When Less is More: Mining Infrequent Events from Medium Sized Datasets

Ana Cristina Bicharra Garcia
Departamento de Informatica Aplicada
Universidade Federal do Estado do Rio de Janeiro
Rio de Janeiro, Brazil
cristina.bicharra@uniriotec.br

Adriana Santarosa Vivacqua
Departamento de Informatica
Universidade Federal do Rio de Janeiro
Rio de Janeiro, Brazil
avivacqua@dcc.ufrj.br

***Abstract*—Data Science has assembled researchers from multiple fields, such as computer science and marketing, to develop methods, algorithms and techniques to discover knowledge from large amounts of data and solve complex problems. The main focus is to find patterns hidden within the data. In general, these patterns must appear frequently to be learned using machine algorithms and reflect the standard, even if tacit, knowledge that guides decisions in the world. The uncovered knowledge may also reflect societal prejudice that will be enforced by the machines. Nevertheless, in many domains, specially when the negative impacts of bad decisions have a high cost, a few instances of patterns within the dataset suffice to warrant further investigation. Considering that decisions will be based on knowledge learned from the data, the challenge lies in determining when fewer appearances of a sequence of events are more important than very frequent patterns. In this context, we have devised an approach that uses a domain ontology to boost these infrequent, but relevant, events. In addition to guiding the search for relevant knowledge, the ontology helps users accept results and further investigate the data. It enables users to create data subsets and views, deriving new attributes from existing ones, guiding the data mining process, and providing a background layer from which even not so frequent patterns stand out and become meaningful. This paper presents our ontology-based data discovery approach, a system developed according to it, and preliminary results of a real life application in the oil production domain.**

***Index Terms*—data science, data mining, ontology, cause-effect relationship, knowledge filter, association rules**

I. Introduction

Knowledge discovery from existing data is still a challenge with ad hoc solutions in specific domains [16]. Data mining researchers have devoted their time to the creation of efficient algorithms to manipulate huge amounts of data to uncover interesting patterns or rules that govern fundamental correlations between uncorrelated data. Cloud computing, parallel programming and data reduction methods have been used by computer scientists to improve data mining techniques. Although a fundamental component, these algorithms play a supporting role withing the knowledge discovery process. Data pre-processing and result post-processing take the starring roles in knowledge discovery [1] [14] [18]. Consequently, data science (DS) professionals need to master the pre- and post-processing activities to guarantee useful results. A successful discovery process requires DS professionals [15] to:

- prepare the data input data by identifying and eliminating "garbage" data, creating new features merging or deriving from existing features and studying and adjusting the dataset before conducting the mining process,
- select and adjust the data mining parameters
- interpret the results by creating feasible hypotheses that explain the findings, often acquiring more information,
- act according to the discovered knowledge.

The process is rarely linear, involving many cycles to reach a conclusion.

Apriori, C4.5, k-Means, SVM, EM, PageRank, AdaBoost, kNN, Naive Bayes, and CART are the most adopted data mining techniques to address the problem of classification, clustering, prediction and association analysis tasks [6]. No matter the technique, they all look for frequent patterns or correlations that may eventually lead to causal relations [3]. These methods address computational challenges concerning time response (how long it takes to reach solutions) and dataset size management (what is the dataset size limit). Additionally, not rarely data mining yields large amounts of patterns or knowledge rules that need to be analyzed, leading DS professionals to increase frequency thresholds.

As a consequence, low frequency patterns are ignored. However, there are domains for which even low frequency occurrences should emerge such as DNA sequencing, to obtain rare genetic variants [4] [13] or high-risk industrial activities, to obtain rare high-impact cause-effect relation [17] [11]. The challenge is to allow low frequency patterns to emerge while controlling the false-positive outcomes.

This paper presents a low-frequency boost (LFB) data mining technique based on domain ontology and the Apriori association rules. Knowledge from the ontology is used to adjust the weights of the occurrences, boosting the low frequency occurrences due to their high semantic importance either in the domain or for the desired task, such as solving a specific problem. We took a design science research method approach [10]: we implemented an information system called DMRISCO that embodied our LFB technique for accidents' cause-effect knowledge discovery in the petroleum production domain. We evaluated the results with actual case scenarios

This project was supported by CNPq and FAPERJ.

and users. Preliminary results indicated relevant low frequent patterns emerged with no significant increase of false positives. We also highlight an added benefit: users' gain a better understanding of the outcomes which eases result evaluation and acceptance.

II. Application Domain: Accidents in petroleum production offshore platform

A. Maintaining the Integrity of the Specifications

Offshore petroleum production is the predominant activity in the Brazilian oil industry, as most reservoirs lay in offshore areas. According to ANP (Brazilian National Agency for controlling the petroleum industry), as of 2018, Brazil has 82 pre-salt offshore, 643 conventional offshore and 7390 onshore oil production wells. Despite the high number, onshore wells produce very little oil. Most of the offshore oil wells are located in the state of Rio de Janeiro. The 2017 oil production was 2,6 million BPD. About 26 companies operate the oil fields in Brazil, however the Brazilian petroleum company responds for about 90% of the oil production. The offshore fields are explored by sixty-four oil platforms, operated by more than twenty thousand workers.

The large numbers of workers, tight working environment, and the dangerous nature of the activities lead to an emphasis on avoiding accidents. Learning from past accidents becomes vital to prevent recurrence. The oil industry safety and regulations require logging any kind of accident or even non-conformity so that whenever necessary events will be traceable. One of these requirements states that any unexpected event must be recorded according to a reporting protocol. Unexpected events' reports include information concerning the people involved, consequences to the processing unit, the set of actions taken to solve the problem and the set of actions taken to prevent recurrence. All petroleum companies operating in Brazil must record any problems that happen during operation, be they non-conformities, accidents, deviations, incidents or unexpected results.

B. Ontology as a unified domain model

Ontologies have been used in computer science since the late 80s [8], for increasing human shared understanding of a domain [9], guiding computer implementations and allowing reuse [12]. We consider an ontology to be a description of a domain negotiated by a community for its specific purposes. Ontologies enable computational reuse and human shared understanding, two important aspects of our system. Therefore, using an ontology was an appropriate solution.

Our ontological description of a domain contains a list of concepts and the relations between them. Each concept has a name, a list of pertinence rules and exceptions. A relation has a name, the number of terms included in the relationship and the description of the behavior resulting from the application of the relation to the concepts. Each concept has a list of possible values (instantiations).

We took into account over 600 accident reports selected from the period of 2006 till 2014, from a Brazilian oil company, considering the relevance of the accident impact. We considered non-conformity, incident, low-impact accidents and high impact accidents. The sample reflected the proportions of the original dataset. Fig. 1 presents a partial view of the accident ontology created for domain understanding, which also functions as the main building block of our LFB method.

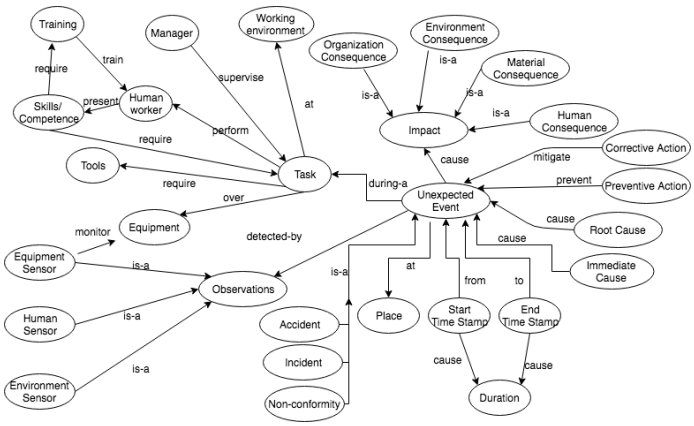

Fig. 1. Accident domain ontology sample.

A description associated to this ontology might read: *"An unexpected event occurs at a place, starts at a date, and finishes at a different date, consequently there is a duration of the unexpected event. The unexpected event occurs while a task is being performed by people, during an activity, which is performed within a working environment. The unexpected event is caused by an immediate cause, but is in reality caused by a root cause. Corrective actions mitigate the unexpected event and preventive actions will prevent the recurrence of the unexpected event. The unexpected event is observed by observations that can be performed by human, equipment or environment sensors."*

The domain ontology includes rules and triggers that indicate how certain concepts behave given circumstances (e.g., what constitutes high financial loss? or, what is considered an environmental cause for accidents?). The system is flexible enough that it allows engineers to configure individual settings (for instance, one value might be high for one business unit and low for another, depending on its total production). Although the ontology may get out of date due to changes in the world it represents, given a certain time frame, it should satisfactorily represent a domain.

The initial ontology created from analyzing the accident reports was adjusted and validate through meeting with accident analysts experts that took 13 full day meetings, spread over about 5 months. A total of 10 people participated in the process, with 6 core participants and 4 occasional participants, who filled in for others or came only when requested. The meetings were conducted in a special meeting room, with the necessary information available, to generate discussion and ensure that corporate rules were taken into account. At each meeting, a large graph of the ontology was available for manipulation, providing an up-to-date view of the ongoing

work. Flipcharts for discussion and an interactive whiteboard where documents could be displayed were also available. All meetings were filmed for later review.

In the first meeting, a brief explanation of ontologies and their usage was given, as well as the reasons for having one in the system. As soon as all participants were in agreement regarding meeting goals, discussion started. Participants created a domain model based on their personal experiences and the views they wanted of the process. This means they defined which ones were the important variables that should be taken into account for the analysis. They also defined the set of values they were interested in. After completion, the ontology was validated through usage on a pilot dataset and revised according to users comments.

III. The Knowledge Cohesion Approach

The Knowledge Cohesion (KC) [7] data mining technique is based on the Apriori algorithm [2]. However, unlike Apriori, KC returns itemsets, not association rules. It executes itemset generation as in the Apriori algorithm, and then calculates each itemsets KC, which is used as a way to rank and prune the result set. The KC model is based on the assumption that existing domain knowledge can be used to mine large datasets. It relies on a domain ontology to calculate two measures for each itemset: semantic distance and relevance assessment. These measurements are then combined to generate an itemsets knowledge cohesion. In this section, we describe the KC data mining method.

The KC method is based on the assumption that domain knowledge can be used to help yield better data mining results. Therefore, it requires that the domain of the dataset be described as an ontology. Ontology is a description of a domain, usually constructed by a group of people who understand the domain and the ontologys usage. An ontology is usually represented a set of concepts and relations between them, in many ways similar to a semantic network. These concepts and relations between them are carefully chosen by ontology engineers to reflect the domain.

In an ontology, concepts are domain specific, but relations are usually not. Generic relations such as is-a, part-of and causes, appear frequently and across domains. Given this observation, a Semantic Distance (SD) value was assigned to each relationship. This table was empirically derived and qualifies each type of relation according to how much new information is acquired by following it. For example, two concepts linked by an is-a relationship, such as dog is-a animal, are probably very similar. Therefore, an itemset that has both dog and animal will probably have high support but is a result that contains little new knowledge (and has low SD, i.e., is less interesting). The semantic distance values are shown in Table I. Note that some relationships are asymmetric, transitive and reflexive, therefore weights are provided for three possible cases.

The semantic distance between a pair of terms is the Euclidean distance between terms, calculated by following the ontological connections in its minimal path, which may be calculated using the Dijkstra algorithm. Thus, for n terms, an n x n matrix containing the terms of the ontology can be built, where each cell contains the calculated SD between two terms. This matrix is used to calculate KC. By assigning values to links in general, instead of assigning weights to relations between specific terms, this model becomes more generally applicable. The values reflect the correlations between each term pair, and were assigned to indicate the level of novelty of a pair of concepts. The second element in KC calculation is a Relevance Assessment (RA) of each element of the ontology. This assessment is a user-provided expression of interest in a given ontological element: it allows the user to determine which pairs of concepts (or even relations) he or she would like to see in or exclude from the resulting itemsets. The RA guides the search, as, during the mining process, it increases the importance of itemsets that contain the elements specified by the user. According to the specification, RA can have one of three values: low (weight = 1), average (weight = 2) or high (weight = 4). The default value is average, while low indicates the user is not interested in itemsets containing these elements and high indicates the user wants to see itemsets containing these elements. Given the values for SD and RA, KC is calculated as a function of both. The KC algorithm initially generates candidate itemsets, as per the Apriori algorithm [2], using support metrics as a cutoff. Then, it calculates KC for each itemset, and compares this calculation with the KC cutoff point. If KC is higher than the cutoff metric, the itemset is inserted in the result set. The KC formula is presented in Equation 1:

TABLE I
Relationship semantic distance values (taken from [7])

Relation	SD(A,B	SD(B,A	SD(A,A
Is-a	0	0.5	0
Part-of	1.5	1.0	0
Is-attribute-of	2.5	2.0	0
Causes	2.5	2.5	0.5
Precedes (time)	2.5	2.5	0
Precedes (space)	2.5	2.5	0
Others	3.0	3.0	0

$$\bar{KC(IS)} = \frac{log(N)}{C*} * \sum_{i=1}^{C*} \frac{RA_(IPair_i)}{SD(IPair_i)} \quad (1)$$

in which:

- IS = itemset = Term$_1$, Term$_2$, Term$_3$, ..., Term$_n$
- N = length(itemset)
- ItemPair denotes the combination of two items from the resulting itemset
- C* = The number of combination of the N elements in the itemset in pairs of 2.
- IPair = Items Pair = a pair of two different items from the itemset. RA is the relevance assessment of the items withing the itemset for the domain SD is the semantic distance between two items defined in IPair$_i$

The KC algorithm is composed of three main phases:

- Generate candidate patterns which include
- For each candidate itemset, (a) Calculate KC (itemset) (b) If KC(itemset) ¿ min KC, then itemset is solution
- For each itemset solution, insert into solution pattern

In this fashion, the KC algorithm identifies patterns by taking into account the domain knowledge and users' interests. We argue that this approach could be improved by adding a contextual layer that affects both SD and RA weights. This is explored in the following section

IV. CKC: Contextual Knowledge Cohesion

KC enables contextual datamining. The domain knowledge maps relations between items in a dataset and steers the datamining process towards results appropriate for that particular domain. However, the semantic distance table contains static information, so its effectiveness is faded to decay with time.

We postulate that, even within a domain, there will be different scenarios or contexts that require data to be treated differently. The role a relation between concepts may vary according to the problem scenario, thus requiring adjustments in the semantic distance table. The relevance assessment may also vary with the problem context. Contextual Knowledge Cohesion (CKC) is an extension of the KC data mining technique that includes the context information to dynamically adjust the semantic table.

A. CKC Method

A traditional way of modeling user interests or expertise is through a technique called overlay modeling [5]. In overlay modeling, a user model that expressed how much a user knows about a certain item is overlaid on top of a domain model. The user model points to the domain model, and places weights on each concept to signify how much is known about it. Through this strategy, it is possible to effect personalization an adaptation of user interfaces. We propose the use of a similar approach to contextualize Data Mining results.

We considered the users' background and preferences and the task specification as guidelines for adjustments in the semantic distance table. The user's background and preferences define the level of expertise and consequently help prune irrelevant knowledge. Task specification refers to the goal to be achieved by the data mining process. We take a data analytics perspective in which there is always a problem to be solved, thus mining has a purpose. Examples of tasks include finding evidences for an intuition or finding causal relation for highly negative/positive outcomes.

To contextualize KC results, we define two context-sensitive multipliers that affect the outcome of SD and RA. The Contextual Semantic Distance (CSD) and the Contextual Relevance Assessment (CRA) are two factors that are multiplied by SD and RA values respectively. Figure 2 shows how contextual models are applied to the domain model in the CKC approach. The Contextual Semantic Distance (CSD) value expresses the relevance of a given relationship in a given context.

The calculation of CRA accounts for novelty and clarity. Novelty reflects the number of different relations connecting the items in the itemset. Assuming itemset composed of concepts A, B and C. If the three of them are related by the "is-a" relation, we assume it brings low degree of novelty. Thus the importance of the itemset should be decreased by a factor of exp(-3). On the other hand, if there is no known relation connecting these three items, there might be some new knowledge worth looking at. Thus their KC value should be boosted to exp(3). Accordingly, concept clarity adjusts the KC value for the itemset. Unfamiliar concepts should boost KC by a factor of exp(N) in which N is the number of unfamiliar concepts. As shown in eq.2 , KC is boosted by CRA.

$$C\bar{RA}(IT) = exp^{(\sum(NNovel+NFuzzy)-(\sum(NKnown+NClear)} \tag{2}$$

in which:

- IS = itemset = Term_1, Term_2, Term_3, ..., Term_n
- N = length(IS)
- NKnown = length(related concepts)
- NNovel = N-NKnown
- NClear = number of concepts the users claim good familiarity
- Fuzzy = N - NClear

Task and domain relevance are the boosters composing the CSD factor. Task relevance reflects the number of items related to the mining goal and the domain relevance concerns the relative importance of a concept related to the entire domain. For instance the concept "Impact" is highly important for the accident domain ontology. Accordingly, having two concepts causally related boost the itemset when the mining goal is to root cause analysis. Equation 3 shows the CSD calculation.

$$C\bar{RA}(IT) = exp^{(\sum(NTask+NDomain)-(\sum(NFreeT+NFreeD)} \tag{3}$$

in which:

- IS = itemset = Term_1, Term_2, Term_3, ..., Term_n
- N = length(IS)
- NTask = length(task related concepts)
- NFreeT = N-NTask
- NDomain = number of core domain concepts
- NFreeD = N - NDomain

Both CSD and CRA require users to provide information at the beginning of the knowledge discovery process. It does not need to be complete, and we assume that this information will appear during the interaction. The application of these multipliers yields the CKC value for itemsets, which are applied in the same manner as with the KC algorithm to prune results.

Each relationship type in Table I has its SD weight multiplied by a contextual factor, generating a CSD weight, which takes into account how relevant the relationship is to the DM process. Thus, the Dijkstra calculation will yield values for pairs of items that reflect this conceptualization. This we call

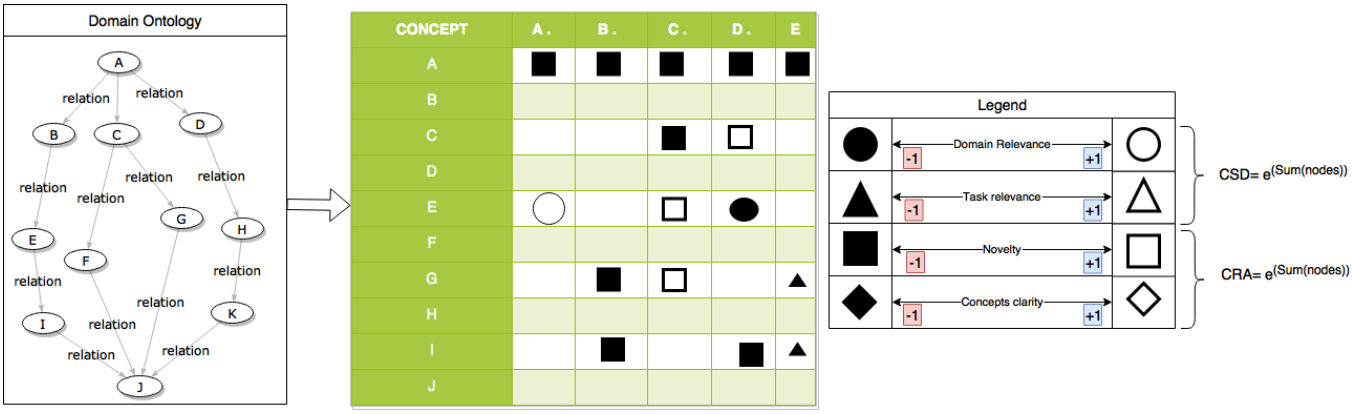

Fig. 2. CKC boost factors calculation. The example shows an ontology with 11 concepts and their relationships. The matrix on the right present the 4 different boost or preventer factor to be used to adjust KC

the CSD value of an itempair. The KC formula is thus altered to include the CRA and CSD factors, as shown in Equation 4

$$KC\bar{(}IS) = \frac{log(N)}{C*} * \sum_{i=1}^{C*} \frac{RA(IPair_i) * CRA(IPair_i)}{SD(IPair_i) * CSD(IPair_i)} \quad (4)$$

in which:

- CSD is the contextual semantic distance of the pair of items considering the task at hand and the users profile
- CRA is the contextual relevance assessment of the pair of items considering the task at hand and the users profile

Having a contextual weight assigned to each of the values allows us to manipulate the results according to context. Thus, given a large amount of data, if we were to assume a plant shutdown situation, information about the plan prior to the shutdown, and about the machinery and equipment involved is more valuable than information about employee performance or production in previous years. In this case, these elements could be emphasized and others deemphasized. Another possibility is a teaching situation: some concept relationships may be obvious to the experienced practitioner, but not for a novice student. Is-a relationships are generally deemphasized in KC calculations, because they add little new knowledge to the result. However, an inexperienced person might need to learn is-a relationships to be able to fully understand more complex ones (and results that include these), so these should be emphasized in a first moment, and, as the novice builds on the knowledge, he or she should be able to move on to more complex ones.

V. CKC APPLIED TO THE PETROLEUM ACCIDENT MANAGEMENT DOMAIN

In our case study domain, accidents are undesired events and corrective actions must be taken to prevent recurrence. Corrective actions might camouflage posterior events, making it difficult to investigate the underlying reasons that lead to root causes of apparently uncorrelated events.

Talking to users during the testing phases, it became clear that some events are so critical that they need to be investigated even though they might happen only a few times in the database. However, reducing the support threshold to mine this data leads to an enormous set of result possibilities. The challenge remains, as it is difficult for users to make sense of a large database. Looking for frequent patterns (high support value) in this scenario may not lead to useful results. On the other hand, low support leads to a set of patterns too large to investigate. We believe the ontology makes the whole difference, helping guide the search for relevant, although infrequent patterns.

Our dataset is not large: about 40000 itemsets, but growing fast. Each itemset structurally reports an accident. A description is composed of 75 attributes (not all are mandatory, so a description will usually feature about 50 attributes). Attributes have an average of 10 possible values, a minimum of 3 and a maximum of 200 possible values. Some attributes are already correlated in the domain ontology. This yields a total of around 3.000.000 items, which means a very large number of possible combinations. Even though the dataset contains 40000 itemsets, results with a support of 10-20 are usually much more relevant than results with greater support. Given that some attributes are correlated, result sets with large support usually lead to known correlations or more abstract rules. When faced with these initial results, users complained the process wasnt generating useful results, and helped us identify certain events that were critical but only appeared a few times.

We built a system and tested it with real accident management analysts for the petroleum domain. As illustrated in Figure 3, the tool offered standard visualization and preprocessing functionalities and the popular data mining techniques such as Apriori and K-means. The novelty was the introduction of KC and CKC technique.

The KC technique greatly reduced the outcomes maintaining the quality when compared with the traditional techniques [7]. We interviewed three senior petroleum company employees and one junior. We videotaped them using the system and captured their reaction to the outcomes. We started out using traditional data mining techniques and then moved to KC. later on, we started showing the results for the CKC. They

evaluated the top 10 outcomes for each method, classifying them as useful or not. One of the CKC outcomes triggered their immediate attention because they said the sequence of events should never happen. It shouldn't have happened, not even once. It might reflect a failure in a accident barrier strategy. They started calling people and the experiment was over.

As illustrated in Figure 4, there are sequences of events that only appeared few times in the dataset, but might be very important. There are many improvements required as far as system usability is concerned. Domain relevance and concept clarity are easy for them to input, while the mining task was somewhat hard for them to understand.

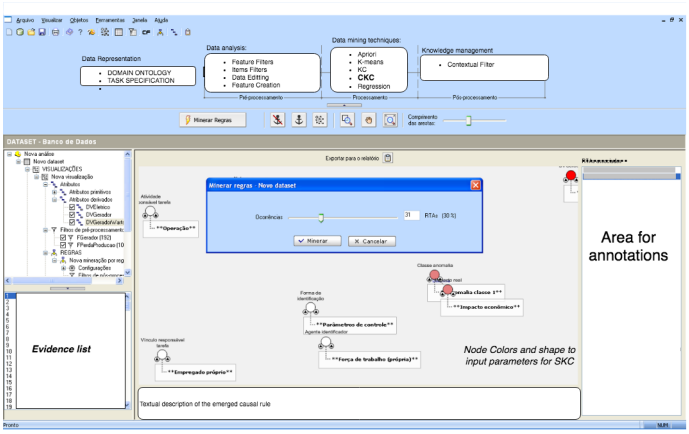

Fig. 3. CKC boost factors calculation. The example shows an ontology with 11 concepts and their relationships. The matrix on the right present the 4 different boost or preventer factor to be used to adjust KC

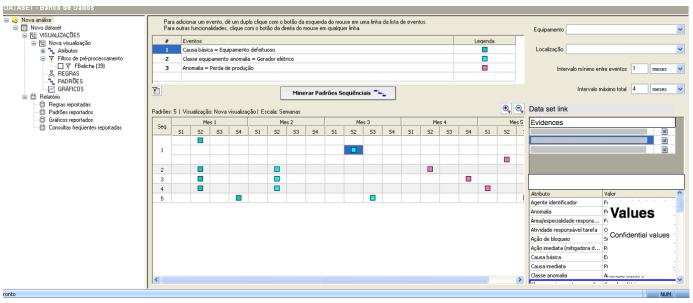

Fig. 4. CKC boost factors calculation. The example shows an ontology with 11 concepts and their relationships. The matrix on the right present the 4 different boost or preventer factor to be used to adjust KC

VI. Conclusion

This paper presents an extension of a data mining technique that handles the problem of boosting low frequent itemsets that matters in the domain. The outcome are association rules and a possibility for a domain ontology expansion. The method requires an initial domain ontology and explicit information concerning the users domain background and goal for the knowledge discovery process. Thus, it takes a data analytics perspective of mining with purpose.

We presented the CKC model and a system that implemented it. The system was evaluated by open focus group interviews in which users evaluated the outcomes for real scenarios and the usability of the tool. They were surprised by some outcomes because they revealed sequences of events that shouldn't happened. The users' evaluation questioned the meaning of the task oriented parameter that adjusts the CKA. They suggest to remove it. More tests are needed to fully evaluate the system and refine the approach. We believe there are many scenarios for which finding sequence of events with low recurrence matters, due to the high impact. CKC addresses these situations.

References

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. *arXiv preprint arXiv:1803.02453*, 2018.

[2] Rakesh Agarwal, Ramakrishnan Srikant, et al. Fast algorithms for mining association rules. In *Proc. of the 20th VLDB Conference*, pages 487–499, 1994.

[3] Yacine Aït-Sahalia and Dacheng Xiu. Principal component analysis of high frequency data. *Journal of the American Statistical Association*, (just-accepted), 2017.

[4] Lorenzo Bomba, Klaudia Walter, and Nicole Soranzo. The impact of rare and low-frequency genetic variants in common disease. *Genome biology*, 18(1):77, 2017.

[5] Peter Brusilovsky and Eva Millán. User models for adaptive hypermedia and adaptive educational systems. In *The adaptive web*, pages 3–53. Springer, 2007.

[6] B. D. Cullity. *Introduction to Magnetic Materials*. Addison-Wesley, Reading, MA, 1972.

[7] Ana Cristina Bicharra Garcia, Inhauma Ferraz, and Adriana S Vivacqua. From data to knowledge mining. *AI EDAM*, 23(4):427–441, 2009.

[8] Thomas R Gruber. Toward principles for the design of ontologies used for knowledge sharing? *International journal of human-computer studies*, 43(5-6):907–928, 1995.

[9] Parvin Hashemi, Ameneh Khadivar, and Mehdi Shamizanjani. Developing a domain ontology for knowledge management technologies. *Online Information Review*, 42(1):28–44, 2018.

[10] Alan Hevner and Samir Chatterjee. Design science research in information systems. In *Design research in information systems*, pages 9–22. Springer, 2010.

[11] Matthias Jakob, John J Clague, and Michael Church. Rare and dangerous: Recognizing extra-ordinary events in stream channels. *Canadian Water Resources Journal/Revue canadienne des ressources hydriques*, 41(1-2):161–173, 2016.

[12] Maulik R Kamdar, Tania Tudorache, and Mark A Musen. A systematic analysis of term reuse and term overlap across biomedical ontologies. *Semantic web*, 8(6):853–871, 2017.

[13] DG MacArthur, TA Manolio, DP Dimmock, HL Rehm, J Shendure, GR Abecasis, DR Adams, RB Altman, SE Antonarakis, EA Ashley, et al. Guidelines for investigating causality of sequence variants in human disease. *Nature*, 508(7497):469, 2014.

[14] Sergio Ramírez-Gallego, Bartosz Krawczyk, Salvador García, Michał Woźniak, and Francisco Herrera. A survey on data preprocessing for data stream mining: current status and future directions. *Neurocomputing*, 239:39–57, 2017.

[15] Jeffrey S Saltz and Nancy W Grady. The ambiguity of data science team roles and the need for a data science workforce framework. In *Big Data (Big Data), 2017 IEEE International Conference on*, pages 2355–2361. IEEE, 2017.

[16] Bharat Tidke and Rupa Mehta. A comprehensive review and open challenges of stream big data. In *Soft Computing: Theories and Applications*, pages 89–99. Springer, 2018.

[17] Pietro Turati, Nicola Pedroni, and Enrico Zio. An adaptive simulation framework for the exploration of extreme and unexpected events in dynamic engineered systems. *Risk analysis*, 37(1):147–159, 2017.

[18] Shichao Zhang, Chengqi Zhang, and Qiang Yang. Data preparation for data mining. *Applied artificial intelligence*, 17(5-6):375–381, 2003.