

Semantic Analysis of Association Rules

Ping Chen

Dept. of Computer
and Mathematical Sciences
Univ. of Houston-Downtown
1 Main St.
Houston, TX 77002

Rakesh Verma*

Dept. of Computer Science
University of Houston
4800 Calhoun
Houston, TX 77004

Janet C. Meininger

School of Nursing
University of Texas
Health Science Center
200 Herman Pressler Dr.
Houston, Texas 77030

Wenyaw Chan

School of Public Health
University of Texas
Health Science Center
200 Herman Pressler Dr.
Houston, Texas 77030

Abstract

When applying association mining to real datasets, a major obstacle is that often a huge number of rules are generated even with very reasonable support and confidence. Among these rules, many are trivial, redundant, semantically wrong, or already known by end-users. Association rule post-processing aims to remove these undesired rules. Existing work mainly focuses on reducing redundant or finding unexpected association rules. In this paper, we propose an innovative method based on semantic network. We semantically divide association rules into five categories: trivial, known and correct, unknown and correct, known and incorrect, unknown and incorrect. Our method can be efficiently integrated with existing rule reduction techniques to construct a concise, high-quality, and user-specific association rule set. We evaluate our approach on a real public-health dataset, the Heartfelt study, and we can prune off 97.81% of association rules that are trivial or incorrect. The remaining rules are confirmed by either health science literature or a high-quality biomedical knowledge base.

Introduction

Association rule mining (Agrawal & Verkamo 1996) has been widely applied to numerous domains, such as analysis of market-basket datasets, text mining, biomedicine, and disease diagnosis. Current mining techniques can efficiently generate association rules that are statistically significant to the source data samples using support and confidence thresholds. However, among these statistically strong rules, some of them are redundant, some are trivial, some are already known by end-users, and some are simply coincident but conflict with common sense or basic domain knowledge. Moreover, the number of frequent rules is often prohibitive for manual analysis. Lack of effective association rule analysis techniques has become a serious obstacle for applying association rule mining to real-world datasets.

There are two major types of approaches to tackle this problem, objective measures based methods and knowledge based methods. Both methods ignore an important objective of association rule post-processing - how to identify statistically significant but semantically incorrect rules,

not simply reduce the number of rules or just search for “unexpected” rules. Association rules generated with current mining methods are only statistically significant to the source dataset. Depending on the quality of the dataset (real datasets are often noisy, incomplete, or biased), the generated rules may conflict with highly confident or common sense knowledge that can not be violated, hence are semantically incorrect. We specify the problem of association rule post-processing as:

How to identify non-trivial, non-redundant, semantically correct, and user-specific association rules?

This paper proposes a new knowledge representation model for association rule analysis. Our model requires minimal user involvement, and provides customized and semantically validated knowledge. Our method can be used with the existing objective measure based methods in a complementary manner to construct a non-redundant and high-quality rule set.

Related Work

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, 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. These methods usually use metrics to evaluate the significance or interestingness of an association rule, such as lift (Bayardo & Agrawal 1999), information-theoretic measure (Blanchard & Briand 2005), statistical hypothesis tests (Webb 2006), etc. To reduce the number of rules that need manual analysis, rules are summarized (Yan X. & D. 2005), generalized (Huang & Wu 2002), clustered (An & Huang 2003), or evaluated as a compression of the original database (Bathoorn & Siebes 2006). These methods investigate relations among rules in order to generate a concise rule set.

Our approach belongs to the knowledge based methods, which usually requires some information from users. In (Wang & Lakshmanan 2003), a user first provides his/her

*Research partially supported by NSF grant CCF-0306475.
Copyright © 2008, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

knowledge in a format similar to association rules. The more different an association rule is from a user's input, the more unexpected it is. Unexpected rules are presented to a user in a dynamic manner. In (Sahar 2002), a user is asked to evaluate interestingness of ancestor rules, which is then used to measure interestingness of rule families. In (Padmanabhan & Tuzhilin 1998) user knowledge is expressed with weighted first-order logic formulas, and association rules deviated from these formulas are marked as unexpected. Our work focuses on how to efficiently represent complex knowledge needed for association rule analysis. Our hierarchical knowledge model lightens users' burden to provide large amount of information and enables an innovative semantic analysis technique. Our method models user and domain knowledge in a more structured way instead of a flat list of rules or beliefs, also we broaden the scope of association rule post-processing.

User and Domain Knowledge Modeling

To organize knowledge in a flexible and scalable way, we choose semantic network to represent knowledge for association rule analysis. Large-scale semantic networks have been implemented in many applications, and usually require huge amount of expensive human power. However, special-purpose semantic networks (to analyze association rules in our case) with a few hundred nodes and several types of relations are relatively inexpensive to create.

Semantic Network

Concepts and ideas in the human brain have been shown to be semantically linked, which motivates the modern research of semantic network (Shapiro 1971). 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 medical concepts and edges are labeled with names of relations. 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.

A Biomedical Knowledge Base - UMLS

In this paper we use Unified Medical Language System (UMLS) from the National Library of Medicine as our domain knowledge base in the case study. UMLS is designed to help an information system "understand" the meanings of concepts and terms and their relationships in biomedicine and health domain (UMLS 2007).

In UMLS 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 contained in the Metathesaurus into semantic types, such as clinical finding, organisms, physical activity, etc. The Semantic Network also defines a set of relationships between the biomedical concepts. These relationships provide the structure for the network. The primary relationship is the "is-a" link, which establishes the

hierarchy within the Network. There is also a set of non-hierarchical relationships, such as, "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)

A Semantic Network for Association Rule Analysis

In our analysis method, we use the following two formats:

1. regular association rules: $v_1 = a_1, \dots, v_n = a_n \rightarrow u = a$, where v_i and u are attributes of a dataset, and a_i and a are their values.
2. attribute-level rules: $v_1, \dots, v_n \rightarrow u$, which means the attributes v_i are semantically relevant to u .

We define a semantic network SN for association rule analysis as a directed graph (Strictly speaking, SN is a hypergraph.), $SN = (V, A, H, S, T)$,

- V is a set of vertices that denote the attributes in the dataset and relevant concepts from its domain, $V = \{v_1, v_2, \dots, v_k\}$;
- A is a set of association edges connecting multiple vertices, $A = \{(v_1, v_2, \dots, v_n, u) \mid v_i, u \in V, (i = 1, \dots, n)\}$. An association edge $v_1, v_2, \dots, v_n \rightarrow u$ denotes an association among attributes, with v_1, v_2, \dots, 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 \rightarrow 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". If we know what values these attributes take, an association edge can represent one or multiple association rules, $v_1 = a_1, v_2 = a_2, \dots, v_n = a_n \rightarrow u = a$;
- 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 child, and u as 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 cardiovascular research, there exists very comprehensive correlation knowledge at least among basic concepts (high-level entities in UMLS 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 \rightarrow 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. The details of this post-processing are beyond the scope of this paper.

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. We use trivial attribute-value pairs to identify trivial association rules.

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 UMLS. These concepts are connected with *associated-with* and causal relations shown as \Rightarrow and *is-a* shown as \rightarrow (dashed line if its label is *KNOWN*, solid line if its label is *BASIC*). If multiple entities affect another entity simultaneously, a combined association edge will be used, e.g., blood vessel feature and heart rate indicate hypertensive diseases circled in Figure 1.

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 (Quillian 1968), and a user does not have to specify every association explicitly as in other existing methods. Here are the three spreading activation methods:

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

Associations are transitive when both $v_1 \rightarrow u_1$ and $u_1 \rightarrow u_2$ are highly confident knowledge.

2. $v_1 \text{ is-a } v_2 \wedge v_2 \rightarrow u \models v_1 \rightarrow u$

The antecedent part of a rule can be specialized, which is called deduction in logic. For example, Tweety is-a bird \wedge bird \rightarrow fly \models Tweety \rightarrow fly (Strictly speaking, this implication is not always valid, which is an interesting topic in default logic). 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* \rightarrow *clinical finding*, *mean artery pressure* \rightarrow *clinical finding*, \dots , instead, one association *observable entity* \rightarrow *clinical finding* will be sufficient.

3. $u_1 \text{ is-a } u_2 \wedge v \rightarrow u_1 \models v \rightarrow u_2$

The consequent part of a rule can be generalized, e.g., fly is-a move \wedge bird \rightarrow fly \models bird \rightarrow move. With this method, all the associations between v and u_2 ’s children can be replaced by a single association $v \rightarrow u_2$. For example, we do not have to specify, *observable entity* \rightarrow *blood vessel finding*, *observable entity* \rightarrow *arterial finding*, \dots , instead, one association *observable entity* \rightarrow *hypertensive diseases* will be sufficient.

Although there are only one antecedent and one consequent in these three activation spreading methods, in some cases they can be extended to multi-antecedent or multi-consequent rules, e.g., both u_1 and u_2 in method 1 can be multi-consequent sets. User and domain knowledge can be efficiently represented and utilized to filter out trivial, semantically incorrect, or user-known rules.

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.

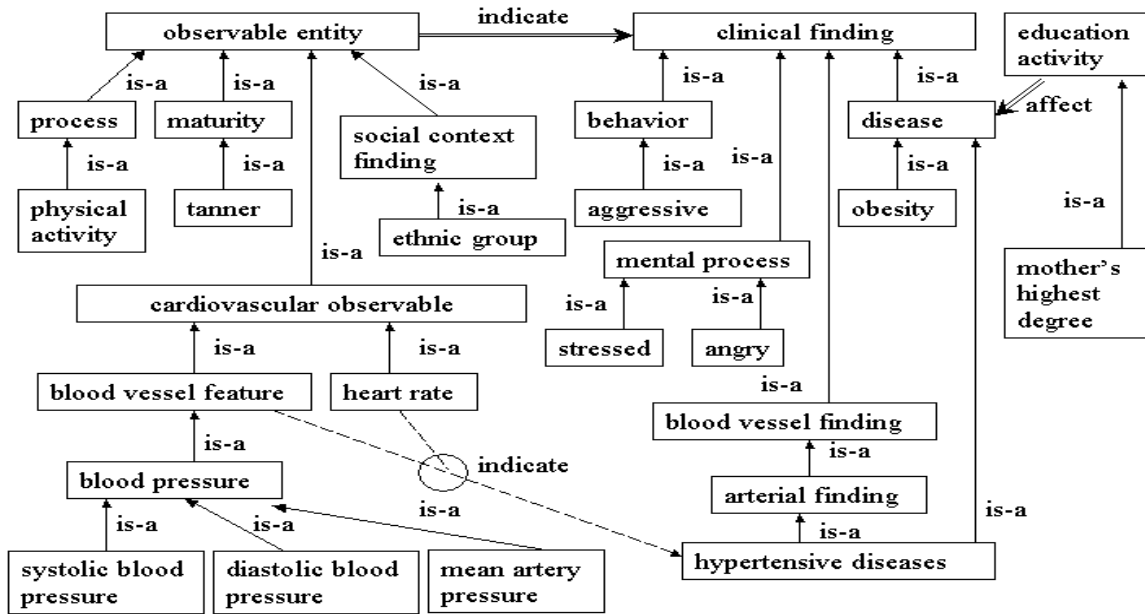


Figure 1: A Fragment of Semantic Network Used in Our Case Study

Rule Matching

Suppose we have an association rule:¹

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

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 $\forall i$

1. $v_i \rightarrow u \in A$ and labeled with *KNOWN*, or
2. $v_i \rightarrow u$ 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 $\forall i$

1. $v_i \rightarrow u \in A$ and labeled with *BASIC*, or
2. $v_i \rightarrow u$ 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:

$$R_1: \text{blood pressure} = \text{high} \rightarrow \text{hypertensive diseases} = \text{yes}$$

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

¹We assume hereafter that the body part of an association rule contains only one item, and our discussion can be easily extended to the case of multiple items.

association “blood vessel feature \rightarrow hypertensive diseases” and generate R_1 . Hence, R_1 is semantically correct.

Let’s look at another example,

$$R_2: \text{heart rate} = \text{high} \rightarrow \text{Mother's degree} = \text{Bachelor}$$

R_2 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 $\exists v_i = a_i \notin T$ ($i = 1, 2, \dots, n$) or $u = a \notin T$; 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,

$$\text{OBESITY}=0 \text{ STRESS1}=0 \text{ ACCOM1}=0 \text{ BORED1}=0 \text{ RUSH1}=0 \rightarrow \text{TAXHYN}=0 \text{ conf:}(0.96)$$

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 $\text{OBESITY} = 0, \text{DISEASE} = \text{YES}$ will be considered interesting even if $\text{OBESITY} = 0$ is in T provided that $\text{DISEASE} = \text{YES}$ 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.

Semantical Rule Grouping

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.

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 testbed for data mining techniques. Therefore we chose a public-health data set collected in the Heartfelt study (Cho 2001 Mar Apr) 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 & Frank 2005) to generate association rules.

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 old, 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 & Portman 1999). These results have been peer-reviewed and published in medical journals, and naturally serve as “gold-standard” to evaluate our method. 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

Building a Semantic Network from UMLS to Analyze the Heartfelt Study

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, which were added into the network as “associated-with” or more specific edges, e.g., “affect”, “indicate”, and labeled with *BASIC*;
4. Ask a user to add additional “associated-with” edges labeled with *KNOWN* and specify trivial attribute-value pairs. This step is subjective. 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.

Rule set	Number of rules	Percentage
Original	1,200,000	-
Trivial	813,906	67.83%
Known Correct	114,409	9.53%
Known Incorrect	184,115	15.34%
Unknown Incorrect	61,429	5.12%
Unknown Correct	26,141	2.19%
Hypotheses	1,920	-

Table 1: Experiment Results.

In our experiment, we set the support as 0.1 and confidence as 0.9 for Apriori algorithm. To generate rules more interesting to medical personnel, we filtered out the "healthy" records (e.g., Obesity = No), and totally 1.2 million association rules were generated.

Experiment Results and Discussion

We have implemented the semantic analysis techniques and analyzed the association rules generated from the Heartfelt dataset. Table 1 summarizes the experiment results. According to the semantic grouping algorithm, we divided the original 1.2 million rules into 5 groups, and *unknown and correct* group saved the 26141 association rules that may be of most interest to end users. The grouping process took 185 seconds.

Assessing quality of association rule analysis techniques is difficult. Usually these methods show how many rules are reduced or discuss a few hand-picked rules from the results, and majority of the results are not analyzed since manual evaluation is not affordable. Due to the extensive research on the Heartfelt study, we are able to evaluate our method using the correlations found by medical researchers, within the *unknown and correct* group we correctly identified 8 out of 10 correlations except the first two involving "sleep quality". We found that the measure for "sleep quality" is the total sleeping time in 24 hours, but the dataset only records whether a subject is sleeping at certain time points during 24 hours, and correlations between "sleep quality" and other attributes do not exist in the original association rule set. We can reduce redundancy among rules in each group by integrating objective measure based methods.

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 goes beyond the scope of existing association rule post-processing techniques, which mainly focus on redundancy reduction using interestingness or unexpected measures. These measures cannot sufficiently measure the usefulness or validity of rules. Our semantic analysis technique can divide association rules into 5 categories and significantly reduce a user's workload. We successfully applied our method to a public-health dataset and obtained promising results.

References

- Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., and Verkamo, A. I. 1996. *Advances in Knowledge Discovery and Data Mining*. AAAI press. chapter Fast Discovery of Association Rules.
- An, A., Khan, S., and Huang, X. 2003. Objective and subjective algorithms for grouping association rules. In *3rd IEEE International Conference on Data Mining*.
- Bathoorn, R., Koopman, A., and Siebes, A. 2006. Reducing the frequent pattern set. In *ICDM Workshop*.
- Bayardo, R. J., and Agrawal, R. 1999. Mining the most interesting rules. In *Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
- Blanchard, J., Guillet, F., Gras, R. and Briand, H. 2005. Using information-theoretic measures to assess association rule interestingness. In *Fifth IEEE International Conference on Data Mining*.
- Cho, S. D., Mueller, W. H., Meininger, J. C., Liehr, P., Chan, W. 2001 Mar-Apr. Blood pressure and sexual maturity in adolescents: the heartfelt study. *Am J Hum Biol* 13(2):227–234.
- Huang, Y., and Wu, C. 2002. Mining generalized association rules using pruning techniques. In *IEEE International Conference on Data Mining*.
- 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).
- Padmanabhan, B., and Tuzhilin, A. 1998. A belief-driven method for discovering unexpected patterns. In *4th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
- Quillian, M. R. 1968. *Semantic Information Processing*. MIT Press. chapter Semantic Memory.
- Sahar, S. 2002. On incorporating subjective interestingness into the mining process. In *the IEEE International Conference on Data Mining*.
- Shapiro, S. C. 1971. A net structure for semantic information storage, deduction and retrieval. In *IJCAI-71*.
- UMLS, 2007. Unified Medical Language System, available at www.nlm.nih.gov/research/umls/,
- Wang, K., Jiang, Y., and Lakshmanan, L. V.S. 2003. Mining unexpected rules by pushing user dynamics. In *9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
- Webb, G. I. 2006. Discovering significant rules. In *12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
- Witten, I. H., and Frank, E. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. MK.
- Yan X., Cheng H., Han J. and Xin D. 2005. Summarizing itemset patterns: a profile-based approach. In *11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*.