Analysis and Application of Formal Concept in Information Science

Weichuan Yang*
Jiangmen Polytechnic, Jiangmen 529090, Guangdong, China

*Corresponding author email: 308713977@qq.com

Abstract: Based on Information Science, new mechanism, two types and basic characteristics of information processing are discussed in the paper. Besides, as a new method of unsteady information processing, the paper not only introduces some basic ideas, concepts, techniques of formal concept analysis (FCA) and Galois liaison but also expounds the feasibility of formal concept analysis for knowledge representation and knowledge acquisition through examples. Finally, the application and research progress of formal concept analysis in information science are further analyzed in the paper.

Keywords: Information science, Formal concept analysis, Application and research.

1. Introduction

1.1 Information Science

(1) On the Concept of Information
The formation and rapid development of information science, which takes information as the main research object, has not been recognized in its research scope. However, it is a comprehensive subject which mainly studies the movement law and application method of information, with the help of computer and other technologies as the main research tools, aiming at further expanding human information function reality. Information science is a new comprehensive science which is composed of information theory, system theory, cybernetics, computer theory, bionics theory, system engineering and artificial intelligence theory. Its pillars are information theory, system theory and cybernetics. In the 1960s, system identification became an important research topic because the complex engineering system needed to control the production process with computer technology. In essence, system identification is to study the behavior and internal structure of the control system through the input and output information, and to express it with mathematical model. Control is to process, transform and utilize information according to system structure and requirements.

(2) Definition of Information
In the sense of ontology, information is the state of things' movement and the way of state change, which marks the attribute of things' existence and relationship. However, such description is difficult to quantify information. In the sense of epistemology, it is the state and way of things' movement perceived or expressed by the cognitive subject. Therefore, Shannon thinks that definition of information is received by the cognitive itself and can eliminate the new content and knowledge of the uncertainty of the cognition of things from the perspective of epistemology, and he also makes a quantitative description about information. According to this definition, the mathematical description of information quantity is given as follows:
Information quantity is a central concept in information theory, which is a physical quantity to measure the amount of information. It reflects the amount of information transmitted when an event with a certain probability occurs. Represented by $H(x)$ as information entropy, it is the average uncertainty of the whole information source, And $I(p)$ is information, it represents the degree of eliminating uncertainty after receiving information from the perspective of information sink, that is, the amount of acquiring new knowledge, so it is not intended to be meaningful until the information entropy sent by the information source is received by the information sink. In the ideal situation of eliminating interference, the signal sent by the source corresponds to the signal received by the sink one by one, $H(x)$ and $I(p)$ are equal.

The measurement of information is related to the randomness of the events it represents or the probability of additional occurrence event. When the probability of occurrence of an event is large and it is easy to judge in advance, the greater the probability of occurrence of the event is, the less the uncertainty is, and the less the amount of information it contains; otherwise, the less the probability of an event occurrence is, the less the probability of an event occurrence is, and the difficulty of an event occurrence is predicted The more difficult it is, the more uncertain it is, and the more information it contains. According to Shannon's formula of information quantity, the concept of information entropy theory statistics is used

$$H(x) = \sum P(x_i) \log \frac{1}{P(x_i)}$$

Mutual information refers to the correlation between two event sets. The common information of the two events and is defined as:

$$I(X, Y) = H(X) + H(Y) - H(X, Y)$$

where $H(X, Y)$ is joint entropy and is defined as $H(X, Y) = -\sum P(X, Y) \log P(X, Y)$

1.2. Type of Information

Since the beginning of the 21st century, scientists have questioned Newton's "linear" theory and paid special attention to the study of nonlinear problems. The non-linearity of information system is manifested in the unsteady nature of information, that is, incompleteness, fuzziness, randomness, roughness, outlier, restatement, time varying and space varying. Therefore, the information can be divided into two categories: steady information and unsteady information.

For the uncertainty of unsteady information system, Bayes theory based on probability theory, D-S evidence theory and possibility theory based on fuzzy set are usually used as mathematical basis. Since the 1990s, scientists domestically and abroad have made great efforts in the research of uncertain information and achieved fruitful results. However, these studies are often limited to the study of fuzzy information and rough information, and the methods used are also targeted to use the fuzzy theory and rough set theory, which has not been broken through in theory; in application, it has not penetrated into the unsteady information system. Compared with fuzzy theory and rough set theory, FCA has at least the following advantages:

(1) The theoretical basis of FCA is lattice theory, which has a solid mathematical theoretical basis for dealing with discrete quantities.

(2) The combination of FCA and Galois theory can reveal the internal relationship between objects and attributes, which makes Galois connection play an important role in knowledge discovery.

(3) Based on concept lattice (Galois lattice) Hasse diagram is easy to construct, and through the use of sub concept and super concept, it can realize the visualization of knowledge variable granularity.

(4) FCA is suitable for knowledge discovery of multivalued database, fuzzy database, approximate database, restated database, temporal database and spatial database.

2. Basic Concepts of FCA

Formal concept analysis is a new mathematical data processing method proposed by Professor Rudolf Wille of Darmstadt University in Germany and his colleagues in the 1980s. It is a powerful
A mathematical tool to simulate concepts and knowledge processing in the physical world. In information system, FCA can be used for data analysis, knowledge representation and information management. In theory, FCA is not only the further development of fuzzy theory and rough set theory, but also enriches and develops the theoretical basis of information processing; besides, it plays an essential role and has a wide range of application fields and prospects in the following ones, such as medical diagnosis, psychological consultation, musicology, natural language processing, library and information science, secondary development of software, economic planning, municipal engineering, marketing, ecological environment and some others.

2.1. Basic Ideas of FCA

The basic ideas of FCA mainly come from domain theory, lattice theory and ontology. The so-called domain theory refers to the idea of the feasibility of approximate calculation from the domain theory of information and infinite structure. It can be concluded that knowledge or information should be limited representation or perception, thus extending the approximate concept. Based on lattice theory, we may know that the approximate concept lattice associated with background is strictly a complete algebra, and a typical formal concept is approximate. In the case of limited formal background, approximate concept and formal concept are consistent. The concept of ontology originated from the field of philosophy. It is a philosophy view of human beings to the nature, namely "ontology", which means knowledge and knowing. With the rapid development of information science, especially computer science in the 1970s-1980s, the representation of natural world cognition and the formalization of knowledge have been represented, interpreted and utilized by computers. This being said domain ontology is a structured domain knowledge, which can be interpreted and utilized by computer. The core concept of FCA is Galois connection, whose duality can be used to describe two types of items related to each other, such as object and attribute, document and item, etc. FCA is made up of formal objects and attributes. Formal objects, attributes and their relationships form formal context. The concept in formal context can be obtained by using the properties and operations of Galois connection induction function. The remarkable feature of FCA is that it can apply the visualization of Galois lattice. The following basic concepts of FCA and Galois contact come from [1] and [9].

2.2. Contact between FCA and Galois

**Definition 1** formal context is a triple $D = (O, I, R)$, where $O$ and $I$ are the set of objects and attributes (items), and $R \subseteq O \times I$ is the binary relationship between objects and items. The data set of data mining can regard $O$ and $I$ as the formal context of non empty finite set. Dual $(o, i) \in R$ indicates that the object $o \in O$ is related to term $i \in I$. For example, table 1 is a formal background with six objects (OIDs) and five items.

Let $D = (O, I, R)$ be the formal context. For $O \subseteq O$ and $I \subseteq I$, the induced function is defined as

$$f : 2^O \rightarrow 2^I$$
$$g : 2^I \rightarrow 2^O$$

$$f(O) = \{i \in I \mid \forall o \in O, (o, i) \in R\}$$
$$g(I) = \{o \in O \mid \forall i \in I, (o, i) \in R\}$$

**Table 1. Data Context.**

| OID | Item sets |
|-----|-----------|
| 1   | A C D E   |
| 2   | A B C     |
| 3   | B C D E   |
| 4   | A C D     |
| 5   | A B C D E |
Function $f(O)$ is the same items set in object set $O$; $g(I)$ is the same set of objects in item set $I$. The concept in context $D = (O, I, R)$ is an ordered pair of $(O,I)$, so that $f(O) = I, g(I) = O$. Among them, $O \subseteq O$, $I \subseteq I$, $O$ is the extension of concept $(O,I)$, and $I$ is the connotation of concept $(O,I)$. For example, in Table 1, if you make $O = \{2356\}, I = \{BC\}$ because of $f(O) = f(\{2356\}) = \{BC\} = I$ and $g(I) = [2356] = \{BC\} = O$, $(\{2356\}, \{BC\})$ is a concept, its extension is the object set $O = \{2356\}$, and its connotation is the item set $I = \{BC\}$; if you make $O = \{25\}, I = \{AB\}$ because of $g(I) = g(\{AB\}) = \{25\} = O$, but $f(O) = f(\{25\}) = \{ABC\} \neq I$, $\{\{25\}, \{AB\}\}$ cannot constitute a concept [8].

**Definition 2** let $(O, I, R)$ be the formal context, call dual $(f, g)$ the inductive function, they are the Galois connection between power set $2^O$ and $2^I$. For $I_1, I_2, I \in I$ and $O_1, O_2, O \in O$, the Galois connection holds for the following properties.

1. $I_1 \subseteq I_2 \Rightarrow g(I_1) \subseteq g(I_2)$
2. $O \subseteq g(I) \iff I \subseteq f(O)$
3. $I \subseteq h(I)$
4. If $I_1 \subseteq I_2$, $h(I_1) \subseteq h(I_2)$
5. If $O_1 \subseteq O_2$, $h(O_1) \subseteq h(O_2)$

If the set of all concepts in context $(O, I, R)$, i.e. $L(O, I, R) = \{(O, I) \in O \times I \mid f(O) = I, g(I) = O\}$, has partial order $(O_1, I_1) \preceq (O_2, I_2)$, then $L(O, I, R)$ is the concept lattice of formal context $(O, I, R)$. Concept lattice is a complete lattice, also named as Galois lattice. The concept lattice and Hasse diagram corresponding to table 1 are shown in Table 2 and figure 1 respectively [2,3,4,5,6].

**Table 2.** Table 1 concept lattice of formal context.

| Concept      | Support degree | Concept      | Support degree |
|--------------|----------------|--------------|----------------|
| $\{\{123456\}, \{\emptyset\}\}$ | 6/6            | $\{\{25\}, \{ABC\}\}$ | 2/6            |
| $\{\{123456\}, \{C\}\}$         | 6/6            | $\{\{145\}, \{ACD\}\}$ | 3/6            |
| $\{\{124\}, \{AC\}\}$           | 4/6            | $\{\{356\}, \{BCDE\}\}$ | 3/6            |
| $\{\{2356\}, \{BC\}\}$           | 4/6            | $\{\{15\}, \{ACDE\}\}$ | 2/6            |
| $\{\{12456\}, \{CD\}\}$          | 5/6            | $\{\{\emptyset\}, \{ABCDE\}\}$ | 0              |
| $\{\{1356\}, \{CDE\}\}$          | 4/6            | $\{\{1345\}, \{CD\}\}$  |                |
| $\{\{1356\}, \{CDE\}\}$          |                | $\{\{125\}, \{ABC\}\}$  |                |
| $\{\{1356\}, \{CDE\}\}$          |                | $\{\{1245\}, \{AC\}\}$  |                |
| $\{\{1356\}, \{CDE\}\}$          |                | $\{\{125\}, \{ABC\}\}$  |                |

**Figure 1.** Hasse diagram of Table 2.
2.3. Galois Contact and Association Rules

**Definition 3** If the support \( \text{Supp}(I) = |\{I\}| \) of items set \( I \) in domain \( D \) is not less than the support threshold \( \text{Min_supp} \in [0,1] \), then item set \( I \) is reported to be frequent, and the multiple item sets in domain \( D \) is recorded as \( F = \{I \subseteq I | \text{Supp}(I) \geq \text{Min_supp}\} \).

Association rule is a partial order implication of form \( r : I_1 \rightarrow I_2 \), where \( I_1, I_2 \subseteq I, I_1 \neq \emptyset \), and \( I_1 \cap I_2 = \emptyset \). The support degree \( \text{Supp}(r) \) and confidence degree \( \text{Conf}(r) \) of association rule \( r : I_1 \rightarrow I_2 \) are respectively defined as:

\[
\text{Supp}(r) = \frac{|\{I \subseteq I_1 | I \cap I_2 = \emptyset\}|}{|\{I \subseteq I_1\}|},
\]

\[
\text{Conf}(r) = \frac{\text{Supp}(I_1 \cup I_2)}{\text{Supp}(I_1)} = \frac{|\{I \subseteq I_1 \cup I_2 | I \cap I_2 = \emptyset\}|}{|\{I \subseteq I_1\}|}.
\]

Association rules in formal context are those rules whose support and confidence are not less than the minimum support and confidence threshold respectively, that is, the set \( AR \) of association rules in \( D \) is:

\[
AR = \{r : I_1 \rightarrow I_2 | I_1 \subseteq I, I_2 \subseteq I, \text{Supp}(I_1 \cup I_2) \geq \text{Min_supp} \land \text{Conf}(r) \geq \text{Min_conf}\}.
\]

If \( \text{Conf}(r) = 1 \), the association rules are accurate; otherwise, they are approximate.

| Item set | \( \text{Supp}(I) \) |
|----------|------------------|
| \{\emptyset\} | 6/6 |
| \{C\} | 6/6 |
| \{AC\} | 4/6 |
| \{BC\} | 4/6 |
| \{CD\} | 5/6 |
| \{ACD\} | 3/6 |
| \{CDE\} | 4/6 |
| \{BCDE\} | 3/6 |

**Definition 4** if item set \( I \subset D \) is both closed and frequent, then it is a frequent closed item set. Frequent closed item set \( FC \) in domain \( D \) is recorded as:

\[
FC = \{I \subseteq I | I = h(I) \neq \emptyset, \text{Supp}(I) \geq \text{Min_supp}\}.
\]

Frequent closed item sets are the largest item sets with a common set of objects. The support degree of the item sets is at least \( \text{Min_supp} \). In Table 1, if \( \text{Min_supp} = 1/2 \) is set, then the frequent closed item set of domain \( D \) is shown in Table 3. For example, \( I = \{AC\} \), because of \( h(I) = f \circ g(\{AC\}) = f(\{1245\}) = \{AC\} = I \), \( \text{Supp}(I) = \frac{|\{AC\}|}{|\{I\}|} = \frac{4}{6} \), \( \{AC\} \) is a frequently closed item set; for \( I = \{AB\} \), because of \( h(I) = h(\{AB\}) = f \circ g(\{AB\}) = f(\{25\}) = \{ABC\} \neq I \), it is not a closed item set.

**Theorem:** The support degree of term set \( I \) is equal to the support degree of its closure (the smallest closed term set including \( I \) ), that is, \( \text{Supp}(I) = \text{Supp}(h(I)) \).

**Prove:** Let \( I \subseteq I \) be the term set, and the support degree of \( I \) in domain \( D \) is \( I \). Because the closure of \( I \), \( \hat{h}(g(I)) = g \circ f(g(I)) = g \circ h(I) = g(I) \), is connected by Galois with properties 3 and 1, \( g(I) \subseteq h(g(I)) \) and \( I \subseteq h(I) \Rightarrow g(h(I)) \subseteq g(I) \). So, \( \hat{h}(g(I)) = g(I) \subseteq h(g(I)) \). And because of \( \hat{h}(g(I)) = g(I) \), so \( g(h(I)) = \hat{h}(g(I)) \). So, \( \text{Supp}(h(I)) = \frac{|\hat{h}(g(I))|}{|\{I\}|} = \frac{|\hat{h}(g(I))|}{|\{I\}|} = \frac{|\hat{h}(g(I))|}{|\{I\}|} = \text{Supp}(I) \).

3. Application of FCA in Information Science

3.1. FCA, Widely Used in IS

It is mainly used in (1) information. FCA supports fine query of information. (2) Knowledge
representation and knowledge discovery. FCA is different from the traditional statistical method of data analysis. It is easy to discuss and develop the conceptual structure. The concept lattice is characterized by external cognition. (3) Logic and artificial intelligence. The concept Diagram (Hasse Diagram) of FCA is similar to the semantic network in AI knowledge representation. It can be used to represent semantic relations. Most natural languages can be translated into concept maps. (4) Software engineering. For the binary relationship between objects and attributes, formal concept analysis is a powerful technique to organize objects and attributes into concept lattice. The technical methods used should depend on the details of three aspects: module reconstruction, design pattern and constraint based conflict analysis [4]. (5) Linguistics. As data analysis and knowledge representation, FCA has the potential application to solve many linguistic problems, including feature recognition and analysis, such as recording and analyzing speech, syntax and grammar characteristics. (6) E-commerce. FCA can be used in concept similarity measurement. Based on collaborative filtering technology, E-commerce recommendation system does not analyze the similarity between goods, but learn the similarity between the purchase behavior of the target user and the historical user, so as to generate recommendation results according to the purchase behavior of similar historical users. [7]

Taking the application of formal concept analysis in software engineering as an example, the paper will further elaborate the universality of the application of formal concept.

3.2. Formal Concept Analysis, Used in Software Engineering

Software engineering is an engineering discipline whose goal is to build a large software system successfully. Formal concept analysis is a mathematical analysis method. When it is applied to specific software design, its main function is to construct feature sets of software projects. It can implement its application [4] through the following steps, as shown in Figure 2. In Figure 2, Firstly, through a system analysis of software requirements analysis and various functions, each specific feature set is abstractly expressed. Secondly, by using the formal concept analysis method, the concrete concepts are constructed and expressed as concrete class forms. Thirdly, by analyzing the relations between concept classes, construct concept lattice and draw concept lattice diagram. Finally, the above analysis is applied to each stage of the system design, and the corresponding work is completed according to the normal development procedure and order of the system, so as to improve the development efficiency of the software project.

(1) Application of Demand Analysis

Demand analysis is a detail analysis aiming at various demands of software. The first step is to construct the feature set of software project. This process is the construction of the formal background of FCA. As shown in Table 4.

| Requirement analysis | Structural design | Detailed design | Code | Test | Comprehensive | Qualification test | Install | Identification support | Software maintenance |
|----------------------|-------------------|-----------------|------|------|---------------|-------------------|--------|------------------------|---------------------|
| debugging            | ×                 | ×               |      |      |               |                   |        |                        |                     |
| concept              |                   | ×               |      |      |               |                   |        |                        |                     |
| Use case             | ×                 | ×               |      |      |               |                   |        |                        |                     |
| Distinguish          | ×                 | ×               |      |      |               |                   |        |                        |                     |
| Reengineering        |                   | ×               |      |      |               |                   |        |                        |                     |
| control              | ×                 |                |      |      |               |                   |        |                        |                     |
| Help                 | ×                 |                |      |      |               |                   |        |                        |                     |

Figure 2. Formal Concepts in Software Engineering.
The concept class is generated by using the concept lattice construction algorithm. The concept lattice construction algorithm can have Christian Linding's fast concept analysis. These algorithms calculate all the concepts and hierarchical relationships of the lattice. A bottom-up calculation method to calculate the concept lattice is used here. Variable description:

**N**: A single object has the maximum number of properties in the background.

**A**: Collection, where the element is a collection of single or multiple attributes in the background.

**Node**: node set. Node (I, J) represents the j-th node on the i-th layer.

**Algorithm:node set. Node (I, J) represents the j-th node on the i-th layer.**

1:Calculate **N**:
   1:For (i = N; i >= 1; i--)
   2:{
      1:Calculate **A**:
      2:For (j = 1; j < size of (A). j++)
      3:{
         1:a = A’(j);
         2:B = A(j);
         3:If (a’=b&&a=b’)
         4:{Node (I, j). attribute list: b;
         5:Node (I, j). object list = a;
         6:Determine whether the attribute list (b) is included in the attribute list of the lower level nodes, and determine the parent-child relationship between nodes;
         7:}
      8:}
      9:}
   10:}

2) Application in Structural Design

On the basis of the demand analysis, the design method is obtained by rationalizing the organization and analyzing the design. In the formal concept analysis method, the concrete concept is constructed by the object attribute set in Table 1. The calculation of structure concept is as follows:

1:K = 1; / / number of current concepts
2:For (i = 1; i <= N; i++)
3:For (j: I; j <= N; j++)
4:If (MJ, FI) ∈ T) MK = MK ∪ MJ; / / if MJ is used by FI, MJ is added to MK
5:}
6:CK = (MK, FI); / / form a concept
7:K++:
8:For (n = I; n >= 1; n 11) / / check whether the concept is repeated or can be combined
9:If (n <= Cn) {delete; Ck, K; break;}
10: {CN = cn ∨ CK; delete; VK, K -- one; break;}
11:}

**Acknowledgement**

Great thanks should be given to the research, supported by the Ministry of education of the people's Republic of China (Project Number: GKA10148)

**Reference**

[1] Ganter B, Wille R. Formal Concept Analysis: Mathematical Foundations[M].New York: Springer-Verlag, 1999.

[2] Changhua Wei. Similarity based approximate information system concept optimization algorithm. *Microcomputer system* [J].2006, 27(4):706-709.

[3] Shuxian He, Guizhi Liu, Shuwen Li. Formal concept analysis and its application progress. *Applied Technology* [J]. 2007 (5): 77-79.
[4] Ping Jiang, Shengbing Ren, Juan Lin. Application of formal concept analysis in software engineering. *Computer technology and development* [J], 2008 (4): 127-130.

[5] Zhiheng Pang, Wei He, Lei fan, Xuelei Cao. An algorithm for constructing concept lattice. *JOURNAL OF CENTRAL UNIVERSITY FOR NATIONALITIES (NATURAL SCIENCE)* [J] 2008, 7 (1): 19-23.

[6] Lifeng Li. A class of fuzzy concept lattices based on RO implication operator. *Journal of Baoji University of Arts and Science (NATURAL SCIENCE EDITION)* [J], 2007, 27 (2): 107-109.

[7] Hongsheng Xu, Ruiling Zhang. Application of art and concept similarity in e-commerce recommendation system. *Computer engineering and application* [J], 2010, 46 (7) 213-217.

[8] Changhua Wei, Min Wei, Weichuan Yang. Mining minimum non redundant association rules based on Galois contact. *Computer engineering and Science* [J], 2007, 02 (2) 93-96.

[9] Information on http://baike.themanage.cn/baike/%e4%bf%a1%e6%81%af%e8%ae%ba