# Personalization of Document Warehouses: Formalization, Design and Implementation

**Kais Khrouf and Hela Turki**

Jouf University, King Khaled Road, Al Jouf, Saudi Arabia

**Summary**
In the decision-making domain, a document warehouse is designed to meet the analysis needs of users who may have a wide variety of analysis purposes. In this paper, we propose to integrate the preferences and interactions of users based on profiles to the concept of document warehouses. These profiles guarantee the integration of personalized documents and the collaborative recommendation of documents between different users sharing common interests.

***Keywords:***
*Document Warehouse, Personalization, Recommendation, Profiles, Similarity degree.*

## 1. Introduction

The rapid development of information technology in recent years has created a huge mass of information, which has caused more difficult for users to retrieve relevant information. More precisely, the problem is no longer the availability of information but rather the ability to access and select information that meets the specific needs and preferences of users.

In the decision-making context, the concept of document warehouse has been proposed in order to analyze information from company documents [2] [3]. However, the decision-makers express very varied needs that the warehouse, as it is designed in a generic way, does not have the ability to satisfy them correctly. Some works are oriented towards the integration of the user (i.e., decision-maker) in the analysis system in order to obtain relevant information and knowledge adapted to his needs. This work is part of the personalization of information [9].

The personalization of information is to provide relevant information to users corresponding to their preferences and needs. For that, it is necessary to collect and store data describing users in the form of profiles. Thus, we propose to define the profiles by referring to domain ontologies associated to document warehouses, in order to guarantee the integration of personalized documents and the collaborative recommendation of documents between different users sharing common interests.

Manuscript received October 5, 2022
Manuscript revised October 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.10.49

This paper is organized as follows. In Section 2, we outline the literature review of works on document warehouses and personalization of data warehouses. We present in next section the formalization for the personalization part. Section 4 describes the meta-model model we propose for document warehouses. In Section 5, we present the proposed formulas for profiles-based collaboration. Finally, we describe the implementation part in order to validate the presented work.

## 2. Literature Review

During the last two decades, the community of data warehouses has paid more attention to the paradigm of document warehouses.

The authors of [1] aims to enhancing the multidimensional model called Diamond Document Warehouse Model with semantics aspects; in particular, it suggests semantic OLAP (On-Line Analytical Processing) operators for querying the Document warehouses.

Data can be heterogeneous and complex: Semi-structured data (Example: XML), Data from social networks (Example: Tweets) and Factual data (Example: Spreading of Covid-19). The authors of [7] propose a generic multidimensional model of document warehouses in order to analyze complex data, according to several dimensions.

By using data analysis technologies, the authors of [6] and [8] propose to apply the multidimensional analysis techniques of data warehouses on data from Excel documents about Coronavirus CoVID-19 evolution and spreading.

Other works have focused on the personalization of data warehouses.

SaaS is used in the different fields (data analysis, information security, etc.). The generalized SaaS platform is unable to satisfy the requirements of enterprise personalization caused by the improvement of the degree of information. The authors of [5] propose to decompose

personalization technology into metadata driver, cloud data placement and mapping mechanism for research.

The Product-Service Systems (PSS) customization entails configuring products with varying degrees of differentiation to meet the needs of various customers. The authors of [4] propose a data warehouse-based recommender system that collects and analyzes large volumes of product usage data.

In this paper, we propose to define the profiles based on domain ontologies for: (1) the personalization of documents and (2) the collaborative recommendation of documents between different users.

## 3. Formalization of Personalization Part

In order to add semantics to textual elements (Summary, Section, Paragraph, etc.), authors use domain ontologies. The ontology can be defined as a set of concepts, as well as relationships between these concepts.

The ontology $O_i$ is defined by $O_i = \{O_i^{Name}, C\}$
- $O^{Name}$: Name of ontology
- $C$: A set of $m$ concepts, noted C = $\{C_1, C_2, \ldots, C_m\}$

A concept $C_i = \{C^{Name}, C^{Weight}, C^{Father}, T\}$
- $C^{Name}$ : Name of concept
- $C^{Weight}$ : Weight of concept (affected by experts)
- $C^{Father}$ : Concept father of $C_i$
- $T$ : It is a set of $k$ terms (Keywords) describing $C_i$, noted $T = \{T_1, T_2 \ldots, T_k\}$. These terms can be synonyms or abbreviations of the concept $C_i$.

The definition of a user profile is to adapt an ontology for needs and preferences of the user. More precisely, the user can: i) select part of the ontology or its totality, ii) add other concepts, iii) add terms to the concepts of the ontology, and iv) delete concepts of the selected sub-ontology. The user profile, defined as follows, is a sub-ontology adapted to his needs and preferences.

A profile, noted $P_i = \{P^{Name}, O_j, C^{Root}, C^{Pi}\}$
- $P^{Name}$: Name of profile
- $O_j$: Ontology selected by the user
- $C^{Racine}$: Name of concept root, $C^{Root} \in O_j.C$
- $C^{Pi}$: A set of concepts for the profile $P_i$, $C_{Pi} \subset O_j.C$

A Concept $j$ of Profile $P_i$, *noted* $C_{Pi,j} = \{C_{Pi,j}^{Name}, C_{Pi,j}^{Type}, C_{Pi,j}^{Weight}, C_{Pi,j}^{Father}, T_{Pi,j}\}$
- $C_{Pi,j}^{Name}$: Name of concept $j$ of profile $P_i$
- $C_{Pi,j}^{Type}$: Type of concept $j$ of profile $P_i$ = {$O$: From the Ontology $P_i.O_j$, $U$: Added by the user}
- $C_{Pi,j}^{Weight}$: Weight of concept, affected by user
- $C_{Pi,j}^{Father}$: Name of concept father
- $T_{Pi,j}$: A set of terms describing the concept $j$ of profile $P_i$

A Term $T_k \in T_{Pi,j} = \{ T_k^{Name}, T_k^{Type}\}$
- $T_k^{Name}$: Name of term
- $T_k^{Type}$: Type of term = {$O$: From the Ontology $P_i.O_j$, $U$: Added by the user}

## 4. Meta-Model of Document Warehouse

The proposed meta-model includes the following components:

1. The list of documents (cf. Fig. 1.a).

2. The structural component: It describes the hierarchical structure of documents. We distinguish two structures:
- Generic structure (cf. Fig. 1.b): Common structure for a set of documents. It is composed into a set of generic elements, described by generic attributes.
- Specific structure (cf. Fig. 1.c): A structure of a document according to a generic structure. It is composed into a set of specific elements, including specific attributes.

3. The content component (cf. Fig. 1.d): It is the content of different parts of the document (elements of the specific structure).

4. The semantic component (cf. Fig. 1.e): It is defined by a set of ontologies that can be decomposed into concepts. These concepts are described by a list of keywords.

5. The profile component (cf. Fig. 1.f): It is defined by a set of profiles decomposed into concepts: Those of the initial ontology (Concept_Ont class) and those added by the user (Concept_Util class). Thus, the user can assign the corresponding weights to the added concepts, according to his needs and preference.

Fig. 1 presents the proposed meta-model of Document Warehouses.

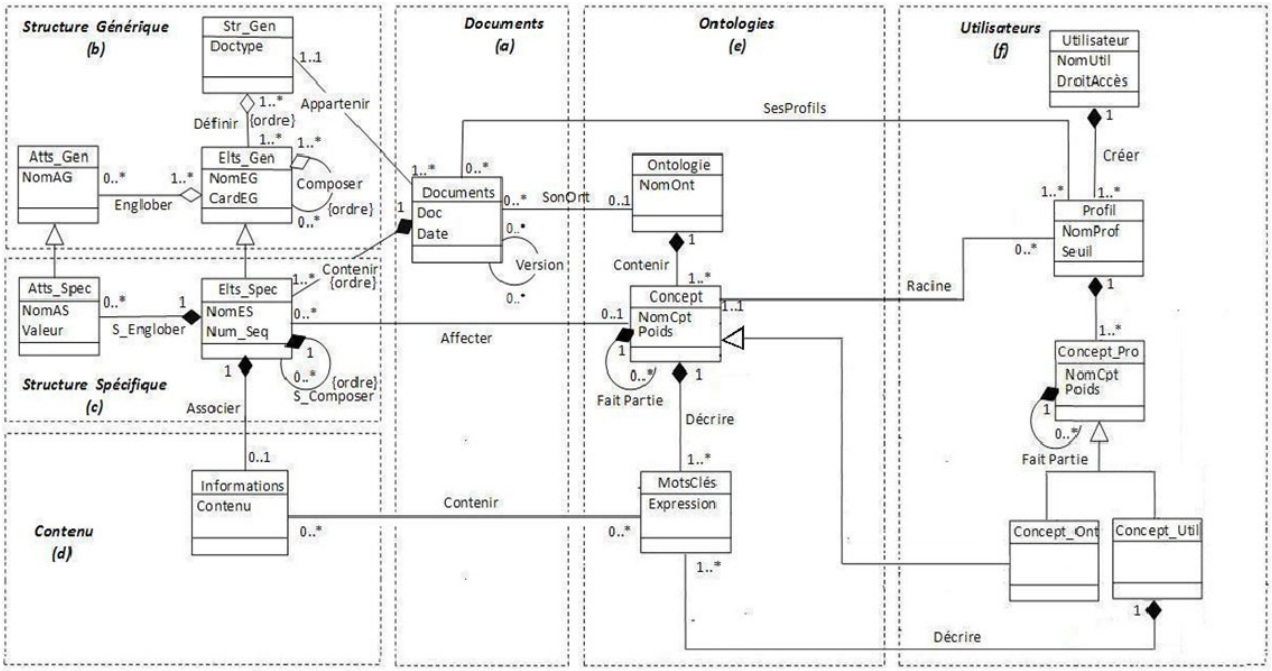


Fig. 1 Meta-Model of Document Warehouse.

## 5. Profile-Based Collaboration

Another objective of this work is to propose a collaborative recommendation between the users of the document warehouse. The idea is to inform users of new documents inserted into the warehouse whenever they are consistent with their centers of interest (profiles). To achieve this goal, we need to compare each new profile created by a user with the existing profiles in the warehouse.

It is in fact a question of determining the resemblance degree between the different profiles created. The common concepts between the different profiles express the similarity of the needs and the preferences shared between the users. Thus, we propose Formula 1.

$$Ress\ (C, P_i) = \frac{\sum_{j=1}^{|P|} freq\ (C, P_j)}{|P|} \qquad [1]$$

With:

- *Ress (C, $P_i$)*: Resemblance degree of the concept *C* of the $P_i$ profile.
- $|P|$ : Number of warehouse profiles.
- *freq (C, $P_j$)*: The frequency of appearance for the concept *C* in the profiles $P_j$.

After the user profiles have been defined, we assign the new documents to the profiles. More precisely, if a user adds a new document *d* in the warehouse, the system automatically prepares a degree of similarity *Sim* (cf. Formula 2) between the document *d* and each of the profiles of all the users.

Later, the recommendation system will inform the user of the presence of the new document *d* and of the similarity degrees of *d* with his profiles. The user will be informed by only the documents having similarity degree *Sim* greater than a threshold.

$$Sim\ (d, P_i) = \frac{\sum_{j=1}^{|Pi|}\left(\sum_{k=1}^{|Cpi,j|} freq\ (d, T_k) * C_{Pi,j}^{Weight}\right)}{\sum_{j=1}^{|Pi|}\left(\sum_{k=1}^{|Cpi,j|} freq\ (d, T_k)\right)} \qquad [2]$$

With:

- *Sim(d, $P_i$)*: Similarity degree of document *d* compared to $P_i$.
- $|C_{Pi,j}|$: Number of terms of concept *j* belonging to the profile $P_i$,
- $|P_i|$: Number of concepts of profile $P_i$,
- *Freq(d, $T_k$):* Occurrence frequency of term $T_k$ of concept $C_{Pi,j}$ in *d*,
- $C_{Pi,j}^{Weight}$: Weight of concept $C_{Pi,j}$.

## 6. Implementation

After the authentication step, the user creates his profiles by giving the Name of the profile and fixing two thresholds (*threshold*$_1$ and *threshold*$_2$) belonging to the interval [0,1] (cf. Fig. 2).

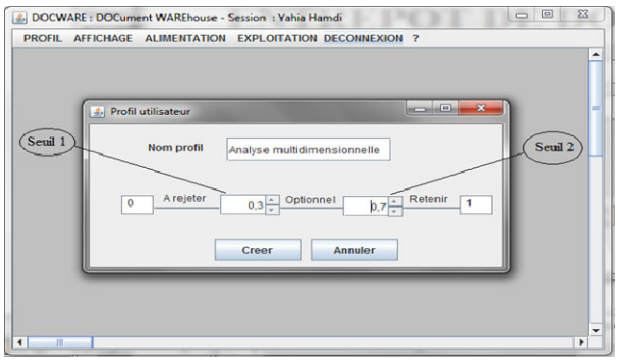


Fig. 2 Interface of Profile Creation.

These two thresholds are then used to assign documents, inserted in the warehouse, to the defined profiles:
- If the similarity degree of the inserted document is between 0 and *threshold*$_1$, the document will not be assigned to this profile. In this case, the profile will not be displayed.
- If the similarity degree of the document is between the two thresholds, the document is optionally assigned to this profile. In this case, the profile will be displayed with the Orange color.
- When similarity the degree of the document is higher than the *threshold*$_2$, the defined profile represents the best profile in order to affect this document. In this case, the profile will be displayed with the Green color.

The user adds a document to the warehouse. The system calculates the similarity degree between this document and the user profiles, then it displays the corresponding profiles with the calculated degrees and the corresponding colors (cf. Fig. 3).

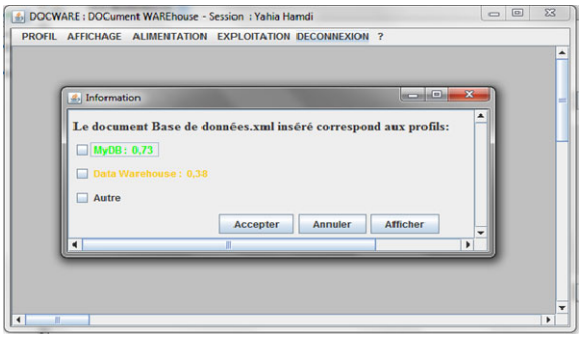


Fig. 3 Interface for adding documents to the warehouse.

To define his profile, the user begins by choosing an ontology of the warehouse, he chooses the root concept for his profile. The system displays graphically the ontology and the user can perform the following operations (adding and deleting concepts, adding and deleting terms and assigning and modifying weights to concepts)

Fig. 4 presents a visualization example of a simplified ontology of the warehouse, from which the user will define his profile.

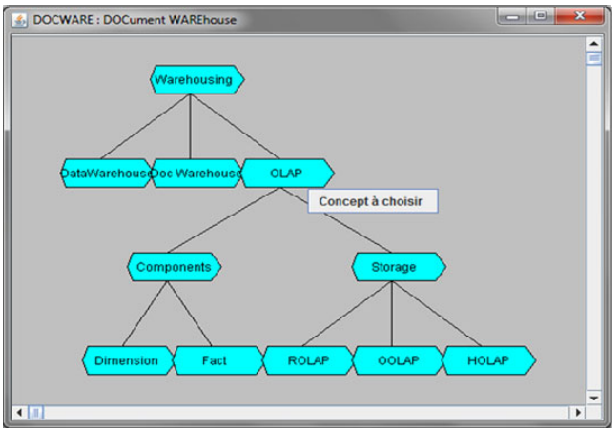


Fig. 4 Graphical visualization of a simplified ontology.

For this example, the user chose the root concept *OLAP*, he removed the concept *Storage* from the profile (its child concepts were automatically removed) and added the concept *Operators* under *OLAP* and the three child concepts: *Slice*, *Dice* and *Rotate*. Fig. 4 presents the user profile built from the simplified ontology of Fig. 3.

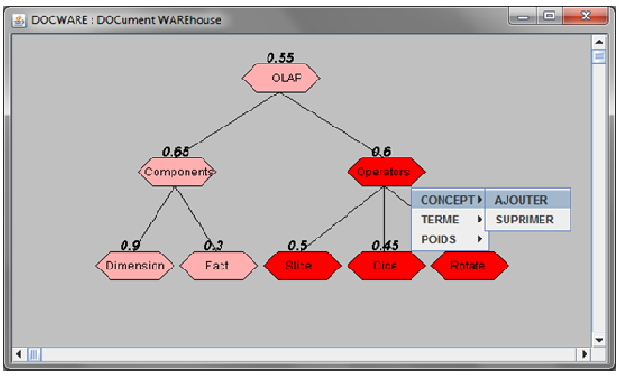


Fig. 4 User profile built from the simplified ontology of Fig. 3.

## 7. Conclusion

In this paper, we proposed a document warehouse meta-model for the personalized integration of documents in the warehouse and the collaborative recommendation of these documents between the different users sharing common interests. To validate this work, we created graphical interfaces for defining user profiles and integrating new documents into the warehouse.

For future work, we plan to take into consideration the evolution of ontologies and their impact on existing profiles, the adaptation to changes in user needs and preferences, such as the reassignment of documents to new profiles. Finally, we will focus on the personalization of multidimensional analysis by adapting the OLAP (On-Line Analytical Processing) operators to personalization.

## Acknowledgment

We would like to express our sincere thanks to Mr Yahia Hamdi for his contribution to this work.

## References

[1] M. Azabou, A. Banjar, J. Feki, "Enhancing the Diamond Document Warehouse Model", *International Journal of Data Warehousing and Mining*, Vol. 16, No. 4, 2020.

[2] I. Ben Messaoud, A. Alshdadi, J. Feki, "Building a Document-Oriented Warehouse Using NoSQL", *International Journal of Operations Research and Information Systems*, Vol. 12, No. 2, 2021.

[3] S. Bouaziz, A. Nabli, F. Gargouri, "Design a Data Warehouse Schema from Document-Oriented database", *Procedia Computer Science*, Vol. 159, 2019.

[4] L. Esheiba, I. Helal, A. M. El-Sharkawi, "A Data Warehouse-Based System for Service Customization Recommendations in Product-Service Systems", *Sensors*, Vol. 22, No. 6, 2022.

[5] Q. Guo, H. Sun, W. Ji, "Design of Personalization Warehouse Management Platform Based on SaaS Model", *IEEE International Conference on Computer Supported Cooperative Work in Design*, p. 1535-1540, 2022.

[6] K. Khrouf, "Multidimensional Analysis of Coronavirus CoVID-19 Spreading: Study in Arabic Countries Context", *International Journal of Computer Science and Network Security*, Vol. 20, No. 6, 2020.

[7] K. Khrouf, H. Turki, "Generic Multidimensional Model of Complex Data: Design and Implementation", *International Journal of Computer Science and Network Security*, Vol. 21, No. 12, 2021.

[8] K. Khrouf, H. Turki, "Data Analysis of Coronavirus CoVID-19: Study of Spread and Vaccination in European Countries", *International Journal of Advanced Computer Science and Applications*, Vol. 13, No. 1, 2022.

[9] J. Liu, C. Liu, N. Belkin, "Personalization in Text Information Retrieval: A Survey", *Journal of the Association for Information Science & Technology*, Vol. 71 No. 3, 2020.

**Dr. Kais Khrouf** was born in Paris, France, 1977. He received the B.S. degree in Computer Sciences from University of Sfax, Tunisia in 1999. He received the M.S. and Ph.D. degrees in Computer Sciences from University of Paul Sabatier, Toulouse, France (2004). From 2005 to 2017, he was an assistant professor at University of Sfax, Tunisia. Since 2017, he is an assistant professor at Jouf University, Saudi Arabia. He is the author of more than 40 articles in international journals and conferences and he is a permanent reviewer in International Journal of Information and Decision Sciences (Inderscience Publishers). His research interests include Decision Support Systems, Data Analysis, Data Warehouses, Personalization, Social Networks and Semi-Structured Data.

**Dr. Hela Turki** received the B.S. and the M.S. degree from University of Sfax, Tunisia in 2012. She received the Ph.D. degree from University of Sfax, Tunisia in 2017. From 2015 to 2017, she was a contractual professor at University of Sfax, Tunisia. Since 2017, she is an assistant professor at Jouf University, Saudi Arabia. She is the author of many papers in international journals and conferences. Her research interest includes Decision Making and Data Analysis.