Summary on the Paper *Hyper-Sequence-Graph Mining*

Naiwei Liu

October 26, 2015

1 Abstract

The era of Big Data brings people with lots of possibilities to improve on many domains, like education, healthcare, economy and so on. To make these possibilities come true, one of the key stages is to extract information from real life datasets. In other words, how to go from data acquisition to data interpretation is a real key for the improvement of the use of Big Data in real life domains. In the previous research, some algorithms aiming at searching and data mining has been investigated in this topic. In this dissertation, the author investigated a structure called hyper-sequence-graph, which represents both sequential relationships and non-sequential relationships. For both types, we would focus on the mining datasets and exploring useful information. For the sequential situation, the dissertation considers about Frequent Pattern Mining (FPM) problem and introduces a new form of the problem called Super-Sequence Frequent Pattern (SS-FPM) problem. Then, a directed weighted graph called sequence graph is generated from the dataset, and a heuristic algorithm using adjacency matrix iteration is proposed. Based on this algorithm, the method could be used in different real world datasets. For the non-sequential attribute site, the dissertation considers the grouping relationship that is readily available in a sequential dataset. To represent such relationship, hyper-graphs are used and the focus is on how to search sub-structures in the large hyper-graph. In this direction, the structured-based indexing from simple graphs to hypergraphs are introduced, as well as the efficient verification method that accelerate the sub-hypergraph matching. Some experiments are used to prove the efficiency and effectiveness of the proposed methods and models.

2 Introduction

2.1 General Objectives

As issued above, the dissertation focused on sequential datasets and explore them from both sequential attribute side and non-sequential attribute side. A sequential dataset is a collection of sequences of ordered elements or events. For the sequential dataset, we could also use sequence graph to describe the relationship. To only express the group relationship for the dataset, we could use hypergraph.

Definition 1. A hypergraph is denoted as $H = (V, E)$, where $V = v_1, v_2, ..., v_n$ is the set of vertices and $E = e_1, e_2, ..., e_m$ is the set of hyperedges. Each hyperedge $(e)_i$ is a non-empty subset of V (i.e., for $\forall v_j \exists e_k$ such that $v_j \epsilon e_k$) and thus $\bigcup_i e_i = V$.

Based on the definition of hypergraph, another structure that could be used to express both sequential and non-sequential relationships between different web pages is introduced. It is defined as Hyper-Sequence-Graph (HSG). It is a combination of sequence graph and hypergraph.

In HSG, we have circles among different vertices. The circles are called hyperedges which can show the relationships among the circled vertices. HSG structure can also represent the sequence relationship between the web pages.

Existing studies have found some ways to analyze the common structures across multiple sequences in a dataset. To determine the interrelations between different subsequences in the data, we must have a new form of sequential pattern mining called super-sequence frequent pattern mining (SS-FPM).

Super-sequences may contain different parts from several frequent sequences. These super-sequences can be used for detecting unusual or even malicious sequence of calls. In the dissertation, graphs will be used to provide a powerful abstraction to represent the above mentioned relationships. As introduced above, we could use the hypergraph to represent the non-sequential grouping relationship among the elements.

Additionally, many sequential datasets can be represented using hyper-sequence-graphs. So, this dissertation would investigate mining and searching from two perspectives.

2.2 Main Contributions of the Dissertation

The new form of sequential mining called super-sequence frequent pattern mining is introduced and a heuristic algorithm using adjacency matrix iteration technique is developed as a solution to SS-FPM.

The SS-FPM is applied on various real-life datasets, which include web log datasets, ADLs and bioinformatics datasets.

A partial clustering problem is investigated based on SS-FPM.

The problem of searching a given sub-hypergraph query on a large hypergraph is studied.

3 Heavy Path Based Super-sequence Frequent Pattern Mining on Sequential Dataset

In this chapter, the author investigated the sequential relationship in a dataset. Mining in this type of datasets could be investigated using Frequent Pattern Mining and some other forms of this strategy. However, this approach sometimes could be limited for the identification of general structures spanning over multiple sequences. As a result, the dissertation introduced a new approach called super-sequential frequent pattern mining (SS-FPM). SS-FPM can be used to determine the super-sequences that can contain the common part of different sequences.

However, finding frequent super-sequence patterns is related to Longest Path in Graph problem, which is NP-hard. So we could transform this problem into a k-hop heaviest problem. This could be solved using sequence matrix method. This method is thought to be efficient especially on large datasets.

We could consider about a table of web log sessions (Table 2.1). The sequence of letters interprets the sequence of users operations. In this example, we could easily see what the most frequent sequence for a user is using FPM algorithms. However, if the sequences could have the meaning of a sequence that a list of events occur, we must take a look among the different sequences in orders to find the order and the relations of the events. In this situation, we must consider the super-sequence pattern with different lengths. For example, we could use SS-FPM to find that the super-sequence of the table would be ABCD. This is helpful because it shows the flow of most sequence of the users or readers. In another word, we could analyze this and make recommendations to the users when they might be in the sequence.

The SS-FPM problem is related to the heaviest (a.k.a. the longest) path problem in directed graphs, which is known to be NP-hard. Accordingly, we first transform the given sequential dataset into a sequence graph, and then search for heaviest paths as patterns. More specifically, we are focusing on finding all the k-hop paths ((k+1)-length super-sequence frequent patterns) that have a larger weight than a given threshold.

To determine the (k+1)-length most frequent super-sequence patterns, we should try to compute all the k-hop heaviest paths that have the total weight more than or equal to the threshold given. For this purpose, the adjacency matrix iteration technique would be used for computing the matrix M_k from $M_k - 1$ and M_1 and this method is called the sequence matrix method.

Based on the dissertation, we could have the algorithm of computing k-hop longest path. This Algorithm could find the most frequent super-sequence with length k+1 starting from ab with small probability of errors when the weight on each link is different from each other. As for the time cost, it would take $O(kn^4)$ in the worst case.

4 SS-FPM Applied on Real World Dataset

Web mining is an important field in data mining area, which is to extract useful information from web logs. In the dissertation, the author introduced some approach to understand the general behavior or flow of users in web usage mining, and classify web pages and users, then make predictions.

In the dissertation, the author take experiments from BMS-WebView-1, in which there are 59602 sessions and 497 web pages in it. As a result, the proposed algorithm finds more paths under the same visiting numbers as the sequence length increases. This is due to the fact that the total weight of a path increases as the sequence length increases. Another result is that there are a lot of paths when path length is relatively long and visiting threshold is relatively small. We can also see that frequent sub-sequences cannot reflect the problems that super-sequences can. Through the super-sequences, we have better understanding of the users behavior flow and different trends of those behaviors.

Monitoring human activities of Daily Living (ADLs) has been researched a lot in recent years. From the existing research, there are methods and algorithms to read the data from the sensor. However, these could be sometimes unreliable since sensor recording is noisy. In the dissertation, the author introduced a new approach by analyzing the data that are collected and generated as the ADLs sequence. From the sequences we could look into the super-sequences of different individuals behaviors and apply the analysis to related applications.

In the last part of this chapter, the author shows an example of using sequence matrix method to search for the protein-protein associations in malaria parasite Plasmodium falciparum with the strongest statistical support. The proposed algorithm enabled the identification of proteins that may serve as central players in parasite life cycle, ranging from DNA repair, DNA replication, transcription, translation, signal transduction, cell cycle progression, parasite invasion and virulence, the key processes that can be targeted toward drug and vaccine development.

5 SS-FP Based Partial Clustering

In this chapter, the author focused on using SS-FPM to partially cluster sequence elements in a dataset.

In the dissertation, the algorithm of SSOPC is introduced, which could be used for generating the evident clusters for web log dataset. It has two major steps, one is primary clustering, and the other is cluster merging.

In primary clustering, the algorithm tries to find out all the clusters based on a pattern and their supporting user sessions according to parameter . Since some super-sequences may have overlaps, we need to merge them in the next step.

In the next step, we need to merge the super-sequences as a group as well as the corresponding user sessions. The author uses the approach of calculating the overlap ratios of each cluster in the super-sequences. If the largest overlap ratio is larger or equal to the value given, there is a merge of these.

The pseudo code is shown in Figure 4.1.

6 Sub-hypergraph Indexing and Matching

As mentioned above, the hypergraph is a type of graph showing relationship for the sequential datasets. From such a hypergraph, one might need to extract some information by searching certain sub-structures. In this dissertation, the author introduced a way to generalize structure-based indexing to hypergraphs and develop a layer-related closure verification method to find matches of a sub-hypergraph in a large hypergraph database.

7 Conclusion

In this dissertation, the author introduced the accomplished work on SS-FPM and sub-hypergraph matching, as well as co-clustering based on hyper-sequence-graph. At first, the author introduced the concept of super-sequences and use that to capture the general structure of a sequential dataset. Some issues on comparing performance among this algorithm and others has also been discussed. Then, the real world datasets also be discussed, as well as the concept of SS-FPM. In the next chapter, it is used for finding the evident clusters in a sequential dataset. Partial cluster can help in seeking the dominant element groups and ignore the noise in the dataset. Then the author introduced the concept of sub-hypergraph indexing and matching.

However, the dissertation showed also some problems that are not solved perfectly. We should focus more on the searching and mining both in sequential and non-sequential relationships together. If we could find a uniform way to solve the mining problems in both situations with high efficiency and effectiveness, it would have a broader use in the future.

References

[1] Yu X, Korkmaz T. Heavy path based super-sequence frequent pattern mining on web log dataset[J]. Artificial Intelligence Research, 2015, 4(2): p1.

[2] Yu X, Korkmaz T. Super-sequence frequent pattern mining on sequential dataset[C]//Big Data, 2013 IEEE International Conference on. IEEE, 2013: 52-59.

[3] Yu X, Korkmaz T. Finding the most evident co-clusters on web log dataset using frequent super-sequence mining[C]//Information Reuse and Integration (IRI), 2014 IEEE 15th International Conference on. IEEE, 2014: 529-536.

[4] Yu X, Korkmaz T. Hypergraph querying using structural indexing and layer-related-closure verification[J]. Knowledge and Information Systems, 2015: 1-29.