International Journal of Multimedia and Ubiquitous Engineering
Vol.10, No.5 (2015), pp.287-296
http://dx.doi.org/10.14257/ijmue.2015.10.5.26

Towards on the MOOCs Knowledge Discovery Based on Concept Lattice*

Junfeng Gao [1], Jingye Qu[2]** and Zhong Xin[3]
[1,2]Beihua University, Jilin City, China
[3], Jilin Technology College of Electronic Information, Jilin City, China
[1]373369834@qq.com, [2]87498073@qq.com [3]

Abstract

This paper puts forward the methods on MOOCs knowledge organization and discovery, adopts the method of Formal Concept Analysis, uses concept lattice as the support tool, and selects the data from "Coursera" to cluster the courses and mine the inner knowledge association among courses, so as to discover the connotative knowledge correlation among MOOCs on the same topic and the structure characteristics among courses. Finally, based on the supergraph of the MOOCs concept lattice, the paper puts forward the visualization navigation method of the courses and provides flexible guiding principles for the learners whose knowledge structure is unusual in various fields when choosing the courses, so as to promote the development and improvement of the MOOCs websites by the knowledge organization and knowledge discovery technology.

Keywords: *Concept lattice; Knowledge discover; MOOCs clustering; Association Rules*

1. Introduction

Since the MOOCs appeared in 2008, it has experienced 6 years from then on. During these 6 years, this new network-based teaching method has made a great impact on the traditional higher education system. MOOCs are the abbreviation of the "massive open online courses". The main characteristic of MOOCs is that it's publicly available: the traditional classroom multimedia materials, lesson plans and exercises are uploaded to the network by binary, and users can register at any time to learn the curriculums they want. In addition, the MOOCs web sites also afford interactive platform, so as to offer facilities for learners to share learning experiences.

To be sure that, MOOCs are different from the distance learning courses collection, which will put the world wide professors and learners together to focus on a serious subject on an open access platform. However, every advantage has its disadvantage. The organization of MOOCs online courses and their appearance has encountered inherent obstacles in information organizing and retrieve activities. The curriculums may be organized in the tree-hierarchical structure, the one-dimensional linear structure or the flat folksonomy structure, but no matter how it is organized; users can not avoid the situation of getting lost in the information flood. By examining the dominated MOOCs websites: "Coursera" and "Edx", the author has found that the knowledge organization architecture of the courses has some flaws and shortcomings existed in varying degrees.

* This paper is the research finding of "Research on Constructing Collegiate Information Ecosystem " ---- the 12th five-year plan research project of the Education Department of Jilin Province .Project No. is [2014]184.and is also the research finding of " Research on the Service Methods of Digital Library Knowledge Based on Concept Lattice" ---- the Jilin Province social science fund project. Project No is 2014B62.

** Jingye Qu is the corresponding author.

ISSN: 1975-0080 IJMUE

For example, the "Edx" courses are organized by alphabetical order. This linear organization architecture of the courses splits the correlations among many related courses. Especially when the learners try to understand the curriculum systems which are incomplete and developing, they can not accurately distinguish between the pilots and the advanced courses. And this kind of knowledge sorting brings more inconvenience for the navigation of the MOOCs courses.

For the above problems，the author will adopt the method of Formal Concept Analysis, using concept lattice as the support tools, and selecting the data from "Coursera" to cluster the courses and mine the inner knowledge association among courses. Based on the supergraph of the MOOCs concept lattice, the visualization navigation method of the courses is put forward for promoting the development and improvement of the MOOCs websites by the knowledge organization technology.

2. Related Research

FCA (Formal Concept Analysis) is the research findings of Rudolf Wille, who was a professor working at German Darmstadt University [1].Data analysis and mining methods based on the FCA theory with concept lattice as its technical supporting tools, are gradually changing the original development ways of the international data analysis fields.

As the core data structure of Formal Concept Analysis, concept lattice is induced by the partical order structure in formal context. Concept lattice will not reduce the complexity of data mining, while the formal context contains all the details of the data to be processed [1].The representative research findings of the concept lattice on data mining include:

Stumme put forward Iceberg concept lattices based on the TITANIC algorithm, which is applied to analyze large and super large database [3].Iceberg concept lattices can show the frequent pattern of association rules without losing any information. At the same time, it gives out the visualization results of the data mining rules. This method is applied to the fields of data analysis, information retrieval, and knowledge discovery, etc.

Carpineto, etc. established text index by concept lattice clustering of the indexed text, built the hybrid navigation systems on the basis of concept lattice, and verified that concept lattice retrieval was equivalent to or even better than the traditional boolean queries [4].The research has shown that, the navigation systems based on the concept lattices have preferable adaptability and favorable retrieval performance.

Kmi, etc. elaborated on the browsing mechanisms of the incremental knowledge acquisition based on the concept lattice with Formal Concept Analysis from domain-specific information retrieval perspectives [5]. The browsing mechanism allows the users to upgrade the organization structure of the domain-specific documents dynamically. As the time goes, multiple users can collaborate to build and maintain the browsing scheme, in order to support the management of the open-ended documents flexibly.

To sum up the above research findings on the knowledge organization based on concept lattice, the author constructs the domain-specific concept lattice of the MOOCs. By describing the data through the content characteristics of MOOCs and mining knowledge correlation rules of the MOOCs multi-dimensionally, we believe that the knowledge characteristics of the topic-related MOOCs will be shown better.

3. Formal Description of the Knowledge Discovery of the MOOCs Based on Concept Lattice

3.1 Clustering of MOOCs Concept Lattice and Knowledge Association Rules Mining

Set triad K: = (G，M，R) as the formal background of the MOOCs, G and M are two sets. G = { g1 ，g2，……gn} is the object set and represents MOOCs unit; M = { m1，m2，……mn} is the attribute set and represents the knowledge points of the MOOCs unit. R is a binary relation between G and M, and it represents the knowledge points included in a MOOC [1]. MOOCs concept lattice is created based on such binary relation.

X and Y are two certain nodes of the given concept lattice L, if X is the parent concept of Y, there will be only one edge between X and Y [6]. The path connected by all the edges between the top node and the node Y is marked as Path (Y). The longest path from the top node to the bottom node passing through node Y is called the Chain of Y. The number of edges from the top node to the node Y in the Chain Y is called the level of Y, and the length of the longest chain of concept lattice L is called the level of L, and the following node next to the top node is the aimed researching aggregated category.

Ream I = { i1，i2，……，im} is the aggregation of the knowledge points of the MOOCS network database TD. The transaction T in the database is an attribute set meeting the requirement of $T \subseteq I$. An association rule can be shown as $T1 \Rightarrow T2$, of which T1，T2 $\subseteq I$，$T1 \cap T2 = \Phi$ [7]. The direct meaning is that learners who are interested in module T1 may be also interested in module T2, T1 and T2haveacademic logical association and are called the pilot and subsequence of the association rule. The support of association rule $T1 \Rightarrow T2$ in the TD can be formulized as support $(T1 \Rightarrow T2) = | T1 \cup T2 | / | TD |$, and the confidence of association rule $T1 \Rightarrow T2$ in the TD can be formulized as confidence $(T1 \Rightarrow T2) = | T1 \cup T2 |/ | T1 |$. The association rules which are bigger than the minimal support and minimal confidence are called strong association rules which mean strong association among courses.

3.2 Technical Route

The research focuses on the partial ordering relation of object and attribute (basic idea of lattices building) without considering threshold value, the structure of concept lattice is not influenced by the threshold value; different association rules will be mined with various support threshold value and confidence threshold value when keeping the same concept lattice structure, so as to study more rules of the knowledge association.

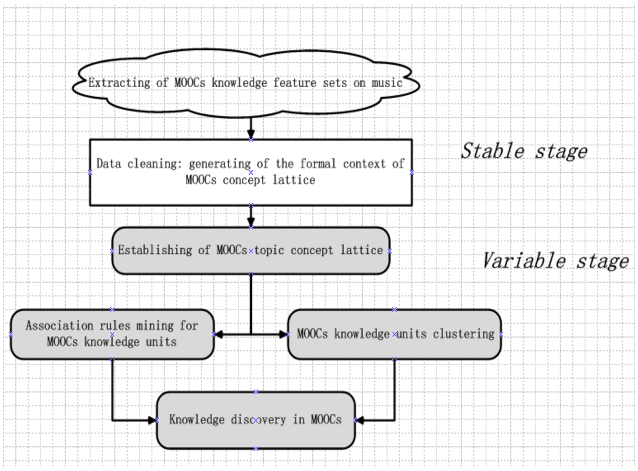

Extracting of MOOCs knowledge feature sets on music
Data cleaning: generating of the formal context of MOOCs concept lattice
Stable stage
Establishing of MOOCs topic concept lattice
Variable stage
Association rules mining for MOOCs knowledge units
MOOCs knowledge units clustering
Knowledge discovery in MOOCs

Figure 1. Technical Route of Knowledge Discovering based on Concept Lattice

The researching approach: based on the “top to bottom and division of concept lattice” data processing idea, taking MOOCs information from 16 topics including “music, movie and audio” from the Web “Coursera” as the object set of the MOOCs concept lattice; analyzing and filtering word frequency of the words in the “courses description” of each MOOCs; building the stop words table including modal particle, conjunction, preposition and reducing the stop words with software “tm”; dividing the filtered words with “Rwordseg”; taking the substantives used to describe knowledge in each course as the candidacy attribute set of the MOOCs concept lattice; counting words frequency with “wordcloud”; making nephogram of high-frequency words of the “music, movie and audio” materials (Figure 2).

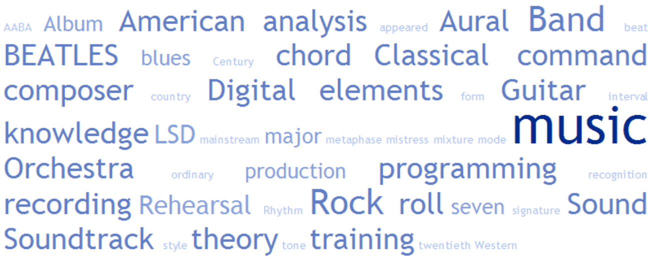
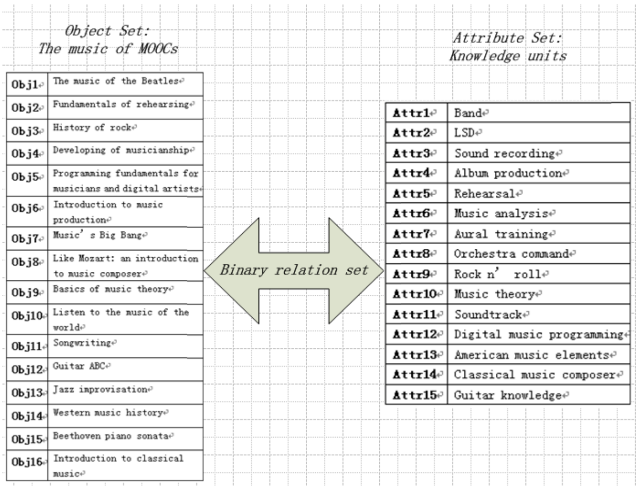


Figure 2. High-frequency Label Words Cloud of MOOCs Knowledge

3.3 MOOCs Concept Lattice Construction

MOOCs topic concept lattice is the latticework made up by nodes of similar resource associations, it formally describes the relations of the MOOCs resources and reflects the consistency of intension and extension [9]. Attribute of concept lattice is not only the key factor to decide the latticework construction time and clustering quality, but also the sole criteria to discover the knowledge in the MOOCs. In reality, through the description of the keywords, the primary coverage of any information resource can be found. The keywords have subtle relations which is the potential concept structure. The concept of the MOOCs concept lattice actually is the set of objects which have the most common attributes, therefore, the selection of the lattice building attribute should be serious. The special words in the words nephogram are regulated with the vocabulary processing software “Wordnet”, and reduction of attribute set is made as shown in Figure 3:



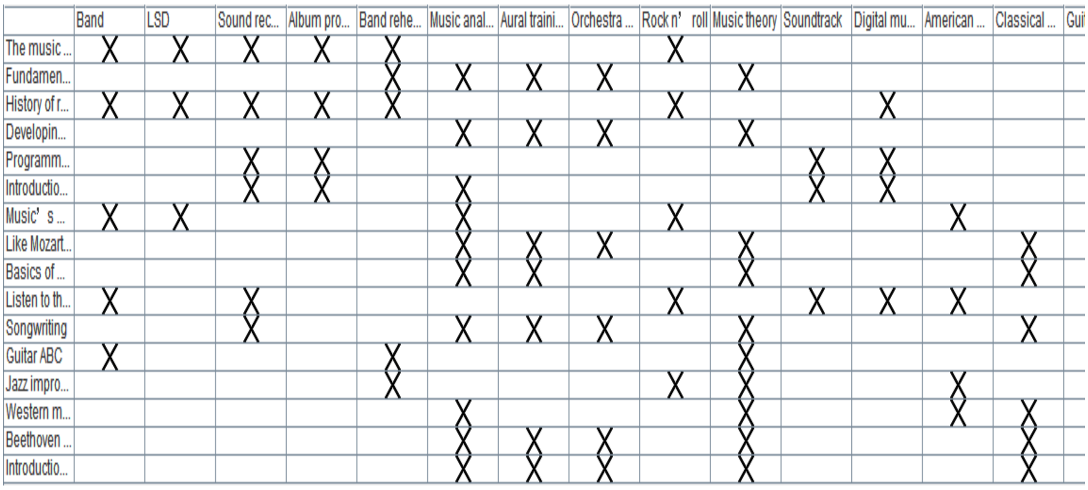
Obj1	The music of the Beatles
Obj2	Fundamentals of rehearsing
Obj3	History of rock
Obj4	Developing of musicianship
Obj5	Programming fundamentals for musicians and digital artists
Obj6	Introduction to music production
Obj7	Music’ s Big Bang
Obj8	Like Mozart: an introduction to music composer
Obj9	Basics of music theory
Obj10	Listen to the music of the world
Obj11	Songwriting
Obj12	Guitar ABC
Obj13	Jazz improvisation
Obj14	Western music history
Obj15	Beethoven piano sonata
Obj16	Introduction to classical music

Attr1	Band
Attr2	LSD
Attr3	Sound recording
Attr4	Album production
Attr5	Rehearsal
Attr6	Music analysis
Attr7	Aural training
Attr8	Orchestra command
Attr9	Rock n’ roll
Attr10	Music theory
Attr11	Soundtrack
Attr12	Digital music programming
Attr13	American music elements
Attr14	Classical music composer
Attr15	Guitar knowledge

Figure 3. “Music, Movie and Audio”, Object Set and Attribute Set of MOOCs Concept Lattice

The formal background of MOOCs concept lattice is a triad K=(A,S,I) which generally shows the relation among coursed/modules, and A is the congregation of MOOCs modules, S is the congregation of knowledge units, I is a binary relation between A and S[2]. The formula I⊆AxS means the knowledge S is included in MOOC A. The formal background can show the binary relation among all modules and all knowledge units of a certain academic subject, and MOOCs concept lattice is derived by such kind of binary relation:

The above graph is part of a complete monodrome formal background, which shows certain character of certain object and represents generalization and specialization of object. The row is the MOOCs resource objects and the line is character vocabulary describing MOOCs resource knowledge, each X means that a certain object has certain character.



	Band	LSD	Sound rec...	Album pro...	Band rehe...	Music anal...	Aural traini...	Orchestra ...	Rock n' roll	Music theory	Soundtrack	Digital mu...	American ...	Classical ...	Gui
The music ...	X	X	X	X	X				X						
Fundamen...					X	X	X	X		X					
History of r...	X	X	X	X	X				X			X			
Developin...						X	X	X		X					
Programm...			X	X							X	X			
Introductio...			X	X		X					X	X			
Music' s ...	X	X				X			X				X		
Like Mozart...						X	X	X		X				X	
Basics of ...						X	X			X				X	
Listen to th...	X		X						X		X	X	X		
Songwriting			X			X	X	X		X				X	
Guitar ABC	X				X					X					
Jazz impro...					X				X	X			X		
Western m...						X				X			X	X	
Beethoven ...						X	X	X		X				X	
Introductio...						X	X	X		X				X	

Figure 4. "Music, Movie and Audio", Formal Background (Part of) of MOOCs Concept Lattice

In the formal background (A,S,I), if any two elements a, h∈A in the case f(a)=f(h),then a=h; any pair of s, n∈S and g(s)=g(n),then s=n. This kind of formal background is seen as pure. Figure 4 is not a pure background, object set "Like Mozart: an introduction to music composer", "Beethoven piano sonata" and "Introduction to classical music" have same attribute distribution, so they should be in the same line.

The knowledge discovering method of MOOCs resource based on the concept lattice is a dynamic constructing process, and the generated MOOCs concept lattice will be improved with the adding of new courses.

Figure 5 is about the visible description of all concepts in the formal background by the concept lattices. The nodes are the reflection of the logical relation of knowledge in the courses of "music, movie and audio" from the "Coursera" website. The concept attribute can be considered as class name or class indication which indicates the characters of class elements. When the formal background is big, the objects and attributes of each node will not be all marked but can be derived from the location of the node according to the character of concept lattice. In a limited concept lattice, each joint includes the object of its next joint as well as all attributes of its above joint. However, concept lattice can not completely demonstrate the implied disciplinary aggregation relations and potential disciplinary curriculum provision associations. Therefore, clustering analysis and association rules should be studied based on concept lattice [10].

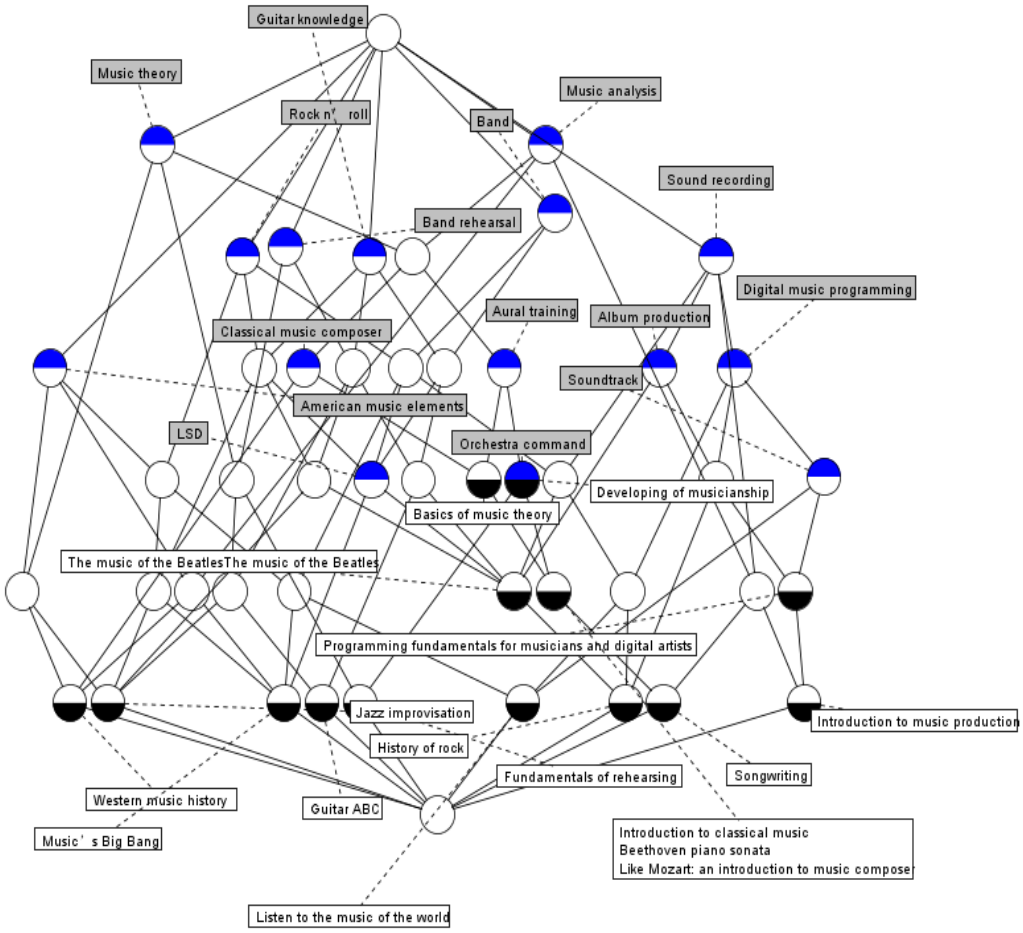


Figure 5. “Music, Movie and Audio” MOOCs Concept Lattice (Label Version)

3.4 Knowledge Clustering based on the MOOCs Concept Lattice

The software “Lattice Miner” is used to do the clustering analysis of the “music, movie and audio” so as to demonstrate explicitly the links among the MOOCs concept lattice.

In the knowledge structure mode of the MOOCs Website, implied poly relations of the line modules and potential knowledge association can not be easily studied through the classification of courses, and the approach of knowledge clustering should be used to do it. In the Hasse of the MOOCs concept lattice in Figure 6, the modules of the “music, movie and audio” topics from the left to the right can be clustered into 8 classes: “Rock n’ roll，Rehearsal，Band，Music theory，American music elements，Sound recording，Music analysis，Guitar knowledge”.

On the surface, the clustering results of the experimental data are not as uniformed as Dewey decimal (DDC)，but in fact, the clustering results based on the concept lattice explain the difference and connection between courses on music[11].

The clustering of the maximal frequent items for labels (label feature) can best reflect the features of the associated labels .The concept lattice of the music MOOCs showed in figure 6 reflects the clustering of the maximal frequent item between objects and attributes, and it can also reveal the structure characteristics of music course on Coursera. This can help to adjust the navigation path of the MOOCs according to retrieval need and establish flexible navigation mechanism of MOOCs, which can facilitate the learners to further mine the knowledge organization characteristics among different MOOCs at different levels.

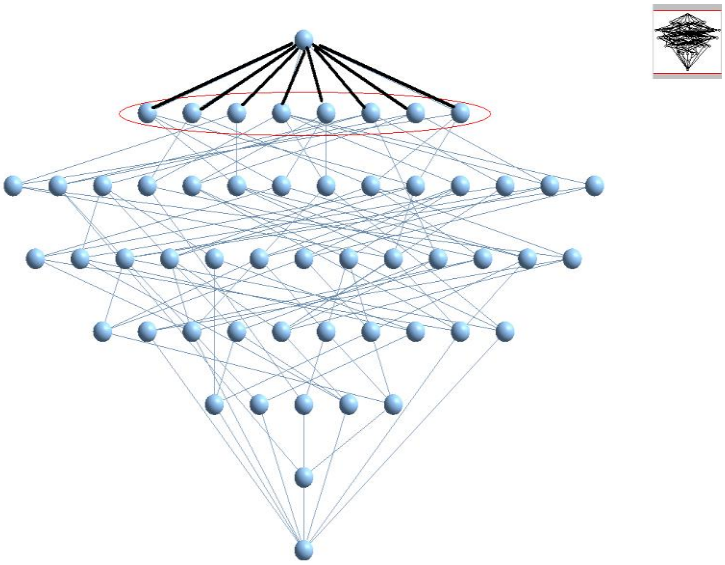


Figure 6. "Music, Movie and Audio" MOOCs Concept Lattice (No Label Version)

We take the subclass with the attribution of "American music elements" for example to analyze in finer granularity, as it's shown in Figure 7:

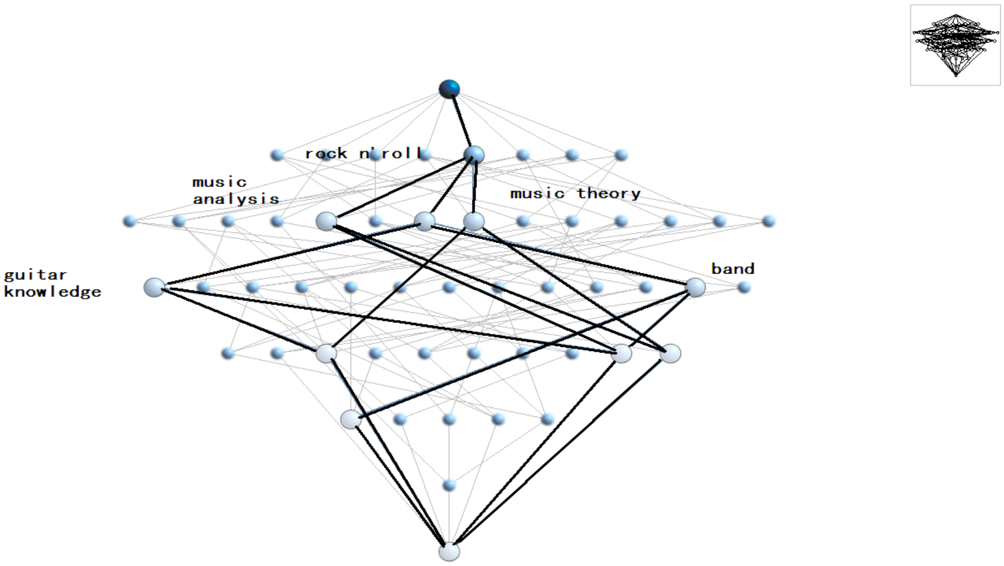


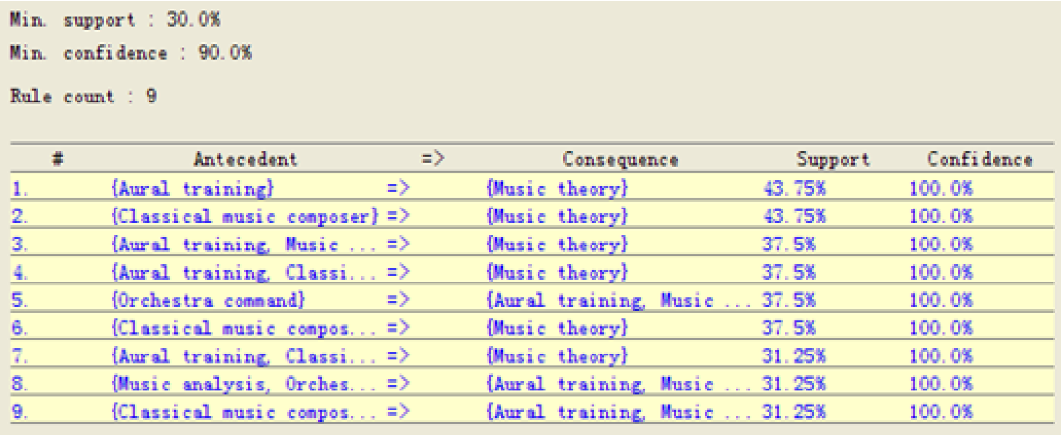
Figure 7. Analysis on Knowledge Granularity of MOOCs Courses based on Concept Lattice

In Figure 7, the subclass "American music elements" can be classified into three sub subclass: the one with "Music analysis and American music elements", the one with "Rock n' roll and American music elements" and the one with "Music theory and American music elements". As sub concept of "American music elements", each of the third subclass refines the concept of the object, owns more concept attributes and keep a clear partial ordering relationship between the objects and their attributes. According to learners' personalized demand, when necessary, the partition granularity of the MOOCs knowledge unit can be further refined. Two fourth subclass or even more detailed subclass can be obtained when necessary. For example, the subclass "guitar knowledge" has the attributes of "Rock n' roll" and "American music elements". Then the related resources corresponding to these subclasses are likely to focus on the knowledge of guitar's role in America rock developments. After all, when it comes to the USA music, it is impossible to avoid the guitar and rock!

Therefore, as for different MOOCs within the same topic, their relationships are neither independent, nor the traditional "the same level course" structure. They are the "relevant courses" with associated relation between each other.

3.5 Mining the Association Rules of MOOCs Knowledge Units

Based on the Hasse Diagram 5 of the MOOCs concept lattice, the association rules of the knowledge units on "music, film and sound" can be mined. In the experiment, the minimum support threshold is set as 30%; the minimum confidence threshold is 90%. The relevant rules can be obtained as Figure 8:



Min. support : 30.0%

Min. confidence : 90.0%

Rule count : 9

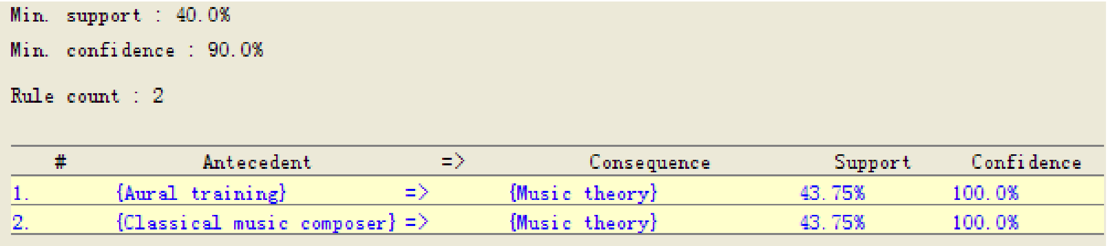
#	Antecedent	=>	Consequence	Support	Confidence
1.	{Aural training}	=>	{Music theory}	43.75%	100.0%
2.	{Classical music composer}	=>	{Music theory}	43.75%	100.0%
3.	{Aural training, Music ...	=>	{Music theory}	37.5%	100.0%
4.	{Aural training, Classi...	=>	{Music theory}	37.5%	100.0%
5.	{Orchestra command}	=>	{Aural training, Music ...	37.5%	100.0%
6.	{Classical music compos...	=>	{Music theory}	37.5%	100.0%
7.	{Aural training, Classi...	=>	{Music theory}	31.25%	100.0%
8.	{Music analysis, Orches...	=>	{Aural training, Music ...	31.25%	100.0%
9.	{Classical music compos...	=>	{Aural training, Music ...	31.25%	100.0%

Figure 8. Association Rule with Support Threshold>=30%, Confidence Threshold>=90%

Taking the association rule R1 in figure 8 for example, "{Aural training} = [43.75%] => {Music theory} [100%] "can be interpreted as: for the courses on the topic of "music, film and sound" , the ones with the knowledge unit of "Aural training" is 43.75% relevant to the courses containing the knowledge unit of "Aural training", and this conclusion is 100% credibility.

The thresholds of the association rules based on the concept lattice can be adjusted by the domain experts by needs after the concept lattice have been finished constructing[12]. So more rules can be obtained flexibly. For example, when the minimum support threshold is adjusted as 40% with confidence threshold unchanged, new valid association rules can be obtained, which is showed in figure 9:

The association rules mining can further indicate that, rich knowledge association exists among the MOOCs courses. From the perspective of the course selection, this knowledge association makes the subsequent courses become a special "knowledge supplement" to the prior courses. Therefore, the association rules mining of the MOOCs knowledge units provide flexible guiding principles for the learners whose knowledge structure is unusual in various fields.



```
Min. support : 40.0%
Min. confidence : 90.0%

Rule count : 2

  #          Antecedent              =>        Consequence          Support      Confidence
1.        {Aural training}         =>      {Music theory}          43.75%       100.0%
2.        {Classical music composer} =>    {Music theory}          43.75%       100.0%
```

Figure 9. Association Rule with Support Threshold>=40%, Confidence Threshold>=90%

4. Conclusion

The author teases out the potential of concept lattice as a powerful data analysis tool on emerging knowledge diffusion mode from two perspectives: one is MOOCs knowledge clustering on the basis of concept lattice and the other is the association rules mining of MOOCs knowledge unit based on concept lattice.

MOOCs knowledge clustering based on the concept lattice contributes to create the knowledge community composed of users (instructors, learners, etc.) who have the similar knowledge background at the same time,

Both the concept generalization and the instantiation refinement functions of the concept lattice make the multi-dimensional expansion of the single retrieval entrance possible, which provides a diversified knowledge service for the MOOCs learners.

The research on the association rules mining of the MOOCs knowledge units contributes to discover the connections among the different online courses under the same topic and find out the developing and evolutional mechanisms of disciplines. This can be used to guide the MOOCs learners to organize the study plans of the courses scientifically from the intercross of many subjects also become one of the hottest research points in the near future.

In conclusion, with the improvement of the concept lattice theory and the continuous deepening of the resources construction for MOOCs, the researches on the construction of the virtual knowledge community and the mining of user preferences based on the concept lattice will be developed accordingly, and it will also become one of the hottest research points in the near future.

Acknowledgements

This paper is granted by the 12th five-year plan research project of the Education Department of Jilin Province “Research on Constructing Collegiate Information Ecosystem ” (No.[2014]184)and is also granted by the Jilin Province social science fund project “ Research on the Service Methods of Digital Library Knowledge Based on Concept Lattice” (No.2014B62). Jingye Qu is the corresponding author of this paper.

References

[1] R. Wille, “Restructuring Lattice Theory: An Approach Based on Hierarchies of Concept”, Proceedings of the7th International Conference on Formal Concept Analysis, **(2009)** October: Berlin, pp. 314-339.

[2] B. Ganter and R. Wille, “Formal Concept Analysis. Science press”, Beijing **(2007)**, pp. 19-22.

[3] G. Stumme, “Efficient Data Mining Based on Formal Concept Analysis”, Proceedings of the13th International Conference on Database and Expert Systems Applications. **(2002)** March: London, pp. 534-546.

[4] C. Carpineto and G. Romano, “Information Retrieval through Hybrid Navigation of Lattice Representations”, International Journal of Human-Computers Studies, vol. 45, no. 5, **(1996)**, pp. 553-578.

[5] M. Kim and P. Compton, “Formal Concept Analysis for Domain-Specific Document Retrieval Systems”, Computer Science **(2001)**.

[6] Y. Zhang and B. Feng, “Tag based User Modeling Using Formal Concept Analysis”, Proceedings of the 8th IEEE International Conference on Computer and Information Technology, **(2008)** October: Sydney pp. 485-490.

[7] F. Masseglia, F. Cathala and P. Poncelet, “The PSP Approach for Mining Sequential Patterns”, Proceedings of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery, **(1998)** September: London, pp. 176-184.

[8] M. Zaki and J. Spade, “An Efficient Algorithm for Mining Frequent Sequences”, Machine Learning **(2001)**.

[9] J. W. Han, J. Pei and B. Mortazavi-Asl, “Free Span: Frequent Pattern-Projected Sequential Pattern Mining”, Proceedings of the6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. **(2000)** September: Boston, pp. 355-359.

[10] J. W. Han, J. Pei and B. Mortazavi-Asl, “refix Span: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth”, Proceedings ofthe17th International Conference on Data Engineering, **(2001)**, pp. 215-224.

[11] Y. K. Kang, S. H. Hwang and Y. K. Kang, “Development of a FCA Tool for Building Conceptual Hierarchy of Clinical Data. Journal of the Korean Society of Medical Informatics **(2005)**.

[12] Wille R. Methods of Conceptual Knowledge Process. Proceedings of the4th International Conference, **(2006)** October: Berlin, pp: 1-29.

Authors

Junfeng Gao, He is a PhD candidate in information science from Jilin University, China and currently a lecturer in information science in Beihua University, China. His research interest is digital library.

Jingye Qu, She received the master degree in information science from Jilin University, China and she is now studying the PhD of information science in Jilin University, China. She is currently a lecturer in computer science in Beihua University, China. Her research interest is information ecology.

Zhong Xin, He received the Master degree in Computer Application Technology from Northeast Dianli University in 2010. He is a currently a lecture in Jilin Technology College of Electronic Information, China. His research interests focus on senor network.