

# Hypothesis Generation and Data Quality Assessment through Association Mining

Ping Chen

Dept. of Computer and Math Sciences  
University of Houston-Downtown  
1 Main St., Houston, TX, USA 77002  
chenp@uhd.edu

Walter Garcia

Dept. of Computer and Math Sciences  
University of Houston-Downtown  
1 Main St., Houston, TX, USA 77002  
garciaw@uhd.edu

## 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 further investigation. 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

## 1 Introduction

Association rule mining [1] has been widely applied to numerous domains, such as analysis of market-basket datasets, text mining, and disease diagnosis. 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 asso-

ciation 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 [13]. 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. 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 mere 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.

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.

This paper is organized as follows. In Section 2 we discuss a knowledge model to represent domain and user knowledge. We present three spreading activation methods in 3. In Sections 4 and 5 we discuss how to assess data quality and generate hypothesis based on association mining. We evaluate our method in Section 6 using a real-world public health dataset. Related work is discussed in Section 7. We conclude in Section 8.

## 2 Association Modeling with a Semantic Network

We define a semantic network  $SN$  for association rule analysis as a directed graph,  $SN = (V, A, H, S, T)$  [3],

- $V$  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,  $A = \{(v_1, v_2, \dots, v_n, u) | v_i, u \in V, (i = 1, \dots, n)\}$ . 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”, etc. If we know what values of these attributes take, an association edge can represent one or multiple association rules,  $v_1 = a_1, v_2 = a_2, \dots, v_n = a_n \rightarrow u = a$ ;

- $H$  is a set of *is-a* edges connecting two vertices,  $H = \{(v, u) | v, u \in V\}$ . 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 in  $A$  can be label 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.

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”. *BASIC* edges are used to identify semantically invalid hypothesis generated in Section 4.

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

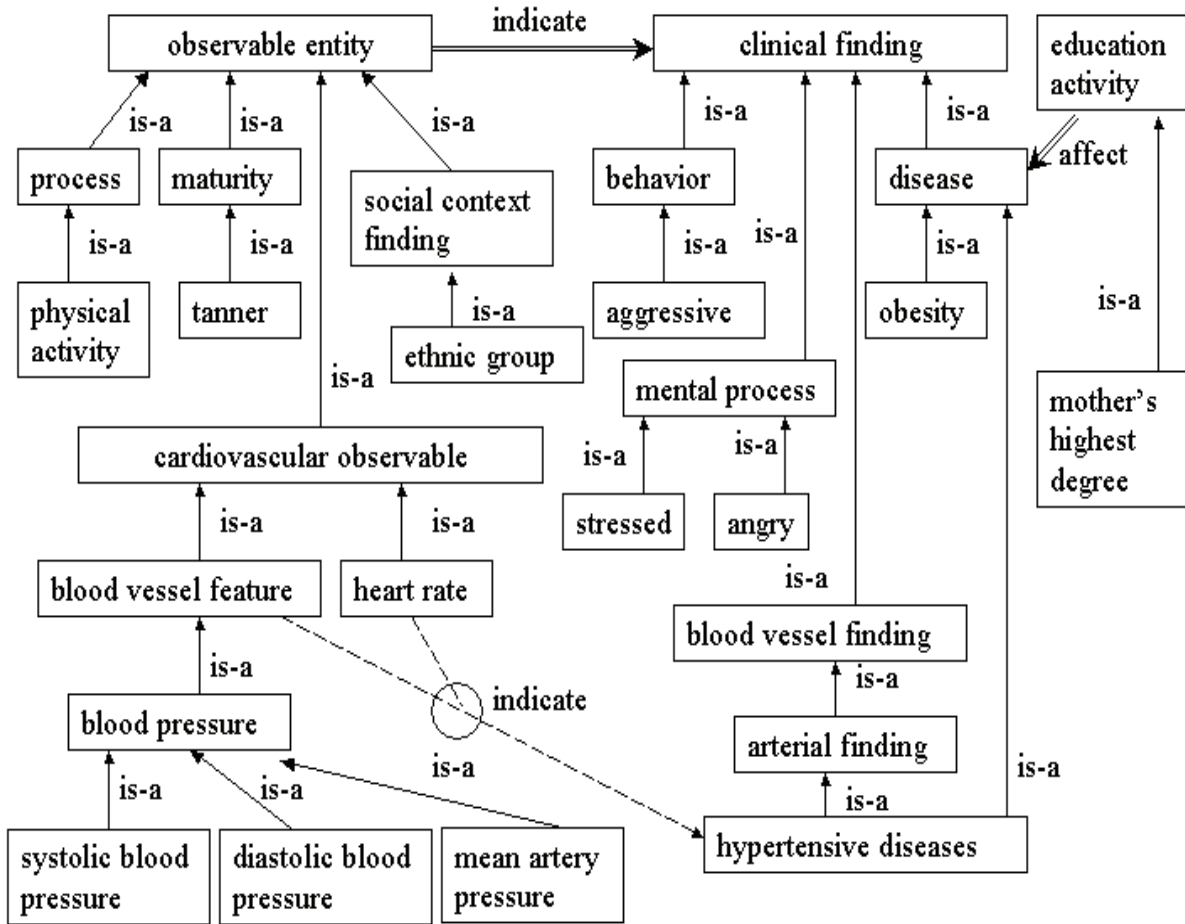


Figure 1: A Fragment of Semantic Network Used in Our Case Study

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  $\Rightarrow$  and *is-a* shown as  $\rightarrow$  (dashed line if its label is *KNOWN*, solid line if its label is *BASIC*).

### 3 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.  $v_1 \rightarrow u_1 \wedge u_1 \rightarrow u_2 \models v_1 \rightarrow u_2$

Generally associations are transitive.

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.<sup>1</sup> 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  $u_1$ 's parents can be replaced by a single association  $v \rightarrow u_1$ . For example, we do not have to

<sup>1</sup>Strictly speaking, this implication is not always valid, which is an interesting topic in modal logic.

specify, *observable entity*  $\rightarrow$  *blood vessel finding*, *observable entity*  $\rightarrow$  *arterial finding*,  $\dots$ , instead, one association *observable entity*  $\rightarrow$  *hypertensive diseases* will be sufficient.

## 4 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,

### Hypothesis Generation Method 1:

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

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.

## 5 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 [12]. 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:**

$$Q_{size} = \frac{N_{KnownCorrect} + N_{UnknownCorrect}}{N_{nontrivial-rules}}$$

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.

## 6 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 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 [19] to generate association rules.

### 6.1 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 [8], association of obesity and poor sleep quality [5], ethnic group differences in moods and ambulatory blood pressure [9], relationship of ambulatory blood pressure to physical activity [4], etc.

### 6.2 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 [16]. 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 SPECIAL-

IST 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”. 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.

### 6.3 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 5,

$$Q_{size} = 0.36$$

$$Q_{attribute} = 0.07$$

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.

## 7 Related Work

Association rule mining has been proved to be very useful in many applications. 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 use metrics to evaluate the significance or interestingness of an association rule, such as lift [2], statistical hypothesis tests [18]. Uninteresting rules will be discarded. However, as shown in [15], 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 [7]. These methods investigate relations among rules in order to present users a concise rule set.
3. Rule ranking. [6] and [20] discussed how to extract top-k significant rules with low redundancy.

## 8 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 National Science Foundation grant CNS 0851984 and Department of Homeland Security grant 2009-ST-061-C10001.

## References

- [1] Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., and Verkamo, A. I., Fast Discovery of Association Rules. In *Advances in Knowledge Discovery and Data Mining*, U. M. Fayyad, et. al., Eds. AAAI press, 1996.
- [2] Bayardo, R. J. and Agrawal, R., Mining the Most Interesting Rules. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, CA, United States, August 15 - 18, 1999.
- [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., 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] Gupta, N. K., Mueller, W. H., Chan, W., and Meininger, J. C., 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.

- [6] Han, J., Wang, J., Lu, Y., and Tzvetkov, P., Mining Top-K Frequent Closed Patterns without Minimum Support. In Proceedings of the IEEE international Conference on Data Mining, 2002.
- [7] Liu, B., Zhao, K., Benkler, J., and Xiao, W., Rule Interestingness Analysis Using OLAP Operations. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, 2006.
- [8] Meininger, J. C., Liehr, P., Mueller, W. H., Chan, W., Smith, G. L., and Portman, R. J., Stress-Induced Alterations of Blood Pressure and 24 h Ambulatory Blood Pressure in Adolescents, *Blood Pressure Monitoring*, 4(3-4), 1999.
- [9] Meininger, J. C., Liehr, P., Chan, W., Smith, G., and Mueller, W. H., 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.
- [10] Padmanabhan, B., and Tuzhilin, A., A Belief-Driven Method for Discovering Unexpected Patterns. In Proceedings of the 4th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1998.
- [11] Padmanabhan, B. and Tuzhilin, A., 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.
- [12] Pipino, L., Lee, Y., and Wang, R., Data Quality Assessment, *Communications of the ACM*, April 2002.
- [13] Quillian, M. R., *Semantic Memory*, *Semantic Information Processing*, M. Minsky, ed., MIT Press, 1968.
- [14] Sahar, S., On Incorporating Subjective Interestingness Into the Mining Process. In Proceedings of the IEEE International Conference on Data Mining, 2002.
- [15] Tan, P. N., Kumar, V., and Srivastava, J., 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.
- [16] Unified Medical Language System. 2007. available at [www.nlm.nih.gov/research/umls/](http://www.nlm.nih.gov/research/umls/).
- [17] Wang, K., Jiang, Y., and Lakshmanan, L. V.S., 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.
- [18] Webb, G. I., Discovering Significant Rules. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, 2006.
- [19] Witten, I. H., and Frank, E., *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd edition, Morgan Kaufmann, San Francisco, 2005.
- [20] Xin, D., Cheng, H., Yan, X., and Han, J., 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.