香 港 資 優 教 育 學 苑
The Hong Kong Academy for Gifted Education

Exploration of Course Enrolment Patterns of HKAGE Student Members using Logical Itemset Mining (LISM) with Answer Set Programming (ASP)

Introduction

Bag-of-items data, such as market baskets in retail, comprise products purchased by a customer in a store visit. Mining of bag-of-items data is an attempt to understand customers' intentions by discovering any (novel) patterns of items bought, from which insights could be obtained (see Figure 1).

Figure 1. A market basket (solid black circle) observed which is composed of items from two logical itemsets (red dotted circles), representing customer intentions.

Logical Itemset Mining (LISM) is a framework of analysis (Shailesh Kumar, Chandrashekar V and C V Jawahar, 2012), in which a specific observed market basket is supposed to be consisted of a number of logical itemsets. Here a logical itemset is simply defined as a set of items that could reflect a customer intent. These logical itemsets are latent in the data and the goal of LISM is to discover them in an unsupervised manner. In brief, the observed bag-of-items data may be best described as a mixture-of and projections-of latent logical items.

In this study, we adopt an analogy to regard the enrollment records of a student member of HKAGE over a specific period concerned simply as a market basket. By using LISM, we try to identify any (stable) learning intentions amongst student members from their enrolment records over the period concerned.

System Data

Student Learning Profile (SLP) is originally intended for student members to review their learning

histories and progresses in HKAGE. It records all learning information (e.g. enrolment, attendance, result, reflection, etc.) of student members once they enroll in programmes/ activities in HKAGE. In this study, Student Learning Profiles (SLP) for the period of 2014/17 were downloaded as at mid-2018, which amounted to around 20 000 records. From each enrolment record, the learning topic corresponding to the course that a student member enrolled over the period concerned could then be obtained. All the learning topics obtained for a student member over the period concerned were then grouped together, which was regarded as a market basket of the student concerned. In this study, totally 87 learning topics (e.g., Algebra and Mathematical Analysis) were identified.

Analysis Methodology: Logical Itemset Mining (LISM)

LISM framework has four stages:

1. *Counting stage* where co-occurrence counts between all pairs of items is computed in one pass through the data.
2. *Consistency stage* where these co-occurrence counts are converted to consistency values, quantifying the statistical significance or information content of seeing each pair of items together vs. random chance.
3. *Denoising stage* where the co-occurrence consistencies are cleaned further to address the mixture-of intents property.
4. *Discovery stage* where logical itemsets in the form of cliques are discovered in the co-occurrence consistency graph to address the projection property.

The steps to achieve the purposes of these stages of LISM are briefly described below.

Step 1: *Counts of co-occurrences*: For each pair of topics, the number of student members that have enrolled both topics are counted ($\psi(\alpha, \beta)$). A co-occurrence matrix can be formed accordingly.

Step 2: *Trim down the co-occurrence matrix*: The co-occurrence matrix is further trimmed down by using a threshold (θ_{cooc}) on the number of counts and a threshold (θ_{cons}) on a measure of co-occurrence consistency, called Normalized Point-wise Mutual Information ($\varnothing_{nmi}$) which is defined and compiled as follows:

$$\psi(\alpha) = \sum_{\beta \in \mathbf{V}, \alpha \neq \beta} \psi(\alpha, \beta), \quad \psi_0 = \sum_{\alpha \in \mathbf{V}} \psi(\alpha)$$

$$P(\alpha,\beta)=\frac{\psi(\alpha,\beta)}{\psi_0}, P(\alpha)=\frac{\psi(\alpha)}{\psi_0}$$

$$\phi_{pmi}(\alpha,\beta)=max\left\{0,\log\left(\frac{P(\alpha,\beta)}{P(\alpha)P(\beta)}\right)\right\}$$

$$\phi_{nmi}(\alpha,\beta)=\frac{\phi_{pmi}(\alpha,\beta)}{-\log P(\alpha,\beta)}\in[0,1]$$

When trimming down the co-occurrence matrix using θ_{cons}, an iterative process is adopted so as to go through the matrix more than one, until the measure of 'quality' converges. Empirically, it is observed that the process converges quickly in two to three iterations. In the study, the values of the thresholds, θ_{cooc} and θ_{cons} were set using a trial and error approach.

Step 3: The thresholded co-occurrence matrix is then binarized and a graph is resulted. Given this graph, a logical ITEMSET, is defined as a set of items L = ($\ell_1, \ell_2, \ldots, \ell_k$) such that each item in this set has a high co-occurrence consistency with all other items. To find the largest logical itemsets, we just have to find ***all maximal cliques*** in the binarized co-occurrence consistency graph. Their definitions are provided below.

Clique: A clique, C, in an undirected graph G = (V, E) is a subset of the vertices, C $\subseteq$ V, such that every two distinct vertices are adjacent.

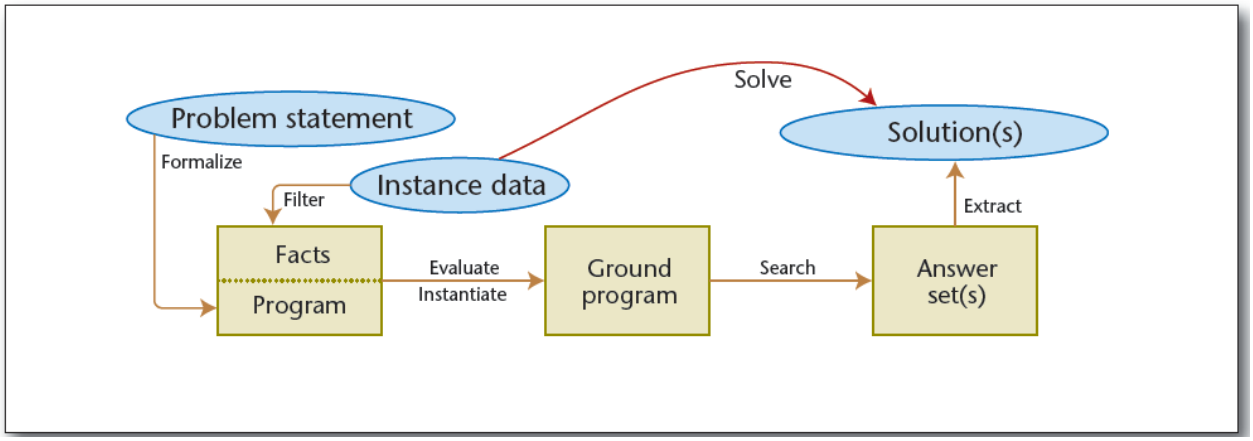
Maximal clique: It is a clique that cannot be extended by including one more adjacent vertex, i.e., it is not a subset of a larger clique.

To find all maximal cliques in the binarized graph, we use a declarative programming approach, ***Answer Set Programming (ASP)*** in this study, instead of using common procedural languages (such as C++ and Java). ASP is a declarative programming paradigm for solving search problems and their optimization variants. In ASP a search problem is modeled as a set of statements (a program) in a logic programming type of a language in such a way that the answer sets (models) of the program correspond to the solutions of the problem. The paradigm was first formulated in these terms by Marek and Truszczynski (1999) and Niemelä (1999). The ASP paradigm has its roots in knowledge representation and nonmonotonic logics research. A recent and more technical overview of ASP has been contributed by Brewka, Eiter, and Truszczynski (2011).

The ASP paradigm provides a general-purpose methodology for solving search and optimization problems encountered in many real-world applications. The key step is to identify and formalize

the problem to be solved. Typically, it consists of the two steps, namely (i) clarify what the potential solutions of the problem are like and set the conditions that solutions should satisfy in the form of rules and constraints; (ii) solving the specific problem on hand means that given the data representing the problem instance, the tailor-made software (ASP solver) could find one or more solutions (answer sets) that satisfy the given conditions (see Figure 2).

Figure 2. Conceptual Model of the ASP Paradigm



In the study, the problem instance is the thresholded and binarized graph, which is represented in the style of logic programming as follows.

```
node(1..87).
edge(2,5). edge(2,27). edge(2,32). edge(2,35). edge(2,40). edge(2,41).
edge(2,52). edge(2,58). edge(2,66).edge(2,82).
……
```

The first programming statement (`node(1..87).`) means that there are 87 nodes in the graph; while `edge(n1, n2)` means that there is an edge from node n1 to node n2. The programming code to find the maximal clips is shown below, which basically follows the definition of a maximal clique.

```
adj(X,Y) :- edge(X,Y). adj(Y,X) :- edge(X,Y).
{clique(X)} :- node(X).

disconnect(X) :-node(X),not clique(X),clique(Y),not X=Y, not adj(Y,X).

:-clique(X1), clique(X2), not X1 = X2, not adj(X1, X2).
```

```
:-node(X), not clique(X), not disconnect(X).
#show clique/1.
```

The programming statement in the first line (`adj(X,Y):-edge(X,Y).adj(Y,X):-edge(X,Y).`) defines the meaning of a node being adjacent to another one (i.e. there is an edge between them). The programming statement in the second line (`{clique(X)} :- node(X).`) is a *choice* statement, which means that a node can be either in the clique or not. The following programming statement defines the conditions of being a clique.

```
:-clique(X1), clique(X2), not X1 = X2, not adj(X1, X2).
```

The reading of the above programming statement is: It cannot be the case that there are two nodes X1 and X2 being in the clique; but they are not adjacent with each other. The following programming statement defines the conditions of a clique being maximal.

```
:-node(X), not clique(X), not disconnect(X).
```

The reading of the above programming statement is: it cannot be the case that there is a node X not in the clique; but is adjacent to any nodes in the clique. Using the ASP program shown above, the student learning patterns (i.e., maximal cliques) were found. To strike for meaningful interpretation and limit the numbers of maximal cliques, only maximal cliques with numbers of nodes being greater than or equal to 5 were derived using the ASP solver, called *clingo*.

```
C:\Users\ericfung\Downloads\clingo-5.2.2-win64>clingo d:\clique-ASP\output_Graph
.lp d:\clique-ASP\solve.lp 0
clingo version 5.2.2
Reading from d:\clique-ASP\output_Graph.lp ...
Solving...
Answer: 1
clique(11) clique(61) clique(71) clique(81) clique(86)
Answer: 2
clique(2) clique(35) clique(41) clique(58) clique(66)
Answer: 3
clique(2) clique(32) clique(35) clique(41) clique(58)
Answer: 4
clique(3) clique(40) clique(54) clique(55) clique(76) clique(79)
Answer: 5
clique(2) clique(40) clique(58) clique(66) clique(82)
Answer: 6
clique(3) clique(4) clique(40) clique(54) clique(55) clique(76)
Answer: 7
clique(3) clique(11) clique(39) clique(55) clique(79)
Answer: 8
clique(3) clique(39) clique(54) clique(55) clique(79)
Answer: 9
clique(61) clique(71) clique(80) clique(81) clique(86)
Answer: 10
clique(61) clique(80) clique(81) clique(86) clique(87)
Answer: 11
clique(2) clique(5) clique(32) clique(35) clique(58)
Answer: 12
clique(2) clique(5) clique(35) clique(40) clique(58) clique(66)
Answer: 13
clique(27) clique(28) clique(81) clique(86) clique(87)
Answer: 14
clique(2) clique(32) clique(35) clique(52) clique(58)
Answer: 15
clique(2) clique(35) clique(52) clique(58) clique(66)
SATISFIABLE

Models       : 15
Calls        : 1
Time         : 0.180s (Solving: 0.06s 1st Model: 0.00s Unsat: 0.03s)
CPU Time     : 0.156s
```

From the answer sets computed, 15 maximal cliques were found from the thresholded graph. They are tabulated in the **Annex 1**.

Findings and Discussions

Totally, 15 learning patterns with numbers of topics being greater than or equal to 5 were found using logical itemset mining (see **Annex 1**) and these patterns are grouped into the following categories.

(i) The three learning patterns #9, #10, and #13, correspond to those members who were primarily interested in languages (e.g., Speaking and Listening, and Writing). In addition, the pattern #13 gears towards the topic of English Literature (English poetry and Literature); while the pattern #9 and #10 gear towards the topics of relationship (e.g., personal and family relationship)

(ii) The seven learning patterns #2, #3, #5, #11, #12, #14, and #15 correspond to those members who were primarily interested in Mathematics (e.g., Algebra, and Geometry and Topology); while amongst them, some students might be also interested in Physics (#2 and #3) and Statistics/Applied Statistics (#5, #11, and #12).

(iii) The four learning patterns #4, #6, #7, and #8 correspond to those members who are primarily interested in STEM topics (e.g., Analytical Chemistry, Mechanics, Introduction to Mathematics, Microbiology, Robotics, and Software Development).

(iv) Finally, there is a learning pattern (#1) which comprises topics from Science, Humanity and Language.

It can be noted that in general, most of the student members did not adopt a multi-disciplinary approach when enrolling various topics for learning during the period concerned. It is especially prominent for those who are primarily interested in Mathematics. The scope of the topics for learning of these student members seems to be quite narrow and is solely related with mathematical subjects. Advices could be provided to these students to encourage them to explore subjects in various areas so that they could enlarge their learning scopes.

Besides, a group of student members, who are particularly concerned with STEM subjects, is observed. This may be due to the emphasis on the importance of STEM recently in various media. On the other hand, those members, who are primarily interested in Language subjects, have not enrolled in any Science/Math courses. As mentioned in the survey results of Needs Assessment Surveys, quite a number of students in Humanities domain were quite interested in Science subjects. In this regard, provision of (introductory) Sciences courses for Humanities student members could be considered.

香 港 資 優 教 育 學 苑
The Hong Kong Academy for Gifted Education

Annex 1

15 Learning Patterns Located Based on the Student Enrollment Records during the Period 2014/17

Number	Pri Concern	Sec Concern	Serial #	Avr num of occurences per edge	Topic/ Student Intention					
	Lang	Leader	#10	49.8	Personal Leadership	Speaking and Listening (Chinese)	Speaking and Listening (English)	Writing (Chinese)	Writing (English)	
	Lang	Leader/Relation	#9	46.6	Personal Leadership	Relationship with Families	Speaking and Listening (Chinese)	Speaking and Listening (English)	Writing (Chinese)	
3	Lang	Lit(Eng)	#13	53.0	English Literature	English Poetry	Speaking and Listening (English)	Writing (Chinese)	Writing (English)	
	Math	-	#14	135.8	Algebra	Geometry and Topology	IMO Training	Mathematical Analysis	Numbers and Arithmetic	
	Math	-	#15	100.5	Algebra	IMO Training	Mathematical Analysis	Numbers and Arithmetic	Probability	
	Math	Phy	#2	102.5	Algebra	IMO Training	Introduction to Physics	Numbers and Arithmetic	Probability	
	Math	Phy	#3	140.0	Algebra	Geometry and Topology	IMO Training	Introduction to Physics	Numbers and Arithmetic	
	Math	Stat/App Stat	#5	91.7	Algebra	Introduction to Mathematics	Numbers and Arithmetic	Probability	Statistics	
	Math	Stat/App Stat	#11	139.5	Algebra	Applied Statistics	Geometry and Topology	IMO Training	Numbers and Arithmetic	
7	Math	Stat/App Stat	#12	89.3	Algebra	Applied Statistics	IMO Training	ntroduction to Mathematics	Numbers and Arithmetic	Probability
1	Multi		#1	43.5	Biomedical Science	Personal Leadership	Relationship with Families	Speaking and Listening (English)	Writing (Chinese)	
	STEM		#4	38.7	Analytical Chemistry	Introduction to Mathematics	Mechanics	Microbiology	Robotics	Software Development
	STEM		#6	40.0	Analytical Chemistry	Applied Mathematics	Introduction to Mathematics	Mechanics	Microbiology	Robotics
	STEM		#8	41.4	Analytical Chemistry	Introduction to Chemistry	Mechanics	Microbiology	Software Development	
4	Chem/Bio	Tech	#7	40.0	Analytical Chemistry	Biomedical Science	Introduction to Chemistry	Microbiology	Software Development	

References

Brewka, G.; Eiter, T.; and Truszczynski, M. 2011. Answer Set Programming at a Glance. *Communications of the ACM* 54(12): 92–103. dx.doi.org/10.1145/2043174.2043195

Marek, V., and Truszczynski, M. 1999. Stable Models and an Alternative Logic Programming Paradigm. *The Logic Programming Paradigm: A 25-Year Perspective*, ed. K. Apt, V. Marek, M. Truszczynski, and D. Warren, 375–398. Berlin: Springer.

Niemelä, I. 1999. Logic Programming with Stable Model Semantics as a Constraint Programming Paradigm. *Annals of Mathematics and Artificial Intelligence* 25(3–4): 241–273.

Shailesh Kumar, Chandrashekar, V. and Jawahar, C.V. 2012. Logical Itemset Mining, *IEEE 12th International Conference on Data Mining Workshops*, December 2012.