Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 6, 244-247 (2022)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Consideration of Ambiguity in the Analysis Phase of Data Warehouses

Djamila Hammouche*,1, Karim Atif [2]

[1]Department of computer science, Faculty of exact sciences and informatics, Hassiba Benbouali University, Chlef, 02180, Algeria

[2]Department of computer systems, Faculty of computer science, University of sciences and technology Houari Boumediene, Bab Ezzouar, Algiers, 16111, Algeria

ARTICLE INFO

Article history:
Received: 03 August, 2022
Accepted: 05 November, 2022
Online: 20 December, 2022

Keywords:
Ambiguity
Server failure
Data warehouse
Decision making
Fuzzy logic
Membership function

ABSTRACT

We are interested in taking into account ambiguity in the analysis phase of data warehouses, using fuzzy logic. We want to offer decision makers the possibility of using natural language in this phase. We created in a previous work the Baccalaureate fuzzy data warehouse which we were able to query with seven natural language terms to which we created seven membership functions. In this work, we present a fuzzy data warehouse for server failures that we created and for which we used the same terms to which we associated seven membership functions too. And, we carried out a comparison at the end of which we concluded that the definition of the values of the membership function differs according to the context of analysis. Our solution is extensible and can be enriched with new natural terms language. The next step is to design a conversational interface that enables a natural language conversation between the decision maker and the fuzzy data warehouse.

1. Introduction

Decision makers need to use natural language in the analysis phase of data warehouses; this allows them to appreciate existing data as they wish. Our field of application is higher education where decision makers need to analyze the failure rate of servers and easily detect the machine that frequently breaks down, the type of failure, the duration in order to be able to make good forecasts.

No machine is 100% reliable. The fault tolerance criterion of the machine is expressed either in average number of hours between failures, or in number of hours of operation before the end of life of the machine. Except this criterion given by the manufacturer, the users of these machines study server failures that occur on a real time basis. It is in this context that our study is written [1].

In this paper, we present the fuzzy server failure data warehouse that we created using Mondrian environment and MDX queries [1].

We are looking for the answer to the questions:

- Which servers have an average failure rate, lowest highest, etc.?
- How to use the terms: medium, high in queries?

The rest of the paper will be organized as follows: We first explain the context of this study, we present an overview of related works, and then we detail the solution based on fuzzy logic and we present the comparative analysis, finally, we finish with conclusion.

2. The study context

The goal of all our work is to show that the use of natural language facilitates the querying of data warehouses. So, we used Fuzzy logic. We defined the membership functions of the predicates that represent the natural language terms most used by analysts. In this context, we have chosen to experiment our solution on two very distinct areas:

- The field of national education to assess the baccalaureate success rate for which we have published the results [2]-[4].
- And the field of server failures to appreciate the server failure rate that we present in this paper.

In this work, we used the same terms to experiment them in the field of server failures and we defined the corresponding membership functions.

3. Related works

In [5]-[8], the author focuses on fuzzy multidimensional data, presenting a solution based on fuzzy logic and using SQL to take

* Corresponding Author: Djamila Hammouche, d.hammouche@univ-chlef.dz

https://dx.doi.org/10.25046/aj070627

into account imperfect values and vague criteria. ReqFlex is an intuitive user interface to the definition of preferences and the construction of fuzzy queries [9]. Also, the study on the medical data warehouse recording the vital parameters (blood pressure) of patients [10]. In [11], the author adds a new fuzzy layer to the existing model without modifying the data warehouse. Linguistic concepts are integrated for the interpretation of the measurements of the fact table. This separation between the fuzzy concepts and the data warehouse constitutes the strong point of this model. This is the reason why we have retained it in all our work despite the fact that this model has not been tested on a complex system.

4. The solution based on fuzzy logic

There are in classical Boolean logic only two states: TRUE or FALSE. In [12]-[14], the author proposed Fuzzy logic in 1965. This logic makes it possible to express different levels and to describe a phenomenon linguistically, then to represent it by a small number of rules.

A fuzzy term is a natural language word. To model it, we use a trapezoidal function for which characteristic parameters are defined.

4.1. The Fuzzy server failure data warehouse model

We designed the star schema of our data warehouse [15]- [21]. In [1], the model is defined and commented. There is one fact table "FactServerFailure" and four dimension tables: "Dim_Department", "Dim_Section", "Dim_Server" and "Dim_Time".

In fact table "FactServerFailure", there are all the primary keys of all the dimension tables related to the fact table and there is the measurement Failure_rate. It is as the metric to calculate and analyze. It indicates the ratio between the number of failure of a server to the total number of failures of the servers.

Failure_rate=number of server failure*100/total number of server failures.

In the other dimension tables, we find the characteristic attributes of these tables and which constitute axes of analysis such as time, department, etc.

The integration of ambiguous terms is done through two fuzzy meta tables that we call Fuzzy_ct table and Fuzzy_mt table.

Fuzzy Fuzzy_ct table is used to store fuzzy classes associated with linguistic terms (absolutely high, average, absolutely low, etc.) and Fuzzy_mt table is used to store the membership degrees of a value to a fuzzy class and the query result expresses the degree of each failure rate to a fuzzy class. Concrete examples of these degrees of membership are illustrated in the various tables which appear in the section some results.

4.2. The membership functions

We integrated seven fuzzy predicates to qualify the failure rate as absolutely high, rather high, somewhat high, average, somewhat low, rather low and absolutely low and we defined the seven corresponding membership functions.

Table 1: The membership functions

The membership function	Values and intervals
Absolutely high	Y=0 in [0, 7], y=1 in [8, 10] and y=x-7in [7, 8].
Ratherhigh	Y=0 in [0, 6] and [9, 10], y=1 in [7, 8], y=x-6 in [6, 7] and y=-9-x in [8,9].
Somewhathigh	Y=0 in [0, 5] and [9, 10], y=1 in [6,8], y=x-5 in [5, 6] and y=-9-x in [8, 9].
Average	Y=0 in [0, 4] and [7, 10], y=1 in [5,6], y=x-4 in [4, 5] and y=7-x in [6, 7].
Somewhatlow	Y=0 in [0, 6] and [9, 10], y=1 in [7, 8], y=x-6 in [6, 7] and y=-9-x in [8, 9].
Ratherlow	Y=0 in [0, 1] and [5, 10], y=1 in [2, 4], y=x-1 in [1, 2] and y=-x+5 in [4, 5].
Absolutelylow	Y=0 in [3, 10], y=1 in [0, 2] and y=3-x in [2, 3].

4.3. Some results

We present in following some results of the realized system.

In this table appear the servers with their characteristics of the University whose failure rates we have studied since their first commissioning until the year 2020.

Table 2: List of servers studied

Server number	Physical server model	RAM	Storage
1	HP Proliant gen 05	8G	512G
2	HP Proliant gen 05	8G	512G
3	HP Proliant gen 05	8G	512G
4	HP Sauvegarde	8G	512G
5	DELL PowerEdge T300	8G	512G
6	DELL	16G	1.7T
7	HP Z420 Workstation	8G	512G
8	DELL PowerEdge T320	16G	1.7T
9	DELL PowerEdge T320	16G	1.7T
10	DELL PowerEdge T320	16G	1.7T
11	HP Proxy1	12G	900G
12	DELL Proxy2	12G	900G
13	HP server Proliant Gen 10	8G	1T
14	DELL PowerEdge T320	16G	1.7T

In this table appear the servers with failure rates

Table 3: List of servers studied with their failure rates

Server number	Physical server model	Server failure rate
1	HP Proliant gen 05	9.09%
2	HP Proliant gen 05	9.09%
3	HP Proliant gen 05	9.09%
4	HP Sauvegarde	9.09%
5	DELL PowerEdge T300	6.06%

6	DELL	5.05%
7	HP Z420 Workstation	9.09%
8	DELL PowerEdge T320	1.01%
9	DELL PowerEdge T320	5.05%
10	DELL PowerEdge T320	6.06%
11	HP Proxy1	9.09%
12	DELL Proxy2	7.07%
13	HP server Proliant Gen 10	9.09%
14	DELL PowerEdge T320	8.08%

In this table appear the servers that recorded an absolutely high rate of failure. The percentage 9.09% and 8.08% are considered absolutely high failure rates with a total degree of membership (100%) and the percentage 7.07% is considered an absolutely high failure rate with only 07% membership degree. And all the percentages of the other remaining servers are not considered absolutely high failure rates because their membership degree is 0%.

Table 4: Result of query absolutely high

Server number	Server failure rate	Membership degree
1	9.09%	100%
2	9.09%	100%
3	9.09%	100%
4	9.09%	100%
7	9.09%	100%
11	9.09%	100%
13	9.09%	100%
14	8.08%	100%
12	7.07%	07%
05	6.06%	0%
06	5.05	0%
08	1.01%	0%
09	5.05%	0%
10	6.06%	0%

In this table appear the servers that recorded an absolutely low rate of failure. The percentage 1.01% is considered an absolutely low failure rate with a total degree of membership (100%) and all the percentages of the other remaining servers are not considered absolutely low failure rates because their membership degree is 0%.

Table 5: Result of query absolutely low

Server number	Server failure rate	Membership degree
8	1.01%	100%
1	9.09%	0%
2	9.09%	0%
3	9.09%	0%
4	9.09%	0%
5	6.06%	0%
6	5.05%	0%
7	9.09%	0%
9	5.05%	0%
10	6.06%	0%
11	9.09%	0%
12	7.07%	0%
13	9.09%	0%
14	8.08%	0%

4.4. The comparative study

We found that for the same terms the meaning remains the same but the membership functions change significantly. For example, for a very high baccalaureate success rate, the percentage of success is between 90% and 100%, while for server failures the highest rate is around 10% only.

Thus, we have established the following result. The definition of the values of the membership function is closely linked to the field of study. For any natural language term, the term retains its meaning all the time but the definition of the values of the membership function which corresponds to it changes according to the context of study.

We are interested in studying the same terms with other data from similar fields to better frame the definition of values of membership functions and to arrive at a single definition for a field of study. Only multiple experiments will prove it. Also, it is interesting to experiment with these same terms in sub-domains of the same domain like studying the failure rate of computers linked to servers.

Table 6: Results comparison

The membership function	Values in Server failure context	Values in Baccalaureate context[2]
Absolutely high	Y=0 in [0, 7], y=1 in [8, 10] and y=x-7in [7, 8].	Y=0 in [0.80], y=1 in [90,100] and y=0.1x-8 in [80, 90].
Ratherhigh	Y=0 in [0, 6] and [9, 10], y=1 in [7, 8], y=x-6 in [6, 7] and y=-9-x in [8,9].	Y=0 in [0.70], y=1 in [80, 90], y=0.1x-7 in [70, 80] and y=-0.1x+10 in [90,100].
Somewhathigh	Y=0 in [0, 5] and [9, 10], y=1 in [6, 8], y=x-5 in [5, 6] and y=-9-x in [8,9].	Y=0 in [0.50] and in [90,100], y=1 in [60, 80], y=0.1x-5 in [50, 60] and y=-0.1x+9 in [90,100].
Average	Y=0 in [0, 4] and [7, 10], y=1 in [5,6], y=x-4 in [4, 5] and y=7-x in [6, 7].	Y=0 in [0.40] and in [70,100], y=1 in [50,60], y=0.1x-4 in [40,50] and y=-0.1x+7 in [60,70].
Somewhatlow	Y=0 in [0, 6] and [9, 10], y=1 in [7, 8], y=x-6 in [6, 7] and y=-9-x in [8,9].	Y=0 in [0.30] and in [60,100], y=1 in [40,50], y=0.1x-3 in [30,40] and y=-

		0.1x+6 in [50,60].
Ratherlow	Y=0 in [0, 1] and [5, 10], y=1 in [2, 4], y=x-1 in [1, 2] and y=-x+5 in [4, 5].	Y=0 in [0.10], y=1 in [20,40], y=0.1x-1 in [10,20] and y=-0.1x+5 in [40,50].
Absolutelylow	y=1 in [0, 20].	Y=0 in [0.50] and in [90,100], y=1 in [60,80], y=0.1x-5 in [50,60] and y=-0.1x+9 in [90,100].

5. Conclusion

We presented the context of our study which concerns the two data warehouses, that of the Baccalaureate and that of Server failures. We were interested in the analysis phase of data warehouses where we were able to formulate MDX queries with natural terms using fuzzy logic and we explained some returned results. In this paper, we carried out a comparative analysis between the membership functions defined for each data warehouse and we found that even if the term of the natural language integrated into the model retains all its meaning, but the definition of the membership function which corresponds to it differs from one field of study to another, which has enabled us to affirm that the definition of the values of the membership function is strongly linked to the field of study.

Some perspectives emerge from our work, namely the insertion of new terms and the design of new query methods such as an expert system or a chatbot. Also the experimentation of the models in other establishments of the same domains to better test the defined membership functions.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] D. Hammouche, K.Atif, M.Loukam, "An An expert system for fuzzy server failure data warehouse analysis," ICIST'20, Lecee, Italy, 2020, doi:10.1504/IJDSS.2021.113767.

[2] D. Hammouche, K.Atif, "An extended solution to recommend fuzzy MDX queries for decision makers by a collaborative filtering profile," International Journal of Decision Support Systems **4**(3), 257-270, 2021, doi:10.1504/IJDSS.2021.113767

[3] D. Hammouche, M. Loukam, K.Atif, K.W. Hidouci, "Fuzzy MDX queries for taking into account the ambiguity in querying the Baccalaureate Data warehouse," Control, Decision and Information Technologies (CoDIT), 4th International Conference, Barcelone, 179-183, 2017, doi:10.1109/CoDIT.2017.8102587.

[4] D. Hammouche, L. Metchat, K.Atif, K.W. Hidouci, "A solution to recommend fuzzy MDX queries for decision makers by collaborative filtering profile in baccalaureate Data warehouse," In proceedings of the 2nd Conference on Computing Systems and Applications, Algiers,190-199, 2016.

[5] A. Laurent, S. Gangarski, C. Mar-Sala, "Cooperation between a fuzzy knowledge extraction system and a multidimensional database management system, " Francophone Meetings on Fuzzy Logic and its Applications, La Rochelle, France, Cepaduèce éditions, 325-332, 2000.

[6] A. Laurent, "Fuzzy multidimensional databases," Advanced database days, Agadir, Maroc, Hermès, 107-117, 2001, doi: 10.1007/978-3-642-10663-7-4.

[7] A. Laurent, FUB et FUB MINER: deux systèmes pour la représentation, la manipulation et la fouille de données multidimensionnelles floues, Information Interaction Intelligence, **3**, 37-83, 2003.

[8] C. Favre, A. Laurent,Y. Pitarch.,P. Poncelet, "Représentation graphique des hiérarchies contextuelles : modèle avec satellites," EDA 2011, Clermont Ferrand, Revue des Nouvelles Technologies de l'Information, Vol. B-7, 23-37, 2011, corpus ID:38627526.

[9] G. Smits, O. Pivert et T. Girault, "ReqFlex: Fuzzy Queries for Everyone,'' In proceedings of the VLDB Endowment, 1206-1209, 2013, doi: 10.14778/2536274.2536277.

[10] Y. Pitarch, A. Laurent, and P. Poncelet," A conceptual model for handling personalized hierarchies in multidimensional databases," In Proceedings of the International Conference on Management of Emergent Digital EcoSystems, France, 107–111, 2009, doi:10.1145/1643823.1643843.

[11] D. Fasel, K.Shahzad "A data warehouse model for integrating fuzzy concepts in meta table structures" 17th ECBS, Oxford, England, 100-109,2010, doi:10.1109/ECBS.2010.18.

[12] D.Dubois, H. Prade, "Using fuzzy sets in flexible querying: Why and how?," In Proc. Workshop on Flexible Query-Answering Systems, 89-103, 1996, corpus ID:33050805.

[13] D. Perez, M. J. Somodevilla, I. H. Pineda. "Fuzzy spatial data warehouse: A multidimensional model," In Eighth Mexican International Conference on Current Trends in Computer Science, 2007, doi: 10.5772/39389.

[14] D. Dubois and H. Prade, Fundamentals of fuzzy sets, **7**, 1-653, 2000, ISBN: 978-1-4615-4429-6.

[15] A.L. Zadeh , "Fuzzy sets," Journal of Information and control, **8**(3), 338-353,1965, doi:10.1016./S0019-9958(65)90241-X .

[16] M. Golfarelli, S. Rizzi., Data Warehouse Design: Modern Principles and Methodologie, Osborne/McGraw-Hill, 2009.

[17] T. Chikawa, M. Hirakawa, "ARES: a relational database with the capability of performing flexible interpretation of queries," In IEEE Transactions on Software Engineering, 624–634, 1986, doi:101109/TSE.1986.6312958.

[18] R. Agrawal., A. Gupta, "Modeling multidimensional databases," In Proceedings. 13th International Conference on Data Engineering, 232–243, 1997, doi:10.1109/ICDE.1997.581777.

[19] R. Bliujute, S. Saltenis, G. Slivinskas, C. Jensen, "Systematic Change Management in Dimensional Data Warehousing," In IIIrd International Baltic Workshop on Databases and Information Systems, Riga, Latvia, 27–41, 1998, corpus ID:10165485.

[20] R. Agrawal, A. Gupta, A. Sarawagi, "Modeling Multidimensional Databases," ICDE'97, 1997, doi:10.1109/ICDE.1997.581777.

[21] C. W. Holsapple, .K.D. Joshi, "Organizational knowledge resources," Decision Support Systems, **31**, 39-54, 2001, doi: 10.1016/S0167-9236(00)00118-4.