we hope to promote our system from a model allowing taking account of the historic of applied corrections and assuring adaptation of the correction process to a particular user.

References

1. Dua, M., Aggarwal, R.K., Virender, K., Dua, S.: Punjabi automatic speech recognition using htk (2012)
2. Aggarwal, R.K., Dave, M.: Acoustic modeling problem for automatic speech recognition system: advances and refinements (part II). I. J. Speech Technology 14(4):309 (2011)
3. Zolnay, A., Schlter, R., Ney, H.: Robust speech recognition using a voiced-unvoiced feature. In: IN PROC
4. Siegler, M.A., Stern, R.M.: On the effects of speech rate in large vocabulary speech recognition systems. In: 1995 International Conference on Acoustics, Speech and Signal Processing, ICASSP '95. Detroit, Michigan, USA, May 08-12, 1995, pp. 612–615 (1995)
5. Zhao, S.Y., Ravuri, S.V., Morgan, N.: Multi-stream to many-stream: using spectro-temporal features for ASR. In: INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association. Brighton, United Kingdom, September 6-10, 2009, pp. 2951–2954 (2009)
6. Rohit, P, Rohit, K, Sankaranarayanan, A, Chen, W., Hewavitharana, S., Roy, M.E., Choi, F., Challenner, A., Kan, E., Neelakantan, A., Natarajan, P.: Active error detection and resolution for speech-to-speech translation. In: 2012 International Workshop on Spoken Language Translation, IWSLT 2012. Hong Kong, December 6-7, 2012, pp. 150–157 (2012)
7. Alex, M., Tom, K., Mari, O., Luke, S.: Using syntactic and confusion network structure for out of vocabulary word detection. In: 2012 IEEE Spoken Language Technology Workshop (SLT). Miami, FL, USA, December 2-5, 2012, pp. 159–164 (2012)
8. Lecouteux, B., Linares, G., Oger, S.: Integrating imperfect transcripts into speech recognition systems for building highquality corpora. Computer Speech & Language 26(2):67–89 (2012)
9. Lecouteux, B., Nocera, P., Linares, G.: Semantic cache model driven speech recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2010, 14-19 March 2010. Sheraton Dallas Hotel, Dallas, Texas, USA, pp. 4386–4389 (2010)
10. Favre, B., Rouvier, M., Béchet, F.: Reranked aligners for interactive transcript correction. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014. Florence, Italy, May 4-9, 2014, pp. 146–150 (2014)
11. Tor, H, Torleiv, K, Levenshtein, V.I.: Error correction capability of binary linear codes. IEEE Trans. Information Theory 51(4):1408–1423 (2005)
12. Antoine, L., Sylvain, M., Téva, M., Paul, D.: Computer-assisted transcription of speech based on confusion network reordering. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2011, May 22-27, 2011. Prague Congress Center, Prague, Czech Republic, pp. 4884–4887 (2011)
13. Bougares, F., Esteve, Y., Deléglise, P., Linares, G.: Bag of n-gram driven decoding for LVCSR system harnessing. In: 2011 IEEE Workshop on Automatic Speech Recognition & Understanding, ASRU 2011. Waikoloa, HI, USA, December 11-15, 2011, pp. 278–282 (2011)
14. Dahl, G.E., Yu, D., Deng, L., Acero, A.: Context dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans. Audio, Speech & Language Processing 20(1):30–42 (2012)
15. Ebru, A., Tara, N., Brian, K., Bhuvana, R.: Deep neural network language models. In: Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT, WLM '12, pp. 20–28. Stroudsburg, PA, USA, 2012. Association for Computational Linguistics (2012)
16. Tomas, M., Martin, K., Lukas, B., Jan, C., Sanjeev, K.: Recurrent neural network based language model. In: INTERSPEECH 2010, 11th Annual Conference of the International

Speech Communication Association. Makuhari, Chiba, Japan, September 26-30, 2010, pp. 1045–1048 (2010)
17. Jeffrey, P., Richard, S., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014. Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, pp. 1532–1543 (2014)
18. Manning, C., Hinrich, S.: Foundations of statistical natural language processing. Cambridge, Mass. MIT Press (1999)
19. Frédéric, B., Benoit, F.: ASR error segment localization for spoken recovery strategy. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013. Vancouver, BC, Canada, May 26-31, 2013, pp. 6837–6841 (2013)
20. Ben Mohamed, M.A., Mallat, S., Nahdi, M.A., Zrigui, M.: Exploring the potential of schemes in building NLP tools for arabic language. Int. Arab J. Inf. Technol. 12(6):566–573 (2015)
21. Satori, H., Hiyassat, H., Harti, M., Chenfour, N.: Investigation Arabic Speech Recognition Using CMU Sphinx System. The International Arab Journal of Information Technology, Vol. 6, No. 2, April (2009)

ISSN 1870-4069

SecuredDW: A Homomorphic Schema to Securely Hosting Data Warehouse in the Cloud

Kawthar Karkouda, Ahlem Nabli, Faiez Gargouri

Sfax University, Miracl Laboratory, Tunisia

Abstract. Currently, cloud computing has become the most popular technologies in the area of IT enterprise. It has many advantages such as computing power, storage, network and software as a service. Moreover, many other benefits have made it attractive. In fact, it is easy to deploy, its technical infrastructure is adaptable to the volume of business activity and its cost is relative to consumption. Whereas building a data warehouse typically necessitates an important initial investment, with the cloud pay-as-you-go paradigm, BI system can benefit from this new technology. However, cloud computing brings its own risks in terms of security. For this purpose, before outsourcing sensitive data to the cloud, the owner must encrypt his data to keep secure. In the particular context of cloud data warehouse, privacy is of critical importance because it contains sensitive data. Cloud provider proposes traditional security solutions to ensure the confidentiality of outsourced data. Unfortunately, those solutions are not practical in the case of data warehouse anymore because they induce a heavy overhead in terms of data storage and query performance. So, a new solution must be proposed for outsourcing data warehouse to the cloud that respects its specification and swings performance and security. In this paper, we propose (SecuredDW) a new sharing schema for securing and querying a data warehouse hosted in the cloud based in a homomorphic algorithm. The integrity of data is also addressed in this paper by proposing two new signatures to verify the correctness of data sent and received from the cloud. Theoretical results show the efficiency of Secured DW in terms of privacy and performance with respect to other solutions.

Keywords: cloud computing, data warehouse, Security, integrity.

1 Introduction

With the booming of cloud computing, people are encouraged to adapt BI system in their companies. Such new delivery model can mitigate the cost of deployment of a data warehouse thanks to its "pay as you go" paradigm. So, company pays just the used resource. It is true that the most attractive advantage of using the cloud is its profitable cost, but also there are many more. Indeed, it is easy to deploy and its technical infrastructure is adaptable to the volume of business activity. The only drawback that prevents the move to the cloud is security. In fact, there are several security issues related to cloud computing. Some of those issues emanated from the traditional

architecture such as network attack, confidentiality of data, availability, authentication and vulnerability exploitation. Because cloud computing evolves rapidly and the push to effective controls to protect data in the cloud is nascent, many security solutions for clouds are presented in the literature. The most used solution is the encryption of data before sending it to the cloud with the symmetric and the asymmetric algorithms. Homomorphic encryption is also used to secure data hosted in the cloud.

In the context of data warehouse, these security problems become tougher to resolve because the high volume of data stocked in the warehouse and because the nature of OLAP query. More precisely, encrypting data warehouse can affect the cost of using the cloud especially in the case of homomorphic encryption that produces a very high volume overhead. Furthermore, symmetric and asymmetric algorithm cannot be a suitable solution for data warehouse because the decryption of data in the cloud can affect the performance of OLAP query and the cost of using the cloud. More than that, such scenario is based on the trust between the owner and the cloud provider, which is not the case.

For this reason, in this paper, we propose SecuredDW as a new sharing schema adapted to the nature of data warehouse. Our proposal is based on the homomorphic privacy presented in [1]. One serious deficiency of this homomorphic privacy is the possibility of being broken by clear text attacks. Thus, our contribution is to make this privacy homomorphism more robust and secure by using data splitting, multi-cloud and perturbation value.

In addition to that, in this work, we enforce data integrity by providing two signatures to verify the correctness of data.

Moreover, when data is encrypted the original order of data is broken. Thus, all fetched data must be decrypted and querying at the owner by the trust tier before to be sent to the client. This operation can affect the performance of range query and some others query when ordering is necessary. For that we will propose a weighted method that reduces the time complexity of such kind of query.

According to my knowledge, this work is one of a few work that provide an environment that takes into account the specifications of a data warehouse while balancing performance and security. It should be noted that it is not our aim to propose a solution as secure as the state-of-the-art encryption algorithms. We rather suggest a technique that provides a considerable level of overall security strength with respect to some performance overheads.

The rest of this article is organized as follows. The second section introduces and discusses the previous research related to our proposal. Then, we present SecuredDW as a new homomorphic schema to securely host data warehouse in the cloud. In the fourth part, we deal with the theoretical and performing results. Finally, the paper ends with a conclusion.

2 Related Works

As encryption is the most used solution to secure data outsourcing to the cloud, we will start by introducing some traditional encryption algorithms that are used in the context

of the cloud. In fact, symmetric encryption is mainly the use of the same key in encryption and decryption. That is to say that he who encrypts the data must share the key with the receiver who decrypts the data. The two most important modern symmetric algorithms are the data encryption standard DES [2] and the advanced encryption standard AES [3]. In opposite to symmetric encryption, asymmetric encryption is by definition the use of two different keys for encryption and decryption. RSA [4], Rabin [5] and ElGamel [6] are the most practical asymmetric algorithms. However, the main threat of using those algorithms is that they are based on trust between the owner and the cloud provider, which is not the case because the cloud provider may not be trustworthy and can fetch into the sensitive data. Although the data and the keys are stocked in the same cloud provider, this scenario can make data subject to external attacks. So, if intruders break the security system of the provider, they can steal the data with the keys and decrypt it easily. The inefficiency of those security techniques is not only in terms of privacy but also in terms of performance. Indeed, decrypting data before processing query is not practical mostly in the case of data warehouse.

To overcome those problems, many works in literature propose running data in ciphertext in the cloud. In this context, homomorphic encryption is presented. Authors in [7,8,9,10,11,12] propose a solution based on a fully homomorphic encryption. The advantage of those algorithms is that they allow addition and multiplication of encrypted data in ciphertext. However, they suffer from high time complexity and high volume overhead.

To perform computations over attributes that are used in the calculation of max and min aggregation functions or attributes that are compared using relational operators, order preserving encryption [13,14] and multivalued Order preserving encryption MV-OPE [15] are proposed.

To apply the power of running data in ciphertext in the cloud, authors in [16] propose to encrypt data warehouse with several encryption techniques depending on the type of attributes. This way, analytical queries can be processed in ciphertext. Yet, this solution suffers from high time complexity of running query and high volume overhead.

The availability of data is also a challenge when entered in the cloud. For this reason, the cloud provider replicates data to ensure its availability. Another solution presented consists in using erasing codes. The advantage of such solution is that it can reduce the volume of replicated data. Facebook, for example, is using this algorithm to ensure the availability of its data warehouse with minimum volume overhead [17].

Multi clouds, cloud of cloud, or inter cloud, are by definition the use of many cloud providers for data storage such as DSky [18], inercloud [19], and NCloud [20]. Authors in [21] use erasure coding to divide the data and stock it in different cloud providers to ensure its availability and to reduce the volume of data. Their approach seems to be good in term of availability and in term of reduction of data volume, but it is not secure.

Authors in [22] present CHARM a multi cloud schema that guarantees the availability of data with minimum cost.

The secret sharing algorithm, when first presented in [23], is very useful in the cloud to ensure confidentiality and availability as in works [24,25,26,27]. The problem with using the secret sharing is the high volume overhead generated after encryption. That's why, authors in [28] try to solve this problem by proposing a new model for sharing

data warehouse inspired from secret sharing. The idea is to split the data into block before encrypting it with a random linear equation. However, this approach suffers from high time complexity of decryption steps. Another problem arises when using this approach is that it cannot resist to collusion attacks.

To work out such a problem, authors in [29] propose S4 as a new schema based on secret sharing for enforcing privacy in Cloud data warehouse. The idea is to store secrets at one single CSP instead of sharing secrets to n CSP's. The privacy in S4 relies on the fact that k-1 splits are stocked in the CSP and the K^{th} splits necessary for reconstructing the secret are stocked in the owner. This way, they can avoid the problem of collusion, but the processing of the query cannot be done totally in the cloud.

Authors in [30] propose HORNS as a sharing schema based on the Residue Number System (RNS) and multi-cloud. The idea in HORNS is to divide the data in small chunks with the RNS and stock those chunks in a multi-cloud. In this concern, time complexity of encryption and decryption steps is reduced. But this scheme suffers from redundant data and collusion attacks.

2.1 Discussion

As presented in the last section, many works try to solve the problem of security when using the cloud. Nowadays, the most appropriate solution is to use symmetric and asymmetric encryption algorithm to encrypt data before sending it to the cloud. Those solutions are based on the trust between the user and the provider. For this reason, they are not secure enough because the provider can fetch into sensitive data. That's why, processing data in ciphertext is improved and new solutions based on homomorphic encryption algorithm, secret sharing, Information Dispersal Algorithm IDA are proposed in the case of database in general and in the case of data warehouses in particular. The problem with those solutions is that they are not practical enough, mainly in the case of data warehouse because it stocked a high volume of data and because of the time complexity of an OLAP query. For example, many homomorphic encryption algorithms are proposed in the literature as described in the previous section. But those algorithms are not a good solution for outsourcing data in the cloud because of the high time complexity of processing data and because of the volume overhead generated after encryption data. So, the famous homomorphic encryption algorithms existing in the literature cannot meet the need for a heavy computing application like the data warehouse. Some authors propose to use the secret sharing for outsourcing data to the cloud. Their choice is based on secret sharing since it has an acceptable time complexity comparing with the homomorphic encryption algorithm. But the problem with adopting this technique in data warehouse is that it generates a high volume overhead. Accordingly, authors in [25] propose a new secret sharing that reduces the volume overhead generated when encrypting data. But this solution suffers from the high time complexity when decrypting data. Information dispersal algorithm (IDA) is also proposed as a solution to outsource data in the cloud. It is known for its low time complexity of encrypt and decrypt data and its low volume overhead. However, this algorithm suffers from its weak security. Therefore, in this paper, we propose a new schema that balances security and performance when outsourcing data warehouse in

the cloud. Our schema is based on the simple privacy homomorphism as described in [1]. The privacy homomorphism will be illustrated as in [1]:

Let p and q be two large secret primes and m= pq the product of such large secret primes. For that m is difficult to factor.

Consider the set of cleartext data T= Z_m, and the set of cleartext operation F= $\{+_m, -_m, \times_m\}$ consisting respectively of the addition, substraction and multiplication module m , with m=pq .

Let the ciphertext data set be $T' = Z_P \times Z_q$. Ciphertext operation F' is the component wise of these in F.

Define the encryption function $\phi(x) = [x \bmod p, x \bmod q]$. Given the two prime numbers p and q and the ciphertext $x_p = x \bmod p$ and the ciphertext $x_q = x \bmod q$, the secret x is decrypted using the Chinese Remainder Theorem (CRT).

We are motivated to use this privacy homomorphism because the latter is based on the modular arithmetic as it is described in the encryption function $\phi(x)$. This is very interesting with regards to volume overhead. In fact, because the data will be divided in a small residue number, the storage space will be reduced. But when it comes to confidentiality, the data will be encrypted with the two prime numbers p and q and will be computed in the range of m = pq. The decryption function of this schema is based on the use of Chinese remainder theorem. This technique is very practical and feasible because of its reasonable temporal complexity. Thanks to the homomorphic characteristic of encryption function, arithmetic operations can be done in a ciphertext.

So, the simple scenario is to encrypt data stocked in the data warehouse with the encryption function $\phi(x) = [x \bmod p, x \bmod q]$. After that, the cipher text data $x_p = x \bmod p$ and the cipher text data $x_q = x \bmod q$ will be sent to the cloud provider with the module m. The two prime numbers p and q will be kept secret in the owner. The data stocked in the cloud will be processed modulo m. So, in this way, the cloud provider cannot decrypt the data with the modulo m because it is hard to factor. Then, the data will be securely stocked in the cloud. Furthermore, with the homomorphic characteristic of modular arithmetic query using arithmetic operation such that {+ ,-, x} will be done in the cloud in a cipher text without decryption. After processing the query in the cloud, the provider sends the result to the owner in a ciphertext. The owner decrypts his data with the two secret prime numbers p and q and the two chunks of encrypted data are received from the cloud using Chinese remainder theorem (CRT).

Unfortunately, this schema can be broken by the cloud provider because it has the two chunks of data and the secret module m. It can infer the two chunks of data and get the two secret parameters p and q. Malicious intruders can also break the security parameters of the cloud provider, get the encrypted data and the modulo m from the cloud provider and decrypt it using the known cleartext attack as described in [18].

There are two factors that threaten the confidentiality of this schema, an internal factor being the cloud provider itself and an external factor being a malicious intruder. Thus, a new way will be suggested in the second section which can reduce the risk of breaking the security parameters of our schema using a multi-cloud and perturbed data.

3 SecuredDW: A New Schema for Securing and Querying Data Warehouse Hosted in the Cloud

This section presents SecuredDW as a new homomorphic schema for hosting and querying data warehouse in the cloud securely. In this schema, we focus on ensuring the three levels of security CIA (confidentiality, integrity and availability) as it is described in figure 1.

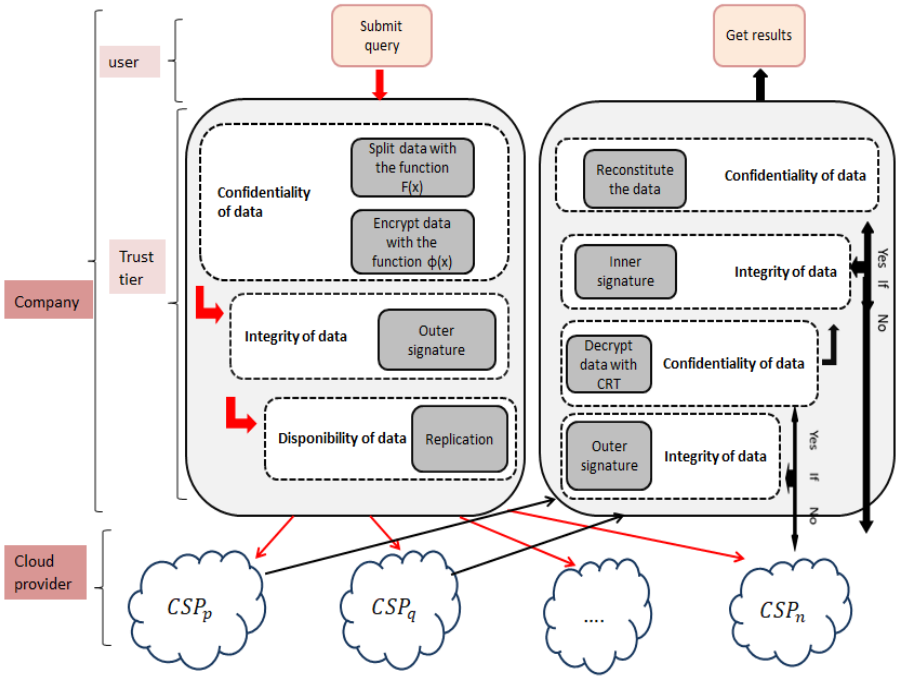


Fig. 1. Global Architecture of SecuredDW.

3.1 Enforcing the Confidentiality of Data

Authors in [28] describe the way how a cryptanalyst can infer the data and get the secret. They argue that this homomorphic privacy presented in [1] can be broken by a known plaintext attack. They illustrate the process of breaking the system as follows:

Suppose x is the integer that will be encrypted and presented by a pair (x_p ,x_q), where $x_p = x \bmod p$ and $x_q = x \bmod q$. Assume that the cryptanalyst has the plaintext, ciphertext pair for some data. They suppose that p' be the gcd$\{x_p - x$ for all data$\}$. In the same way, they suppose that q' be the gcd$\{x_q - x$ for all data$\}$. After that, it tests that $p=p'$ and $q= q'$and if this is the case, the cryptanalyst can decrypt all ciphertext. They prove that when ciphertext (x_p ,x_q) is specifically given, the cryptanalyst can find x' such as, $x' \equiv x_p \bmod p'$ and $x' \equiv x_q \bmod q'$.

So, it is clear that if the two chunks of data (x_p ,x_q) or one of them is kept secret from the cryptanalyst, the probability of inferring the data and breaking the system with the known plaintext attack will be reduced.

Therefore, as a first modification, we propose to split the secret data into a k small chunk with a random function F(x) like this:

$$F(x) = \sum_{i=1}^{k} x_i \quad \text{with} \quad x_i \in Z_m. \tag{1}$$

After splitting the original data with the random function F(x), we encrypt each chunk of data with the homomorphic function ϕ(x) :

$$\phi(x) = [\, x \bmod p \,,\, x \bmod q \,].$$

Splitting data and encrypting each chunk of data separately with the homomorphic function ϕ(x) is not enough to secure sensitive data. To achieve this, we propose to stock each share of secret data in a different cloud provider. In this way, we can reduce the probability of inferring data and breaking the encryption function because each provider has only one chunk of the secret data. So, the problem of intern risk will be decreased. Likewise, it is difficult for malicious users to break the security parameter of two cloud providers at the same time and get the chunks of secret data necessary to reconstruct the original data. Hence, the risk of breaking the system by an external intruder will be diminished. This way, our model will be more secure in terms of confidentiality towards the cloud providers as well as from external attack.

Besides, when p, q, $m = pq$ are very large integers, a small value x is very likely to have the same representation over Z_m, Z_p, and Z_q, that is $x \bmod m = x \bmod p = x \bmod q$ if $x < \min(p, q)$. This is an undesirable feature because the homomorphic function $\phi(x)$ leaves the cleartext unencrypted (trivial ciphertext). To overcome this drawback, we propose to multiply secret data with two secret values r_p and r_q such that $r_p < p$ and $r_q < q$.

So our new homomorphic encryption function will be:

$$\begin{aligned} \phi(x) = (\, & [x_1 \times r_p \bmod p,\, x_1 \times r_q \bmod q\,], \\ & [x_2 \times r_p \bmod p,\, x_2 \times r_q \bmod q\,], \quad \text{with} \quad k \geq 2 \\ & \ldots\ldots \\ & [x_k \times r_p \bmod p,\, x_k \times r_q \bmod q\,]\,). \end{aligned} \tag{2}$$

After encrypting data with the homomorphic function ϕ(x), k pair of data will be produced (x_{kp}, x_{kq}) with $x_{kp} = x_k \times r_p \bmod p$ and $x_{kq} = x_k \times r_q \bmod q$. After that, the chunks of data $(x_{1p}, x_{2p,\ldots} x_{kp})$ will be sent in the CSP_p and $(x_{1q}, x_{2q,\ldots} x_{kq})$ will be sent in the CSP_q.

3.2 New Homomorphic Integrity Function

To ensure the integrity of data, we introduce in this paper new signatures named outer signature to verify the integrity of share and inner signature to verify the integrity of original data.

Outer signature

Outer signature is a new homomorphic function based on the modular approach.

For $c = x \bmod n$ is equivalent to $x-k= n*c$, k is the rest of division.

So the signature of each share designed " Sgn_x" will be computed with the homomorphic function as:

$$H_s(x) = x \bmod n = \text{Sgn\_x}, \tag{3}$$

and we will compute the key of each signature as:

$$key\_x = x - (Sgn\_x \times n). \tag{4}$$

In our case, to verify the integrity of our chunks of data, we will use the two equations (3) and (4). After that, each pair of data will be sent to a CSP.

In the cloud, each provider will verify the correctness of its share by computing:

$$Sgn\_x = share \bmod n,$$

and it verifies that the key received is:

$$key\_x = share - (Sgn\_x \times n).$$

Similarly, when the owner receives data from the CSP's, he can verify his data with the two equations (3) and (4).

The main advantage of our integrity function is in its homomorphic characteristic. This is very useful in the case of data warehouse when using the aggregation function.

Inner signature

To enhance the integrity of our schema we propose an Inner signature. The latter is a self-checked signature based on the homomorphic propriety of modular approach. It works by computing the equivalence between the chunks of data generated after splitting the original data with the random function F(x) and the share of data received from the cloud. Algorithm 1 is used to verify the integrity of the original data in the owner after decrypting the share of data x_{1p}, x_{1q}, x_{2p}, x_{2q},... x_{kp} ,$x_{kq}$with the CRT and getting the original chunk of data x_1, x_2,… x_k.

Algorithm 1

Inner signature $(x_1, x_2, x_k, x_{1p}, x_{2p}, x_{kp}, x_{1q}, x_{2q}, x_{kq})$

{

$S_p = |x_{1p} + x_{2p} + \ldots + x_{kp}|_p$

$S_q = |x_{1q} + x_{2q} + \ldots + x_{kq}|_q$

$Chunk_p = |x_1 + x_2 + \ldots + x_k|_p$

$Chunk_q = |x_1 + x_2 + \ldots + x_k|_q$

```
If ((S_p=Chunk_p) and (S_q=Chunk_q ))
 Than
     Write ("Data is correct ")
 Else
     Write ("Data is not correct ")
}
```

After computing the inner signature, whether it is correct that the trust tier reconstitutes the original data X with the function F(x), or else it asks the cloud provider to get other share of data.

3.3 Data Availability

In this section, we want to show the robustness of our solution if we use data replication as solution to ensure the availability of data. In this concern, we propose to replicate each chunk of data three times. So, nine chunks of data will be produced if k =3 is the number of data splitting. After that, the chunks of data (x_{1p},$x_{2p,...}x_{kp}$) will be sent to the CSP_p, (x_{1q},$x_{2q,...}x_{kq}$) will be sent in the CSP_q and the other replicated chunks will be sent to the two other clouds . This way, we can guarantee the availability of data if one or two cloud providers are not available. Moreover, we can eliminate the dependency of a single cloud provider.

With our method, data overhead cannot exceed the volume of original data warehouse twice. So, if we replicate the entire data warehouse three times, the volume overhead generated will be six times the volume of the original data warehouse. This is not very practical in the case of data warehouse, but it does not exceed the volume overhead generated with the other solutions.

4 Sharing Data Warehouse

In this section we will demonstrate how data warehouse will be shared in the cloud with our schema. Our new sharing model is based on two initial steps. The first step is a data sharing process, and the second step is a data reconstruction process. We will also delineate how the query will be processed and finally we will present a new method to process the range query in ciphertext.

4.1 Data Sharing Process

The data sharing process describes how original data will be processed before being sent to the clouds. It consists of the following steps:

- The trust tier person who is responsible for data security in the company proposes two secret prime numbers p and q and two secret values r_p and r_q, such that $r_p < p$ and $r_q < q$. He also calculates the module $m = pq$.

- He chooses the parameter k which is the number of chunks of data generated after splitting the original data.

- He affects each prime number p and q to a specific cloud provider. This is very important for maintaining the coherence of secret data.

- He splits the original data with the random function F(x) presented in equation (1).

- After that, he encrypts the data with the homomorphic function $\phi(x)$ presented in equation (2).

- The trust tier computes the signature of each chunk of data generated after encryption with the new homomorphic integrity function Hs presented in equation (3) and the keys of each signature with equation (4).

- Finally, the trust tier replicates all the chunk of data with its signature and its key and sends them to the corresponding cloud provider.

The scenario of data sharing process is presented in figure 2:

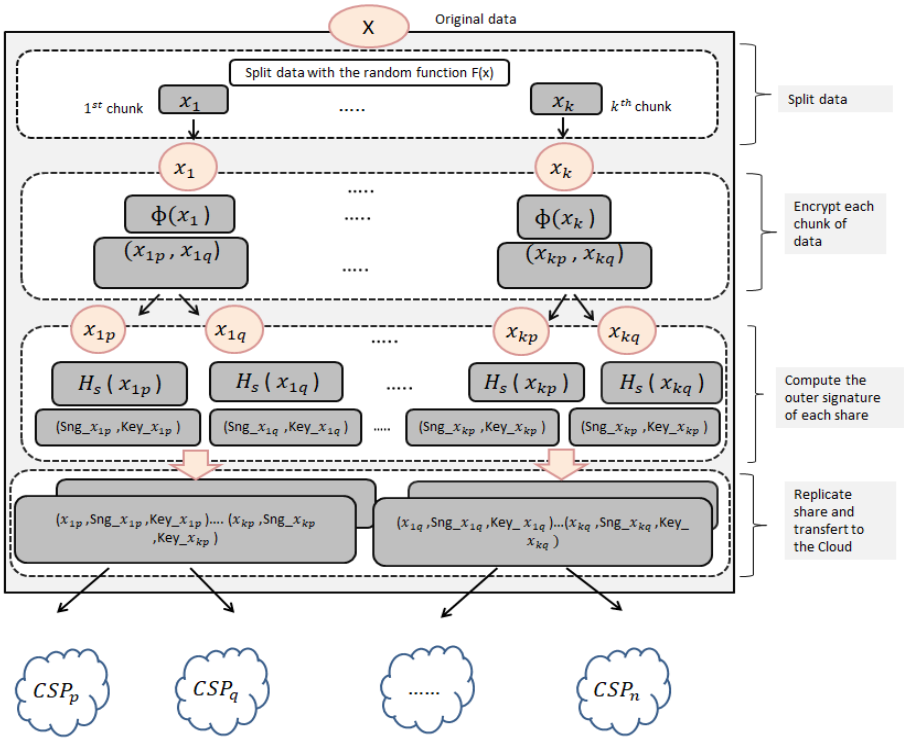


Fig. 2. Data sharing process.

4.2 Data Reconstruction Process

The data reconstruction process describes how original data will be reconstituted after being received from the clouds. It consists of the following steps:

-The trust tier asks two cloud providers CSP_p and CSP_q to get the shares of corresponding data X (x_{1p}, Sgn_x_{1p}, key_x_{1p}) ,(x_{2p} , Sgn_x_{2p} , key_x_{2p}) and(x_{kp}, Sgn_x_{kp},

key_x$_{kp}$) from CSP$_P$, (x_{1q} , *Sgn_x*$_{1q}$, *key_x*$_{1q}$) , (x_{2q} , *Sgn_x*$_{2q}$, *key_x*$_{2q}$) and (x_{kq}, *Sgn_x*$_{kq}$, *key_x*$_{kq}$) from CSP$_q$.

- He verifies the correctness of each share with the signature and the key. In case of errors, the trust tier can ask CSP to get a new share.

-He computes the scalar product of each share (x_{1p} , x_{1q}) , (x_{2p} , x_{2q}) and (x_{kp}, x_{kq}) by (r_p^{-1} *mod p*, r_q^{-1} *mod q*) to retrieve (x_1 *mod p*, x_1 *mod q*), (x_2 *mod p*, x_2 *mod q*) and (x_k *mod p*, x_k *mod q*) .

-After that, the trust tier decrypts the data using the Chinese remainder theorem with the two secrets parameters p and q and with the shares of data (x_1 *mod p*, x_1 *mod q*), (x_2 *mod p*, x_2 *mod q*) and (x_k *mod p*, x_k*mod q*) .

- He verifies the integrity of the original data with the inner signature.

-If the inner signature is correct, the trust tier computes the original data with the function F^{-1}(x). Otherwise, he asks the cloud provider to get another share of data.

The scenario of data reconstruction process is presented in figure 3:

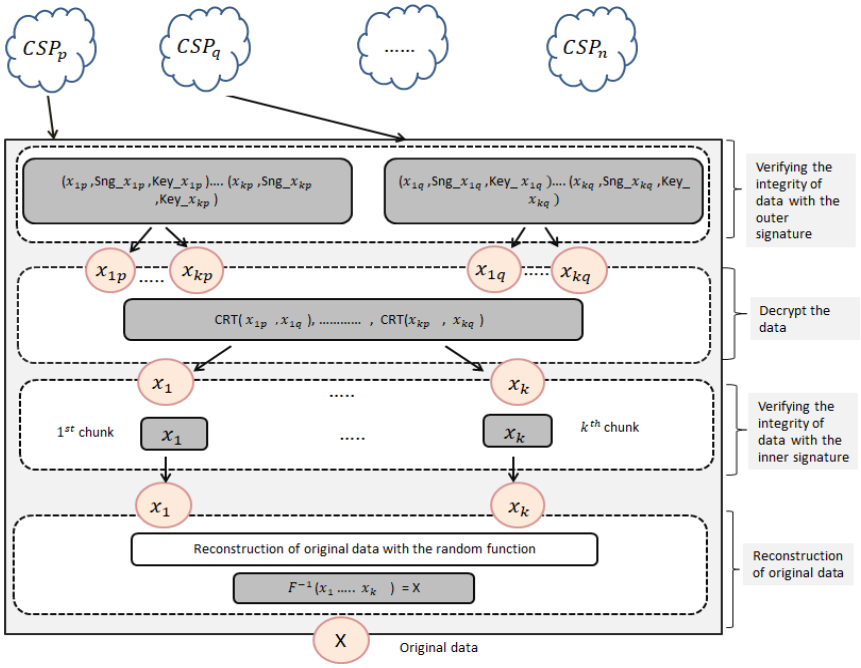


Fig. 3. Data reconstruction process.

To illustrate our schema, we present the example of data sharing and reconstruction process of the integer x= 17,

k = 3,

F(17) = 5+ 8 + 4 ,

P= 5, q= 7, r_p =3, r_q=2, m=35 , n = 2,

ϕ (5) = (5× 3 mod 5, 5× 2 mod 7),

ϕ (8) = (8 × 3 mod 5, 8 × 2 mod 7),
ϕ (4) = (4 × 3 mod 5, 4 × 2 mod 7),
ϕ (5) = (0, 3), ϕ (8) = (4, 2), ϕ (4) = (2, 1),

Compute the signature of each chunk of data as:

sign_x_{1p} = $H_s(0)$,sign_x_{1q} = $H_s(3)$,sign_x_{2p} = $H_s(4)$,sign_x_{2q} = $H_s(2)$,sign_x_{3p} = $H_s(2)$ and sign_x_{3q} = $H_s(1)$
sign_x_{1p} = $H_s(0)$ = 0 mod 2 = 0 and the key is key_x_{1p} = 0- (2* 0) = 0 ,
sign_x_{1q} = $H_s(3)$ = 3 mod 2 = 1 and the key is key_x_{1q} = 3 - (2* 1) = 1 ,
sign_x_{2p} = $H_s(4)$ = 4 mod 2 =2 and the key is key_x_{2p} = 4 - (2* 2) = 0 ,
sign_x_{2q} = $H_s(2)$ = 2 mod 2 =1 and the key is key_x_{2q} = 2- (2* 0) = 0 ,
sign_x_{3p} = $H_s(2)$ = 2 mod 2 =1 and the key is key_x_{3p} = 2- (2* 1) = 0 ,
sign_x_{3q} = $H_s(1)$ = 1 mod 2 = 0 and the key is key_x_{3q} = 1- (2* 0) = 1 ,

After that all data will be sent to the cloud provider.

CSP_p is identified with the secret parameter p and CSPq is identified with the secret parameter q.

For reconstruction the integer x =17,
The trust tier gets all data from the CSP's and verifies the correctness of data:
He verifies that
sign_x_{1p}= 0mod2 = 0 and thatkey_x_{1p} = 0 – (0×2) = 0
sign_x_{1q}= 3 mod 2 = 1 and that key_x_{1p} = 3 – (1×2) = 1
sign_x_{2p}= 4 mod 2 = 2 and that key_x_{1p} = 4 – (2×2) = 0
sign_x_{2q}=2 mod 2 = 1 and that key_x_{2q} = 2 – (1×2) = 0
sign_x_{3p}=2 mod 2 = 1 and that key_x_{3p} = 2 – (1×2) = 0
sign_x_{3q}=1mod 2 = 0 and that key_x_{3q} = 1 – (0×2) = 1

If it is the case, he computes:
P_p =7 , P_q= 5 ,b_p=3 ,b_q=3,x_{1p} = 0, x_{1q} = 3 ,x_{2p} = 4,x_{2q} = 2 ,x_{3p} = 2 ,x_{3q} = 1 , m=35
r_p^{-1} mod p = 3^{-1} mod 5 = 2,
r_q^{-1} mod q = 2^{-1} mod 7 = 4 ,

After that we compute: (0× 2 mod 5, 3×4 mod 7) = (0, 5) ,(4× 2 mod 5, 2×4 mod 7) = (3, 1) ,(2× 2 mod 5, 1×4 mod 7) = (4, 4).

Using the CRT we can compute:

x_1= CRT (0 , 5) = (0×7×3 + 5×5×3) mod 35 = 5
x_2= CRT(3 , 1) = (3 ×7×3 + 1×5×3) mod 35 = 8
x_3= CRT (4 , 4) = (4×7×3 + 4×5×3) mod 35 = 4

to verify the correctness of each chunk of data, we compute:

$S_p = |0 + 3 + 4|_5 = 2$
$S_q = |5 + 1 + 4|_7 = 3$
$Chunk_p = |5 + 8 + 4|_5 = 2$
$Chunk_q = |5 + 4 + 8|_7 = 3$
$S_p = Chunk_p$ and $S_q = Chunk_q$, so the chunks of data is correct.

After that, we compute the original data X = 5 + 4+ 8 = 17.

4.3 Sharing Data Warehouse

The whole table of a shared data warehouse is stored in a relational database at a given CSP's (two initial CSP's in our case) and each attribute value in each record is encrypted independently as described in the data sharing process except the primary keys and the foreign keys. Figure 4 gives an example of star schema data warehouse that is shared among two CSP's. Each shared model of data warehouse stands for the same schema as the original data warehouse, except that two other attributes are added to store signatures and keys.

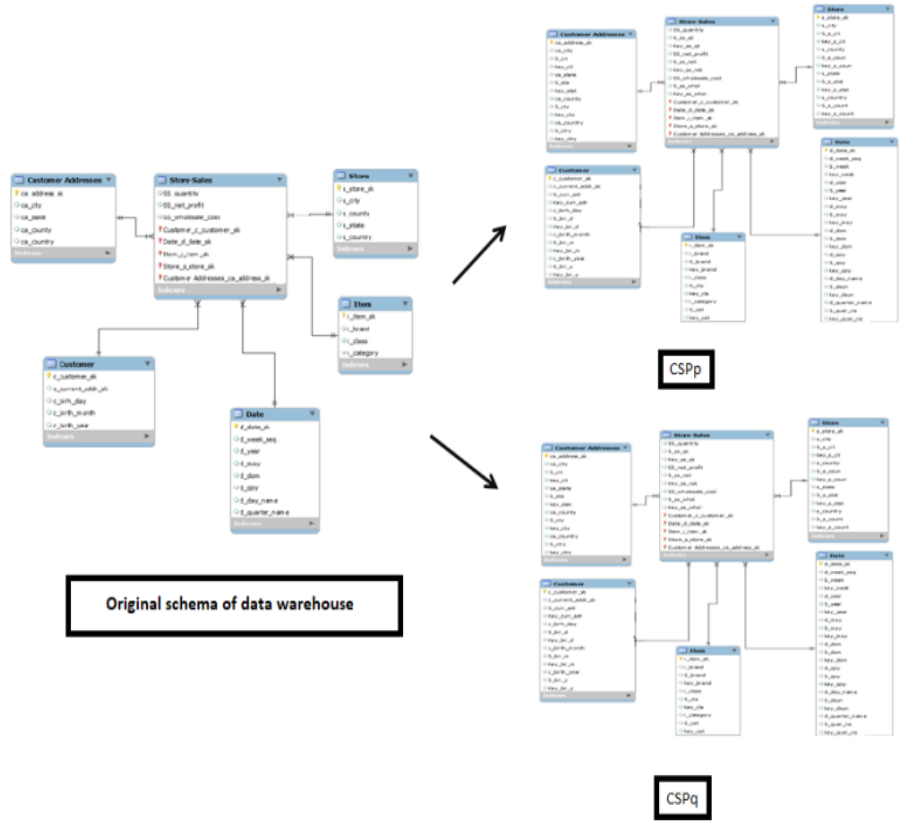


Fig. 4. Sharing schema of data warehouse in each CSP.

4.4 Querying Data Warehouse Hosted in the Cloud

Our schema can directly support some basic OLAP operations at the CSP's through SQL operations and aggregation function. For example, simple select-from queries can be directly applied in the cloud. However, when expressing a condition in a where or having clause, the trust tier must rewrite the query and post processing some operation in the company because the MOD operator is non- injective. Given that for X MOD Y = Z, the same output Z, considering Y a constant, can have an undetermined number of possibilities in X as an input which will generate the same value Z when applying the operator. Figure 5 describes the scenario of processing a select query.

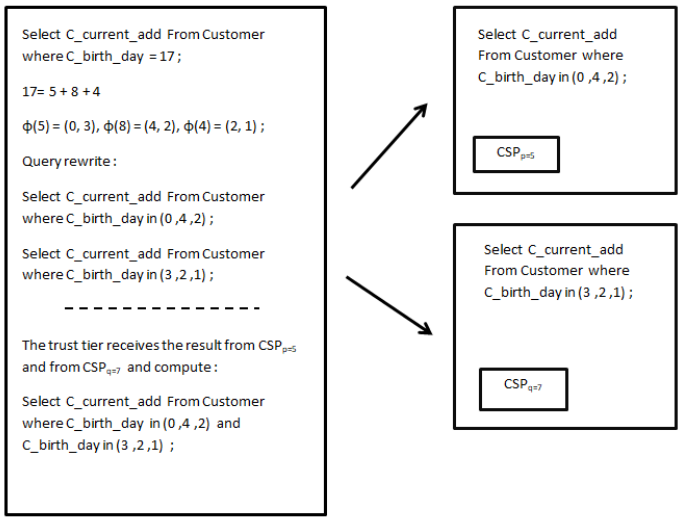


Fig. 5. Scenario of processing Select query.

This routine works for many comparison operators (=, ≠, EXISTS, IN, LIKE…) and their conjunction. Arithmetic operation and aggregation function such as sum, avg, count can be computed in ciphertext by the trust tier after eliminating the error rows.

4.5 Weighted Method for Answering Range Query

When ordering is necessary, as in ORDER BY clauses and many comparison operators (>, <, ≥, ≤, BETWEEN…), it can no longer be applied since the original order is broken when sharing data. Thus, all fetched data must be decrypted and quered at the owner by the trust tier before being sent to the client. This operation is very expensive in terms of time computation. To infer the performance of such kind of query, we propose a weighted method that reduces the time complexity of running range query.

We will also propose two weights, the first one is to encrypt data with modulo p, while the second one is to encrypt data with modulo q, so:

$$Wp = \frac{|\ P^{-1}|_p}{p}, \tag{5}$$

$$Wq = \frac{|\ Q^{-1}|_q}{q}, \tag{6}$$

with $P = \frac{m}{p}$ and $Q = \frac{m}{q}$.

After that, for comparing two integer A ((a_{1p}, a_{1q}), (a_{2p}, a_{2q}), (a_{kp}, a_{kq})) and B ((b_{1p}, b_{1q}), (b_{2p}, b_{2q}), (b_{kp}, b_{kq})) encrypted with the homomorphic function we will compute :

$$Reslt_A = |a_{1p} \times Wp \times \frac{1}{r_p} + a_{1q} \times Wq \times \frac{1}{r_q} + a_{2p} \times Wp \times \frac{1}{r_p} + a_{2q} \times Wq \times \frac{1}{r_q} + a_{kp} \times Wp \times \frac{1}{r_p} + a_{kq} \times Wq \times \frac{1}{r_q}|_1, \tag{7}$$

$$Reslt_B = |b_{1p} \times Wp \times \frac{1}{r_p} + b_{1q} \times Wq \times \frac{1}{r_q} + b_{2p} \times Wp \times \frac{1}{r_p} + b_{2q} \times Wq \times \frac{1}{r_q} + b_{kp} \times Wp \times \frac{1}{r_p} + b_{kq} \times Wq \times \frac{1}{r_q}|_1. \tag{8}$$

After that, if Reslt_A>Reslt_B so A > B, else A < B. Consequently, using this method, there will be no need for decrypting all data when querying range query and the operation of comparison will be done in ciphertext in the company.

However, some range queries can be transformed and performed in the cloud if the comparison range is known. For example, the query "Select C_current_add From Customer where C_birth_day between 17 and 19" would be transformed to "Select C_current_add From Customer where C_birth_day in (0 , 4, 2) or in (3 , 0, 1) or in(3 , 0, 2) " at $CSP_{p=5}$, where(0 , 4, 2) , (3 , 0, 1) ,(3 , 0, 2) are the shares of 17, 18 and 19 at $CSP_{p=5}$, respectively for $CSP_{q=7}$.

5 Security Analysis and Performance Evaluation

This section is devoted to illustrate the relevance of our approach along two axes. The first axis is about the security features of our scheme, while the second axis is about the performance of our schema in terms of time complexity and volume overhead when using the pay-as-you go paradigm.

5.1 Security Analysis

In this section we will deal with the security analysis of our schema in terms of confidentiality and integrity. We will also show the efficiency of our proposal to overcome the problem of collusion.

Confidentiality of data

To protect data from plaintext attacks and from malicious cloud providers, we propose to split data into k chunks with the random function F(x) and encrypt each chunk with

the homomorphic function ϕ(x) presented in equation 2. Then, all encrypted chunks will be stocked in two cloud providers.

The objective is to keep minimal information about data and parameters among the cloud provider. As a result, we can reduce the risk of inferring the data and breaking the system.

The role of trust tier as a middle tier between the user and the cloud is an effective solution which guarantees the confidentiality of the two secrets p and q.

Proof:
If the cloud provider predicts the p and the q or gets the p and the q (worst case), he can predict x as:

$X = y \times p + xp$ and $X = y \times q + xq$ such that $x < m$.

We can conclude that even if the two secret parameters are discovered by the cloud provider, he cannot identify whether the chunks of data correspond to the parameters p or q or not. This ambiguity can disrupt the work of inferring the data. Furthermore, since factoring is hard, inferring the k chunks of data without known whether the chunks of data correspond to the parameters p or q cannot be done in the case of a huge volume of data as in the data warehouse.

In the worst case, if the cloud provider infers the data, decrypts the entire share and gets original chunk, the security of our original data cannot be breached because the cloud provider has just a chunk of data and not all the original data. So, we can deduce that splitting data is essential to the security of our schema. That's why the parameter k will be chosen carefully and should be $k \geq 2$.

Similarly, it can be argued that the confidentiality of our schema is better with this new sharing strategy. In fact, even if the malicious intruder gets the two secret parameters p and q, it is hard to break the security of the two cloud providers and get the chunks of secrets at the same time. Even if he does this and decrypts the k chunks of data, he cannot reconstruct the original data because he does not know the random function F(x).

Integrity of data

Outer signature

In our schema, the processing of data is done in ciphertext. Thus, there is no need to use a complex integrity function to ensure the security of data. Our goal is to propose a simple method which allows the verifying of the data sent and received from the cloud with minimum time complexity.

Our integrity function Hs (x) is homomorphic. This is very practical in the case of data warehouse. There is no need to verify each chunk of data separately if we calculate the aggregations functions.

Inner signature

The inner signature is a self-checked signature. It is based on the homomorphic characteristic of arithmetic modular. So, if there are some mistakes that cannot be detected with the outer signature, the inner signature can detect it.

Collusion

With this schema the risk of collusion is small because data is split randomly with the random function F(x). Consequently, it is hard for the two clouds providers to predict how data is split if it colludes.

5.2 Performance Evaluation

Volume overhead

Our encryption function is based on the MOD operator which divides the data in a small residue number. So, there is no overhead volume when encrypting initial data. The original data is split into k chunks and each chunk will be encrypted with the Mod operator two times. The volume of the encrypted data cannot exceed twice the volume of the original data.

Temporal complexity

In our scheme for the encryption phase, we need just O(n) operation because the encryption function needs only modular operation.

The decryption function is based on the CRT. So we just need O (lglg n) operation for the decryption phase.

Comparing our schema with the existing related approaches

In this section, we compare our schema with the approaches presented in our state of the art in terms of security and performance. Table 1 synthesizes the features of some approaches discussed above.

As it is described in table 1 the main advantage of our schema is that ensure three levels of security confidentiality, integrity and availability. More than that, it resists to collision attack.

Also, compared to other solutions our schema is very performing in term of time complexity of encryption and decryption phase. This makes it very suitable in the case of data warehouse. To all those benefits, are ensured with a reasonable volume overhead and a capacity of answering query processing totally or partially in the cloud.

Table 1. Synthesis of the characteristics of some approaches discussed.

		[24]	[31]	[28]	[27]	Our schema
Confidentiality		yes	yes	yes	yes	yes
Integrity	Outer signature	No	No	No	yes	yes
	Inner signature	No	No	No	yes	yes
Availability		yes	No	No	yes	yes
Collusion		yes	No	No	yes	No
OLAP query		yes	yes	No	yes	yes
Range query		No	yes	No	No	yes
Aggregation function		yes	yes	No	yes	yes
Data volume		6n	-	6n	3n	6n (three replicas)
Time complexity of encryption phase		O(n)	-	O(n)	O(n)	O(n)
Time complexity of decryption phase		$O(nlg^2 n)$	-	$O(nlg^2n)$	$O(nt^2)$	O(lglg n)

6 Conclusion

This paper presents SecuredDW as a system for securely hosting and querying data warehouse in the cloud. SecuredDW uses a homomorphic encryption algorithm to ensure the confidentiality of data. This homomorphic encryption algorithm reveals a serious weakness as it can be deciphered by ciphertext attacks. To overcome this weakness, we propose a new sharing method of using this homomorphic privacy based on splitting data, multi cloud providers and perturbation values.

With this new sharing schema, we can reduce the risk of breaking the security of the system by both cloud providers and malicious intruders. Two new signatures are

suggested to ensure the integrity of data sent and received from the cloud inner signature and outer signature.

The weakness of our schema lies in the processing of range queries in the owner after decrypting all the data. This operation can take a lot of time in the case of data warehouse because of the huge volume of data that will be decrypted before processing range query. That's why, we propose a weighted solution to this situation in order to reduce time consumption in the decryption phase. Our schema SecuredDW can be a promising solution for hosting data warehouse in the cloud that balances security and performance.

We eventually endeavour to evaluate our schema in a real cloud provider.

References

1. Rivest, R.L., Dlemam, L.A., Dertouzos, M.L.: On data bank and privacy homomorphisms. Foundations of secure computation. Academia Press, pp 169–177 (1978)
2. National Bureau of Standards: Data Encryption Standard. U.S. Department of Commerce, FIPS Publication 46, Washington, D.C., January (1977)
3. Anderson, R., Biham, E., Knudsen, L.: Serpent: A proposal for the advanced encryption standard. AES proposal: National Institute of Standards and Technology (NIST). In: W.K. Chen (ed.), Linear Networks and Systems (Book style). Belmont, CA: Wadsworth, 1993, pp. 123–135 (1998)
4. Rivest, R.L., Shamir, A., Adleman, L.: A method for obtaining digital signatures and public-key cryptosystems. Communications of the ACM, 21(2):120–126 (1978)
5. Rabin, M.O.: Digitized signatures and public-key functions as intractible as actorization. Technical Report LCS/TR-212, MIT Laboratory for Computer Science (1979)
6. ElGamal, T.: A public-key cryptosystem and a signature scheme based on discrete logarithms. IEEE Transactions on Information Theory, IT-31(4):469–472, July (1985)
7. Jyh-Haw Yeh. A Secure Homomorphic Encryption Algorithm over Integers for Data Privacy Protection in Clouds. SmartCom (2016)
8. Zhao, F., Li, C., Chun, F.L.: A cloud computing security solution based on fully homomorphic encryption. In: 16th international conference on advanced communication technology IEEE (2014)
9. Yu, Y., Niua, L., Yang, G., Mu, Y., Susilo, W.: On the securit of auditing mechanisms for secure cloud storage. Future Generation Computer Systems, vol. 30, pp. 127–132 (2014)
10. Wei, L., Zhu, H., Cao, Z., Dong, X., Jia, W., Chen, Y., Vasilakos, A.V.: Security and privacy for storage and computation in cloud computing. Information Sciences, vol. 258, pp. 371–386 (2014)
11. Lopez-Alt, A., Tromer, V., Vaikuntanathan, E.: On-the-Fly Multiparty Computation on the Cloud via Multikey Fully Homomorphic Encryption. Proceedings of the forty-fourth annual ACM symposium on Theory of computing, pp. 1219–1234 (2012)
12. Brakerski, Z., Vaikuntanathan, E.: Efficient fully homomorphic encryption from (standard) LWE. SIAM Journal on Computing 43(2):831–871 (2011)
13. Agrawal, R., Kiernan, J., Srikant, R., Xu, Y.: Order preserving encryption for numeric data. In: Sigmod 2004, June 13-18 (2004)
14. Hore, B., Mehrotra, S., Canim, M., Kantarcioglu, M.: Secure multidimensional range queries over outsourced data. The VLDB Journal pages, pp. 333–358 (2012)

15. Kathen, H., Amagasa, T., Kitagawa, H.: MV-OPES: Multivalued-order preserving encryption schemes: A novel scheme for encrypting integer value to many different values. IEIC Trans (2010)
16. Cruz-Lopes, C., Cesário-Times, V., Matwin, S., Rodrigues Ciferri, R., Dutra de Aguiar-Ciferri, C.: Processing OLAP Queries over an Encrypted Data Warehouse Stored in the Cloud. In: DaWaK 2014, LNCS 8646, pp. 195–207 (2014)
17. Rashmi, K.V., Shah, N.B., Gu, D., Kuang, H., Borthakur, D., Ramchandran, K.: A Solution to the Network Challenges of Data Recovery in Erasure-coded Distributed Storage Systems: A Study on the Facebook Warehouse Cluster. In: HotStorage'13. San Jose, CA, June 27-28 (2013)
18. Bessami, A., Correia, M., Quaresma, B., André, F., Sousa, P.: DepSky: dependable and secure storage in the cloud-of-clouds. In: Proceedings of the sixth conference on computer systems. ACM, pp. 31–46 (2011)
19. Caclin, C., Haas, R., Vukolic, M.: Dependable storage in the intercloud. IBM rechearch, vol. 3783, pp. 1–6 (2010)
20. Alsolami, F., Chow, C.E.: N-cloud: improving performance and security in cloud storage. In: high performance switching and routing (HPSR), IEEE (2013)
21. Xu, H., Bhalerao, D.: A Reliable and Secure Cloud Storage Schema Using Multiple Service Providers. In: ICSE. DOI: 10.18293/SEKE2015-045 (2015)
22. Zhang, Q., Li, S., Liy, Z., Xingz, Y., Yang, Z., Dai, Y.: CHARM: A Cost-efficient Multi-cloud Data Hosting Scheme with High Availability. IEEE Transactions on Cloud Computing (2015)
23. Adi, S.: How to share a secret. Communication of the ACM, vol. 22 (1979)
24. Pundkar, S.N., Narendra-Shekokar, D.J.: Cloud Computing Security in Multi-clouds using Shamir's Secret Sharing Scheme. In: International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) (2016)
25. Chattopadhyay, A.K., Nag, A., Majumder, K.: Secure Data Outsourcing on Cloud Using Secret Sharing Scheme. International Journal of Network Security 19(6):912–921 (2017)
26. Butoi, A., Tomai, N.: Secret sharing scheme for data confidentiality preserving in a public-private hybrid cloud storage approach. In: 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing (2014)
27. Hadavi, M.A., Jalili, R., Damiani, E., Cimato, S.: Security and searchability in secret sharing-based data outsourcing. In: Int. J. Inf. Secur 14(6) (2015)
28. Varunya, A., Harbi, N., Darmont, J.: A novel multi secret sharing approach for secure data warehousing and On-Line analysis processing in the cloud. In: IGI (2015)
29. Moghadam, S.S., Darmont, J., Gavin, G.: S4: A New Secure Scheme for Enforcing Privacy in Cloud Data Warehouses. In: 7th International Conference on Information Systems and Technologies (ICIST 2017). Dubai, United Arab Emirates, pp. 9–16 (2017)
30. Mahadewan, A.G., Kamesh, T.: Horns: A homomorphic encryption schema for cloud computing using residue number system. In: IEEE 45th Annual Conference on Information Sciences and Systems (2011)
31. Rivest, R.L., Dlemam, L.A., Dertouzos, M.L.: On data bank and privacy homeomorphisms. Foundations of secure computation. Academia Press, pp. 169–177 (1978)

Electronic edition
Available online: http://www.rcs.cic.ipn.mx