A Semantic Approach to Uncovering Implicit Relationships in Textual Databases

Dildre G. Vasques and Solange O. Rezende
ICMC
University of Sao Paulo
Sao Carlos, Brazil
Email: dildre.vasques@usp.br; solange@icmc.usp.br

Paulo S. Martins
School of Technology
University of Campinas
Limeira, Brazil
Email: paulo@ft.unicamp.br

Abstract—The discovery of knowledge in textual databases is an approach that basically seeks for implicit relationships between different concepts in different documents written in natural language, in order to identify new useful knowledge. To assist in this process, this approach can count on the help of Text Mining techniques. Despite all the progress made, researchers in this area must still deal with the large number of false relationships generated by most of the available processes. A semantic approach that supports the understanding of the relationships may bridge this gap. Thus, the objective of this work is to support the identification of implicit relationships between concepts present in different texts, considering the verbal semantics of relationships. To this end, analysis based on association rules were used together with metrics from complex networks and a verbal semantics approach. Through a case study, a set of texts from alternative medicine was selected and the different extractions showed that the proposed approach facilitates the identification of implicit causal relationships.

Keywords—concept map; semantic analysis; text mining; knowledge acquisition; complex networks; association analysis.

I. INTRODUCTION

The amount of knowledge accumulated in documents written in natural language represents a potential source for new discoveries. Due to the advancement of digital media, this knowledge can be stored and made available to those who need it [1]. However, researchers must deal with a significant number of publications, the overload of information, and the lack of data structure. This scenario promotes many challenges related to the acquisition and analysis of this content [2].

To deal with these challenges, different models, processes, and methodologies have been proposed. Among these is the Knowledge Discovery in Text (KDT) [3][4], which may be understood as a version of the Knowledge Discovery in Database (KDD) [5]. KDT basically differs from KDD in the data preprocessing phase, which must first undergo a structuring process. In short, KDT is an approach used to find latent rules and patterns in texts that result in useful knowledge [4][6][7].

Due to the potential of the KDT process, it is fundamental to develop models that facilitate the process of discovering implicit relationships. To make this process viable, faster and more efficient, one can count on the support from Text Mining (TM) techniques [8][9][10]. TM fundamentally relates to the discovery of useful information from unstructured or semi-structured documents. Its techniques enable automatic analysis of a large textual set with little manual intervention.

However, the area surrounding KDT must still deal with the complexity inherent in the overall structure of manipulated texts, since these documents are composed of complex and heterogeneous data. Therefore, it is essential to identify and consider existing relationships in the dataset [11] to better understand the general context of the processed texts.

Association analysis is one of the several TM techniques presented in the literature [12], which is responsible for highlighting indirect relationships and making explicit potentially useful connections between the words that make up a set of documents. The main advantage of an associated set of words is the fact that it relates to the context of the document [13], and it contributes to the maintenance of its representativeness. However, some of the discovered relationships are potentially false, because they may happen simply by chance. Thus, the results of an association analysis must be carefully interpreted, since inference by an association rule does not necessarily imply causality. Causality requires the knowledge of cause-and-effect attributes in the data [11].

Despite the available technology, the extraction of knowledge from texts remains an arduous task [14]. Textual documents have a structure that requires the application of specialized techniques due to the implicit meaning assigned to each word in the human language [15] and between the relationships in which those words are involved. We note, however, that several researchers are only committed to the meaning of the word and its syntactic function in a sentence, or else, they only focus on the meaning of words and their hierarchical (taxonomic) relationships [16][17].

In most cases, the methods frequently used in the KDT process disregard the contextual semantic of the processed texts and, therefore, they provide few elements corresponding to the general content of the text. The contextual semantic is related not only to the order (i.e. syntax) in which the words are distributed in a sentence but also to the semantic role that these words play in that sentence and how the set of all sentences intertwine to form the general context of a document. Therefore, semantic research techniques focused on the verb can help in the specification of this type of
information. Verbs lead to the specification of words that fit in a sentence, the morphological forms of these words, and the order in which they should appear in the sentence [18][19][20][21] as well as their cause-effect relationships [22]. However, in addition to the problem of textual data complexity, the KDT process must also deal with the large number of words present in a dataset. To deal with the issue, statistical measurements from complex networks may be used. There is a wide range of quantitative measures in the complex networks approach available for the characterization of the topological properties of a linguistic network. These metrics can aid in the process of assigning weights to the main words and relationships present in the data. In this sense, metrics are considered as a type of weight assignment, in order to select a subset of more representative words. By using this approach, we can apply a filter to select valid and useful relationships for the discovery of knowledge.

The interest in modeling and analyzing text as complex networks has been on the rise in recent years and a considerable amount of research in this area has already been accumulated [23][24]. These investigations demonstrate that several aspects of natural language systems may be represented as complex networks. Their vertices depict linguistic units (such as words, morphemes, and phonemes), while links shape their relationships [25][23]. In this context, the goal of this work is to apply a TM technique based on association analysis to process the data, complex network metrics to the post-processing data phase and verbal semantics to allow a visualization of causal relationships and the human inference of implicit relationships. This approach allows us to add the advantages of the quantitative and qualitative analysis models to the TM processes. The extracted relationships may contribute to the discovery of implicit relationships by the user assisting in the KDT process (as discussed in Section III).

In essence, our methodology is based on rules of association that enable the identification of frequently associated items (words) in the entire corpus (set of processed texts). These items and their relationships are then represented in the network format and are analyzed using measures from complex networks, that provide a ranking of the concepts (nouns) that stand out in the set. With this ranking, we obtain the main concepts of the network. With these main concepts and their explicit relationships, we model a causal conceptual map with the aid of linguistic techniques based on verbal semantics. This map enables us to manually infer implicit relationships. The main contribution is the application of a process aimed at understanding the semantic aspects of the causal relationships between the concepts of a domain.

This article is structured as follows: Section II discusses related work. Section III shows the proposed approach. Section IV addresses the implementation and the results obtained. Finally, Section V presents the conclusion.

II. RELATED WORK

The explosive growth of the Internet has substantially increased the available amount of textual information. Thus, automatic text analysis applications have also grown [25], which stimulated the development of different research in the field of KDT and TM. Bhardwaj and Khosla [12] point out that in TM, one of the main components to be considered in the treatment of textual data is context analysis. According to the authors, the techniques of association rules extraction stand out in obtaining this type of information. The association process is responsible for evidencing indirect relationships, with the purpose of explaining potentially useful connections between concepts.

In this sense, Ruiz et al. [26] use association rules to merge information extracted from various datasets in order to decrease the volume of data, their distribution and volatility. The model proposed by the authors produces meta-association rules, that is, rules in which the antecedent or the consequent may also contain rules, to find joint correlations between the trends found individually in each dataset. According to the authors, this approach produces a set of more manageable rules for human inspection. It also allows the incorporation of contextual information into the mining process, expressed in a more human readable format.

d’Amato et al. [27] demonstrated that the use of association rules can also help in the coupling between ontologies and statements that may be out of sync. The authors proposed a method to discover a multi-relational relationship through the use of association rules, considering intentional knowledge. Moreover, the rules discovered could be directly integrated into the ontology, enriching its expressive power and increasing the assertive knowledge that could be derived. The discovered rules were also able to suggest new axioms to be added to the ontology.

Rai et al. [28] used association rules extracted from a textual database to provide additional information that could be extracted from future input documents. This improved the recovery of the underlying extraction system. Thus, the author developed a project to transform unstructured and semistructured digital data into structured data with the help of Information Extraction and KDT techniques. The results of this study were presented through the application of these techniques to an employment news corpus from the Internet.

Amin et al. [29] used association rules to assist in the development of a dengue control system in Pakistan. Through the rules obtained, the authors demonstrated that if a virus attacks somewhere, it is possible to predict its next geographic target. This is because the dengue virus will be spread from one place to another through contaminated water and the carrier mosquito of the disease. Therefore, by using the above-mentioned data sources, performing some preprocessing with techniques such as transformation, filtering, stemming and indexing of documents, and then applying association rules extraction, the system was able to help identify patterns of geographic spread of viruses, and suggest the next geographic location where the virus is most likely to develop.

In addition to work that use association rules extraction, we also find in the literature work from complex networks for the exploration of the contextual aspects of a document. Wachs-
Lopes and Rodrigues [25] developed a study that proved that complex networks also provide a useful tool to aid in the understanding of context-related issues. In their study, the authors modeled two complex networks, the first being in English and the second in Portuguese, and they presented the study of the dynamics of these two networks. They showed the behavior of the small world and the influence of the hubs, thus suggesting that these databases have a high degree of modularity, which indicated specific contexts of words. In addition, they presented a method for computing the clustering coefficient and extracted other features such as the proportion of reciprocal connections and mean connection density for both networks.

Ke et al. [30] also used complex networks to develop a method to analyze the quality of nonlinear Chinese-language text. For the development of the study, the authors used texts produced by university students in China. These texts were then represented as free-scale networks (word adjacency model), from which typical network resources were obtained, such as clustering coefficient and network dynamics. The results revealed that complex network features of different text qualities can be clearly revealed and used in potential applications for other text analysis.

Xu et al. [31] developed Knowle, an online news management system, in which they introduced a semantic link network model. The central elements of Knowle are Web news events linked by their semantic relationships. Knowle is a hierarchical data system, composed of three different layers formed by concepts, resources and events. In short, this system provides the various semantic relationships between these layers. In their case study, the system was used to organize and mine health news, and it showed its potential to analyze large databases in the health field.

All of these works reveal the importance that TM techniques and complex networks approach assume in the context of KDT. However, for the most part, the methods used still have little emphasis on contextual semantics, thus impairing the performance in terms of the representativeness of explicit relationships and the discovery of implicit relationships. Thus, we note that a large part of the recent KDT and TM research has focused on reinforcing more advanced semantic applications. Our work aims at the syntactic and semantic exploration of sentences for more detailed information and more meaningful analysis of the extracted relationships [32][33][34][35].

The process of extracting implicit relationships used in our work, in addition to having techniques for extracting association rules and metrics from complex networks, uses as attributes the semantics contained in the verbs that make up the existing relationships between concepts. This type of semantic analysis allows the understanding of causal relationships existing between the concepts of the texts analyzed. Thus, it seeks to maintain, as far as possible, the contextual semantic load, in order to use the relationships extracted in knowledge discovery.

It may be seen from the literature that there is early work that uses association rules, complex networks, and work that combines both. To our knowledge, no prior work has combined association rules, complex networks and conceptual maps, structured around verbal semantics, to support the task of uncovering implicit relationships in textual databases.

III. PROPOSED MODEL

Much of the information produced and made available by digital means is in textual format. The texts do not present data organized in rows and columns and, therefore, they are composed by a set of unstructured data. Because it does not have a formal structure, this type of information becomes difficult to analyze. To assist in the process of extracting relationships from this type of information, we can rely on a set of techniques from different areas, which may contribute to a more advanced TM.

However, the KDT process is not trivial, mainly due to the lack of methodologies that emphasize the contextual aspect in which the extracted relationships are involved. Thus, the present work proposes a process for discovering implicit relationships present in a text corpus, using three different techniques: association analysis, complex networks metrics, and analysis based on verbal semantics. Fig. 1 illustrates the phases (and their respective stages) that make up the proposed process for uncovering implicit relationships. These phases enable the interconnection of textual content represented by concepts in a domain of analysis with the purpose of supporting the process of knowledge discovery.

To test the feasibility of the proposal, we apply the process to a set of ten health texts dealing with alternative medicine, which are processed and represented in a network format in order to enable the exploration of data from different points of view.

We highlight that, one of the major challenges of TM is the high dimensionality of the words. A set of documents may have hundreds or even thousands of words. However, many of these words are irrelevant and do not contain useful information for the mining task. An option to deal with this issue is to analyze the words in terms of their relationships. In this type of approach, the first step in the analysis is to find relationships between the different words to then execute the remaining analysis using these relationships, rather than analyze the words themselves [11].

Fig.1 shows the phases of the proposed process. The first phase is to prepare the data and derive patterns that summarize the underlying relationships in the data (preprocessing and data mining). The second phase performs the weighing of words to reduce the dimensionality, i.e. filter out unrelated words and keep the relationships that play an important role in the model. In this sense, the metrics may be considered as a type of weight assignment, in order to select a subset of more representative words (post-processing). The third and last phase is used as a technique to add verbal semantics to the relationships (use of knowledge). This phase aims to extract and explain the results obtained through the visualization, considering that a different view of textual data may reveal important features.
Thus, the representation model used to visualize the data is a causal conceptual map, since it is able to capture and reveal important aspects of the relationship between the words (i.e., concepts) under analysis. The next subsections address each phase and the corresponding steps.

A. Association Analysis

Association analysis is used to find patterns that describe highly associated characteristics within the data [11]. Discovered relationships can be represented in the form of association rules. Thus, association rules [36] extract the relationships between items (words) in a database, in which $A \rightarrow B$, meaning that when a term $A$ occurs, a term $B$ also tends to occur. The strength of an association rule is measured in terms of support and confidence. Support determines the frequency with which a rule is applicable to a particular dataset. Confidence, in turn, determines how often items in a set appear in transactions that contain another set of items [11].

Transactions are sets of words (items) under the same context of analysis. These transactions are typically obtained from sentences, paragraphs, or a sliding window. In summary, association rules for textual documents encode important information about the relationships between items (i.e., words, documents, etc.) and, ultimately, are able to identify patterns, themes, and context [37]. Thus, in this work, we use the association analysis with the objective of: (i) mining the data (extracting patterns); (ii) identify the words that are most frequently related; (iii) decrease the volume of data; (iv) provide pairs of related words for the construction of a complex network.

The required steps to extract association rules are described below:

- **Textual Documents Selection.** This first step consists in selecting the textual documents that will compose the process input dataset.
- **Preprocessing of Data.** Preprocessing transforms documents written in natural language into a viable data format, which is then used to extract interesting relationships and information relevant to the KDT process. To improve the quality of mining standards and the time required for mining, data cleaning techniques must be applied in the preprocessing phase. This cleanup aims to remove special characters (digits, punctuation and accent marks, and line breaks), plus some stop-words (articles, prepositions, conjunctions).
- **Mapping the Textual Document into Transactions.** After the preprocessing steps, the set of text is mapped into transactions (different sets of items). In association analysis, a collection of items is called an item set. The mapping of the textual document into transactions allows the extraction of association rules. To obtain the interactions in this work, we opted for the sliding window technique. Thus, the first transaction contains only the first word, the second transaction contains the two first words, and so on, until the window contains the defined number of words [38].
- **Extraction of Association Rules.** For the extraction of association rules from transactions previously defined, we may set in the tool the value of the threshold of automatic minimum support. This value considers the mean frequency of words in transactions, thus exempting the user from knowing the characteristics of the document or the collection of documents, and avoiding the definition of a very low minimum support value, which could generate a very large amount of rules [38].

The execution of these three steps (i.e., preprocessing, mapping and extraction) is carried out with the support of the Features gEnerator tool [38].

B. Weighing of Items

Metrics can be considered as an automatic weight assignment type for the objects that make up the dataset [11]. This weighing is an alternative used to select the words that must be maintained or eliminated from the set of texts. Heavily connected words must be selected because they play an important role in the overall context of documents. Thus, the selected words (items) and their relationships will compose a subset of data. This subset enables us to reap the full benefits of reducing the number of words.

In this context, the approach based on complex networks can provide important metrics to assist in the selection of words. In this work, we use complex networks to: (i) perform the post-processing of the data; (ii) model a complex network with the
association rules extracted in the previous step; (iii) extract metrics from that network; (iv) use some measures to weigh items; (v) rank the core items of the network; (vi) extract a subset of items; (vii) decrease the number of items. To execute this phase, we perform the following steps:

- **Construction of the Complex Network.** With the association rules extracted in the previous step, we may build the complex network. In a simplified way, the concept of a complex network is that of a set of elements that are connected to each other [39]. A complex network can be described as a graph and as such inherits its conceptual properties. A graph is defined as a mathematical structure consisting of two sets: \( V \) (vertices) and \( E \) (edges), resulting in the formula \( G = (V, E) \) [40]. A certain vertex is adjacent to another if there is an edge that joins both. Thus, graphs are able to model textual content and may be useful in TM steps [41]. A network representation can increase the number of links that make up the intermediate path between different words, allowing the extraction of all the words involved in a relationship. Thus, a network is able to represent different types of objects and different types of relationships, obtaining better results in comparison with traditional algorithms of knowledge representation, based on the space-vector model [42]. In this phase of development of the work, we use a non-directed graph because the direction of the relationships will be considered only in the Semantic Analysis phase.

- **Extraction of Metrics.** Complex networks have a set of definitions that can describe a series of behaviors related to a network, such as the degree of a vertex, its hubs, the average connectivity, the shortest path, its diameter, clustering coefficient, closeness centrality, betweenness centrality, among others. To this end, we use the Gephi Tool [43], a free software available at [https://gephi.org](https://gephi.org). It enables us to view and analyze networks by providing their various metrics. For the development of this work, we used the following metrics: 1) the average degree of a vertex (number of connections); 2) hubs (vertices with greater intensity of connections); 3) betweenness centrality (quantifies the participation of a vertex \( i \) in paths of minimum length), and 4) clustering coefficient (number of connections between the neighbors closest to a vertex) [44].

  These metrics are used because they emphasize the most connected nodes. These nodes are closer to the underlying context of interest. Due to their larger connectivity, they may correspondingly uncover a larger number of implicit relationships than the nodes that are less connected. On the other hand, a remote node would reveal far less implicit relationships and these could be more outside the context of interest.

- **Selection of Concepts.** The metrics obtained in the previous step may help with the identification of the main concepts (nouns) of the network. The Gephi tool generates a ranking of the words (vertices) for each of the measures of interest, providing a list in descending order of corresponding values, which effectively contributes to the identification of the main concepts of interest, that is, the most representative concepts of the network built with association rules.

C. Semantic Analysis of Relationships

Meaningful words or sentences embedded in any paragraph of a full text can be processed to extract logical relationships. With a chain of logical connections, hidden relationships can be identified. From the identification of semantic relationships between words, it is possible to identify the logical meaning of such relationships, without the need for explicit words such as “cause” or “effect” [45]. This idea was also explored by Vasques et al. [22] who, with the help of verbal semantic analysis, developed a process called Verbka that is capable of extracting causal relationships from a text. To this end, a basic distinction between a participant \( X \) (Cause) and one (or more) participant(s) \( Y \) (Effect) is necessary.

In this phase of work we: (i) select the main concepts (main nouns) ranked in the previous step; (ii) search the original textual dataset for the sentences in which those concepts appear; (iii) divide the sentences into linguistic blocks using Verbka to extract logical relationships; (iv) create a causal conceptual map to enable the visualization of the mined result; (v) add verbal semantics to the explicit relationships; (vi) make inferences and find implicit relationships. The steps needed for the execution of this phase are detailed as follows:

- **Selection of Concepts.** Based on the word ranking obtained in the previous phase, it is possible to make a selection of two (or more) key concepts. The selected concepts serve as the basis for the construction of a subset of data.

- **Capture of Sentences.** With the selection of key concepts performed in the previous step, we capture the sentences present in the original dataset that contain at least one of the previously selected key concepts.

- **Structuring of Sentences.** Verbka decomposes a sentence into a set of minimal linguistic blocks formed by concepts. This is possible by applying a sequence of steps that input the preselected sentences and output a structured table of concepts and their relationships.

- **Representation of relationships.** The table obtained in the previous step is then mapped into a conceptual map, which enables the visualization of the explicit relationships from the data subset. Conceptual maps graphically illustrate the relationships between the concepts of a knowledge domain [46]. Each concept is positioned within a circle or box (vertices) and relates to other concepts by means of arcs (edges) which represent the connections (verbs, prepositions). Concepts connected by an arc give rise to a proposition, which is the particular characteristic of conceptual maps [47].

- **Extraction of Semantic relationships.** After the construction of the map, it is possible to assign semantics to the
verbal relationships existing between the concepts, and to create a causal conceptual map. The verbal relationships are divided into three categories which are represented by colored arrows: red arrows represent action verbs, i.e. causal relationships where the action starts from a “cause” concept and affects an “effect” concept, which in turn is modified by this action (e.g. she breaks the vase); blue arrows indicate reflexive relationships, that is, actions that depart from a cause concept and affect themselves (e.g. she washes herself, she feels happy); black arrows represent non-action verbs, indicating the absence of action and, therefore, absence of a causal relationship (e.g. she stays at home).

- Identification of Implicit relationships. Based on the visualization of the map, we can then proceed to the qualitative analysis. This analysis should be performed by humans as it requires an overview of the context. Therefore, based on this map, we look for connections that are not explicit in the network, but can be inferred from the knowledge that already exists there. It is the cause-effect semantics present in the map and the visual aid (layout of information) that it provides that facilitate the identification of implicit relationships by humans.

IV. IMPLEMENTATION AND RESULTS

In this section, we address the experiments that were carried out using the methodology described in Section III. Initially, we decided to use a set of ten textual documents ranging from alternative medicine ("human gestation" and "vitamin" domains), in order to extract the implicit relationships in these texts. The texts were taken from the Cochrane website 1, which is a global and independent network of researchers, practitioners, patients, caregivers, and people interested in health, who work to produce reliable health information. The texts had an average of 550 words and the combination of all texts had 5575 words.

After selecting the texts, we proceed to the data pre-processing phase and the extraction of association rules with the support of the FEATuRE tool [38]. The texts were pre-processed separately, since the frequency of a given word within the whole collection may be negligible, whereas if processed separately, it is not. We opted for the removal of stop words to decrease the amount of words. The texts were not subject to stemming in order to keep the semantic quality (related to the meaning) of the extracted words. In this work, we chose word frequency instead of stemming to generate transactions.

The documents were then mapped into a set of transactions with sliding windows of size five and step one (this size was chosen since it represents the usual number of connected words in a phrase). Sliding windows were chosen instead of sentences since they generate more transactions. The sliding windows correspond to excerpts from textual documents and they extract sets of words that are related to a specific context of the document. Table I provides an example of the transactions extracted.

![Table I: Examples of transactions extracted from a text](http://brazil.cochrane.org/)

For the extraction of association rules from the transactions previously obtained, we configure the tool for operation in automatic support mode. This mode takes into account the average frequency of words in transactions, thus exempting the user from knowing the characteristics of the collection of documents. This avoids the definition of a low minimum support value, which could generate a large number of rules [38]. Table II shows some examples of extracted rules, where the first value represents the support and the second the confidence values.

![Table II: Example of association rules composed of two words extracted from transactions](http://brazil.cochrane.org/)

We considered the association rules only for sets of two words in order to preserve a common structure for the construction of the complex network in the next phase. In possession of these rules, we start to build the complex network using the Gephi tool. For the network modeling we consider each set of rules as related items. Thus, the network was created by mapping items as vertices of the network and the relationship between them was represented as non-directed edges. We emphasize that in this phase, the order of the relationship (edge sense) between the items does not interfere with the measures of interest. Upon execution, the tool generated a network as a non-directed graph of 161 vertices.

Networks have several properties that are part of their topology. Having built the network, we move on to the calculation of the metrics. After performing these calculations, it was possible to generate a ranking of items based on the metrics, i.e. degree of a node, clustering coefficient, hubs and betweenness centrality. Notice that the verbs must be excluded
from the ranking, since 1) they represent the connections between the items that compose the causal conceptual map in the next phase, and 2) in this phase we wish only to capture the main nodes (items, i.e. nouns) that should compose the causal concept map. In a later stage, these verbs need to be retrieved back to complete the map.

Table III shows the ranking of the average degree and hubs, and Table IV shows the ranking of the betweenness centrality and clustering coefficient. We emphasize that the first items of the ranking are the most significant and are more likely to have implicit relationships due to their importance to the network.

### Table III

**Ranking of Vertices Using A) Average Degrees and B) Hubs**

<table>
<thead>
<tr>
<th>Vertices</th>
<th>Avg. Degree</th>
<th>Vertices</th>
<th>Hubs</th>
</tr>
</thead>
<tbody>
<tr>
<td>vitamin</td>
<td>12</td>
<td>premature</td>
<td>8.87E+0.3</td>
</tr>
<tr>
<td>labor</td>
<td>6</td>
<td>risk</td>
<td>8.87E+0.3</td>
</tr>
<tr>
<td>pregnancy</td>
<td>5</td>
<td>evaluated</td>
<td>2.31E+0.3</td>
</tr>
<tr>
<td>supplement</td>
<td>4</td>
<td>studies</td>
<td>8.14E+0.2</td>
</tr>
<tr>
<td>studies</td>
<td>4</td>
<td>hypothyroidism</td>
<td>8.14E+0.2</td>
</tr>
<tr>
<td>women</td>
<td>4</td>
<td>vitamin</td>
<td>0.575</td>
</tr>
<tr>
<td>orofacial</td>
<td>4</td>
<td>supplements</td>
<td>0.284</td>
</tr>
<tr>
<td>premature</td>
<td>3</td>
<td>pregnancy</td>
<td>0.236</td>
</tr>
<tr>
<td>cocoa</td>
<td>3</td>
<td>prenatal</td>
<td>0.227</td>
</tr>
<tr>
<td>pressure</td>
<td>3</td>
<td>labor</td>
<td>0.110</td>
</tr>
<tr>
<td>obstrictrics</td>
<td>3</td>
<td>studies</td>
<td>0.106</td>
</tr>
<tr>
<td>risk</td>
<td>3</td>
<td>folate</td>
<td>0.090</td>
</tr>
<tr>
<td>cleft</td>
<td>3</td>
<td>premature</td>
<td>0.089</td>
</tr>
</tbody>
</table>

### Table IV

**Ranking of Vertex Using A) Betweenness and B) Cluster, Coeff.**

<table>
<thead>
<tr>
<th>Vertices</th>
<th>Betweenness Centrality</th>
<th>Vertices</th>
<th>Clustering Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>vitamin</td>
<td>117.5</td>
<td>prenatal</td>
<td>1</td>
</tr>
<tr>
<td>labor</td>
<td>834</td>
<td>malformation</td>
<td>1</td>
</tr>
<tr>
<td>risk</td>
<td>798</td>
<td>risk</td>
<td>0.333</td>
</tr>
<tr>
<td>pregnancy</td>
<td>487.5</td>
<td>premature</td>
<td>0.333</td>
</tr>
<tr>
<td>supplements</td>
<td>470.5</td>
<td>cleft</td>
<td>0.333</td>
</tr>
<tr>
<td>women</td>
<td>420</td>
<td>supplements</td>
<td>0.166</td>
</tr>
<tr>
<td>folate</td>
<td>323</td>
<td>orofacial</td>
<td>0.166</td>
</tr>
<tr>
<td>baby</td>
<td>270</td>
<td>labor</td>
<td>0.066</td>
</tr>
<tr>
<td>studies</td>
<td>171</td>
<td>vitamin</td>
<td>0.015</td>
</tr>
<tr>
<td>premature</td>
<td>58</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>incidence</td>
<td>58</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>malformation</td>
<td>24</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>cleft</td>
<td>20</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>orofacial</td>
<td>15</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>obstrictrics</td>
<td>15</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>models</td>
<td>10</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>control</td>
<td>10</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>cocoa</td>
<td>9</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

An analysis of these tables shows that some of the most repeated concepts in all rankings are “vitamin”, “labor”, “risk”, “premature” and “pregnancy”. The most common approach to dealing with the quantity of words is to select a small subset of these words (usually two) for the display. This technique constitutes a type of dimensionality reduction [11]. For the sake of illustration, we select two concepts that henceforth are called key concepts: “vitamin” and “premature”. Note that this step admits some flexibility in that the user (analyst) may also pick up other concepts that he or she may be interested other than the ones selected here.

The next steps of the process are focused on obtaining all the sentences of the input corpus that have one or both key concepts. All the sentences that contain the two concepts, chosen for analysis, are selected from the set of input documents (original texts). These sentences are now into linguistic blocks, as proposed by the Verbka process (Section III-C). In Verbka, a number of steps are applied corresponding to some linguistic adjustments in the sentences, such as removal of articles, separation of coordinated sentences, and transformation of sentences written in passive voice to the active voice. For the compilation of the table that structures the selected sentences in linguistic blocks (Table V), one must select all the verbs present in these sentences and then ask them some questions (who?, what?, for whom?, where?, when?). These questions reveal the subjects/nouns and the verbal complements (direct/indirect objects and accessory terms, i.e. verbal adjuncts of each sentence) (for more information on the Verbka process, the reader is referred to [22]).

With this set of selected sentences structured in tables, we move on to the construction of the concept map. According to Novak [46] and Novak and Cañas [46], an important criterion to the construction of concept maps is the ideal number of key concepts needed. This number must vary from 15 to 25 concepts to facilitate the reading and understanding of the map by humans.

In this network, the subjects and the verbal complements of each sentence are called concepts, and they constitute the vertices. The relationship between these vertices is given by either the verb or by prepositions, and are represented by the edge (or directed arrow). This arrow starts from the subject concept and reaches the object concept(s) (Fig. 2). The key concepts are highlighted in the network in orange. The concepts highlighted in yellow are the ones considered to be most important from a first qualitative interpretation of the network, since they have a strong relationship with the key concepts.

With this first constructed network model, we move to the verbal semantic aggregation phase of this network, to construct a causal conceptual map (Fig. 3). Thus, each of the edges was colored according to the verbal typology in which it was inserted: red for cause and effect relationships, blue for reflective relationships and black for static ones (i.e. those with no causal affectation). We emphasize that the network obtained in this application did not present reflexive verbal relationships (i.e. blue edges). The information based on cause and effect facilitated the understanding of the causal sequence that reaches the key concepts, thus allowing the identification, through inferences made by the researcher, of some implicit relationships, as shown in Fig. 3 (dashed edges). The dashed arrows show that “vitamin D combined with calcium” may generate “excess calcium” and may increase the “risk of premature birth.” This is possible to see, starting from the analysis of the concept “vitamin D” and following a line of...
TABLE V
FRAGMENTED SENTENCES IN LINGUISTIC BLOCKS EXTRACTED FROM THE SET OF INPUT TEXTS

<table>
<thead>
<tr>
<th>P</th>
<th>Agent - NS</th>
<th>Patient - VS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>subject</td>
<td>verb</td>
</tr>
<tr>
<td>1</td>
<td>vitamin A</td>
<td>reduces</td>
</tr>
<tr>
<td>2</td>
<td>vitamin A</td>
<td>does not help prevent</td>
</tr>
<tr>
<td>3</td>
<td>levothyroxine</td>
<td>reduces</td>
</tr>
<tr>
<td>4</td>
<td>vitamin D combined with calcium</td>
<td>increases</td>
</tr>
<tr>
<td>5</td>
<td>liver oil</td>
<td>produces</td>
</tr>
<tr>
<td>6</td>
<td>mushroom</td>
<td>produces</td>
</tr>
</tbody>
</table>

Fig. 2. Conceptual map

reasoning based on cause and effect: (vitamin D → keeps → calcium balance ← may harm ← vitamin D with calcium). If vitamin D is able to maintain calcium balance, it means that it keeps enough levels (balance) in the body under normal circumstances of ingestion of some foods. Therefore, the intake of vitamin D combined with extra calcium (given that the levels of calcium are either normal or above normal levels) may generate an “excess of calcium” (additional vertex inserted in the network), i.e. it may harm calcium balance and thus increase the risk of premature labor during pregnancy. Note that this is a hypothesis that needs to be corroborated with further scientific investigation. Nevertheless, the inference through the map was able to elicit this relationship.

This example was chosen for its simplicity as a didactic way of illustrating the process. Clearly, other implicit relationships may be uncovered that may be more difficult to grasp using a simple and conventional approach (e.g. through reading a text). Notice that, at best, the uncovering of this implicit relationship may be relatively easy obtained through the proposed approach (mainly once the tools are all integrated), in comparison to a conventional approach that would require the analyst to read more than 5000 words spread into ten different texts to achieve similar result. At worst, either the analyst would not complete this task within a reasonable time, or would not accomplish it at all given the volume of information. Thus, we argue that the approach proposed in this work reached the proposed goal: to extract implicit information through text mining techniques and complex networks in conjunction with an approach based on verbal semantics.

V. SUMMARY AND CONCLUSION
Considering the complex behavior of natural language, in this work a model capable of generating a contextualized net-
work was constructed. This model was based on the extraction and interpretation of textual data relationships. By mapping the explicit relationships in a new view space, it was possible to reveal important characteristics of the set of texts that were used to extract implicit relationships. We chose to develop a process that facilitated the visualization and observation of the main explicit relationships existing in a given dataset and, consequently, aided in the inference of new relationships that were implicit in the same set. This was made possible by the techniques used to decrease the amount of words and the application of a semantic approach that created a new set of higher-level features from the original data.

The process was able to handle the various stages of the mining process, from data preprocessing to post-processing and use of the results. With the application of association analysis in the data mining phase, it was possible to extract the most interesting patterns from the set of texts and decrease the amount of words. With the rules obtained, we were able to build a complex network and automatically extract its metrics in the post-processing phase. Thus, it was possible to assign weights to the items (words) and rank them. With this ranking, we obtained a subset of items, further reducing its amount, and thus eliminating characteristics that were irrelevant for the task of extracting implicit relationships.

The process was satisfactory in dealing with two major challenges of text mining: the complexity and heterogeneity of textual data (word diversity) and the high amount of its objects (words). With the reduction of the number of words (using complex networks) we were able to select a subset (key concepts) to model a causal conceptual map of the relationships obtained through the application of an analysis based on verbal semantics. With the reduction of the number of items (also using complex networks), we obtained a more understandable model, since the reduction of the number of words allowed a better visualization of the data in a map. This map contributed to the visualization of relationships and enabled us to analyze a relatively large amount of contextual information, thus allowing the uncovering of relationships that were implicit in the network.

Future work will focus on the automation of the knowledge acquisition process based on verbal semantics. This would allow the exploration of larger knowledge bases. Furthermore, other metrics may also be tried, including the analysis of nodes in the network with lower degree.

REFERENCES