# Natural Language *Why-Question* Answering System in Business Intelligence Context

Rahma DJIROUN[1], Meriem Amel GUESSOUM[1], Kamel BOUKHALFA[1], and El hadj BENKHELIFA [2]

[1]Department of Computer Science. LSI/USTHB, Algiers, Algeria
[1](rdjiroun,mguessoum,kboukhalfa)@usthb.dz
[2]Cloud Computing and Application Research Lab. Staffordshire University, UK
[2]e.benkhelifa@staffs.ac.uk

***Abstract*—Business Intelligence is the key technologies that ensures effective decision making through extracting relevant information and providing adapted systems as the Data Warehouses. To access decisional information, the decision maker should express his requirements in Natural Language interfaces without any technical skills, avoiding the *IT-Designer* intervention. Often, the decision maker's requirements are expressed as WH-questions ("*What, Who, Where, etc.*") or Keyword-like questions. In this paper, we emphasize on a *"Why-Question"* asked in Business Intelligence context. This question has not been well dealt in the literature in terms of produced answers. Indeed, to respond this type of question, it is necessary to provide explanations. These explanations are determined by identifying causal relationships between the phenomenon highlighted in the *Why-Question* and factors that can influence this phenomenon. In this context, we propose an approach on which a system can address a causality problem related to answering a decisional *Why-Question*. To validate our approach a tool called *"BI Why Q/A"* is developed. In order evaluate our proposal in terms of efficiency and relevance, a set of experimental studies is carried out and presented.**

***Index Terms*—Business Intelligence, Data warehouse, Natural language, *Why-Question*, Question Answering, Causality.**

## I. Introduction

The decision making process in enterprises is often complex when the decision maker is confronted with choices that require stakes and options analysis leading to the final decision. In this perceptive, Business Intelligence (BI) systems had been developed to ease the decision making process as much as possible. BI systems provide decision makers with an overview of the enterprise's activities in order to make effective decisions. The BI system relies on a data repository known as the Data Warehouse (DW).

To make a decision through DW's exploration, the decision maker analyses a set of *"measures"* capturing the enterprise's activities regarding several analysis axes (*"dimensions"*). In this context, this data can't always be accessed easily by the decision maker without a minimal mastery of formal languages such as SQL and MDX. Hence, the decision maker solicits the *IT-Designer*. Indeed, the decision maker expresses his requirements in Natural Language (NL) then the *IT-Designer* translates these requirements into formal queries (SQL, MDX). This querying scenario engenders certain issues as the dependence of the decision maker of the *IT-Designer* causing mobility constraints. To palliate these problems, the *IT-Designer's* intervention can be replaced by Question Answering systems (Q/A) as proposed in [1]–[6] or search engines allowing to find data cubes as in [7]. Indeed, nowadays, BI technologies are moving towards self-service solutions (modern BI) [8] where Q/A systems serve to assist analytical conversation. In addition, researchers have moved recently towards integrating chatbots applications based on a NL Dialogue flow in order to interact with dashboards [9]. These applications enable decision-makers to ask NL queries and receive instant responses instead of navigating in the dashboard [9]. This approach has good properties such as speed, accessibility, compatibility, and interactivity over the traditional BI dashboard [9]. Notably, chatbots-driven AI technology has the capabilities to interact and communicate with users and generate human-like responses such as ChatGPT which has achieved a momentous change and made substantial progress in natural language processing [10].

In BI context, Q/A systems handle decision makers' requirements expressed as NL questions in free syntax and without any technical skills. NL questions can be categorised as WH-questions (*"What, Where, When", etc.*). Sometimes, the decision-making need can take the form of a *Why-Question* ($WQ$) such as:"*"Why has the number of accidents increased in 2019?"*. This question type is interesting when it's about knowing the origin of a phenomena observed on an activity in enterprise/organisation (decrease in sales, increase in recourse, etc.). It allows decision makers to understand some decision-making indicators such as *"cause or origin of a trend"*.

In this context, we have proposed an approach that deals with NL decisional *Why-Question*. This approach aims to provide answers that could bring help in the decision making process. These answers were a set of observations provided regarding only analysis axis, for example: regarding the *Why-Question* ($WQ$), one provided answer has been *"the human factor"*, more precisely *"the age of the driver"*; the decision that may be taken is to review more closely the conditions for obtaining the driving license for the youngest. Unfortunately, this answer proves insufficient when the explanations of the phenomenon's origin depend on other phenomenon. These explanations can be determined by studying the relations between factors influencing the phenomenon. These relations are interpreted as possible *"causal correlations"* (cause and effect relations), for example; According to the *Why-Question* ($WQ$), a question that can be raised: *"How much can climate disturbances influence the increase of the number of road accidents"*. However, to answer a decisional *Why-question*, it is necessary to go through assessing causal influence between factors related to a phenomenon. Discovering this causal knowledge can be engendered between factors located in the DW and factors extracted from external sources such as Open Data (OD) (climatological Data).

Causality has been studied extensively in a wide range of disciplines including Psychology, Philosophy and Computer Science [11]. Discovering causality is a challenging and important task that aids in planning and decision making in several fields [12]. For example in medicine, determining the cause of a disease helps in the prevention and the treatment [12].

In computer science, causality analysis continues to remain one of the fundamental research questions and the ultimate objective for a tremendous amount of scientific studies [13]. It is driven by the instinctive desire of knowledge and has been considered one of the fundamental studies regardless of the research area in a broad sense [13].

In the literature, several approaches have been proposed to deal with causality's issues in Information Retrieval (IR) field related to *Why-Question* answering systems as in [14]–[34]. Nevertheless, these approaches prove unfortunately insufficient when it comes to dealing with decisional questions. Researchers report that a NL *Why-Question* is qualified as complex [1] for which the expected results require particular methods to provide them. The most appropriate model to use must take into consideration the multidimensional aspect characterizing a DW. Actually, when a decision maker has a decisional need, he generally expects a mean that helps him to make quick analyses for an effective decision making. These analyses are generally carried out on the basis of the DW's multidimensional aspect that reveals the DW's concepts as well as the relations between these concepts (facts, measures, dimension, hierarchies, dimension's levels).

To the best of our knowledge, no work and even our approach [35] have addressed the causality analysis problem related to a decisional NL *Why-Questions* in BI context. To cope with this issue, we propose in this paper, a much more robust approach than our preliminary one [35]. This approach targets to provide quite satisfactory *Why-Question's* answers to the decision maker extracted from the DW and external data sources. These answers are attached to a set of causal factors having an impact on the phenomenon highlighted in the asked NL *Why-Question*. The core of this approach, is, first, a model [36] that describes our causality perception in a BI context by emphasising on the concept *"event"* and then a statistical method that we have proposed in [36], that fits this model. This proposal aims at evaluating the causal influence between factors [36] related to quantitative data located in the DW as well as to qualitative/quantitative data extracted from external sources.

The paper's structure is as follows. In the section II, we present a motivation example useful in the approach's unfolding. The related works are outlined in section III. In section IV, some definitions necessary to understand our proposal are presented. The proposed approach is described in section V. The implementation and experimental study details are given in section VI. Finally, in section VII, we draw some conclusions and define future lines of research.

## II. Motivation Example

In order to illustrate our approach, we present in this section a case study that will be used in the rest of this paper. In this case study, we present (1) the DW that we use to unfold our approach; (2) a set of NL *Why-Question* that can be asked by decision makers regarding the target DW; and (3) examples of external data sources interesting to answer the *Why-Question*. More details are presented in the following.

1) **"Microsoft Adventure Work 2020 Data Warehouse"** We have considered the *Microsoft Adventure Works-DW 2017* [1]. It's schema is illustrated in figure 1. This schema concerns a DW designed and fuelled to cover sales, purchases, products, customers and some human resources. This DW includes several measures that can be analysed according several perspectives. The measures are related to two main activities: *"internet sales"* as well as *"reseller sales"*. These measures are: *"sales amount, tax amount, freight-transport, order Quantity, discount amount"*. The Microsoft AdventureWorks DW allows to analyse the activity *"Internet sales"* according to the dimensions: *"customer, product, date, territory, currency, promotion"*. The *"reseller sales"* activity is concerned by the dimensions: *"employee, product, date, territory, currency, promotion, reseller"*. In addition, the decision maker may be interested in the *"product inventory"*, for which the measures are *"unit cost, unit balance"* and the dimensions are *"date, product"*.
2) **"Why-Questions' Basis"** After analysing the conceptual DW schema presented above, we have built a *Why-Question's* basis accessible at https://wq-bi.jimdo.com/. This questions' basis is produced from several combinations made between the different DW's components and according to questions those can be asked in company environment such as *"why the company hasn't evolved this year ?"*.
   A decisional *Why-Question* is composed mainly of a set of *"measures and dimensions ( dimension's levels and members)"*. A NL decisional *Why-Question* can be classified into two categories: explicit and implicit. In the first and the second category, we focus on the *"measure"* component whether is explicit i.e. expressed in the NL *Why-Question* or not i.e. implicit. Both of these two categories are divided into two sub categories, depending on the presence of the "dimension" component in the *Why-Question* (implicit or explicit). The corresponding

[1] https://github.com/microsoft/powerbi-desktop-samples

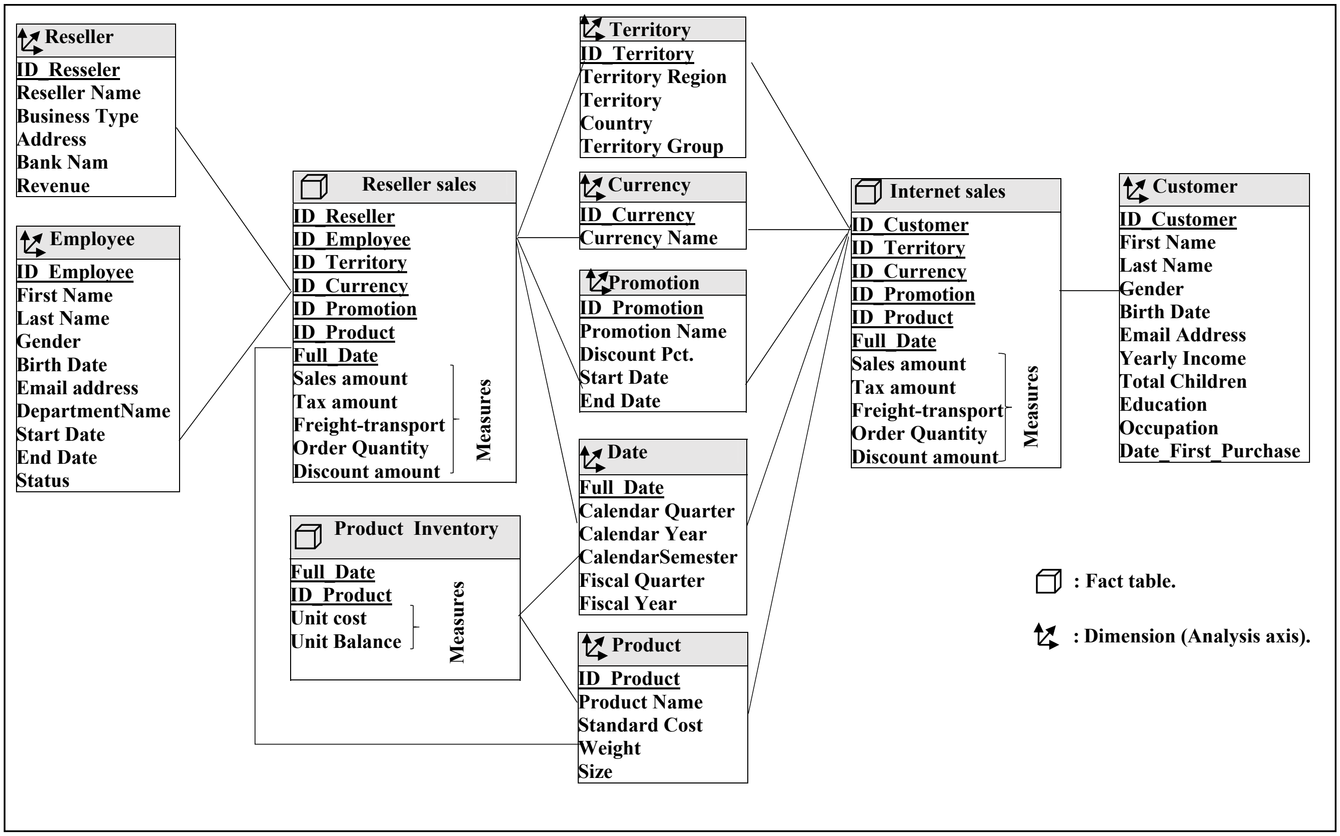


Fig. 1: Microsoft Adventure Works Data Warehouse-2020 schema.

TABLE I: Decisional *Why-Question* classification

| Category | Sub-category | | Examples |
|---|---|---|---|
| 1. Explicit decisional *Why-Question*. | 1. Explicit measure | 1. Implicit dimension | - Why has **internet sales amount** decreased?<br>- Why have **internet sales and reseller's sales** decreased? |
| | | 2. Explicit dimension | - Why has **internet sales amount** decreased during the *years* comprised between 20017 and 2019?<br>- Why has **internet sales amount** increased in *USA*? |
| 2. Implicit decisional *Why-Question*. | 1. Implicit measure | 1. Explicit dimension | - Why have *customers* become more and more demanding?<br>- Why the *product p1* is sold more than the *product p2*?<br>- Why do *employees* resign? |
| | | 1.Implicit dimension | - Why the company hasn't evolved ? |

examples are presented in table I, in which, the measures and the dimensions are respectively illustrated with bold and italic fonts.

Let us assume that the decision maker asks the *Why-Question*: *" Why has the internet sales amount decreased?"* ($Q_1$) (see table I). To answer this *Why-Question*, it is necessary to provide the decision maker with indications that aim at making him understand the phenomenon *"Internet sales decrease"*. In other words, the decision maker expects plausible explanations that expose the causes of the phenomenon *"Internet sales decrease"*. An eventual answer for the *Why-Question* $Q_1$ would be a *"Internet sales amount has decreased because a significant decrease in the internet sales order quantity has been observed"*. Consequently, it becomes necessary to dig into the DW's historical data in order to reveal the causal correlations existing in this DW. This correlation can be determined by assessing the influence of factors on the phenomenon *"Internet sales decrease"*. However, exploring only the DW to extract answers provides partial answers and sometimes limited ones. Indeed, the answers can be external to the DW as the *"meteorological data, external events, etc."*. In this case, it will be necessary to evaluate the influence of factors located in external sources to the DW on the phenomenon as the *"Internet sales decrease"* as exposed in the following.

3) **"External Data Examples"**

In order to identify the influence of certain factors not

TABLE II: External Data Examples

| Data Source | Description | Why this data source? | URLs |
|---|---|---|---|
| Climate Change Knowledge Portal, developed by the World Bank Group | This dataset contains historical climatological data as the temperature and the precipitation (rainfall) data for several countries and basin levels | - The weather dictates consumer behaviour. Thanks to Climate Business Intelligence, a global vision of the impact of weather conditions on your business can be acquired. - Climate Business Intelligence is the study and understanding of the sensitivity of an activity compared to the weather in order to optimize its strategy | climateknowledgeportal.worldbank.org |
| Election Calenders | Calendar of key elections dates and deadlines in USA co and France countries | During periods of political events as "election's events", several activities can be affected, for example internet activities or e-commerce. | For instance, for the France country, presidential elections informations for 2012 and 2017 are published on the opendata web site : https://opendata.paris.fr/ explore/dataset/elections-presidentielles-2017-1ertour *Other URLs:* data.amerigeoss.org/dataset/election-event-calendar data.oregon.gov/Administrative/Oregon-Elections-Calendar data.montgomerycountymd.gov/Elections/Election-Event-Calendar |
| Holidays Calenders | Calender of religious holidays dates | Calendar events such as religious holidays (Christmas) are events that allow e-merchants to realise a significant turnover from theirs online sales. | www.interfaith-calendar.org/ |

located in the DW, we assume in this paper that a domain expert specifies the data relating to these factors. This information may be of interest to decision makers and can be found in external data sources as described in the table II.

By analysing the table II, the questions that can arise in this case study according to the *Why-Question* $Q_1$ are: *"can climate disturbances, political and religious events cause a decline in the internet sales amount"?* and *"how much the climatic conditions and events (political and religious holidays) can influence the internet sales activity"?*. Hence, the decision maker needs a solution that brings relevant answers to his *Why-Question*, provided on the basis of causal correlations. These causal relations can be determined by assessing the influence between a set of factors (located in the DW and in external sources) and an observed phenomenon in the DW.

## III. Related Works

In this section, we present a literature overview that exposes a summary of some related works addressing the problem of dealing with NL *Why-Question*. These works concern two research communities. The first one is related to works those focused on approaches proposed to extract causal knowledge to deliver answers for a NL *Why-Question* in the Information Retrieval IR field as in [14]–[30], [32]–[34], [37]. The second community concerns approaches that handle a NL *Why-Question* in Business Intelligence BI context [35].

In the first community, most of works aim at proposing approaches in order to develop *Why-Question* answering systems. Answering a *Why-Question* is usually a challenging task because the asking point can not be simply mapped to the defined knowledge [20]. Answering a *Why-Question* in the IR field (e.g., *"Why are tsunamis generated?"*) consists to retrieve concise answers (clauses, sentences or paragraphs) from textual documents ( [14]–[19], [21]–[24]). However, in [20], the author deals with a formalized *Why-Question* (*"Why X is important to Y"*) with respect to a knowledge base in Biology domain. The expected *Why-Question's* answers are usually explanations provided by recognizing automatically the cause and effect relations, expressed with explicit cues as *"because"* [22] or not in textual passages. These causal relations can be identified using causal relations extraction techniques as in ( [23]–[30], [32], [33]), whether are non-statistical techniques (linguistic and semantic pattern matching, connetctive methods) or statistical and machine learning techniques (pattern classification, supervised or non-supervised machine learning techniques (neural networks)) [11]. Indeed, in [14] and [21], authors propose approaches that detect automatically causal relations in English as well as Japanese texts. Those methods are based on a set of lexico-syntactic patterns that relied on the existence of causal expressions such as: *Tsunamis* ***are caused by*** *the sudden displacement of huge placement of water*.

In [16] and [37], authors propose an approach based on the analysis of the discourse of English and Arabic documents respectively. The discourse analysis approach is performed on the basis the *"Theory of Rhetorical Structure"* (TSR). This method generates a tree that describes causal relationships existing between parts of the text and explains the coherence by postulating a hierarchical structure. In [16], it is reported that 75 % among the instances of the relations expressed in these answers, represent explanations and arguments. While in [37], the *"Lemaza"* Q/A system provides results for which the recall and precision are 72.7 and 79.2, respectively.

In [19], the authors present an approach applied to a Japanese text corpus. This approach is based on specific patterns inspired by the observation that a *Why-Question* and their answers often follow the fact that if something desirable or undesirable happens, its reason is also desirable or undesirable respectively. The authors combine this principle based on sentiment analysis, with the semantic word classes

idea using clustering algorithms. This approach leads to capture *associations* between word's classes expressed in the *Why-Question* and the words located in the expected answers. Authors in [26] propose an approach based on the technique of *"Word-embedding"* as well as on lexico-syntactic models. This approach aims to extract answers according to the causal relations that appear in contexts close to that of the *Why-Question*, with minimal supervision.
In [23], a causal graph is designed on the basis of a linguistic pattern to visualise explanations related a *Why-Question*. This question is emitted by ordinary people on community web broad for diagnosing plant disease problems.
Authors in [24], propose an approach for the Japanese Q/A system *"NAZEQA"*. This approach enables to acquire automatic lexico-syntactic models from corpora, annotated by causal relations. These models are used to create characteristics related to causality. These features are lexical, syntactic and semantic such as n-grams or morphemes and syntactic dependencies, allowing to improve the selection and the classification of relevant answers.
In [22], authors propose an approach that answers *Why-Question*, by recognizing causal relations expressed in Japanese text archives. Authors propose the mechanism of *"Causality attention"* and a neural network model to extract causal structures.

In BI community, we have proposed in [35] a *Why-Question's* model that aims to formalize this question in terms of components and constraints. On the basis of this model, an approach is built to procure answers to decision makers. To provide these answers, a mathematical model has been proposed (*"trend's function"*). The answers have been a set of observations produced according to the DW's axis analysis to bring help in the decision making process.

We summarize all the works presented above as shown in table III. We compare these works according to the following criteria:

- *Approach:* this criterion refers to the types of techniques used in each approach for extracting answers to a question *Why-Question*. To this end, the authors have proposed solutions, which are at the intersection of *linguistic approaches*, *data mining* and *machine learning* techniques. However, in the work of *Baral et al.* [20], the extraction of responses is done by querying a knowledge base.
- *Input:* this criterion refers to the *Why-Question* entered in the Q/A system as well as to the corpus queried by this system. We have found that in some works like [20], the *Why-Question* is formalized according to a model while in several works the *Why-Question* is not formalized. In most of these works, the targeted corpus is a *document* with the exception of [20] where the approach extracts answers from a knowledge base.
- *Output:* represents the result returned by the Q/A system. The results refer to cause and effect semantic relationships. These relationships are captured in an automatically generated causal graph or in concise answers (clauses, sentences or paragraphs). These last represent the causes having provoked the topic/event on which the user wonders in his *Why-Question*. These causes are, in general, identified by explicit causal expressions such as *"because, caused by, etc."* and sometimes implicit as raised in [22]. In this case, the answers do not follow any formalism while in [20], the answers are formalized.
- *Objective:* represents the main goal of each approach in terms of performance and relevance.

After analysing all the works presented above, we elucidate what follows:
- (a) Authors propose to answer a *Why-Question* in the IR field, on the basis of a causality analysis process. This process aims to extract cause and effect relationships from text.
- (b) The approach proposed in BI context [35] has not well dealt with a decisional NL *Why-Question* in terms of provided answers. Indeed, authors have not studied thoroughly all possible correlations reflecting *causal relations* contributing to providing significant answers supporting decision making process.
- (c) In BI context, the *Why-Question's* answers should be provided according to:
- (i) The DW's multidimensional representation (fact, measure, dimension, etc.). Therefore, the decision maker must have the DW decisional indicators in the answers of his question.
- (ii) An approach carried out on the basis of a causality analysis, aiming to extract causal relationships as proposed in the solutions that deal with a *Why-Question* in the IR field.
- (d) Thus, addressing a decisional NL *Why-Question* in BI context is quite different than treating a *Why-Question* asked for IR purposes because the nature of the questions, corpus and required answers are different. In addition, the approaches proposed in the IR field do not take into account the multidimensional aspect characterizing DWs. Hence, we can't fully adopt the approaches proposed in the IR field in our context.
- (e) To the best of our knowledge, we can highlight that no work has been proposed in order to answer a decisional NL *Why-Question* on the basis of a causality analysis approach.

To analyse a causality problem related to a *Why-Question* asked in BI context, we have proposed in [36] a model that aims to capture the crucial concepts allowing the extraction of causal knowledge in our context. In this model, we foucs on the concept *"event"*. To deal with a causality analysis problem related to our context, we have proposed in [36] a statistical method that quantifies causal influence between events extracted from DW (measures) and external sources. On the basis of this model and the causality analysis method [36], we propose in this paper, an approach that deals with a NL *Why-Question* and enables extracting answers built around factors influencing the phenomena specified in this asked question.

## IV. Preliminary Definitions

In this section, we present the most important definitions necessary to understand our approach.

TABLE III: Related Works

| Criteria | | | | | Works | | | | | | | | | | | | Our app |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | [14] | [15] | [37] | [24] | [27] | [18] | [20] | [19] | [21] | [26] | [24] | [22] | |
| Approach | Linguistic | | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | |
| | Text mining | | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | |
| | Automatic Learning | | | | ✓ | | | | | | | ✓ | ✓ | ✓ | ✓ | ✓ | |
| | Novel | | | | | | | | | | | | | | | | ✓ |
| Input | *Why-Question* | Formalised | | | | | | | | | ✓ | | | | | | ✓ |
| | | Not Formalised | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | |
| | Corpus | Documents | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | |
| | | Knowledge base | | | | | | | | | ✓ | | | | | | |
| | | DW, Open Data | | | | | | | | | | | | | | | ✓ |
| Output | Causal relations | Answers | Formalised | | | | | | | | ✓ | | | | | | ✓ |
| | | | Not Formalised | Paragraphs | | ✓ | ✓ | ✓ | ✓ | ✓ | | | | | ✓ | | |
| | | | | Clauses, phrases | ✓ | | | | | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | |
| | | Causal Graph | | | | | | | | | | | | | ✓ | | |
| Obj | Performance | | | | | | | | | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ |
| | Relevance | | | | | | | | | | ✓ | | | | | | |

***Definition 1.*** A DW is modelled in a snowflake or fact's constellation schema. It is composed of a set multidimensional elements ($ME$). The $ME$ are a set of fact tables ($F$) composed of a set of measures ($M$) $/M = \{m_i\}$ where $i = 1..n$, a set of dimensions ($D$) $/D = \{D_j\}$ where $j = 1..m$. Each $D_j$ is described via a set of attributes ($A$) $/A = \{a_k\}$ and $k = 1..p$. A dimension $D_j$ is provided or not with a level of hierarchy ($L$) $/L = \{l_t\}$ where $t = 1..s$, we note so a dimension as: $D_j[l_t^+[a_k]]$.

***Definition 2.*** A dimension can reference temporal information and non temporal ones, necessary to carry out decisional analysis. We note a *"temporal dimension"* $Dt_j$ such as the dimension *"date"* and the *"non temporal dimension"* $D_j$ as *"Customer, Product"*.

***Definition 3.*** A decisional *Why-Question* ($Q$) is composed of a set of DW multidimensional elements $ME$. $Q$ must comport at least one *measure* $m_i$. The referenced *dimensions* are the *temporal dimension* $Dt_j$ to specify the time and the *non temporal dimensions* $D_j$.

The model capturing the decisional NL *Why-Question's* components is illustrated in the figure 2. ***Definition 4.*** The notion of *"trend"* ($Tr$) is included in a *Why-Question* $Q$. A *trend* is a changing observed on an activity during a given *period* such as: *decrease, increase, high, low, stagnation, change, stability*, etc.

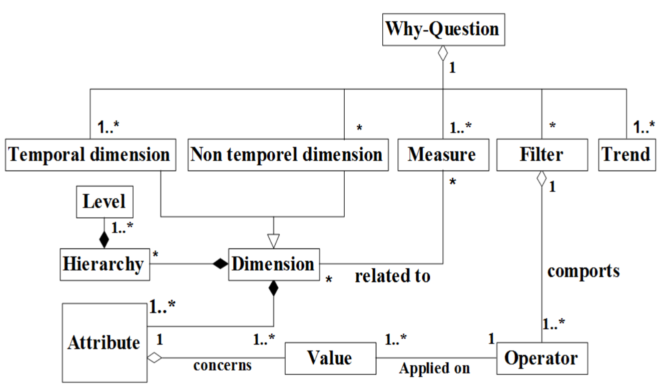

Fig. 2: Decisional *Why-Question* model.

***Definition 5.*** The *Why-Question* can or not comport filters ($f$). A *filter* $f$ consists to apply a restriction on the values ($V$) of the attributes of a dimension such as $V = \{v_1, ...v_e, ...v_r\}/e = 1..r$. A filter $f$ is defined according to a set of operators ($OP$) such as $OP \in \{$ *equals*, *between*, *less than*, *more than*, etc. $\}$. We note a filter $f$ : $f[OP][D_j[l_t^+[[a_k[v_e]]]$.

**Example 1.** *"Why has the internet sales amount decreased during the years between 2016 and 2019"* where $OP$ = *"between"* and *"2016, 2019"* is a filter to apply on an attribute of the dimension *"year"*.

***Definition 6.*** A phenomena ($ph$) designates a set of events ($e$) varying over time perceived by a conscious subject. A trend $Tr$ can be interpreted as a phenomena $ph$ related to the enterprise's activity. $Tr$ is deduced from the variations of the DW's measure $m_i$, for example "Internet sales amount decrease" where *"Internet sales amount"* is a measure $m_i$ and *"Decrease"* is the related trend ($Tr_{m_i}$).

***Definition 7.*** An event $e$ is something that happens, especially when it is unusual or important. The events describe all the things that happened in a particular situation over a temporal interval ($I$). An event $e$ can belong to a trend $Tr$ $/e \in Tr$ or not (located in external sources to the DW), for examples: ($e_a$) as *"lower peak in the order quantity decrease"* and ($e_b$) as *"Presidential election"* respectively.

***Definition 8.*** Identifying factors affecting a phenomena $ph$ consists in finding an event $e$ or set of events having a causal influence on this phenomena $ph$.

***Definition 9.*** Every phenomena $ph$ has a cause ($c$). Indeed, from a determined cause $c$ necessarily results an effect ($fc$); and conversely, if no specific cause $c$ is given, it is impossible for an effect $fc$ to occur.

***Definition 10.*** A causal relation ($Rc$) consists in finding the correlation between a cause $c$ and effect $fc$ where $c$ and $fc$ are two distinct events. This correlation is determined the by bias of the causal influences quantification between $c$ and $fc$.

***Definition 11.*** A *Why-Question's* answer ($A$) is a response provided on the basis of the detection of the causal relation $Rc$ between the phenomena $ph$ requested in the *Why-Question* $Q$ and events $e$.

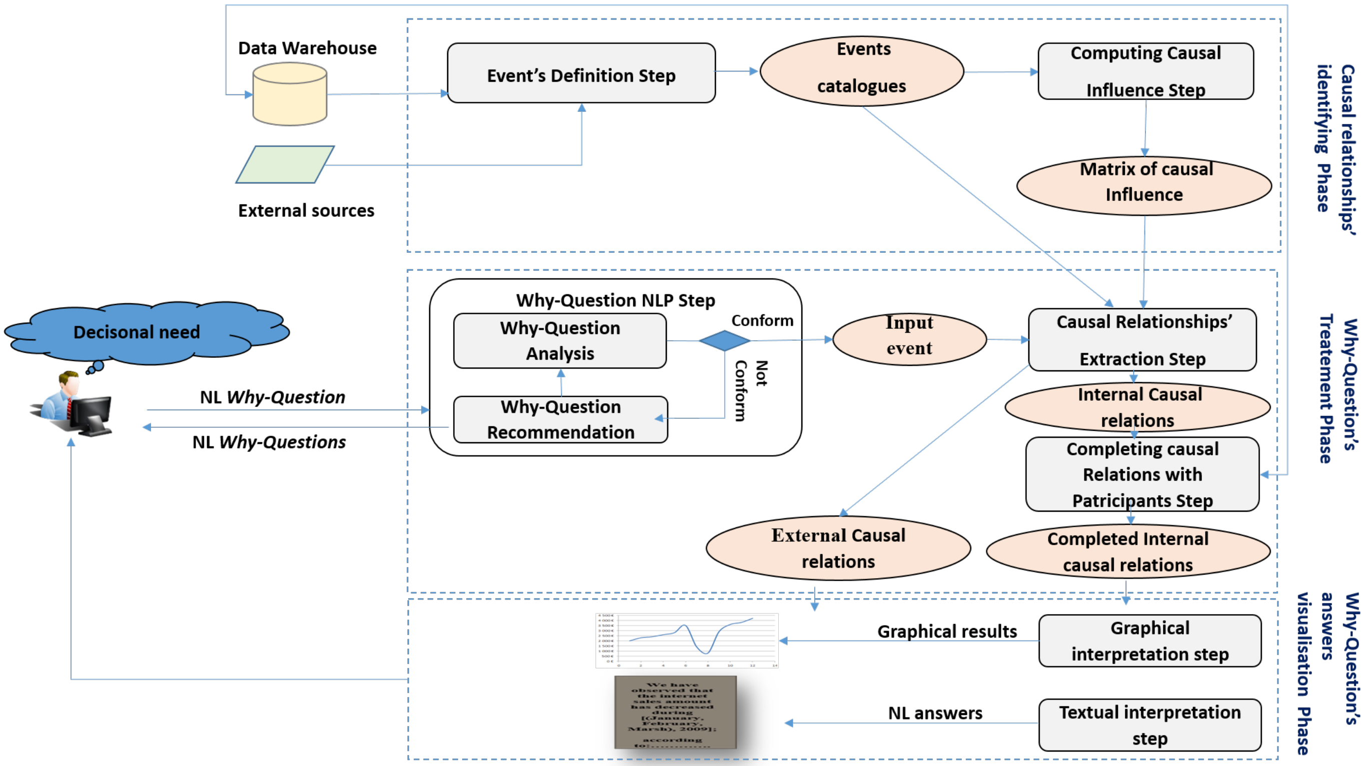


Fig. 3: General architecture of our approach.

## V. Answering NL Decisional *Why-Question* Approach Description

In this section, we present our approach that allows to answer a NL *Why-Question*. Answering a decisional NL *Why-Question* consists into pin down relevant leads contributing to the decision-making process. To deliver the decision maker with these leads, it is necessary to dig into the DW and the external data to pick out eventual causal correlations. At the core of this approach, we manipulate the concept *"event"*. Our approach consists into three main phases: (1) the causal relationships' identifying phase, (2) the *Why-Question's* treatment phase and (3) the *Why-Question's* answers visualization phase.

1) The first phase consists of a set of processes which aims at defining events $e$ in order to calculate the causal influence between these events (causes $c$ and effects $fc$), from which causal relationships $Rc$ are identified. It is worth noting that the analysis of causality is a process related in general to historical data, which fits quite well the DW context
2) The second phase takes as input the decisional NL *Why-Question* according to which it performs an NLP tasks in order to attempt causal knowledge extraction.
3) The third phase interprets the potential causal relations as NL answers and graphical results and returns them to the decision maker.

The details of each phase are presented in the remainder of this section. The general architecture of the proposed approach is depicted in figure 3.

### A. Causal Relationships' Identifying phase

Discovering causal relationships is a challenging task, because there is not a general acceptable definition of causal relationships. Causality is more a philosophical phenomenon, and it may have different meanings in different areas. This makes it difficult to expound causality in a unified form [38]. Therefore, it is necessary to set a strategy that reflects how the causality is perceived in the targeted context and what are the needs and the reasons that incite to the discovery of causality. In this perspective, we propose in this phase, to prepare the necessary data in order to mine easily the causal relations from DW and external sources. This phase consists in two steps:(a) an event's definition step and then (b) computing causal influence step. The details of each step are presented below.

*1) Events Definition Step:* In order to discover causal relations from DW and external sources for answering a *Why-Question*, we have, first, proposed in [36], a model that enables to define our causality perception in BI context. This model is inspired from the event models proposed in [39] called the *"causality pattern"* and the *"participant pattern"*. Therefore, we have emphasised, in our model, on the concept *"event"* which its perceptions, in the real world, heavily depends on the context and point of view of the observer [39].

- The causality pattern defines the concept *"event"* classified into *"cause"* and *"effect"* and the concept *"jus-*

*tification"*. This pattern explicitly expresses the causal relationship between the cause and the effect under the justification of some theory. A theory might be an opinion, a scientific law, or not further specified, for example, during a heavy storm, a power outage might occur caused by a snapped power pole. The Justification of this causal relationship is the laws of physics [39].
- The participant pattern enables to formally express the participation of *"objects"* in events ( person, place, designed artifact). This pattern defines the concept *"situation"*. This concept includes the *"event"* being described and the *"objects"* participating in this event [39]. In this pattern,the authors define the concept *"parameter"*. This concept refers to *"time parameter"* that describes the general temporal region when the event happened. It parametrizes a time interval, for example, one can state that the house re happened on June 13, 2006 [39].

In the patterns presented above, four main concepts captivated us: *"event, participant, time parameter and situation"* concepts. Thus, we consider these concepts to propose a UML model. This model as well as an example of its instance is depicted in the figure 4:

Our causality perception model is described as follows:

1) An *"event"* $e$ is classified into *"cause"* $c$ and *"effect"* $fc$.
2) The event effect $fc$ is related to a trend ($Tr_Q$) requested in the *Why-Question* $Q$.
3) We propose to specify that an event cause $c$ can be internal and external in BI context.
   - An internal event ($e_n$) is an event identified in the DW, extracted from a *"trend"* $Tr$ observed on a DW's measure $m_i$.
   - An external event ($e_x$) is an event that comes from external sources such as the climatological open data.
4) The events whether are internal $e_n$ or external $e_x$ engender a *"situation"* ($S$). This situation $S$ is triggered by a *"trend"* ($Tr_Q$) related to a measure ($m_Q$), requested in a *Why-Question* $Q$.
5) An event $e$ has as parameter *"the temporal constrain"*. This constraint is related to the temporal dimension $Dt_j$. This dimension refers to an analysis axis, omnipresent in a DW.
6) The concept *"Participant"* ($P$) is related to an internal event $e_n$ and influences in a situation $S$. In BI context, the participants are objects representing the instances of the non temporal dimensions $D_j$. A participant $P$ is specified by the filter $f : f[OP][D_j[l_t^*[[a_k[v_e]]]$ (see *Definition.5* presented in section IV).
   In this model, we consider only the participants taking part in internal events $e_n$.
7) In this model, we propose to add the property *"qualifier"* ($Q_f$) to the event cause $c$. A qualifier is an adjective that allows to specify a characteristic of a thing. This latter enables to have a qualitative view of the events for an effective decision making process. A qualifier concerns an internal and external event. For an internal event $e_n$, the qualifier $Q_f$ is attributed according to the observed trend $Tr$ such as *"significant"*, *"average"* and *"weak"* increase. For the external event $e_x$, the qualifier $Q_f$ depends on the event's typology and the external data source such as: worldwide pandemic (*Covid-19*), *cultural* (musical concert), *political* (presidential election, war), *economic* (drop in oil price), *scientific manifestation* (International conference, symposium), *natural disaster* (earthquake, avalanche), *sports event* (football world cup), *social* (back to school), *religious* (religious celebration) and *historic events* (fall of the Berlin wall, Algiers' battle), etc.

**Example 2.**
We assume in the instances model (see figure 5)that the decision maker asks the $WQ_2$: *"Why has internet sales amount decreased in 2020?"* with respect to the DW *"Microsoft adventure work 2020"* [2]. In this question, the requested trend is *"decrease"* regarding the measure *"internet sale amount"*. To answer $WQ_2$, we need to know the events that cause the *"internet sales decrease"*. To this end, we have to extract a set of events regarding the trends observed in 2020 such as *"decrease of the order quantity of internet sales"* and from external events like the political events and storm located in external sources. The participants in the decrease of internet sales amount refer to the instances of the non temporal dimensions (*"product"*).

Thus, on the basis of the model presented in figure 4, we propose to define the *events* involved for the purpose of causal knowledge extraction.
The event's definition step takes as input the DW and the external sources to generate as output a set of events' catalogues. Thus, we have to collect all necessary information representing events as we perceive them in BI context. To this end, we define an event $e$ according to the following criteria:

1) We enumerate two types for the needed data to formalize events: *quantitative data* and *qualitative*.
   - The *quantitative data* is the numerical data. It can be additive or semi additive such as the values of the DW's measures or the numerical data that we can extract from external sources ($M^{'}$) such as *"rate rainfall"* and *"temperature"*.
   - The *qualitative data* is the non numerical data, obtained from the external sources. This data describes events that occur at precise dates such as: *"presidential election"*, *"religious holidays"*.
2) Internal events $e_n$ belong to trends $Tr$ observed on all the DW's measures $M$. Indeed, DW's measures $M$ are the clues for providing significant answers related to a decisional NL *Why-Question*.
3) We formalise two types of external events as quantitative external events($e_{x_{QN}}$) and qualitative external events($e_{x_{QN}}$).

[2]https://github.com/microsoft/powerbi-desktop-samples

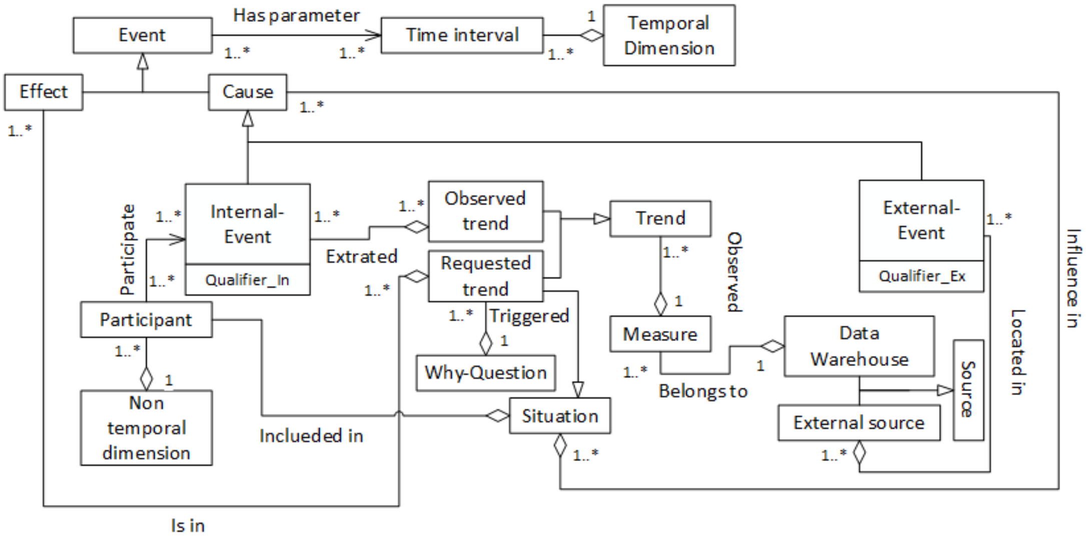


Fig. 4: Our Causality perception model in BI Context [36]

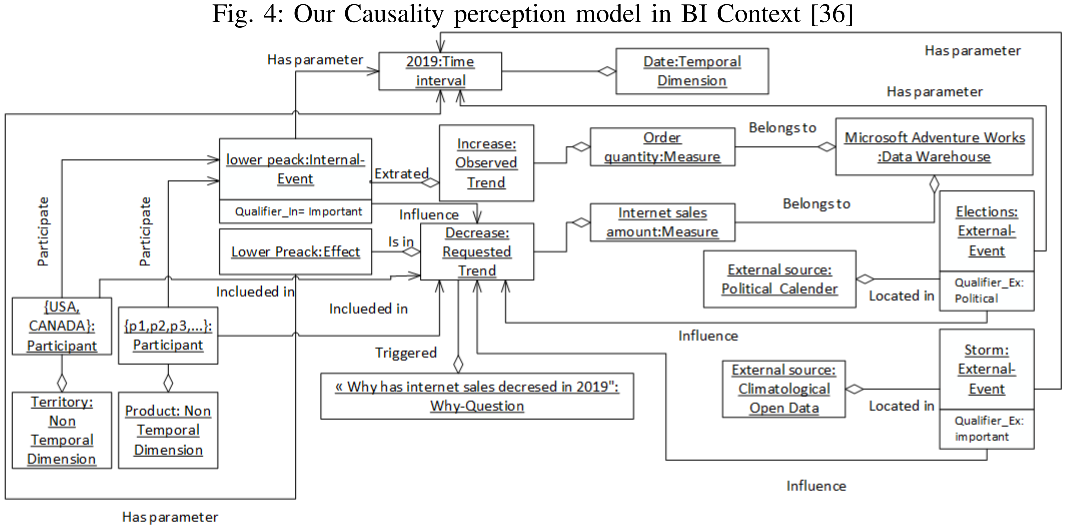


Fig. 5: An instance of this model [36].

- $e_{x_{QN}}$ are events extracted from trends $Tr$ observed according to numerical data $M^{'}$ located in the external sources.
- $e_{x_{QL}}$ are events obtained from the qualitative data that comes from the external sources.

In the remainder of this section, we explain in details the event's definition step. This step aims at transforming the data extracted from the DW and the external sources into a set of events $e$ whether are internal $e_n$ or external $e_x$. Each event is thereafter stored in the suitable catalogue which depends on the event type. This step consists into four sub steps:(1) trend's analysis, (2) trend's classification, (3) event's extraction and (4) storing events. The architecture of this step is depicted in figure 6. More details are given below.

1) **Trend's analysis**:
To capture the trends $Tr$, we have proposed in [35] a mathematical representation for the DW's measures $M$ and the external source's numerical data $M^{'}$. This representation consists into synthesizing a qualitative perception of the numerical data values (aggregated $M[V]$) and (aggregated $M^{'}[V^{'}]$) in order to generate the overall trends $Tr$. In other words, this representation is about transforming all the aggregated instances $M[V]$ for all the DW's measures $M$ and $M^{'}[V^{'}]$ according to a period ($T$) i.e the temporal dimension $Dt_j$, into a set of qualitative trends $Tr/\ Tr = \{Tr_1, ...Ttr_d..., Tr_v\}, d = 1..v$ such as: *"an increase, a decrease and a stagnation"*.
To identify these trends $Tr$, we rely on a mathematical model called *"trend function"* that we have proposed in [35]. This function describes mathematically the data $(Xi, Yi)$, where $X$ is the time period $T$ defined according to $Dt_j$ and $Y$ is $M[V]$ or $M^{'}[V^{'}]$. This mathematical model is based on the principle of non-linear regression ( [40]), for which the curve does not necessarily go

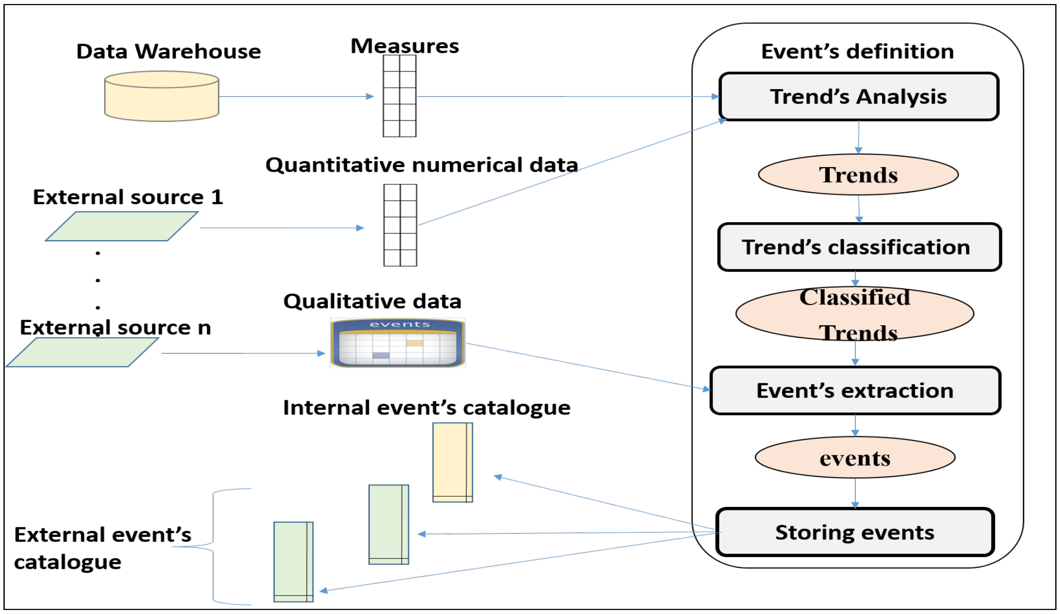


Fig. 6: Architecture of the Events Definition Step.

through all the coordinates $(xi, yi)$ but approaches them as much as possible. It allows to perform a good descriptive approach and to obtain the desired precision without being hindered by the multiple local oscillations of the numerical values $M[V]$ or $M^{'}[V^{'}]$.
The *"trend function"* can take one of these forms: polynomial, logarithmic, exponential, sinusoidal[3]:$f(Xi) = Yi$.
We have to look then for the value of $R$ (the relative error), defined as:

$$R = \sqrt{\frac{\sum_{i=0}^{k}(Y_i - f(Xi))^2}{k}} \tag{1}$$

Where $k$ is the number of the coordinates $(Xi, Yi)$ .
In this paper, we consider the polynomial function defined as as :

$$f(x) = P^n(x) = \sum_{j=0}^{n} a_j x^j \tag{2}$$

The ideal function $f(x)$ is obtained, when $R$ reaches its minimum value. This is performed, when the partial derivatives of R vanish simultaneously:

$$\left[\frac{\partial R}{\partial a_0} = 0, ..., \frac{\partial R}{\partial a_j} = 0, ...., \frac{\partial R}{\partial a_n} = 0\right] \tag{3}$$

This equation's system leads us to fix the parameters $\{a_j\}$.
To look for the order $(n)$ of the function $f(x)$, we observe the value's evolution of $(\Omega \times R)$ calculated for each value of $n$:

$$\Omega = \sqrt{\frac{\sum_{i=1}^{N-1} \Delta Y^2}{N-1}} \tag{4}$$

Where:
$N$: is the number of central points $P_i(x_i, y_i)$.
$\Delta\ Y$: is the value calculated at the central point located between $p_i(x_i, y_i)$ and $p_{i+1}(x_{i+1}, y_{i+1})$.
The optimal order $n$ will be fixed when $\Omega \times R$ reaches a minimum value.

**Example.3**
Let suppose the values $M[V]$ of the measure *"internet sales amount"* for the time period= [01/01/2017,01/01/2020], as presented in the table IV.
From these values, we build the polynomial trend function $f(x) = P^n(x)$. To do this, we look for the ideal order $n$ which allows us to have a $f(x)$
By applying this approach, we found according to the values $M[V]$ that the ideal order $n$ is 13. The trends obtained thanks to these results are as illustrated in the figure 7.

2) **Trend's classification:**
Once the trend function determined, it becomes now possible to perform a standard function's study that highlights the different aspects as: variations *"high peak, low peak, decrease, increase and stagnation"* with the respective intervals $I$ and the amplitude of the variations $\delta Y$.
In order to distinguish one trend from another, we first

[3]https://onlinehelp.tableau.com/current/pro/desktop/fr-fr/trendlines-model.html

TABLE IV: Internet sales amount values $M[V]$

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Date | 01/01/2017 | 01/02/2017 | 01/03/2017 | 01/04/2017 | 01/05/2017 | 01/06/2017 | 01/07/2017 | 01/08/2017 | 01/09/2017 | 01/10/2017 |
| Internet sales amount | 13128,84 | 16706,08 | 16874,12 | 17726,41 | 17961,33 | 21487,57 | 20914,32 | 19993,36 | 19452,29 | 21409,13 |
| | 01/11/2017 | 01/12/2017 | 01/01/2018 | 01/02/2018 | 01/03/2018 | 01/03/2018 | 01/04/2018 | 01/05/2018 | 01/06/2018 | 01/07/2018 |
| | 23653,17 | 19584,97 | 20102,34 | 16570,66 | 14667,14 | 12366,09 | 11803,78 | 16565,39 | 15605,21 | 15835,59 |
| | 01/08/2018 | 01/09/2018 | 01/10/2018 | 01/11/2018 | 01/12/2018 | 01/01/2019 | 01/02/2019 | 01/03/2019 | 01/04/2019 | 01/05/2019 |
| | 17148,11 | 16427,32 | 16074,98 | 19210,39 | 24023,26 | 28408,49 | 31084,45 | 35171,38 | 35505,06 | 51830,33 |
| | 01/06/2019 | 01/07/2019 | 01/08/2019 | 01/09/2019 | 01/10/2019 | 01/11/2019 | 01/12/2019 | 01/01/2020 | | |
| | 46454,50 | 49105,58 | 48882,88 | 48586,65 | 59393,21 | 63248,08 | 26673,27 | 1635,24 | | |

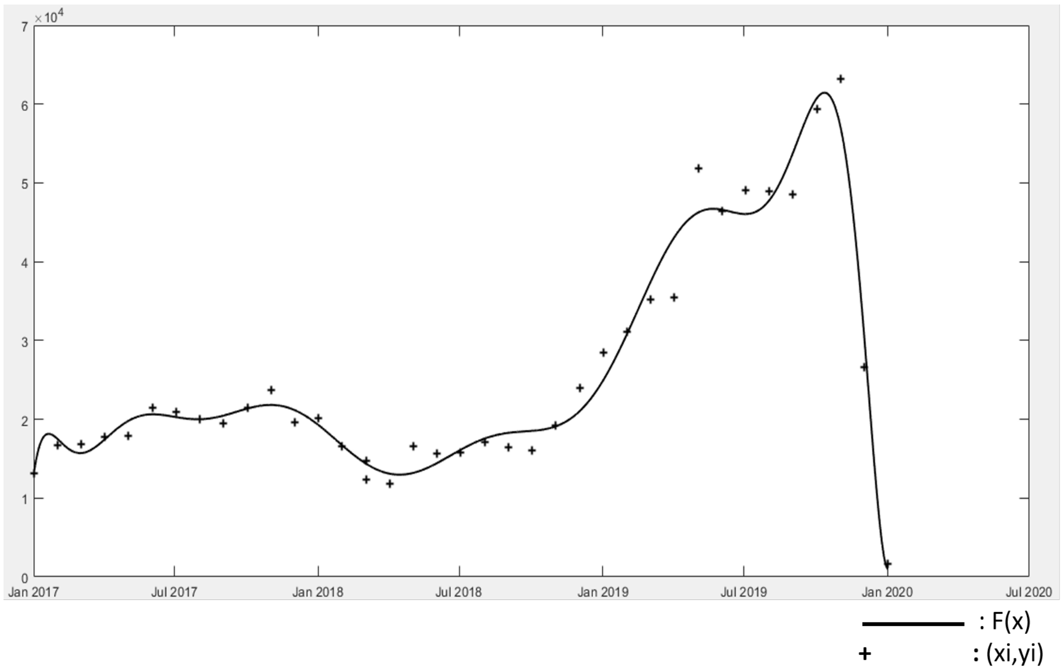


Fig. 7: The trends observed for the values $M[V]$ of the measure "Internet sales amount".

classify the trends according to the amplitude $\delta Y$ that identifies an "increase, decrease and stagnation". Then, to differentiate between two trends defined according to $\delta Y$, for example, $< Tr_1, I_1 = "increase" >$ and $< Tr_2, I_2 = increase >$, we propose to categorise these trends by assigning three qualifiers $Q_f$ as *"significant, average and weak"*.

We have partitioned the main qualifications $Q_f$ according to the following criterion, chosen arbitrarily in this paper: $Q_f =$ *"significant"* is attributed to the variations of intensity included between the maximum variation and its two-thirds. The $Q_f =$ "average" is assigned for the variations of intensity comprised between one third of the maximum recorded variation and its two-thirds. The rest of the variations will be classified with $Q_f =$ "weak".

**Example 4.**

We apply the process explained above, for the trends observed for the values $M[V]$ of the *"Internet sales amount"* measure, as illustrated in figure 7. Thereafter, we assign qualifiers to these trends according to which five categories of trends are identified: *"important increase", "average increase","weak increase","important decrease"* and *"weak decrease"* (see table V).

TABLE V: Example of classification of trends observed on $M[V]$

| Trend | Amplitude | Qualification | Start-Date | End-Date |
|---|---|---|---|---|
| *Increase* | 5193,241263 | *Weak* | 01/01/2017 | 01/20/2017 |
| *Decrease* | -2450,18973 | *Weak* | 01/20/2017 | 03/01/2017 |
| *Increase* | 4924.157105 | *Weak* | 01/03/2017 | 02/06/2017 |
| *Decrease* | -629.9123758 | *Weak* | 06/02/2017 | 07/30/2017 |
| *Increase* | 1820.019746 | *Weak* | 07/30/2017 | 10/31/2017 |
| *Decrease* | -8849,301495 | *Important* | 10/31/2017 | 04/15/2018 |
| *Increase* | 33741,51173 | *Important* | 04/15/2018 | 05/22/2019 |
| *Decrease* | -662.0074002 | *Weak* | 05/22/2019 | 07/02/2019 |
| *Increase* | 15395,5727 | *Average* | 02/07/2019 | 12/10/2019 |
| *Decrease* | -60472,92065 | *Important* | 12/10/2019 | 01/01/2020 |

3) **Event's extraction:**
In this step, we proceed to the extraction of events from the classified trends $Tr$ obtained in the previous step, as well as from the qualitative data of the external sources.
a) From the classified trends, we can enumerate seven categories of trends as: *" 1-significant increase; 2-*

*average increase;3- low increase; 4-significant decrease; 5-average decrease; 6-low decrease; 7- stagnation (very weak decrease)"*. An event $e_n$ or $e_{x_{QN}}$ can belong to these trend's categories, for instance: $e_n$ =*"low peak"* $\in Tr$ = *"significant decrease of internet sales amount"*. We set the following properties for an extracted event $e_n$:
$e_n =< Identifier_{e_n}, Measure_{name}, Trend_{name}, Q_f, Date_{starting}, Date_{end} >$.
An event $e_{x_{QN}}$ is formalized as:
$e_{x_{QN}} =< Identifier_{e_{x_{QN}}}, M'_{name}, Trend_{name}, Q_f, Date_{statring}, Date_{end} >$
b) We collect automatically the external qualitative events $e_{x_{QL}}$ by setting the following properties:
$e_{x_{QL}} =< Identifier_{e_{x_{QL}}}, Name_{event}, Q_f, Date_{starting}, Date_{end} >$.
**Example 5.** By applying the processes presented above for the measure *"Internet sales order quantity"*, we have found that this measure comprises four event categories: *"significant increase", "weak increase", "significant decrease" and "weak decrease"*.
**Example 6.**
For the trends captured in Table V, we extract a set of internal events $e_n$ as presented in Table VI.

TABLE VI: Example of events extracted from the trends observed on $M[V]$

| Event | Date | Optimum |
|---|---|---|
| Low peak of weak increase | 01/20/2017 | 18156,94467 |
| High peak of weak decrease | 01/03/2017 | 15706,75494 |
| Low peak of weak increase | 02/06/2017 | 20630,91205 |
| High peak of weak decrease | 07/30/2017 | 20000.99967 |
| Low peak of weak increase | 31/10/2017 | 21821,01942 |
| High peak of significant decrease | 04/15/2018 | 12971.71792 |
| High peak of weak decrease | 05/22/2019 | 46713.22965 |
| Low peak of average increase | 02/07/2019 | 46051.22225 |
| High peak of significant decrease | 12/10/2019 | 61446,79495 |

According to Table VI, we identified an $e_n$ = *"High peak"* extracted from $Tr$=*"important decrease of internet sales amount"*, where *"High peak"* represents the starting event of the trend *"important decrease of internet sales amount"* at the time $Date$=12/10/2019.

4) **Storing event's:**
This step aims at saving all the prior extracted events as $e_n, e_{x_{QN}}, e_{x_{QL}}$ in three structures called respectively *"internal event's catalogue (catalogue$_{e_n}$), external event's-QN catalogue (catalogue$_{e_{x_{QN}}}$) and external event's-QL catalogue (catalogue$_{e_{x_{QL}}}$)"*. Each entry in these catalogues is an event set with the proprieties defined above. The purpose of setting up these catalogues is on one hand to permanently save all the observed events regarding the DW's and external data source, until a major update is carried regarding these data sources [4]. On the other hand, this structure facilitates mining causal relationships due to the complete view that we obtain about all the extracted events, allowing us to gain in terms of processing time.
Once the event catalogues structures have been built, it becomes now possible to evaluate the causal influence between the events saved in these catalogues as presented in the following section.

[4]This depends on the periodic supplying of the DW as well as on the exploitation of new external data sources

*2) Computing Causal Influence Step:* This step targets to compute the causal influence between all the events, defined in the previous step. To this end, we rely on the causality analysis that we have proposed in [36]. This step takes as input all the events $e$ gathered in the events catalogues and calculates the causal influence between distinct events. Two events $e$ are distinct if they belong to different catalogues whether are internal $e_n$ or external $e_x$. Two events, whether are $e_n$ or $e_{x_{QN}}$ are distinct if they belong to two distinct trends $Tr$ according to different numerical measures, for example: $e_1$ = *"The beginning of a significant decrease of the internet sales amount"* and $e_2$= *"the end of an average decrease of the order quantity"*.

A causal relation $Rc$ is the relation between an event (*the cause*) $c$ and a second event (*the effect*) $fc$, where the second event $fc$ is understood as a consequence of the first $c$ [41]. In other words, the cause $c$ is the producer and the effect $fc$ is the result [42].

By definition, a cause $c$ should always occur before the effect $fc$, i.e., if an event ($e_1$) causes an event ($e_2$), $e_1$ should occur before than $e_2$ [28]. Thus, if we want to know what is the probability that the event $e_1$ causes the event $e_2$, then statistically this probability can be $Prob(e_i, e_j)/$:

$$Prob(e_i, e_j) = \frac{follow(e_i, e_j)}{Occ(e_j)} \quad (5)$$

Where $(follow)$ is the number of occasions where $e_i$ is followed by the event $e_j$.
$(Occ)$ is a causal property that represents the number of occurrences of the event $e_j$.
**Example 7.**
Let assume the example illustrated in figure 8. We consider two distinct events $e_1$ and $e_2$ and a temporal period $T$ equal to 12 months. Hence, the corresponding $Prob(e_1, e_2) = \frac{3}{5}$.

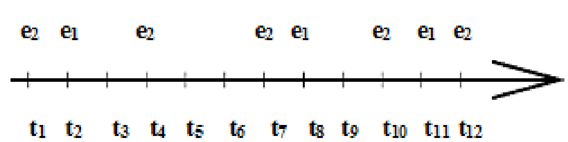


Fig. 8: Cause and effect example

**Example 8.**
If we analyse the figure 9, we notice that the event $e_1$ happened $(n)$ times and the event $e_2$ occurred only once after. Thus, we have to go through $n$ times the events $e_1$ (significant number of events) to find that the event $e_2$ has occurred once.

In this case, we can affirm that the event $e_1$ causes *surely* the event $e_2$. However, if we analyse only the events $e_1$ temporally close to the event $e_2$ in the past, we can quickly deduce that the event $e_1$ causes, certainly, the event $e_2$. Therefore, it proves necessary to consider the parameter *"Time"* in the analysis of causality. Indeed, temporal assumptions have an important impact on the judgement of causality.

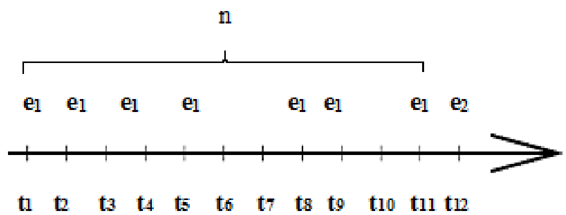


Fig. 9: Considering "Time" aspect in the causality analysis

**Example 9.** In general, the causes of economic events are expected in the near past. Thus, more an event is distant, it should be less considered as a cause.

To consider the parameter *"Time"* in causality analysis, we propose to consider a function called *"Temporal Exclusion Function"* ($TEF$). This function allows us to reduce the importance of *distant events cause* $c$ in the past and to privilege the *closer ones* [36].
The $TEF$ value will be set closer to *"1"* when the event cause $c$ occurs close enough before the event effect $fc$. This event can be considered as relevant; and close to *"0"* when the event cause $c$ was too far in the past from the event effect $fc$. In this case, the event cause $c$ is excluded [36]. A decreasing exponential function fits this use ($TEF = e^{-a*x}$ with $a \in ]0, +\infty[$). Indeed, $TEF$ must have a form, through which, periods during which the cause is expected are privileged; and intervals according which it is less likely to locate the cause $c$ are gradually excludes [36]. Thus,

$$TEF : \psi(\Delta t) = e^{-\eta.\Delta t} \tag{6}$$

where:

- ($\eta$): is as parameter which quantifies the temporal gap between an event $e$ and another (cause $c$ and effect $fc$).
- ($\Delta t$) represents the temporal gap between an event ($e_j$) and an event ($e_i$) that would be one of the probable causes to consider.
- $TEF$ is parametrized by the reference coordinates: $(t_0, \psi_0)$ according the following principle:
- Let suppose that $\bar{\Delta t}$ is the average of the temporal intervals $I$ of the reappearance (recurrence) of the event $e_j$.
- Therefore, we judge that the event $e_i$ beyond $t_0$ should be excluded by more than for example 90% that we note it as an exclusion rate ($E$). $E$ is an adjustable parameter depending on the domain and the opinion of the expert. Thus, $t_0$ is calculated as follows:

$$t_0 = \bar{\Delta t} = \frac{\sum_{i=1}^{n} (t_{e(i+1)} - t_{ei})}{Occ_e - 1} \tag{7}$$

Where $Occ_e - 1$ is the number of the temporal intervals $I$ of the appearance of the event $e_j$.
- Since, $\psi_0 = e^{-\eta.t_0}$. Therefore, we note:

$$\eta = \frac{-ln(1-E)}{t_0} \tag{8}$$

In order to assess causality influence between events, we have proposed in [36] a measure called *"Degree of Causal Influence"* ($D_{CI}(e_i, e_j)$). We define the measure $D_{CI}(e_i, e_j)$) as:

$$D_{CI}(e_i, e_j) = \frac{\sum_{k=1}^{Occ_{e_j}} e^{-\eta_{e_j}.(t_{e_{j_k}} - t_{e_{i_k}})}}{Occ_{e_j}} \tag{9}$$

Where $t_{e_j}$ and $t_{e_i}$ are the dates of starting of the events $e_i$ and $e_j$ respectively.

$D_{CI}(e_i, e_j)$ is the average of possible causes between the events $e_i$ and $e_j$ according to all occurrences of $e_j$. The possible causes between $e_i$ and $e_j$ is measured by the temporal exclusion function $TEF$. This function allows to reduce the importance of *distant events cause* $c$ and to privilege *close events cause* $c$ in the past [36]. Thus, we propose $D_{CI}(e_i, e_j)$ as a mean that extracts the degree of causality based on the principle of recurrence of close events in the past.

In order to compute the $D_{CI}$ between distinct events $e$. We propose a square matrix called *"Matrix of Causal Influence"* ($MCI$) that contains the $D_{CI}$ values calculated between all pairs of distinct events (formula 9). We define $MCI$ as follows:

$$[MCI] = \begin{array}{c} \\ e_1 \\ e_2 \\ \vdots \\ e_n \end{array} \begin{array}{c} \begin{array}{cccc} e_1 & e_2 & \cdots & e_n \end{array} \\ \begin{pmatrix} 0 & DI_{1,2} & \cdots & DI_{1,n} \\ DI_{2,1} & 0 & \cdots & DI_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ DI_{n,1} & DI_{n,2} & \cdots & 0 \end{pmatrix} \end{array} \tag{10}$$

**Example 10.**
To find the $[MCI]$ regarding the example illustrated in the figure 8, we proceed as follows:
- We start, for example, by assigning an exclusion rate $E$= 90 %. This means that we penalize the event $e_1$ by 90 % when this event is temporally distant from $e_2$ with a temporal interval greater than $t_0 = \frac{9}{2} = 4.5$ ($e_1$ appears on average every four and a half months). We should calculate then $\eta_{e_1} = \frac{-ln(1-E)}{t_{0_{e_1}}} = 0.51$.
- Thereafter, we calculate $D_{CI}(e_1, e_2) = \frac{0.69}{5} = 0.16$. We apply the same procedure to calculate $D_{CI}(e_1, e_2)$. We obtain thus: $[MCI] = \begin{array}{c} \\ e_1 \\ e_2 \end{array} \begin{array}{c} \begin{array}{cc} e_1 & e_2 \end{array} \\ \begin{pmatrix} 0 & 0.16 \\ 0.60 & 0 \end{pmatrix} \end{array}$

**Example 11.**
The table VII represents a part of the matrix $[MCI]$, obtained according to the events extracted from the DW *"MAW-2020"*.

TABLE VII: Example of the $[MCI]$ content with interpretation

| Causal relationship (event ***causes*** event) | Degree of influence |
|---|---|
| Significant-Decrease-Order Quantity Sales ***causes*** Significant-Decrease-Internet Sales Amount | 0.952 |
| Significant-Increase-Tax Amount ***causes*** Average-Decrease-Internet Sales Amount | 0.884 |
| Significant-Increase- Internet sales Order Quantity ***causes*** Significant-Increase-Internet Sales-Amount | 0.835 |
| Significant-Decrease-Order Quantity Reselle rSales ***cause*** Average-Increase-Internet Sales Amount | 0.738 |
| Christmas ***causes*** Significant-Increase-Internet Sales-Amount | 0.62 |
| Significant Increase-Order Quantity Resller sales ***causes*** Significant-Decrease-Internet Sales Amount | 0.534 |
| Low-Increase-Order Quantity Sales ***causes*** Average-Decrease-Internet Sales-Amount | 0.532 |
| Significant-Temperature-Increase ***causes***Average-Increase-Internet Sales-Amount | 0.37 |
| Low-Increase-Order Quantity Reseller Sales ***causes***Low-Decrease-Internet Sales-Amount | 0.36 |
| Low-Increase-Order Quantity Reseller Sales ***causes***Significant-Increase-Order Quantity Sales | 0.230 |
| Election ***causes***Average-Decrease-Internet Sales Amount | 0.22 |
| Low-Decrease-Reseller Sales ***causes*** Low-Increase-Internet Sales Amount | 0.172 |

### *B. Why-Question's treatment phase*

In this section, we present the *Why-Question's* treatment phase. This phase takes as input the decision maker's NL *Why-Question*, the events catalogues and the matrix $MCI$ to produce as outputs the causal relations related to this question. This phase consists of three steps: (a)*Why-Question* NLP, (b) causal relations extraction and (c) completing internal causal relations with participants. The details of each step are given in the following sections.

*1) Why-Question NLP step:* An effective analysis of the *Why-Question* is essential since from the obtained information characterizing this question we can identify the cause and effect relationships. Indeed, as proposed in the causality model (see figure 4), the *Why-Question* triggers the situation $S$ in which the cause and effect events are included.
The Natural Automatic Language Processing (NLP) step comprises two sub steps: a) *Why-Question* analysis and b) *Why-Question* recommendation.

1) *Why-Question* Analysis
In our approach, we consider that an input NL *Why-Question* $Q$ must be conform to the model that we have proposed in [43] (see section IV). Thus, a *Why-Question* $Q$ must comport at least one *measure* $m_Q$ and one trend indicator ($TI_Q$) that references the requested trend $Tr_Q$ (increase, decrease, low, high). In addition, the *temporal dimension* ($Dt_Q$) must be referenced in the *Why-Question*, whether or not it is specified by the decision maker in the question. Filters $f$ related to the non temporal dimension ($D_Q$) can be or not specified in the *Why-Question*. To capture these constraints, we have proposed a grammar in [43], [44] that fits the *Why-Question* model, on which the *Why-Question* analysis is based. This grammar aids in exacting the most important elements describing the *Why-Question*. These elements form the input event $e_Q$ which corresponds actually, to the event *"effect"* $f_c$ (see figure 4). We note as $e_Q$: $e_Q =< m_Q, Tr_Q, f[Dt_Q], (f[D_Q])^+ >$. According to the event $e_Q$, we aim to discover the eventual causal relations regarding the other events gathered in the events catalogues.
*Example 12.*
Let suppose that we have as input the *Why-Question* $Q_1$: *"Why has the internet sales amount decreased decreased in 2019?"*.
The analysis of the *Why-Question* $Q_1$ produces the following information: { requested trend $Tr_Q$= *decrease* }, { Measure $m_{Q_1}$= *the internet sales amount* }, { Temporal dimension $Dt_{Q_1}$ = *2019* }.
The corresponding input event is $e_{Q1}$ is described as : $e_{Q1}$ = { decrease of internet sales amount } provided with temporal specification such as: {2019}.
Once the *Why-Question* analysis achieved, we can determine what we need to focus to proceed with the causal knowledge discovery process. Otherwise, a recommendation process is triggered as explained in the following.
2) ***Why-Question*** **Recommendation** Once the *Why-Question* analysis is performed, two cases are possible: (1) the *Why-Question* is conform to the *Why-Question* model (see section IV ). In this case, the *Why-Question* is considered in order to trigger the process of extracting the causal relationships; (2) the *Why-Question* is not conform to the model (a *Why-Question* without measure, trend or temporal dimension). To cope with the second case, we have proposed a *Why-Question's* recommendation approach in [43]. This approach takes as input a NL *Why-Question* model and generates as outputs a set of recommended NL *Why-Questions* ranked according to their relevance. The decision maker can then select the most closest *Why-Question* to his need in order to launch the *Why-Question* treatment phase.

*2) Causal Relationships' Extraction Step:* In this step, we show how we extract the eventual causal relations existing between the input event $e_Q$ and the others gathered in the events catalogues. This step passes per two sub-steps: (1) a mapping step and (2) extracting internal and external causal relations step. The details of each sub step are exposed in the remainder of this section. The architecture of this step is illustrated in figure 10.

1) **The Mapping**
This step aims at the verification and the validation of the existence of the input event $e_Q$. Indeed, in order to deliver the decision maker with significant answers as a concrete means for decision support, we have to valid that the input event $e_Q$ corresponds to a real phenomena observed in the enterprise. To this end, we look for the input event $e_Q$ in the *internal events catalogue* $catalogue_{e_n}$, to inspect its existence in the DW. Thus, we have to verify the information characterizing the event $e_Q$ as follows:

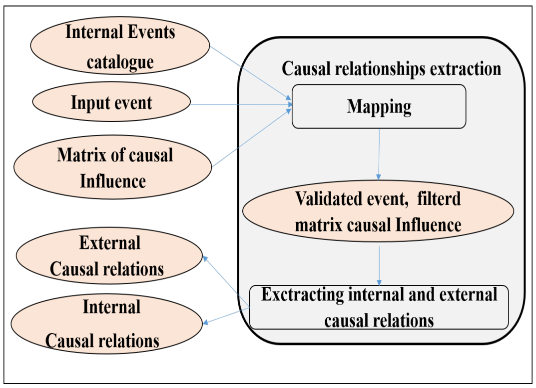


Fig. 10: Architecture of the causal relationships' extraction step

- Fist, we have to check that the requested trend $Tr_Q$ related to the measure $m_Q$ corresponds to an observed trend $Tr$ in the $catalogue_{e_n}$.
- Once we find that event $e_Q$ exists in the $catalogue_{e_n}$, we have to assert that temporal information $D_{t_Q}$ issued in the *Why-Question* $Q$ corresponds partially or completely to the temporal interval $I$ specified in the $catalogue_{e_n}$.
- If filters $f$ of non temporal dimensions $D_{j_Q}$ are specified in the *Why-Question* $Q$, then we generate automatically technical SQL queries to query the DW. This operation aims at validating whether the non temporal dimension's specifications $f$ are actually participants $P \in e_Q$ or not.

Once the mapping process is performed, the step of extracting the causal relationships will be triggered. To this end, we need the matrix of causal influence $[MCI]$ prepared previously in the causal relationships' identifying phase.
In order to meet the requirements specified in the *Why-Question* $Q$, we must adjust the matrix $[MCI]$ according to these specifications. Among these specifications, we focus on the temporal information. Indeed, the matrix has been built on the basis of the DW's history i.e. from the earliest date with the most recent one. Consequently, if the decision maker asks his *Why-Question* for a precise date, therefore we propose to filter the matrix $[MCI]$ until on this date. We note a filtered matrix as $[MCI^{'}]$. If no temporal information is specified in the *Why-Question* $Q$, we operate with the whole matrix $[MCI]$ ($[MCI^{'}]$ = $[MCI]$).

2) **Extracting Internal and External Causal Relations Step**
This step targets to identify the events causing the input one $e_Q$. To achieve the extraction of this knowledge, we exploit the filtered matrix $[MCI^{'}]$. The input event $e_Q$ is represented by a state vector ($S_{e_Q}$) as follow:

$$S_{e_Q} = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ \vdots \end{pmatrix}$$

Where "1" is the probability that the event $e_Q$ really happened.

To retrieve the measure $D_{CI}$ of the events influencing $e_Q$, we proceed by the multiplication of the matrix $[MCI^{'}]$ and the vector $S_{e_Q}$ as follows: $[MCI^{'}] \times S_{e_Q} = (V_{e_Q})$

$$[MCI^{'}] \times S_{e_Q} = \begin{pmatrix} D_{CI_1} \\ D_{CI_2} \\ \vdots \\ D_{CI_n} \end{pmatrix}$$

By analysing the resulting $V_{e_Q}$, the most relevant causal relationship $Rc$ represents the relationship having the most significant $D_{CI}$. The causal relationship $Rc$ can be internal or external. $Rc$ is internal if the event that causes $e_Q$ is an internal event $e_n$. $Rc$ is external if the event that causes $e_Q$ is an external event $e_x$. In the case where the cause of the cause would be known, we have only to carry out another multiplication between the resulting vector $V_{e_Q}$ and the matrix $[MCI^{'}]$.
*Example 13.*
According to the *Why-question* $Q_1$: *"Why has the internet sales amount decreased?"*, we show some results:
1- A significant decrease of *"internet sales order quantity"* causes a significant decrease of *"internet sales amount"* with a $D_{CI}$= 0.95.
2- A significant decrease of "temperature" causes a weak increase of "internet sales amount" with a $D_{CI}$= 0.37
3- "Political election" in France (political event) causes an average decrease of "internet sales amount" with a $D_{CI}$= 0.22.
4- In contrast, "Christmas holidays" (religious event) causes an important increase of "internet sales amount" with a $D_{CI}$= 0.62.

*a) Completing Internal Causal Relations with participants Step:* In order to deliver the decision maker with quite satisfactory *Why-Question's* answers $A$, we have to complete the internal causal relations extracted previously with the set of participants $P$. To perform this task, we generate automatically a set of formal queries (*SQL or MDX*), regarding the DW's non temporal dimensions $D_j$ related to the internal events $e_n$ influencing the input one $e_Q$. More precisely, through these formal queries, we project the temporal intervals $I$ of the event $e_n$ on the non temporal dimensions $D_j$ regarding their attributes values $a[v]$. To display to the decision maker the most important dimension's instances that should appear in the answers, we give him the possibility to select the dimension's attributes (this depends on his analysis need).
The external causal relations are not concerned with the DW's non temporal dimensions.

**Example 14.**
Let suppose the *Why-Question* $Q_1$: *"Why has internet sales amount decreed in 2019?"*. For this question, we found that the decrease in *"Order Quantity- Internet Sales"* influences the decrease in *"internet sales amount"* in the interval [01/12/2019, 02/26/2019 ]. One of the automatically generated SQL queries that enables retrieving the participants $P \in$ to the dimension *"Sales Territory"* is as follows:

```
SELECT top(5) SUM (OrderQuantity-IS) As Order
Quantity, SalesTerritory.Country,
FROM Internet Sales IS,
Where Date.DateKey=IS.DateKey and
SalesTerritory.Id-Territory= IS.Id Territory and
12/01/2019<=Date.FullDate=<=26/02/2019
Group by f.Territory,
Order by Order Quantity desc;
```

Where 01/12/2019 and 02/26/2019 represent the start and end dates of the trend *"Order Quantity-Internet sales decrease"* respectively.

### C. Why-Question's Answers Visualisation Phase

Once the treatment phase is performed, we interpret the obtained results in a textual and graphical format, as exhibited in the remainder of this section.

*1) Textual interpretation Step:* We provide a set of NL answers $A$ on the basis of a set of templates. We have defined four templates:
(1) the templates ($T_1, T_2, T_3$) capture the answers those represent the causal relations between the input event $e_Q$ and the internal $e_n$ and external events $e_x$ ( $e_{x_{QN}}$ and $e_{x_{QL}}$ );
(2) with the fourth template ($T_4$), we want to provide more details to the decision maker, which correspond to the completed internal causal relations with the participants $P$.
In these templates, we use the conjunction *"because of"* to refer to the clause that stands for the cause $c$. These templates are as follows:
$T_1$ = A $< m_Q >< Tr_Q >$ is observed, during $< Dt_Q >$ because of $< M >< Tr >$ during $I$.
$T_2$= A $< m_Q >< Tr_Q >$ is observed, during $< Dt_Q >$ because of $< M' >< Tr >$ during $I$.
$T_3$= A $< m_Q >< Tr_Q >$ is observed, during $< Dt_Q >$ because of $< Name - event >$ during $I$.
$T_4$= A $< m_Q >< Tr_Q >$, during $< Dt_Q >$ because of $< m' < Tr >$, according to: $< P >$.
To avoid cluttering these answers with the details of the attributes values $a_k$ of a non temporal dimension $D_j[l_t^+[a_k]]$, we render the participants $P$ as the top n dimensions $D_j[l_t^+[a_k]]$. More tuples of $P$ can be viewed according to the decision maker's demand.

**Example 15.**
According to the *Why-Question* $Q_1$: *"Why has internet sales decreased in 2019?"*, the returned answers are as follows:

- An important decrease of internet sales amount is observe during [12/10/2019,01/01/2020] **because of** an important decrease of internet sales order quantity during [06/10/2019,31/12/2019];
- A weak decrease of internet sales amount is observed during [22/05/2019, 02/07/2019] **because of** a weak of decrease of internet sales order quantity during [20/04/2019,30/06/2019];
- A weak decrease of internet sales amount is observed during [22/05/2019, 02/07/2019] **because of** a weak increase of temperature during [15/03/2019,20/05/2019];
- An important decrease of internet sales amount is observed during [12/10/2019,01/01/2020] **because of** an important decrease of internet sales order quantity during [06/10/2019, 31/12/2019], according to: United States, Australia, United kingdom, Canada;

*2) Graphical Interpretation Step:* To support the NL answers, we generate graphical representations for the extracted causal relations, for example trend's curves are generated to represent internal correlations. The trend's curve shows clearly the temporal intervals according to which the trend analysis is performed and the corresponding variations those illustrate the overall trends (increase, decrease).

## VI. Implementation and Experimentation Study

We present in this section, the tool that we develop to implement our proposed approach as well as some experiments to validate it.

### A. Implementation

Let us consider the case study presented in our motivating example (see section II): Microsoft Adventure Work DW 2020 with the external data sources (climatic data, political calender and religious calender).

We present some details about our experimental environment. We have implemented our approach and run the experiments on an Intel(R) Core(TM) i5-4210U CPU @ 1.70 GHZ 2.40 GHZ machine with 4G RAM memory. We have used the *MATLAB 2018* environment for the numerical computing, the *JAVA* language in the *NetBeans IDE 8.0.2* environment to *Why-Question's* analyser tool development and the *Microsoft SQL*

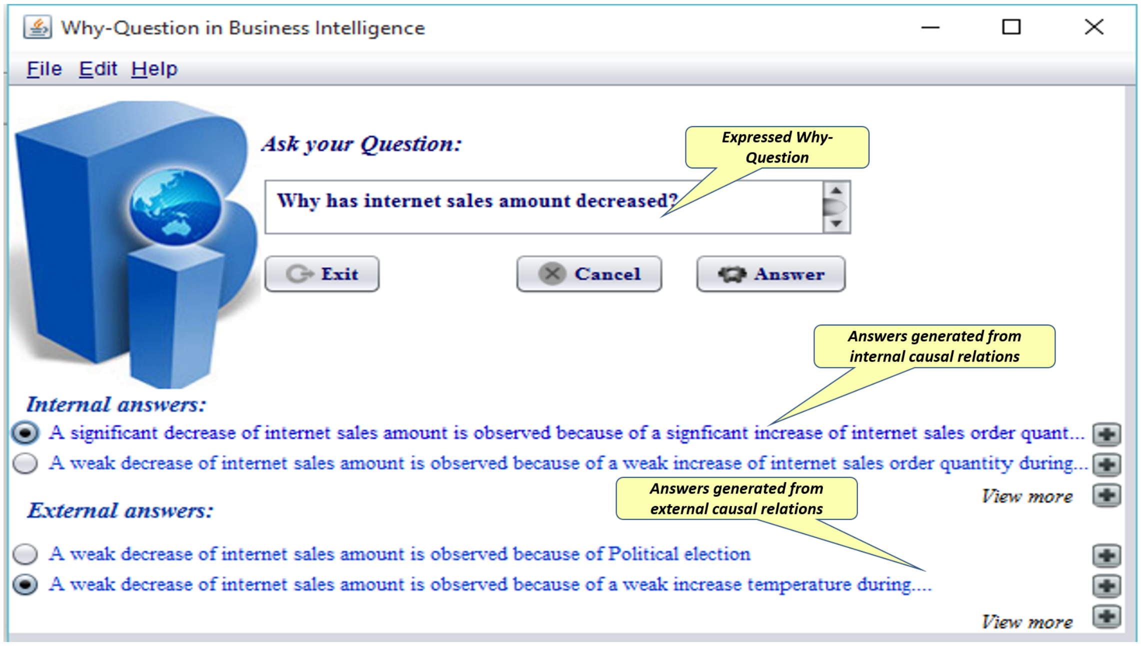


Fig. 11: *"BI Why Q/A"* screen shot (NL answers).

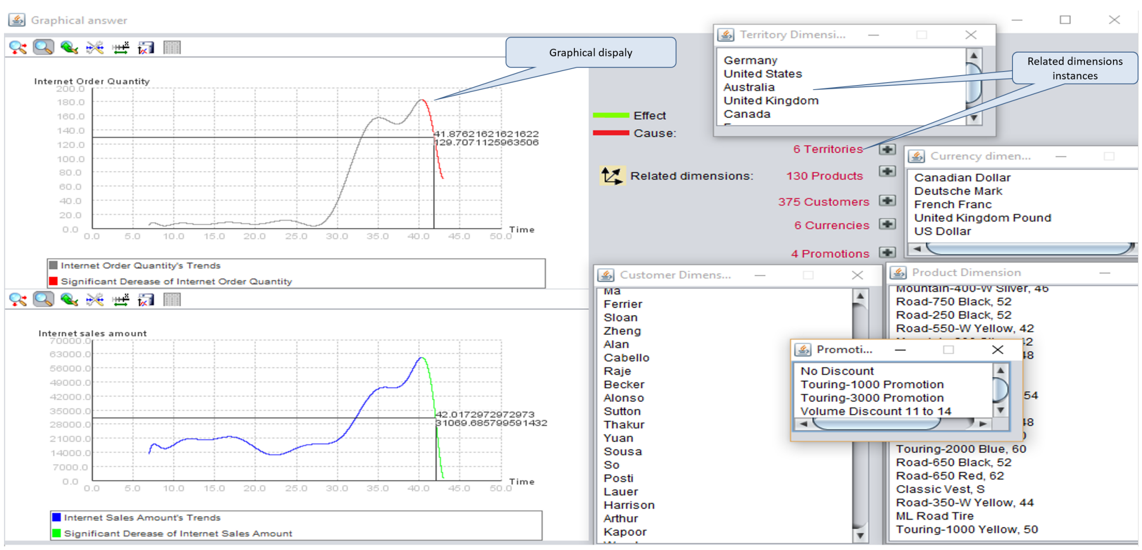


Fig. 12: *"BI Why Q/A"* screen shot (graphical results).

*Server 11.0.2100* to exploit the *Microsoft AdventureWorks-DW 2020*.

To test our approach, we have designed and developed a tool called *"BI Why Q/A"*, provided with a graphical interface. This tool, takes for example as input the *Why-Question* $Q_1$:*"Why has internet sales amount decreased?"*. It visualises a set of NL answers (see figure 11) and graphical representations (see figure 12). The decision maker can interact with the graphical answers to analyse values as well as dimension's instances.

## B. Experimental Study

In this section, we carry out a set of experiments to show the relevance and the effectiveness of our system . More details are presented below.

*1) Relevance evaluation:* In order to validate our proposal in term of relevance, we evaluate the returned causal relations. To this end, we compare our method with other techniques existing in the literature [45]–[47], that enable causality analysis. The most used methods for the discovery of causal relationships are: (a) the Granger causality test, (b) the causal Bayesian networks and (c) the association rules mining algorithms. For this evaluation purpose, we adapt each method in our context except the causal Bayesian network CBN one

[46] [5]. Adapting these methods in our context means applying each method's definition to discuss then the obtained results. Through this experimentation, we want to show how and until what limits we can adapt the existing methods in our context, by assessing whether the provided results are satisfactory or not.

*a) **Comparison with Granger Causality tests**:* the causality of Granger is a well known technique used to analyse causality. It was introduced by *Granger* in 1969 [45]. In the *"Granger"* sense, a variable $X$ causes the variable $Y$ if the past values of $X$ have a statistical impact on the current or future value of $Y$. In the *"Granger"* sense, $Y$ causes $X$ with a period's delay. Indeed, The dependence of a variable $Y$ on another variable $X$, is rarely instantaneous. Very often, $Y$ responds to $X$ with a lap time called a *lag* ($l$). The idea is so to look at the evolution of certain variables represented in *time series* (only numerical data). If a variable seems to precede another variable in terms of evolution according to a lag $l$, we can conclude a causal relation $RC$. The Granger's causality is based on the fact that a time series ($x_t$) causes another time series ($y_t$) if the prediction of $y_t$ conditionally to its past is improved taking also into account the past of $x_t$. The Granger analysis is based on the Vector autoregressive model ($VAR$) regarding a lag $l(VAR(l))$ [48].

The *"Granger's"* approach consists into testing two hypothesis:

$H_0$: $X$ Does not Granger cause $Y$ (probability Prob $> 5\%$).
$H_1$: $X$ Granger causes $Y$ (probability Prob $< 5\%$).

The test of these two hypotheses is established using the standard *"Fisher"* statistical test [49]. In the case where both hypotheses are accepted, a retroactive loop is obtained ($X$ causes $Y$ and $Y$ cause $X$).

In order to carry out a *Granger* causality test, we use the *EViews 11 application* [6]. We apply the *"Granger causality"* only for numerical data as: the Microsoft AdventureWork DW's measures $M$ and the temperature data [7] $M'$, extracted from climatic *csv* files. We compare then the causal influence probabilities obtained with *"Granger"* analysis and the ones provided with our method. To this end, we follow the principle described below:

- For the *"Granger"* tests, time-series (numerical data) are manipulated while our method is oriented events $e$ (qualitative and quantitative data). In order to compare the results obtained according to different types of variables, we opt for the evaluation of the ranking of the causal relations $Rc$ provided by both methods.

- On one hand, let suppose that we have three variables (DW's measures values observed during same temporal period $T$) $m_1$, $m_2$ and $m_3$. To determine causality between pairwise of these variables, using *Granger* tests, we have to analyse the provided probability $Prob$. The ranking of the causal relations $Rc$ is carried out on the basis of the order of the $Prob$ values. For example, $m_1$ causes $m_2$ with $Prob_1$ and $m_1$ causes $m_3$ with a $Prob_2$. If $Prob_1 > Prob_2$ then we interpret that there is a causal relation between $m_1$ and $m_2$ stronger than $m_1$ and $m_3$.

- On the other hand, let suppose that we have three events: $e_{n1} \in m_1, e_{n2} \in m_2$ and $e_{n3} \in m_3$. To rank the causal relations for these events, we analyse the provided degree of causal influence $D_{CI}$ between pairwise of events.

- Finally, to assert that Granger test and our methods provide similar results concerning causal relations we have to find for example that the event $e_{n1} \in m_1$ causes $e_{n2} \in m_2$ with a $D_{CI}(e_n1, e_n2)$ higher than $D_{CI}(e_n1, e_n3) \in m_3$.

On the basis of the principle presented above, we begin by showing the *"Granger"* analysis results obtained with *EViews application* for the input variables: *"internet sales amount, internet order quantity, resellers sales amount, reseller order quantity and temperature"* for a period from 2012 to 2017.

To choose a lag $l$ is better in general to use more rather than fewer lags, since the theory is couched in terms of the relevance of all past information. It should be preferable to pick a lag length ($l'$) that corresponds to reasonable beliefs about the longest time over which one of the variables could help predict the other. Hence, in order to carry out a relevant *"Granger"* analysis it is necessary to select the optimal lag length $l'$. There is no hard and fast rule on the choice of the lag length. It is basically an empirical issue. In our experimental study, we fix the optimal lag length $l'$ on the basis of the *Akaike information criterion* (AIC) [50]. Consequently, the selected lag length is $l' = [1-8]$.

The obtained results of the *"Granger's"* analysis are as captured in the table VIII.

We interpret the pairwise *"Granger"* Causality tests as follows:
- When the probability $Prob$ is $<$ 0.05, means that it exists a causal relation $Rc$ between variables (rejected null hypothesis). Otherwise, when $Prob$ is $>$ 0.05, the null hypothesis is accepted i.e. the variables are not causally correlated. Thus, we have found that:
- It exists two bidirectional causal relations (retroactive loops) between *"reseller sales amount"* and *"reseller order quantity"*; and *"internet sales amount"* and *"internet sales order quantity"*.
- The *"temperature"* influences the *"reseller sales amount"* with Prob = 0.0465 and affects *"reseller order quantity* with Prob=0.0366.

We continue this experimental study by comparing the probabilities $Prob$ discussed above with the causal influence degrees $D_{CI}$ obtained by our proposal for the same input variables, with a high exclusion rate $E = 90\%$. We have obtained 6 event's categories: (SI/AI/WI/SD/AD/WD) where each category stand respectively for *"significant increase,*

[5]Causal Bayesian network CBN requires an initial causal network (variable and value parameters) defined by a domain expert [46]. Unfortunately, the CBN can not be adapted in our context

[6]https://www.eviews.com/home.html

[7]Since the data available in the DW concern the duration [2012,2017], then we extract the average temperature for sales territories, recorded in this period.

TABLE VIII: Granger Analysis tests

| **Pairwise Granger Causality Tests**<br>**Lags:8** | | | |
|---|---|---|---|
| **Null Hypothesis:** | **obs** | **F-Statistic** | **Prob** |
| *INTERNET SALES AMOUNT* ***does not Granger Cause*** *INTERNET ORDER QUANTITY*<br>*INTERNET ORDER QUANTITY* **does not Granger Cause** *INTERNET SALES AMOUNT* | 30 | 1.05389<br>1.70304 | **0.0173**<br>**0.0465** |
| *RESELLER ORDER QUANTITY* does not Granger *Cause INTERNET SALES AMOUNT*<br>*INTERNET SALES AMOUNT* does not Granger Cause *RESELLER ORDER QUANTITY* | 28 | 1.43789<br>0.86846 | 0.2820<br>0.5687 |
| *RESELLER SALES AMOUNT* does not Granger Cause *INTERNET SALES AMOUNT*<br>*INTERNET SALES AMOUNT* does not Granger Cause *RESELLER SALES AMOUNT* | 28 | 0.91480<br>0.58035 | 0.5385<br>0.7751 |
| *TEMPERATURE* does not Granger Cause *INTERNET SALES AMOUNT*<br>*INTERNET SALES AMOUNT* does not Granger Cause *TEMPERATURE* | 30 | 0.18953<br>3.09650 | 0.3457<br>0.9853 |
| *RESELLER ORDER QUANTITY* does not Granger Cause *INTERNET SALES ORDER QUANTITY*<br>*INTERNET SALES ORDER QUANTITY* does not Granger Cause *RESELLER ORDER QUANTITY* | 28 | 1.31377<br>0.42213 | 0.3293<br>0.8846 |
| *RESELLER SALES AMOUNT* does not Granger Cause *INTERNET SALES ORDER QUANTITY*<br>*INTERNET SALES ORDER QUANTITY* does not Granger Cause *RESELLER SALES AMOUNT* | 28 | 0.48119<br>0.41880 | 0.8458<br>0.8867 |
| *TEMPERATURE* does not Granger Cause *INTERNET SALES ORDER QUANTITY*<br>*INTERNET SALES ORDER QUANTITY* does not Granger Cause *TEMPERATURE* | 30 | 0.67497<br>1.67314 | 0.7060<br>0.1967 |
| *RESELLER SALES AMOUNT* **does not Granger** Cause *RESELLER ORDER QUANTITY*<br>*RESELLER ORDER QUANTITY* **does not Granger** Cause *RESELLER SALES AMOUNT* | 28 | 3.31802<br>7.63375 | **0.0346**<br>**0.0015** |
| *TEMPERATURE* **does not Granger Cause** *RESELLER ORDER QUANTITY*<br>*RESELLER ORDER QUANTITY* **does not Granger** Cause *TEMPERATURE* | 28 | 3.25923<br>1.40172 | **0.0366**<br>0.2950 |
| *TEMPERATURE* **does not Granger Cause** *RESELLER SALES AMOUNT*<br>*RESELLER SALES AMOUNT* does not Granger Cause *TEMPERATURE* | 28 | 3.01874<br>1.93002 | **0.0492**<br>0.1544 |

*average increase, weak increase, significant decrease, average decrease and weak decrease"*. Thus, for the input variables, 20 events have been provided as: (SI/AI/WI/SD/WD) of the internet sales amount, (SI/WI/SD/WD) of internet sales order quantity, (SI/AI/WI/SD/AD/WD) of reseller sales amount, (SI/AI/WI/SD/WD) of reseller order quantity and (SI/SD/WD) of temperature. The results are gathered in table IX. The corresponding $D_{CI}$ is the average $D_{CI}$ of all the events $\in m_i$ according all events $\in m_j$.

TABLE IX: Our Causal Influence Method Results.

| Measure/ events | ISA | ISO | RSA | RSO |
|---|---|---|---|---|
| Internet sales amount (ISA) (SI/AI/WI/SD/WD) | 0 | 0.32 | 0.28 | 0.23 |
| Internet sales order quantity (ISO) (SI/WI/SD/WD) | 0.58 | 0 | 0.21 | 0.18 |
| Reseller sales amount (RSA) (SI/AI/WI/SD/AD/WD) | 0.24 | 0.21 | 0 | 0.41 |
| Reseller order quantity (RSO) (SI/AI/WI/SD/WD) | 0.22 | 0.27 | 0,61 | 0 |
| Temperature (SI/SD/WD) | 0.2 | 0.17 | 0.28 | 0.3 |

By analysing the table IX and figure VIII, we elucidate what follows:

- On one hand, we judge that our method produces relevant results in terms of causal influence probabilities because it provides the same conclusions of the *Granger's* tests as shown in table X. We note that more the *Granger's* prob is $< 0.05$ more a causal relation is relevant while with our method more $D_{CI}$ is high then it is more likely that the causal relation is relevant (see table X).

- On the other hand, we notice that our method reveals all possible causality's relations between the *"internet sales and reseller sales activities"* and *"temperature"*. This causal influence is assessed with a $D_{CI}$ that varies from 0.17 to 0.28. However, *Ganger's* analysis tests show that there is no causal correlations between these variables (Prob $>$0.05). Indeed, the Granger analysis studies the mathematical evolution of numerical values between variables. However, the VAR model and the Fisher test (Prob$<$0.05) are so rigorous for inspecting the dependence between variables to reach prediction purposes that sometimes certain possible causes can be neglected. While our method seeks for influence between all events. Thus, our method highlights all the events capable of being possible causes even with minimal probabilities. Indeed, on the basis of the recurrence of close events in the past, this method does not neglect any event $e$ even if it is far in the history. Since our method is oriented events, it approximates as much as a possible human reflection and can help in the decision making process, for examples: (1) *A significant decrease of "temperature" (winter season) causes an average increase of the "internet sales amount" with a $D_{CI}$=0.18* (On one side, in winter season customers prefer to shop on line. On the other side, climate changes can cause technical connection problems); (2) *an average decrease of "internet sales amount" for the product "all purpose bike stand and front brakes" causes an average increase of "reseller sales amount" for the same products with a $D_{CI}$= 0.28.*

TABLE X: Granger's test and our method conclusions.

| Causal relation | Granger's Prob | Our method $D_{CI}$ |
|---|---|---|
| Reseller order quantity *influences* Reseller sales amount | 0.0015 | 0.61 |
| Internet sales order quantity *influences* Internet sales amount | 0.0173 | 0.58 |
| Reseller sales amount *influences* reseller order quantity | 0.0346 | 0.41 |
| Temperature *influences* reseller order quantity | 0.0366 | 0.3 |
| Internet sales amount *influences* Internet sales order quantity | 0.0492 | 0.32 |
| Temperature *influences* reseller sales amount | 0.0465 | 0.28 |

*b) **Comparison with Association Rules Algorithms**:* The most used data mining algorithms for causal mining in databases are the associations rules algorithms. Indeed, discovering causal relationships using associations is a norm in the literature [47]. Causal relationship discovery in data, using association rules algorithm, is to find a short list of rules that are most likely causal on the basis of the beforehand set metrics *support* ($minsupp$) and the *confidence* ($minconf$). These causal rules represent a small set of statistically reliable relation that is likely to embed cause and effect relations.

This experimental study consists in transforming the problem of causal relation $Rc$ extraction between events $e$ into an association rule mining issue in our context. The objective of this study is to assess the relevance of the extracted association rules when this rule means a causal relation $Rc$. In order to mine these $Rc$, we rely on the *Apriori* and the *FP-Growth* algorithms.

For the exploration of the association rules, we start first by defining the transaction's base ($B_{Tr}$). To this end, we consider all historical events $e$ captured in catalogues (internal and external) as items ($It$). We define a transaction ($Ts$)$\in B_{Tr}$ as a sequence of events $e$ that occur in a time window ($Tw$). To specify $Tw$, we refer to the DW's temporal dimension $Dt$ (date) (see figure 1). In this study, we set $Tw = "quarter"$, the minimal support $(minsupp) = 0.1$ and the minimal confidence $(minconf) = 0.5$. Thereafter, we extract all the rules with strong confidence $(cf)$ i.e $cf >\geq minconf$

To analyse the generated rules, let suppose that we want to answer the *Why-Question* $Q_1$:*"Why has internet sales amount decreased?"*. We have then to study the provided rules according to the events: $e_{n1}$=*"significant decrease of internet sales amount"* and $e_{n2}$=*"weak decrease of internet sales amount"*. The analysis results are illustrated in figure 13.

Through this experimental study, we notice that:

- On one side, both algorithms generate a small set of causal rules compared to what we should have expected, for which some of the rules are redundant. In addition, some causal rules are missed by the *Apriori* algorithm.

- On the other side, in the mining procedure, events $e$ are considered correlated when they occurred in the same transaction frequently. However, we call into question the confidence's value of the causal rules generated by the *Apriori* and *FP-Growth* algorithms (0.5 and 0.76, respectively). Actually, causality has not been used as an interestingness criterion to mine causal rules. Indeed, we have found for both algorithms that some causal rules resulting from transactions $Ts$ in which the events effect $fc$ ( $e_{n1}$ and $e_{n2}$) occur before a cause $c$. For the *Apriori* and the *FP-Growth* algorithms, 2 and 3 transactions $Ts$, receptively, include spurious causal rules.

- In this study, we have computed the average of the causal influence degree $D_{CI}$ for the causal rules generated by the *Apriori* algorithm (0.45) and the *FP-Growth* algorithm (0.51) (see figure 13). Interpreting association rules as causal relations requires justification [47]. Indeed, we have unfortunately intervened to judge the correctness of the generated causal rules. However, it would be inefficient to find causal rules in a large collection of association rules as a secondary discovery process, and new approaches are required for causal rule discovery [38]. While, our causality analysis method scans the historic to highlight the events cause $c$, on the basis of the principle of recurrence in the past, without any prerequisite and minimal probabilities ($minsup$ and $minconf$).

*2) Effectiveness study:* In this section, we carry out an experimental study that shows the effectiveness of the whole approach proposed to answer a decisional NL *Why-Question*.

*a) **Approach Utility Study**:* In this section, we assess the usefulness of our approach by the same users involved in the user effort study. We have asked for these users to inspect the *Why-Question's* answers returned by our approach. We invite them to judge the interestingness of the results (ranked internal and external causal relations), in the decision making process, in terms of *expected* (obvious), *unexpected* and rejected responses. We collect the results of this study in the table XI.

In this experimental study, the users agree on average with 66% with the internal answers and with most of our external answers (89.6%), generated by the *"BI Why Q/A"* tool. Nevertheless, they reject 13.33% of the internal responses while no external answer has been rejected. However, the users judge that 22.05% of the internal answers are obvious while only 10.41% of the external answers are suspected.

- On one hand, in summary, the users find the internal answers quite useful in the decision making process. Indeed, these answers are harder to detect with another analysis tool. This is interesting especially when the DW is voluminous in terms of measures and dimensions with which the analysis becomes fastidious. However, the users appreciate the obvious answers when these latter are completed and rendered with the dimensions properties. Nevertheless, some of the answers are rejected. These answers concern precisely some events that belong to measures of distinct fact tables with a high value of $D_{CI}$.

- On the other hand, we notice that the users are quite

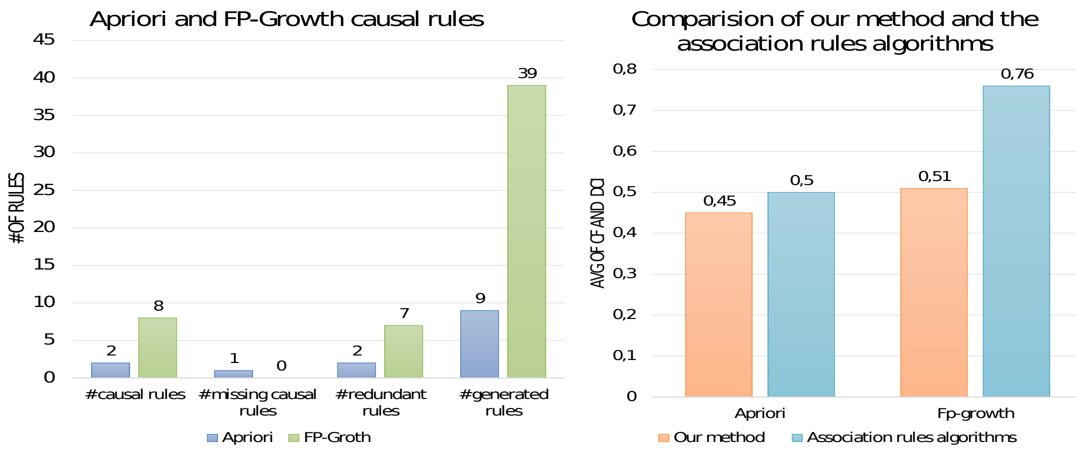


Fig. 13: Comparison with the Association Rules Algorithms.

TABLE XI: Approach utility study

| *Why-Question* | Users | Internal answers | | | External answers | | |
|---|---|---|---|---|---|---|---|
| | | Expected | Not expected | Rejected | Expected | Not expected | Rejected |
| *"Why are internet sales amount not stable?"* | U1 | 20% | 73.48% | 10.76 % | 0% | 100% | 0% |
| | U2 | 24.61% | 60% | 15.38% | 18.75% | 81.25% | 0% |
| | U3 | 21.54% | 64.61% | 13.84% | 12.5% | 87.5% | 0% |

satisfied by the provided external answers. Indeed, external sources can enrich the decisional analysis to support effective and well-informed decisions.

*b) **User Effort Study**:* In this section, we measure the time taken by users if they want to try to answer a *Why-Question* using existing tools as: (a) SQL queries in Microsoft SQL Server and (b) *Microsoft SQL Server Reporting Services*. We notice that these tools can't be readily applied to obtain possible answers without a time needed to analyse in a naive way and manually row data, pivot tables or even curves. To carry out this experimental study, we have invited three real users of the *Microsoft adventureWork DW 2017* ( members of our researcher laboratory), familiar with *SQL Server Reporting Services* and have skills in programming SQL queries.

To make this study manageable, we assume that the users know a priori the eventual causal relationship, for example: if a *Why-Question* is asked regarding the measure *"Reseller sales amount"* then the users will be interested in analysing the measure *"Reseller order quantity"*. Thus, these users will retrieve the *"Reseller order quantity"* instances in table form and BI report per each related dimension. They analyse manually and separately these tuples for a doublet $<$ measure, dimension $>$ (only 60 tuples in this human effort study) to detect an attracting anomaly that can help in the decision making process. Thereafter, we measure for each user the total time spent in finding *Why-Question's* answers.

We reported the results in table XII. In summary, the users spent significant time to find eventual answers to their *Why-Question*. On average, users take 19.8 minutes with SQL queries and 17.7 minutes with the SQL server reporting tool. In contrast, our *"BI Why Q/A"* tool takes 0.47 seconds to render NL *Why-Question's* answers and graphical representations in a single interface.

TABLE XII: User Effort study (min).

| *Why-Question* | Analysis mean | U1 | U2 | U3 |
|---|---|---|---|---|
| *"Why has reseller sales amount decrease?"* | SQL Queries | 17.4 | 19.6 | 22.4 |
| | Reporting service | 15.3 | 17.5 | 20.3 |

## VII. Conclusion and Future Works

In this paper, we have proposed an approach that addresses NL *Why-Question* answering problem in BI context. This approach delivers to the decision makers a set of answers provided thanks to processes that enable discovering potential causal relationships between the events highlighted in the *Why-Question* and those observed in the DW as well as in external sources.

Our proposal is based, first, on a model that captures our causality perception, focusing on the concept of *"event"* in BI context [36]. This model leads us to define the events for causality analysis purposes. Secondly, our approach is mainly performed on the basis of a causality analysis method, oriented events [36]. This method computes causal influence between events. It is performed with respect to the principle of recurrence of close events in the past and temporal assumptions. Indeed, it considers a *"Temporal Exclusion Function"* that aims to reduce the importance of *distant events causes* in the past and to privilege the *closer ones*.

In order to validate our approach, we have developed the *Why Q/A BI* tool. This tool allows a decision maker to express his need in the form of a NL *Why-Question* and to provide him a set of NL answers and graphical interpretations for an effective decision making.

A set of experimental studies has been made. On one side, we have assessed the effectiveness of our approach by involving users in judging its utility and to show its performance. Within this study, we have noticed that, on one hand, the users found that the answers returned on the basis of internal causal relations .i.e. only from the DW, are helpful in the decision-making process. Indeed, because it is more difficult to detect these responses with a data restitution tool and manual efforts such as querying or reporting. Additionally, these responses are automatically completed and rendered with the Top-n dimension values which is interesting when the DW is voluminous with which the analysis becomes tedious. On the other hand, the answers related to external sources can enrich the decision analysis by offering new points of view, useful in the decision-making.
On the other side, we have evaluated the relevance of our proposal. To support the obtained results, we have compared this approach with existing techniques used for causality analysis as the *"Granger's"* causality tests as well as the association rules mining algorithms. In this study, the causality analysis method that we propose looks for causal influence between distinct events extracted from the DW as well external sources. In this optic, our proposal highlights all events likely to be possible causes even with reduced probabilities. This method being event-oriented and providing responses with dimension values, can help as much as possible in decision-making. However, we found that certain causal relationships, do not really reflect causes such as the causal relationship between *"Freight-Transport"* and *"Tax amount"*. These two measures certainly have consequences on *"Internet sales amount"* but are they really linked by a cause and effect relationship?. At this stage, we consider that the intervention of an analyst becomes necessary in order to attest that *"Freight-Transport"* and *"Tax amount"* are not causal relationships. From such a survey, we will be able to improve the proposed approach in terms of returned answers. However, if some correlations persist even with further DW's feeding, then these relationships come closer to causality presumption.

Currently, we are working on how: (1) to enhance the event's definition using fuzzy logic; and (2) to prune spurious answers by refining them and investigating about the participants through using learning as well as requirements classification techniques [51].

The proposed approach explores the DW and a set of prior selected external sources to extract answers. We intend so, in the near future, to integrate reliable external data sources into the DW. The most suitable source will be selected on the basis of the causal influence method, for further decisional analysis.

As future work, we plan to extend our proposals to handle the parameter *"spatial constraint" or "location"*, to produce more relevant answers.

## REFERENCES


[1] N. Kuchmann-Beauger, "Question answering system in a business intelligence context," Ph.D. dissertation, Ecole Centrale Paris, 2013.

[2] M. A. Naeem, S. Ullah, and I. S. Bajwa, "Interacting with data warehouse by using a natural language interface," in *Natural Language Processing and Information Systems: 17th International Conference on Applications of Natural Language to Information Systems, NLDB 2012, Groningen, The Netherlands, June 26-28, 2012. Proceedings 17*. Springer, 2012, pp. 372–377.

[3] F. Popowich, M. Mosny, and D. Lindberg, "Interactive natural language query construction for report generation," in *Proceedings of the Seventh International Natural Language Generation Conference*. Association for Computational Linguistics, 2012, pp. 115–119.

[4] J. Saias, P. Quaresma, P. Salgueiro, and T. Santos, "Binli: An ontology-based natural language interface for multidimensional data analysis." *Intelligent Information Management*, vol. 4, no. 5, 2012.

[5] N. Kuchmann-Beauger and M.-A. Aufaure, "A natural language interface for data warehouse question answering," in *Natural Language Processing and Information Systems: 16th International Conference on Applications of Natural Language to Information Systems, NLDB 2011, Alicante, Spain, June 28-30, 2011. Proceedings 16*. Springer, 2011, pp. 201–208.

[6] A. Sangroya, P. Saini, M. Rawat, G. Shroff, and C. Anantaram, "Natural language business intelligence question answering through seqtoseq transfer learning," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2019, pp. 286–297.

[7] R. Djiroun, K. Boukhalfa, and Z. Alimazighi, "Designing data cubes in olap systems: a decision makers requirements-based approach," *Cluster Computing*, vol. 22, pp. 783–803, 2019.

[8] C. Imhoff and C. White, "Self-service business intelligence," *Empowering Users to Generate Insights, TDWI Best practices report, TWDI, Renton, WA*, 2011.

[9] V. Vashisht and P. Dharia, "Integrating chatbot application with qlik sense business intelligence (bi) tool using natural language processing (nlp)," in *Micro-Electronics and Telecommunication Engineering*. Springer, Singapore, 2020, pp. 683–692.

[10] M. Al-Hawawreh, A. Aljuhani, and Y. Jararweh, "Chatgpt for cybersecurity: practical applications, challenges, and future directions," *Cluster Computing*, vol. 26, no. 6, pp. 3421–3436, 2023.

[11] N. Asghar, "Automatic extraction of causal relations from natural language texts: A comprehensive survey," *arXiv preprint arXiv:1605.07895*, 2016.

[12] S. Mani and G. F. Cooper, "Causal discovery using a bayesian local causal discovery algorithm." in *Medinfo*, 2004, pp. 731–735.

[13] H. Hassani, X. Huang, and M. Ghodsi, "Big data and causality," *Annals of Data Science*, vol. 5, pp. 133–156, 2018.

[14] R. Girju, "Automatic detection of causal relations for question answering," in *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering-Volume 12, Association for Computational Linguistics*, 2003, pp. 76–83.

[15] S. Verberne, "Developing an approach for why-question answering," in *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*. Association for Computational Linguistics, 2006, pp. 39–46.

[16] S.Verberne, "Paragraph retrieval for why-question answering," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2007, pp. 922–922.

[17] S. Verberne, L. Boves, N. Oostdijk, and P.-A. Coppen, "Evaluating discourse-based answer extraction for why-question answering," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2007, pp. 735–736.

[18] V. Moriceau, X. Tannier, and M. Falco, "Une étude des questions complexes en question-réponse," in *Actes de la Conférence Traitement Automatique des Langues Naturelles (TALN 2010, article court), Montréal, Canada*, 2010.

[19] J.-H. Oh, K. Torisawa, C. Hashimoto, T. Kawada, S. De Saeger, J. Kazama, and Y. Wang, "Why question answering using sentiment analysis and word classes," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, 2012, pp. 368–378.

[20] C. Baral, N. Ha Vo, and S. Liang, “Answering why and how questions with respect to a frame-based knowledge base: a preliminary report,” in *LIPIcs-Leibniz International Proceedings in Informatics*, vol. 17. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2012.

[21] J.-H. Oh, K. Torisawa, C. Hashimoto, M. Sano, S. De Saeger, and K. Ohtake, “Why-question answering using intra-and inter-sentential causal relations.” in *ACL (1)*, 2013, pp. 1733–1743.

[22] J.-H. Oh, K. Torisawa, C. Kruengkrai, R. Iida, and J. Kloetzer, “Multi-column convolutional neural networks with causality-attention for why-question answering,” in *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. ACM, 2017, pp. 415–424.

[23] C. Pechsiri, “Explanation based why question answering system,” in *Proceedings of the International MultiConference of Engineers and Computer Scientists*, vol. 1, 2016.

[24] R. Higashinaka and H. Isozaki, “Automatically acquiring causal expression patterns from relation-annotated corpora to improve question answering for why-questions,” *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 7, no. 2, p. 6, 2008.

[25] S. Tirunagari, “Data mining of causal relations from text: Analysing maritime accident investigation reports,” *arXiv preprint arXiv:1507.02447*, 2015.

[26] R. Sharp, M. Surdeanu, P. Jansen, P. Clark, and M. Hammond, “Creating causal embeddings for question answering with minimal supervision,” *arXiv preprint arXiv:1609.08097*, 2016.

[27] S. Vazquez-Reyes and W. J. Black, “Evaluating causal questions for question answering,” in *Computer Science, 2008. ENC’08. Mexican International Conference on*. IEEE, 2008, pp. 132–142.

[28] E. Blanco, N. Castell, and D. I. Moldovan, “Causal relation extraction.” in *Lrec*, 2008.

[29] A. Sorgente, G. Vettigli, and F. Mele, “Automatic extraction of cause-effect relations in natural language text.” *DART@ AI* IA*, vol. 2013, pp. 37–48, 2013.

[30] A. Ittoo and G. Bouma, “Extracting explicit and implicit causal relations from sparse, domain-specific texts,” in *International Conference on Application of Natural Language to Information Systems*. Springer, 2011, pp. 52–63.

[31] C. Khoo, S. Chan, Y. Niu, and A. Ang, “A method for extracting causal knowledge from textual databases,” *Singapore journal of library & information management*, vol. 28, pp. 48–63, 1999.

[32] Q. X. Do, Y. S. Chan, and D. Roth, “Minimally supervised event causality identification,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011, pp. 294–303.

[33] C. S. Khoo, S. Chan, and Y. Niu, “Extracting causal knowledge from a medical database using graphical patterns,” in *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2000, pp. 336–343.

[34] Q. X. Do, Y. S. Chan, and D. Roth, “Minimally supervised event causality identification,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011, pp. 294–303.

[35] M. A. Guessoum, R. Djiroun, and K. Boukhalfa, “Dealing with decisional natural language why-question in business intelligence,” in *8th International Conference on Information and Communication Systems (ICICS)*. IEEE, 2017, pp. 52–57.

[36] M. Guessoum, R. Djiroun, and K. Boukhalfa, “Causality analysis method and model related to why-question answering in business intelligence context,” in *International Conference on Computing Systems and Applications*. Springer, 2022, pp. 15–26.

[37] A. M. Azmi and N. A. Alshenaifi, “Lemaza: An arabic why-question answering system,” *Natural Language Engineering*, vol. 23, no. 6, pp. 877–903, 2017.

[38] Z. Jin, J. Li, L. Liu, T. D. Le, B. Sun, and R. Wang, “Discovery of causal rules using partial association,” in *Data Mining (ICDM), 2012 IEEE 12th International Conference on*. IEEE, 2012, pp. 309–318.

[39] A. Scherp, T. Franz, C. Saathoff, and S. Staab, “F–a model of events based on the foundational ontology dolce+ dns ultralight,” in *Proceedings of the fifth international conference on Knowledge capture*. ACM, 2009, pp. 137–144.

[40] M. Katell, “Méthode danalyse de données en régression non linéaire,” *Hall, Hoboken, NJ, USA*, pp. 7–8, 2013.

[41] J. R. Hobbs, “Toward a useful concept of causality for lexical semantics,” *Journal of Semantics*, vol. 22, no. 2, pp. 181–209, 2005.

[42] J. Kim, “Causes and counterfactuals,” *The Journal of Philosophy*, vol. 70, no. 17, pp. 570–572, 1974.

[43] M. A. Guessoum, R. Djiroun, K. Boukhalfa, and E. Benkhelifa, “Natural language why-question in business intelligence applications: model and recommendation approach,” *Cluster Computing*, pp. 1–24, 2022.

[44] M. A. Guessoum, R. Djiroun, and K. Boukhalfa, “Towards decisional natural language why-question recommendation approach in business intelligence context,” in *2019 International Conference on Networking and Advanced Systems (ICNAS)*. IEEE, 2019, pp. 1–6.

[45] C. W. Granger, “Investigating causal relations by econometric models and cross-spectral methods,” *Econometrica: Journal of the Econometric Society*, pp. 424–438, 1969.

[46] J. Pearl, “Causal inference,” *Causality: objectives and assessment*, pp. 39–58, 2010.

[47] K. Karimi and H. J. Hamilton, “Timesleuth: A tool for discovering causal and temporal rules,” in *Tools with Artificial Intelligence, 2002.(ICTAI 2002). Proceedings. 14th IEEE International Conference on*. IEEE, 2002, pp. 375–380.

[48] M. B. Shrestha and G. R. Bhatta, “Selecting appropriate methodological framework for time series data analysis,” *The Journal of Finance and Data Science*, vol. 4, no. 2, pp. 71–89, 2018.

[49] W. G. Cochran, “Approximate significance levels of the behrens-fisher test,” *Biometrics, JSTOR*, vol. 20, no. 1, pp. 191–195, 1964.

[50] V. K.-S. Liew, “Which lag length selection criteria should we employ?” *Economics bulletin*, vol. 3, no. 33, pp. 1–9, 2004.

[51] R. Wang, J. Liu, Q. Zhang, C. Fu, and Y. hou, “Federated learning for feature-fusion based requirement classification,” *Cluster Computing*, pp. 1–20, 2023.