Adaptive Study Design through Semantic Association Rule Analysis

Ping Chen, University of Houston-Downtown, USA

Wei Ding, University of Massachusetts-Boston, USA

Walter Garcia, University of Houston-Downtown, USA

ABSTRACT

Association mining aims to find valid correlations among data attributes, and has been widely applied to many areas of data analysis. In this paper we present a semantic network based association analysis model including three spreading activation methods, and apply this model to assess the quality of a dataset, and generate semantically valid new hypotheses for adaptive study design that is especially useful in medical studies. We evaluate our approach on a real public health dataset, the Heartfelt study, and the experiment shows promising results.

Keywords: Association rule mining, Semantic network, Hypothesis generation, Data quality assessment, Adaptive study design

1. INTRODUCTION

Association rule mining has been widely applied to numerous domains, such as analysis of market-basket datasets, text mining, and disease diagnosis (Agrawal et. al. 1996). Association rules whose support and confidence are above user-specified thresholds are considered statistically significant and presented to end-users. While these objective measures are effective to reduce rule redundancy, incorporation of subjective and domain-specific knowledge is still a critical challenge for association analysis, and these knowledge should be represented in a more structured way to maximize its usage. Hence, we choose semantic network to represent knowledge for association analysis. Semantic network has been implemented in many knowledge bases. Concepts and ideas in the human brain have been shown to be semantically linked, which motivates the modern research of semantic network (Quillian, 1998). On recent development on human memory study was described in Widrow et al. 2010), and a general cognitive knowledge representation model was described in Ramirez 2010. Numerous knowledge representation models for more specific areas are also proposed recently to tailor and model interesting aspects of knowledge (Chen et al. 2009).

A semantic network represents knowledge as a directed graph, where vertices represent concepts and edges represent semantic relations between the concepts. Figure 1 shows a sample semantic network whose vertices represent concepts and edges are labeled with names of relations. This semantic network was created in the case study (Section 8) when we examined the Heartfelt adolescent health study. Concepts are organized into a hierarchical structure by is-a edges, and other edges show causal relations, e.g., observable entity diagnose disease or syndrome, stressed is a mental process, diseases can be result of mental process. Comparing with other knowledge representation models, a semantic network has the following advantages:

1) Easy to use. A user needs little training or computer background to build semantic networks. Semantic networks are easy to understand and its explanation is usually straightforward.
2) Flexible, incremental, and easy to update. Building a semantic network does not require a user to have a complete or perfect understanding at the beginning. Instead, the building processing can be incremental, and knowledge can be updated locally as a user gets more familiar with a domain.
3) Generative. A semantic network is not a merely static structure, instead it has a vertex-firing mechanism called spreading activation. Firing or activation of a vertex sends activation to its semantically connected neighbor vertices. Spreading activation only accesses local neighbor vertices, so its time complexity does not grow with the size of the network.

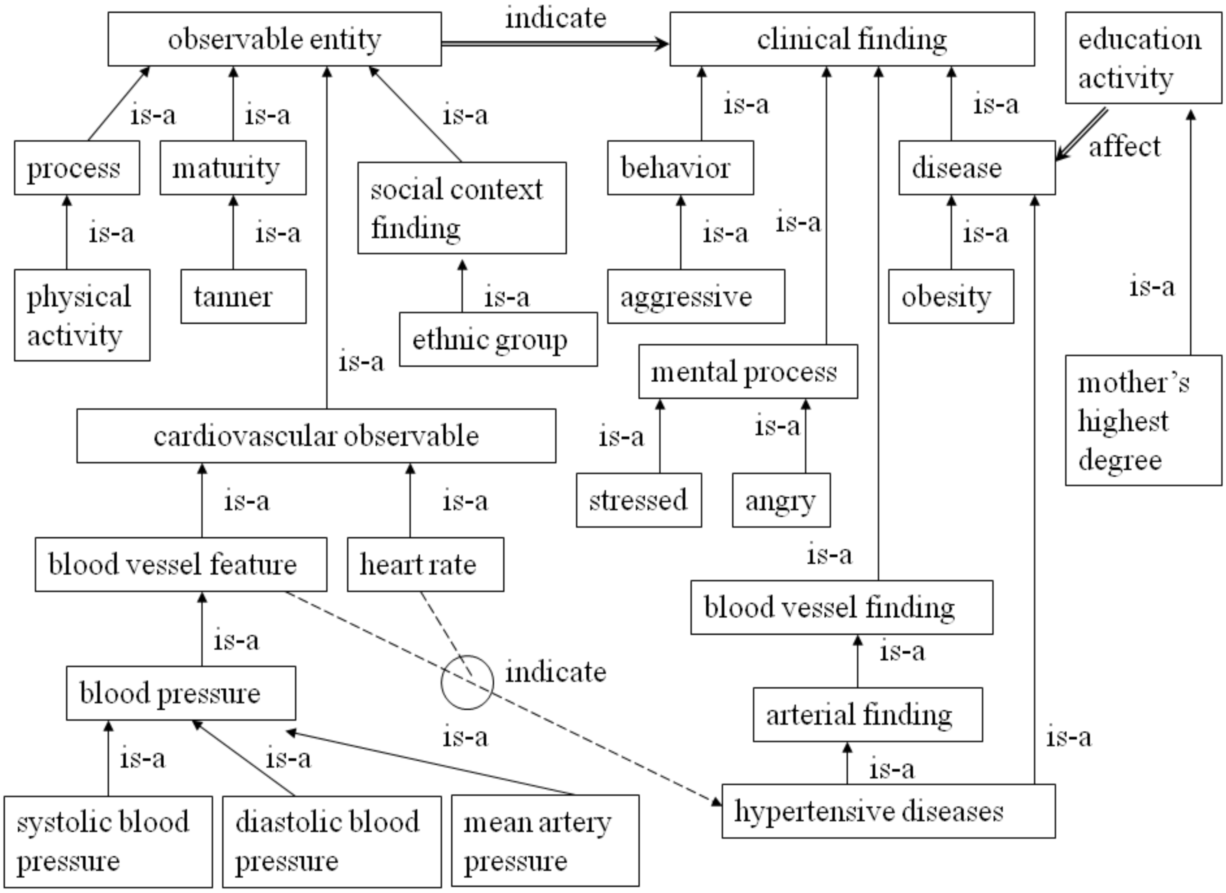

Figure 1 A Fragment of Semantic Network Used in Our Case Study

In this paper we will discuss a semantic network-based association analysis model. With this model we will provide the following analysis techniques:

1) Hypothesis generation. New hypotheses are generated through generalization and inference from the association rule set, and give end-users directions for further investigation.
2) Data quality assessment. A dataset is just an imperfect and incomplete reflection of a real-world object or scenario. By analyzing association rules we can assess the quality of original dataset.

Our work is closely related to cognitive informatics, which is a transdisciplinary field emerging from Cognitive Science, Computational Intelligence, Artificial Intelligence, Formal Semantics, and Human-Computer Interaction. A set of cognitive models for causation analyses and causal inferences was proposed in Wang 2011, which formalized causal inference methodologies to simulate subtle aspects of human reasoning.

This paper is organized as follows. In Section 2, we provide some background knowledge on related technologies including ontology, concept algebra, semantic computing, and computing with words (CWW). In Section 3 we discuss a knowledge model to represent domain and user knowledge. We present three spreading activation methods in Section 4. In Section 5 we discuss how to model semantic analysis on association rules. In Sections 6 and 7 we discuss how to assess data quality and generate hypothesis based on association mining. We evaluate our method in Section 8 using a real-world public health dataset. Related work is discussed in Section 9. We conclude in Section 10.

2. BACKGROUND

Semantic network has been applied to many fields since its introduction in Artificial Intelligence in 1960's. Many techniques in the general field of symbolic computing have been studies recently. Here we provide a brief review of several critical fields.

Ontology study focuses on formally representing knowledge as a set of concepts and relations among these concepts (Ganter et al. 2005). The goal of an ontology is to provide a shared and interchangeable vocabulary for the modeling of a domain. Hundred of ontology systems have been created for numerous domains, such as Basic Formal Ontology for scientific research, BioPAX for Biology, Customer Complaint Ontology for e-Business, Cell Cycle Ontology for medicine.

Formal concept analysis aims to building an ontology from a domain of object and entities automatically along with their relations and properties. The theoretical foundation of this filed include applied lattice and order theory (Gruber 1993).

Semantic computing focuses on understanding of meaning/semantics in a general computing environment. It studies the following critical problems (Mendel 1999):

1) User intention analysis and automatic processing;

2) Data semantics analysis and processing;

3) User goal understanding.

Computing with words (CWW) was proposed by Zadeh in 1996 using his Fuzzy Logic theory to activate and convert words into a mathematical representation (Zadeh 1996). Fuzzy set is adopted as the machinery to transform input words to output words, and then back to users. Much work has been done since then, please refer to (Lawry 2001& 2003, Mendel 2001, Mendel 2002, Mendel 2003, Mendel 2007) for more details. CWW has many potential applications, e.g., Internet search engines, summarization, information extraction.

3. ASSOCIATION MODELING WITH A SEMANTIC NETWORK

We define a semantic network SN for association rule analysis as an extended directed graph, SN = (V, A, H, S, T) (Chen 2008),

- V is a set of vertices that denote the attributes in the dataset and relevant concepts from its domain, $V = \{v_1, v_2, \cdots , v_k\}$;

- A is a set of association edges connecting multiple vertices, $A = \{(v_1, v_2, \cdots , v_n, \rightarrow u) \mid v_i, u \in V, (i = 1, \cdots , n)\}$. An association edge $v_1, v_2, \cdots , v_n \rightarrow u$ denotes an association among attributes, with $v_1, v_2, \cdots , v_n$ as the antecedent part of an association (also called the body), and u as the consequent part (also called the head).

For example, the association blood vessel feature, heart rate → hypertensive diseases is shown in Figure 1, which involves three vertices. Semantically an association edge means associated-with. In practice an edge often can be labeled with more specific relations, such as result-of, indicate, etc. If we know what values of these attributes take, an association edge can represent one or multiple association rules, $v_1 = a_1, v_2 = a_2, \cdots , v_n = a_n \rightarrow u = a$;

- H is a set of is-a edges connecting two vertices, $H = \{(v, u) \mid v, u \in V\}$. An edge v is-a u denotes a subclass-superclass relation, with v as the child, and u as the parent;

- S is a label set, S = {KNOWN, BASIC}. An association edge can be labeled with KNOWN, BASIC, or both. KNOWN labels are specified by end-users. A KNOWN association edge means that this association is already known by the user.

An experienced user knows a lot about his/her domain, and may label many KNOWN tags. So relatively less UNKNOWN knowledge will be extracted. A novice user may label only a few KNOWN tags, and a large amount of knowledge will be classified as UNKNOWN, and this is exactly what this user needs to learn.

The goal of our method is not to always incorporate all existing knowledge about a domain and make genuine discoveries, instead we aim to generating unknown knowledge customized for a specific user and improve his/her understanding about the domain. Whether this unknown knowledge is unknown to the whole domain is left to users for further analysis. Probably some new knowledge can be discovered. A BASIC edge can be obtained from a user or other knowledge sources. BASIC association edges represent highly confident principles about a domain, e.g., observable entity indicates clinical finding. There are two ways to specify BASIC labels, closed scheme and open scheme. In closed scheme, BASIC association edges exhaustively list all valid associations among vertices, and by default, any other associations are not allowed. In open scheme, a BASIC association edge means that an association among connected vertices is not allowed, and by default, all other associations are allowed, although they may or may not hold in practice. Basically whether to choose open or closed scheme is determined by the development of a domain. For a well-established domain, such as cardio-vascular research, there exists very comprehensive correlation knowledge at least among basic concepts (high level entities in UMLS (UMLS 2007) Semantic Network). In this case, a closed scheme can be adopted. The open scheme will be more suitable for an emerging field. BASIC edges are used to identify semantically invalid association rules. For example, the rule Gender = Male → Mother's Highest Degree = Master is generated in our case study, but it is not a valid rule since there is no association between Gender and Mother's Highest Degree. In the rest of our paper, the closed knowledge assumption is adopted. In the case of open assumption, the invalid rules can be further processed to identify contradictions to the given knowledge and shown to the user in order to identify interesting and useful exceptions. For a well-explored domain, our method is still useful. Knowledge generated from our technique can be used to verify and validate existing knowledge obtained with other types of techniques, especially knowledge based on personal direct or indirect experience. This is why in our semantic network we have BASIC and KNOWN labels. If a BASIC or KNOWN labeled edge is violated many times, its validity should be further examined.

- T is a set of attribute-value pairs, and $T = \{v_i = a_i \mid v_i \in V\}$.

These pairs are provided by users as not interesting or trivial instances. For example, in public health domain, Obesity = No is usually not interesting, but Hypertension = Yes is interesting.

Creation of such a semantic network can be highly automated if there exist electronic domain knowledge sources. Figure 1 shows a fragment of semantic network built for our case study. The vertices are medical concepts from a dataset. These concepts are connected with associated-with and causal relations shown as ⇒ and is-a shown as → (dashed line if its label is KNOWN, solid line if its label is BASIC).

4. SPREADING ACTIVATION METHODS

To create a high-quality semantic network, often we have to acquire many association edges and their labels from end-users and other knowledge sources. However, the hierarchical design of our semantic network can greatly lighten the burden of knowledge acquisition, and many associations can be generated by spreading activation, and a user does not have to specify every association explicitly as in other existing methods. Here are the three spreading activation methods:

1. $v_1 \rightarrow u_1 \land u_1 \rightarrow u_2 \models v_1 \rightarrow u_2$

Generally associations are transitive.

2. v_1 is-a $v_2 \land v_2 \rightarrow u \models v_1 \rightarrow u$

The antecedent part of a rule can be specialized, which is called deduction in logic. For example, Tweety is-a bird ∧ bird → fly |= Tweety →fly.

With this method, all the associations between v_2's children and u can be replaced by a single association $v_2 \rightarrow u$. For example, we do not have to specify, heart rate → clinical finding, mean artery pressure → clinical finding, ⋯, instead, one association observable entity → clinical finding will be sufficient.

3. u_1 is-a $u_2 \land v \rightarrow u_1 \models v \rightarrow u_2$

The consequent part of a rule can be generalized, e.g., fly is-a move ∧ bird → fly |= bird → move. With this method, all the associations between v and u1's parents can be replaced by a single association $v \rightarrow u_1$. For example, we do not have to specify, observable entity → blood vessel finding, observable entity → arterial finding, ⋯, instead, one association observable entity → hypertensive diseases will be sufficient.

5. SEMANTIC ASSOCIATION RULE ANALYSIS

Association rules are statistically supported by data, but not matter how massive the data is, it is just a sampling of bits and pieces at discrete times about an object or scenario, and often contains noise and erroneous information. Inevitably, rules generated from such data can be simply coincidence or even wrong. With semantic analysis, we are able to detect trivial or known association rules, weed out invalid association rules that conflict with common sense or domain knowledge, and generate a semantically validated association rule set. During this process, the basic operation is to match an association rule with the association edges in the semantic network, which will be discussed first.

Suppose we have an association rule:

$R : v_1 = a_1, \cdots , v_n = a_n \rightarrow u = a$

where v_i and u are attributes of a dataset, and the a_i and a are their values.

Definition 1. A rule R is known to a semantic network SN iff $\forall i$

$v_i \rightarrow u \in A$ and labeled with KNOWN, or $v_i \rightarrow u$ can be generated by applying the three spreading activation methods on KNOWN association edges in A.

Otherwise, a rule R is unknown to SN.

Definition 2. An association rule is semantically correct iff $\forall i$

$v_i \rightarrow u \in A$ and labeled with BASIC, or $v_i \rightarrow u$ can be generated by applying the three spreading activation methods on BASIC association edges in A. otherwise, it is semantically incorrect.

Suppose we have the following rule:

R1: blood pressure = high → hypertensive diseases = yes

In the semantic network shown in Figure 1, there does not exist a BASIC association edge between blood pressure and hypertensive diseases. But according to the spreading activation method 2: the antecedent part of a rule can be specialized, we can specialize blood vessel feature in the association blood vessel feature → hypertensive diseases and generate R1. Hence, R1 is semantically correct.

Let's look at another example,

R2: heart rate = high → Mother's degree = Bachelor

R2 is simply a coincidence, and cannot be validated by the semantic network and is semantically incorrect.

Definition 3. A rule R is non-trivial to a semantic network SN if $\exists v_i = a_i \in T$ ($i = 1, 2, \cdots, n$) or $u = a \in T$; otherwise, R is trivial.

If all attribute-value pairs in a rule are uninteresting, this rule is classified as trivial. This definition proposes a new semantic interestingness measure. A trivial rule may be correct or incorrect, but a user has little interest in it. For example, here is a rule from our case study,

OBESITY=0 STRESS1=0 ACCOM1=0 BORED1=0 RUSH1=0 → TAXHYN=0 conf: (0.96)

This rule means, if a person is not obese, does not feel stressed, accomplished, bored, or rushed, then with 96% confidence he/she does not have hypertensive diseases. Such a rule may be correct since it does not violate any common sense or domain knowledge, but it is not interesting to physicians. In practice, there may be many such trivial rules, and it is important that they are separated from interesting rules. Note that a rule is interesting provided there is at least one attribute-value pair that is not in the set T. This means that a rule like OBESITY = 0, DISEASE = YES will be considered interesting even if OBESITY = 0 is in T provided that DISEASE = YES is not in T. Also note that we do not propose to delete any transactions from the database based on T, we only use T to classify the generated rules.

As shown in Figure 2, using the semantic network described in the previous section, we group association rules into 5 semantic categories: *trivial, known and correct, known and incorrect, unknown and incorrect, and unknown and correct.* This group process is straightforward by matching association rules with labeled edges (trivial, BASIC, KNOWN) in our semantic network. Generally a user will be interested in the last category: unknown and correct. Some users may also be interested in the known and incorrect category, which indicates the contradictory knowledge from users and other domain sources.

Closed scheme requires a complete list of all valid associations (labeled as BASIC), which may look unrealistic in practice. However, in an established field, usually we have exhaustive knowledge about properties and relations of at least high-level concepts. For example, UMLS list totally 6864 associations among 189 high-level concepts (called Semantic Types), and it is unlikely that there still exist any unknown relations among them. Spreading activation methods can be used to generate associations among more specific concepts.

The quality of semantic network plays an important role in the grouping process. The more domain knowledge is incorporated and the better understanding a user has of the dataset, the unknown and correct categories will be more concise and precise. Then objective measure based methods can be applied to this group and filter out redundant association rules. By integrating objective methods with our approach, we can successfully identify non-trivial, non-redundant, semantically correct, and user-specific rules.

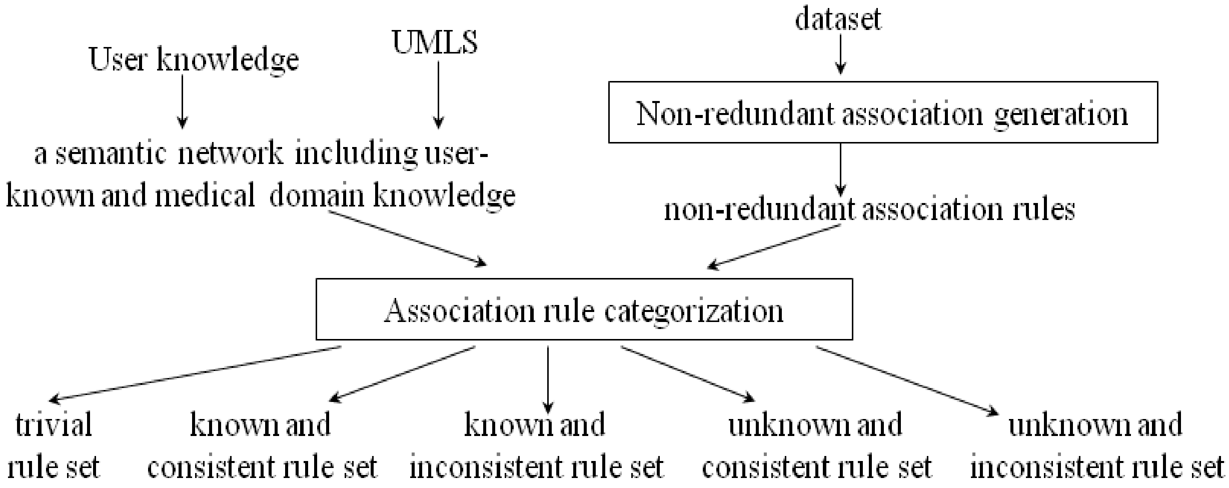
User knowledge
UMLS
dataset
Non-redundant association generation
a semantic network including user-known and medical domain knowledge
non-redundant association rules
Association rule categorization
trivial rule set
known and consistent rule set
known and inconsistent rule set
unknown and consistent rule set
unknown and inconsistent rule set

Figure 2 Semantic Association Rule Analysis System

6. HYPOTHESIS GENERATION

Generating high-quality new hypotheses is very important for knowledge discovery in scientific study. With concepts semantically organized and correlated in a semantic network, the intuition for generating hypotheses is that if two concepts are associated, maybe their semantically connected neighbors (children and siblings) are also associated. We have the following hypothesis generation methods (Chen et al. 2010),

Hypothesis Generation Method 1:

{v's child}→ u |= v → u

This is called induction in logic. If v's child is associated with u, likely v is also associate with u. Induction is useful when the direct observation of v is difficult or impossible when v is an abstract concept.

Hypothesis Generation Method 2:

$\{v\text{'s sibling}\} \rightarrow u \models v \rightarrow u$

Analogy is another technique used by human beings to generate hypotheses.

If these generated hypotheses already exist in the rule set, they will be discarded, and only new hypotheses are kept. Hypotheses are not necessarily facts, but they are more likely to be true than random guess, and they provide directions for further investigation. Additional constraints can reduce the number of hypotheses and keep only highly plausible ones, e.g., using only immediate children and siblings.

7. DATA QUALITY ASSESSMENT

A dataset is just a sampling of a real-world object or scenario at different spatial and temporal points or intervals. Naturally we want to assess the quality of a dataset, that is, how precisely they reflect reality. Data quality is a multi-dimensional concept including completeness, appropriate amount, amount of errors/missing values, objectivity, believability (Padmanabhan & Tuzhilin 1998 and 2000). Among these properties, what directly affect the quality of association rules are:

1) whether the amount of collected data is appropriate.

If the collected data is not enough to approximate the true scenario precisely, we will get many wrong or coincident rules (false negative).

2) whether the set of data attributes is semantically coherent.

An association rule is valid only if the attributes in the rule are semantically relevant. Rules generated by semantically isolated attributes will likely be coincident instead of valid. A poorly designed experiment with many isolated attributes will miss many useful and interesting rules (false positive).

Let N_x denote the number of rules of type x. To measure these two factors, we propose the following metrics,

Data Quality Metric 1:

$$Q_{size} = \frac{N_{KnownCorrect} + N_{UnknownCorrect}}{N_{nontrivialrules}}$$

We calculate the ratio of the number of semantically valid rules to the number of nontrivial rules. The intuition is that the larger a dataset is, the more closely it should reflect the basic domain principles, and the less semantically incorrect rules will be generated.

Data Quality Metric 2:

$$Q_{attribute} = \frac{N_{NewHypotheses}}{N_{UnknownCorrect} + N_{NewHypotheses}}$$

Since the hypotheses are generated by replacing original attributes with semantically similar attributes (children, siblings) in the unknown and correct rules, the more new hypotheses we get, the more semantically incomplete the original attribute set is.

8. A CASE STUDY

Public health monitoring and analysis is very important to national policy makers and general public. Public health data is generally of large volume, noisy, and high-dimensional, which is an ideal test bed for data mining techniques. Therefore we chose a public health data set collected in the Heartfelt study as our case study. All experiments were performed on a Pentium 4 3.0GHz PC running Windows XP. We used the Apriori algorithm implemented in Weka 3.4 (Witten 2005) to generate association rules.

8.1 THE HEARTFELT STUDY

In 1999, the Heartfelt study was conducted to collect data on adolescent health. The target population for this study was African, European, and Hispanic American adolescents, aged 11 -16 years, residing in a large metropolitan city in southeast Texas with an ethnically diverse population. 383 adolescents were recruited, and the collected data included totally 105 attributes and 16912 records. The attributes include age, gender, ethnic/racial group, physical maturity, resting blood pressure and heart rate, ambulatory blood pressure, heart rate and moods reported at 30-minute intervals, body mass index, fat free mass, psychological characteristics such as anger and hostility. Numerous findings have been reported based on bio-statistical analysis of the Heartfelt study, such as stress-induced alterations of blood pressure (Meininger 1999), association of obesity and poor sleep quality (Gupta et al. 2002), ethnic group differences in moods and ambulatory blood pressure (Meininger 2001), relationship of ambulatory blood pressure to physical activity (Eissa et al. 2001), etc. Here are a few findings that have been reported in medical literature:

1) sleep quality *associated-with* obesity

2) ethnicity, age, body mass index, height, maturity *associated-with* systolic blood pressure

3) fat mass, percent body fat *associated-with* heart rate

4) mood, ethnicity, maturity, gender *associated-with* systolic blood pressure, diastolic blood pressure

These associations were found with bio-statistical techniques, and are different from association rules generated by Apriori algorithm. Some transformations are necessary for evaluation, for example, association 4 can be mapped to the following association rules:

- ethnicity = African American, Maturity = high, mood = neutral, gender = boy → systolic blood pressure = high
- maturity = low, mood = rushed → diastolic blood pressure = high
- ethnicity = Hispanic American, Maturity = high, mood = neutral, gender = girl → diastolic blood pressure = high

8.2 BUILDING A SEMANTIC NETWORK FROM UMLS TO ANALYZE THE HEARTFELT STUDY

Unified Medical Language System (UMLS) is designed to help an information system understand the meanings of concepts and terms and their relationships in biomedical and health domain (UMLS 2007). The UMLS Knowledge Sources are multi-purpose, and can be used to create, process, retrieve, integrate, and aggregate biomedical and health information. UMLS divides medical ontology knowledge into three sources: the SPECIALIST lexicon, the Metathesaurus, and the Semantic Network. The SPECIALIST lexicon is designed to provide lexical information for the SPECIALIST Natural Language Processing System. The Metathesaurus is a multi-lingual vocabulary database that contains definitions of biomedical terms, their various names (such as synonyms and abbreviations), and the relationships among them. The Semantic Network categorizes all concepts in the Metathesaurus into semantic types, such as clinical finding, organisms, physical activity, etc. The Semantic Network also defines a set of relationships between biomedical concepts. These relationships provide the structure for the Semantic Network. The primary relationship is the is-a link, which establishes the hierarchy within the Semantic Network. Besides, there are also a set of non-hierarchical relationships, e.g., associated-with, affect, functionally related to. Here are a few examples,

- C0002871|CHD|C0002891|is-a|MSH

Neonatal (encoded by C0002891) has is-a relations to Anemia (C0002871)

- C0002871|RO|C0002886|clinically associated with |

CCPSS Megaloblastic anemia due to folate deficiency has clinically associated with relationship to Anemia (C0002871)

Using UMLS we created a semantic network for the Heartfelt dataset as follows (a fragment of the semantic network is shown in Figure 1):

1) Analyze the attributes in the Heartfelt dataset, assign the attributes that are semantically similar to the same vertex, e.g., age of subject in years and age of subject in months are assigned to one vertex, and totally we obtain 39 vertices;
2) Extract parent and child concepts (totally 162) of the original attributes from UMLS, and

add these new concepts and their is-a relations into the semantic network. As shown in Figure 1, majority of concepts are organized into the observable entity tree and clinical finding tree;

3) Find the semantic type of each attribute using UMLS. Different concepts can have the same semantic type, and we found totally 9 semantic types. UMLS provides 49 relations among these semantic types, and they were added into the network as associated-with or more specific edges, e.g., affect and indicate, and labeled with BASIC ;
4) Ask a user to add additional associated-with edges labeled with KNOWN and specify trivial attribute-value pairs. In our experiment, we add associated-with edges that should be known by general public, such as body mass index is associated with obesity, age is associated with sexual maturity, etc. Trivial attribute-value pairs are generally not interesting to medical personnel, such as obesity = no, blood pressure = normal, etc.

It took us about two hours to set up this semantic network. Although actual time can vary from one dataset to another and from one user to another user, once the semantic network is set up, it can be reused by other users and revised to analyze similar datasets.

8.3 EXPERIMENT RESULTS AND DISCUSSION

We applied our hypothesis generation method to the association rules generated from Heartfelt dataset, and totally we generated 1920 new hypotheses for further investigation. These hypotheses point out new attributes that a user may collect in future experiments. These hypotheses introduced new attributes (siblings and children of original attributes, excluded if they already exist), we do not have any real values for these attributes. Instead, these hypotheses describe possible correlations among semantically relevant attributes. For example,

ZBMI is associated with Maternal obesity syndrome.

ZBMI is the z-score of body mass index that measures obesity, and it is reasonable that ZBMI relates to maternal obesity syndrome. These hypothesis should be of high quality since they are based on the rules generated from real data and validated by the basic biomedical principles specified in our semantic network.

We calculated Q_{size} and $Q_{attribute}$ according to two metrics proposed in Section 6,

$$Q_{size} = 0.36$$

$$Q_{attribute} = 0.07$$

The value of Q_{size} is low and indicates that the dataset is small, which is common in biomedical field due to the prohibitive data collecting cost. A small $Q_{attribute}$ shows that the attributes in the dataset are semantically self-closed since not many hypothesis can be generated, which indicates that the Heartfelt study was very carefully designed.

9. RELATED WORK

Association rule mining aims to detect relationships or associations between specific values of categorical variables in large data sets. Association rule mining has been proved to be very useful in many applications. Various techniques have been adopted. For example, one feature-based technique was proposed using rough set theory (Liu 2010). One major obstacle in practice is how to identify correct, interesting, user-specific rules from a huge number of redundant, wrong, or trivial rules. Recently association rule post-processing has become a very active research area. Based on whether external knowledge sources are used, we can divide the existing methods into objective measure based methods and knowledge based methods.

Objective measure based methods do not require any domain information besides the rule set itself, and can be used by both domain experts and novice users. However, lack of domain knowledge makes it impossible to detect wrong rules that are just coincidence and do not make sense, and lack of user input results in presenting many rules already known by users. Based on the analysis tasks this type of methods can be further divided into:

1) Metric-based rule evaluation. This type of approaches uses metrics to evaluate the significance or interestingness of an association rule, such as lift, statistical hypothesis tests. Uninteresting rules will be discarded. However, as shown in (Wang et al. 2003), each metric has different properties and may be useful only for some specific domains and applications and choosing the right metric is often difficult.

2) Rule summarization and generalization. To reduce the number of rules that need manual analysis, rules are analyzed with their context (Liu et al. 2006). These methods investigate relations among rules in order to present users a concise rule set.

3) Rule ranking. Han et al. 2002 and Xin et al. 2006 discussed how to extract top-k significant rules with low redundancy.

10. CONCLUSION

In this paper, we discussed how to model domain knowledge with a semantic network and apply it to association rule analysis. Our semantic association rule analysis can generate semantically valid hypothesis, and assess data quality. We successfully applied our method to a public health dataset and obtained promising results.

ACKNOWLEDGMENTS

This work is funded by United States National Science Foundation grant CNS 0851984 and United States Department of Homeland Security grant 2009-ST-061-C10001.

REFERENCES

[1] Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., and Verkamo, A. I., (1996). Fast Discovery of Association Rules. In Advances in Knowledge Discovery and Data Mining, U. M. Fayyad, et. al., Eds. AAAI press, 1996.

[2] Chen, P., Garcia, W., (2010). Hypothesis Generation and Data Quality Assessment through Association Mining, The 9th IEEE International Conference on Cognitive Informatics, July 7-9, 2010, Beijing, China.

[3] Chen, P., Verma, R., Meininger, J. C., and Chan, W., (2008). Semantic analysis of association rules. In Proceedings of the International FLAIRS Conference, FL, USA, 2008.

[4] Eissa, M., Meininger, J. C., Nguyen, T., and Chan, W., (2001). The Relationship of Ambulatory Blood Pressure to Physical Activity in a Tri-Ethnic Population of Obese and Nonobese Adolescents. American Journal of Hypertension, Volume 20, Issue 2, Pages 140-147

[5] Ganter, Bernhard; Stumme, Gerd; Wille, Rudolf, eds. (2005), Formal Concept Analysis: Foundations and Applications, Lecture Notes in Artificial Intelligence, no. 3626, Springer-Verlag, ISBN 3-540-27891-5

[6] Gruber, Thomas R. (1993). A translation approach to portable ontology specifications. Knowledge Acquisition 5 (2): 199–220, June, 1993

[7] Gupta, N. K., Mueller, W. H., Chan, W., and Meininger, J. C., (2002). Is Obesity Associated with Poor Sleep Quality in Adolescents? American Journal of Human Biology : the Official Journal of the Human Biology Council, 14(6), 2002.

[8] Han, J., Wang, J., Lu, Y., and Tzvetkov, P., (2002). Mining Top-K Frequent Closed Patterns without Minimum Support. In Proceedings of the IEEE international Conference on Data Mining, 2002.

[9] Lawry, J., (2001). An alternative to computing with words, Int'l. J. of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 9, Suppl., pp. 3–16, 2001.

[10] Lawry, J., Shanahan, J., and Ralescu, A. (Eds.), (2003). Modeling With Words, Lecture Notes in Artificial Intelligence 2873, Springer, New York, 2003.

[11] Liu, B., Zhao, K., Benkler, J., and Xiao, W., (2006). Rule Interestingness Analysis Using OLAP Operations. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, 2006.

[12] Liu, Y., Jiao, L., Bai, G., & Feng, B. (2010). Feature Based Rule Learner in Noisy Environment Using Neighbourhood Rough Set Model. International Journal of Software Science and Computational Intelligence (IJSSCI), 2(2), 66-85.

[13] Meininger, J. C., Liehr, P., Mueller, W. H., Chan, W., Smith, G. L., and Portman, R. J., (1999). Stress-Induced Alterations of Blood Pressure and 24 h Ambulatory Blood Pressure in Adolescents, Blood Pressure Monitoring, 4(3-4), 1999.

[14] Meininger, J. C., Liehr, P., Chan, W., Smith, G., and Mueller, W. H., (2001). Developmental, Gender, and Ethnic Group Differences in Moods and Ambulatory Blood Pressure in Adolescents, Annals of Behavioral Medicine: a Publication of the Society of Behavioral Medicine, 28 (1), 10-9.

[15] Mendel, J. M., (1999). Computing with words, when words can mean different things to different People, in Proc. of Third International ICSC Symposium on Fuzzy Logic and Applications, Rochester Univ., Rochester, NY, June 1999.

[16] Mendel, J. M., (2001). Uncertain Rule-Based Fuzzy Logic Systems: Introduction and New Directions, Prentice-Hall, Upper-Saddle River, NJ, 2001.

[17] Mendel, J. M., (2002). An architecture for making judgments using computing with words, Int. J. Appl. Math. Computer. Sci., vol. 12, no. 3, pp. 325–335, 2002.

[18] Mendel, J. M., (2003). Fuzzy sets for words: a new beginning, Proc. FUZZ-IEEE 2003, St. Louis, MO, pp. 37–42, 2003.

[19] Mendel, J. M., (2007). Computing with words and its relationships with fuzzistics, Information Sciences, vol. 177, pp. 988–1006, 2007.

[20] Mendel, J. M., and R.I. John, (2002). Footprint of uncertainty and its importance to type-2 fuzzy sets, Proc. 6th IASTED Int'l. Conf. on Artificial Intelligence and Soft Computing, Banff, Canada, pp. 587–592, July 2002.

[21] Ramirez, C., & Valdes, B. (2010). A General Knowledge Representation Model for the Acquisition of Skills and Concepts. International Journal of Software Science and Computational Intelligence (IJSSCI), 2(3), 1-20.

[22] Sheu, P., H. Yu, C.V. Ramamoorthy, A. Joshi and L.A. Zadeh, (2010). Semantic Computing. IEEE/Wiley, 2010

[23] Padmanabhan, B., and Tuzhilin, A., (1998). A Belief-Driven Method for Discovering Unexpected Patterns. In Proceedings of the 4th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1998.

[24] Padmanabhan, B. and Tuzhilin, A., (2000). Small is Beautiful: Discovering the Minimal Set of Unexpected Patterns. In Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Boston, Massachusetts, 2000.

[25] Pipino, L., Lee, Y., and Wang, R., (2002). Data Quality Assessment, Communications of

the ACM, April 2002.

[26] Quillian, M. R., (1998). Semantic Memory, Semantic Information Processing, M. Minsky, ed., MIT Press, 1968.

[27] Sahar, S., (2002). On Incorporating Subjective Interestingness Into the Mining Process. In Proceedings of the IEEE International Conference on Data Mining, 2002.

[28] Tan, P. N., Kumar, V., and Srivastava, J., (2002). Selecting the Right Interestingness Measure for Association Patterns. In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 2002.

[29] Unified Medical Language System. (2007). available at www.nlm.nih.gov/research/umls/.

[30] Wang, K., Jiang, Y., and Lakshmanan, L. V.S., (2003). Mining Unexpected Rules by Pushing User Dynamics. In Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington D.C., August, 2003.

[31] Wang, Y. (2011). On Cognitive Models of Causal Inferences and Causation Networks. International Journal of Software Science and Computational Intelligence (IJSSCI), 3(1), 50-60.

[32] Webb, G. I., (2006). Discovering Significant Rules. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, 2006.

[33] Widrow, B. C., & Aragon, J. (2010). Cognitive Memory: Human Like Memory. International Journal of Software Science and Computational Intelligence (IJSSCI), 2(4), 1-15.

[34] Witten, I. H., and Frank, E., (2005). Data Mining: Practical Machine Learning Tools and Techniques, 2nd edition, Morgan Kaufmann, San Francisco, 2005.

[35] Xin, D., Cheng, H., Yan, X., and Han, J., (2006). Extracting Redundancy-Aware Top-k Patterns. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, 2006.

[36] Zadeh, L. A., (1996). Fuzzy logic = computing with words, IEEE Trans. on Fuzzy Systems, vol. 4, pp. 103-111, 1996.