Pattern Mining: Current Challenges and Opportunities

Philippe Fournier-Viger\textsuperscript{1}, Wensheng Gan\textsuperscript{2}, Youxi Wu\textsuperscript{3}, Mourad Nouioua\textsuperscript{4}, Wei Song\textsuperscript{5}, Tin Truong\textsuperscript{6}, and Hai Duong\textsuperscript{6}

\textsuperscript{1} Shenzhen University, Shenzhen, China
\textsuperscript{2} Jinan University, Guangzhou, China
\textsuperscript{3} Hebei University of Technology, Tianjin, China
\textsuperscript{4} University of Bordj Bou Arreridj, Bordj Bou Arreridj, Algeria
\textsuperscript{5} North China University of Technology, Beijing, China
\textsuperscript{6} Dalat University, Dalat, Vietnam

philfv@szu.edu.cn, wsgan001@gmail.com, wuc567@163.com, mouradnouioua@gmail.com, songwei@ncut.edu.cn, tintc@dlu.edu.vn, haidu@dlu.edu.vn
Introduction

- **Pattern mining:**
  - using algorithms to (semi) - automatically discover *interesting* and *useful* patterns in data
  - patterns can be easily interpretable and used for decision-making, clustering, classification, etc.

- The **first studies** focused on finding frequent patterns in shopping data and clickstream data.

- Over the years,
  - the focus has changed to other data types, pattern types.
  - and more efficient algorithms, with more features

→ **Current challenges and opportunities?**
We invited 7 researchers to write about a key challenge:

1) **Mining patterns in complex graph data**  
   Philippe Fournier-Viger

2) **Targeted pattern mining**  
   Wensheng Gan

3) **Repetitive sequential pattern mining**  
   Youxi Wu

4) **Interactive pattern mining**  
   Mourad Nouioua

5) **Heuristic pattern mining**  
   Wei Song

6) **Mining Interesting patterns**  
   Tin Truong

7)  
   Hai Duong
1) Mining Patterns in Complex Graph Data

• **A trend**: algorithms to analyze **complex data** such as temporal data, spatial data, and graphs.

• **Graphs**: social networks, road network, etc.

• Many papers on **frequent subgraph mining**

• The traditional problem is too simple!
  • The graphs are **static**
  • Vertices and edges have a **single label**
  • At most a single edge between two vertices
1) Mining Patterns in Complex Graph Data

Handling more complex types of graphs

- A directed graph
- A weighted graph
- An attributed graph
- A multi-labeled graph
- A multi-relational graph

Handling dynamic graphs

Multi-modal data
Specialized types of patterns
• Optimized algorithms
• Trees, paths, stars, cliques, etc.

Novel pattern types
• New criteria to select patterns
• Statistically significant patterns

Custom solutions for applied problems
2) Targeted Pattern Mining

- Traditional pattern mining: find all interesting patterns using several thresholds.
- But a huge number of discovered patterns may be uninteresting.
- **Targeted pattern mining**
  - **Aim**: filter out redundant information and obtain concise results
  - **How?** The user can input one or several targets and discover/query only the patterns containing a target.
    - eg. Find only the patterns that contain milk and/or bread.
  - **Challenging**: How to efficiently find only those patterns?
2) Targeted Pattern Mining

- Targeted frequent itemset and association rule mining
  - Itemset-Tree (Kubat et al.) to query frequent itemsets and rules, can be updated incrementally for new transactions.
  - Guided FP-Growth: Based on FP-Growth, can query multiple itemsets at the same time.

- Targeted sequential pattern mining (SPM)
  - Chueh et al. [8] targeted SPM with time intervals.
  - Chand et al. [5] SPM with recency and monetary constraints
  - [7] goal-oriented algorithms to extract transaction activities before losing a customer.

- Targeted utility-driven mining
  - TargetUM [26] mining high utility itemsets with target items
  - Zhang et al. [44] targeted high-utility sequence querying; utilize some specialized data structures.
2) Targeted Pattern Mining

Open problems

1. Other types of data:
   space, time, events, web, text, etc.

2. Other types of pattern or knowledge:
   graph, sequence, rules, etc.

3. More effective data structure.
   to store, index and search information more efficiently

4. More powerful strategies:
   to reduce the search space more effectively

5. Different applications.
   classification, clustering, etc.

6. Visualization.
   Interactivity, interpretability, ease of use...
3) Repetitive Sequential Pattern Mining

- **Sequential pattern mining (SPM):** finding subsequences that appear frequently in a set of sequences.

  ```plaintext
  INPUT: minsup = 3
  
  Sequence database
  
<table>
<thead>
<tr>
<th>Sequence</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>{{a, b}, {c}, {a}}</td>
<td></td>
</tr>
<tr>
<td>{{a, b}, {b}, {c}}</td>
<td></td>
</tr>
<tr>
<td>{{b}, {c}, {d}}</td>
<td></td>
</tr>
<tr>
<td>{{b}, {a, b}, {c}}</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
  OUTPUT: all sequential patterns:
  
  - {{a}} support = 3
  - {{b}} support = 4
  - {{c}} support = 4
  - {{a}, {c}} support = 3
  - {{b}, {c}} support = 4
  - {{a, b}, {c}} support = 3
  ```

- **Limitations:**
  - Often unnecessary to deal with sequences of itemsets. e.g. COVID genome: ACGATAAAA...
  - SPM ignores repetitions of a patterns in sequences.
  - Reptitions are important!
3) Repetitive Sequential Pattern Mining

Repetitive sequential pattern mining

• Can handle repetitions.

• Several variations:
  • **Without gaps** (items must appear consecutively) [6]: easier problem, but may lose important information
  • **With self-adaptive gap** [41]: can find more patterns but result set can be large.
  • **With gap constraint**: the user has to predefine gap [M,N]. Then, each occurrence must satisfy this gap constraint. Difficult problem with several types:
    • no condition [27],
    • one-off condition [21],
    • nonoverlapping condition [40]
    • and disjoint condition [28].
3) Repetitive Sequential Pattern Mining

• Key challenges:
  • What are the computational complexities of calculating the supports under different conditions?
  • How to design effective mining algorithms for these conditions?
  • If the dataset is dynamic or a stream database, how to design effective mining algorithms?
  • For a specific problem like classification, there are many approaches to solve it (e.g contrast pattern using different conditions). However, what is the best approach?
4) Interactive Pattern Mining

• Traditional pattern mining algorithms like Apriori and FP-Growth are **batch algorithms**.
• Thus, user needs to run again the algorithm to get new results even if the database is slightly changed.
• How to deal with **databases that are dynamics**?
  • **Incremental pattern mining**: update the set of patterns when the database is updated
  • **Stream Pattern Mining**: handle databases that are updated in real-time
4) Interactive Pattern Mining

- **Interactive pattern mining:** Handle dynamic databases by injecting users preferences, users feedback or user targeted queries, into the mining process.
  - **Targeted querying based approaches:** the user search for patterns containing specific items by sending targeted queries.
  - **Users feedback based approaches:** progressively address feedback sent by users during the mining process.
  - **Visualization based approaches:** Various visualization techniques for different forms of pattern.
5) Heuristic Pattern Mining

- Traditional pattern mining problems
  - can have a high computational cost
  - for many applications like recommendation, unnecessary to find all patterns

- Heuristic pattern mining
  - Algorithms to find an approximate subset of all patterns within a reasonable time.
  - Based on genetic algorithm (GA) [9], particle swarm optimization (PSO) [24], artificial bee colony (ABC) [33], crossentropy (CE) [35], and bat algorithm (BA) [34], etc.
5) Heuristic Pattern Mining

Key challenges:

• **Identifying the appropriate objective.**
  • Define appropriate objective functions for difficult problems (e.g. top-k, closed).
  • Multi-objective functions

• **Speed-up the mining process**
  • How to narrow the search space?
  • Using length, and keep track of invalid combinations
  • New data structures

• **Diversifying the results**
  • To avoid falling in local optima, increasing the diversity of results is important
5) Heuristic Pattern Mining

Key challenges:

• Designing a general framework
  • integrating all the objectives, processes, and results into a general

• Other pattern types
  • Not just itemsets...
  • Graphs, sequential patterns...
6) Mining Interesting Patterns

• **Traditional pattern mining** relies on the **support function** to identify interesting patterns.

• Support is not enough for many other applications.

• **Utility functions** are popular to find **profitable patterns in quantitative**

• **Key challenges:**
  • Utility functions generally do not respect the downward-closure property
  • Thus, it is necessary to define **upper-bounds** or **weak-upper bounds** on **utility** functions to be able to reduce the search space.
  • Designing upper-bounds is not easy but key to performance!
6) Mining Interesting Patterns

• High utility itemset mining
• High utility sequential pattern mining
  • various utility functions:
    • $U_{\text{min}}$: pessimistic utility function
    • $U_{\text{max}}$: optimistic utility function

• Generally:
  • Difficult to find upper-bounds. It took 8 to 10 years to find good upper-bounds for some problems.
  • **Danger:** not proving mathematically the properties of upper-bounds may lead to inexactness of the algorithms...

• How to propose a genetic framework to design upper-bounds and search space pruning strategies?
Conclusion

• The field of pattern mining is changing
• Six key challenges
• The full paper can be downloaded on the workshop website.