

# 1 Literature Survey

## 1.1 Lattice-theoretic Concept Analysis

The importance of formal models has been widely acknowledged for nearly two decades [?, ?, ?, ?, ?]. Wille’s pioneering paper, *Restructuring Lattice Theory: An Approach Based on Hierarchies of Concepts* (1982), received wide attention and many researchers have reported successful applications of his lattice-theoretical approach to information retrieval and data mining [?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?]. To formalize the notion of a concept, he operates with a *context*.

**Definition 1** A context is defined as a triplet  $(G, M, I)$ , where  $G$  and  $M$  are sets and  $I \subseteq G \times M$ .

The elements of  $G$  are objects, and the elements of  $M$  are attributes describing the objects. Without loss of generality, the attributes can be thought of as boolean variables: if an object possesses a property, then the relevant attribute is *True* for that object, otherwise it is *False*. The expression  $(g, m) \in I$  reads “object  $g$  has attribute  $m$ .” Table 1 shows an example of the context  $(G, M, I)$ , where  $G = \{N_1, N_2, N_3, \dots, N_8\}$  is a set of students, while the set of attributes is  $M = \{H\text{-black}, H\text{-blond}, H\text{-red}, E\text{-blue}, E\text{-brown}\}$ <sup>1</sup>.

Wille defines a concept as a pair [extension, intension]<sup>2</sup>. *Extension* is a set  $A \subseteq G$  of objects, and *intension* is a set  $B \subseteq M$  of attributes shared by objects in  $A$  (note that this definition of intensions is somewhat constrained as compared to machine-learning research). Let  $A \subseteq G$  and  $B \subseteq M$ , and let

$$\beta(A) = \{m \in M \mid gIm, \forall g \in A\},$$

$$\alpha(B) = \{g \in G \mid gIm, \forall m \in B\}.$$

The functions  $\beta$  and  $\alpha$  have been shown to define a *Galois connection* between the power sets of  $G$  and  $M$  [?]. Intuitively,  $\beta(A)$  is the maximal set of attributes shared by all the objects in  $A$ , whereas  $\alpha(B)$  is the maximal set of objects possessing all attributes in  $B$ .

**Definition 2** A concept in the context  $(G, M, I)$  is a pair  $(A, B)$  such that  $\beta(A) = B$  and  $\alpha(B) = A$ , where  $A \subseteq G$  and  $B \subseteq M$ .

Only certain subsets of  $G$  and that of  $M$ , correspond to concepts. The subsets that define concepts are called *feasible sets*.

---

<sup>1</sup>The context from Definition 1 is identified with the training set in machine learning.

<sup>2</sup>Although Wille himself speaks about “extents” and “intents”, we will follow the terminology used in machine learning: “extensions” and “intensions.”

Table 1: Example of a Context

Student	Hair Color			Eye Color	
	Black	Blond	Red	Blue	Brown
$N_1$			×		×
$N_2$		×		×	
$N_3$		×		×	
$N_4$		×		×	
$N_5$		×			×
$N_6$	×				×
$N_7$	×				×
$N_8$	×				×

**Definition 3** A set  $B \subseteq M$  is called feasible if  $\beta(\alpha(B)) = B$ , i.e.  $B \subseteq M$  is the intension of the unique concept  $(\alpha(B), B)$ . Similarly, a set  $A \subseteq G$  is called feasible if  $\alpha(\beta(A)) = A$ . In this case,  $A$  is the extension of the unique concept  $(A, \beta(A))$ .

The intension of an atomic concept is a conjunction of attributes that each object in the extension must possess. Let  $\mathcal{C}(G, M, I)$  denote the set of all concepts of the context  $(G, M, I)$ . An order relation on  $\mathcal{C}(G, M, I)$  can be defined as follows. If  $(A_1, B_1)$  and  $(A_2, B_2)$  are concepts in  $\mathcal{C}(G, M, I)$ , then  $(A_1, B_1) \leq (A_2, B_2)$  iff  $A_1 \subseteq A_2$  (or equivalently  $B_1 \supseteq B_2$ ), and we say that  $(A_1, B_1)$  is a *subconcept* of  $(A_2, B_2)$  and  $(A_2, B_2)$  is a *superconcept* of  $(A_1, B_1)$ . The *fundamental theorem* proved by Wille [?] establishes that  $\mathcal{L}(G, M, I) = (\mathcal{C}(G, M, I), \leq)$  is a lattice; more specifically,  $\mathcal{L}(G, M, I)$  is a complete lattice called the *concept lattice* of the context  $(G, M, I)$ . The theorem not only establishes that any concept lattice is complete but also that for every complete lattice,  $\mathcal{L}$ , there exists a concept lattice isomorphic to  $\mathcal{L}$  [?]. Figure 1 shows the concept lattice for the atomic concepts of the context shown in Table 1.

The definition of context given above can be easily adapted to information retrieval or association mining domains. For example, for the association mining domain, we define a context as follows:

**Definition 4** An association mining context is defined as a triplet  $(T, \mathcal{I}, R)$  where  $T$  is a set of transactions,  $\mathcal{I}$  is a set of items and  $R \subseteq T \times \mathcal{I}$ .

An association mining context is a formal definition of a transaction database. The set  $T$  is the set of all transactions in the database and the set  $\mathcal{I}$  is the set of all items in the database. For  $t \in T$  and  $i \in \mathcal{I}$  we write  $(t, i) \in R$  to mean that the transaction  $t$  contains the item  $i$ .

Formal concept analysis can be used for many processes in data mining as well as information retrieval (e.g. for classification, for the analysis of attribute dependencies, and

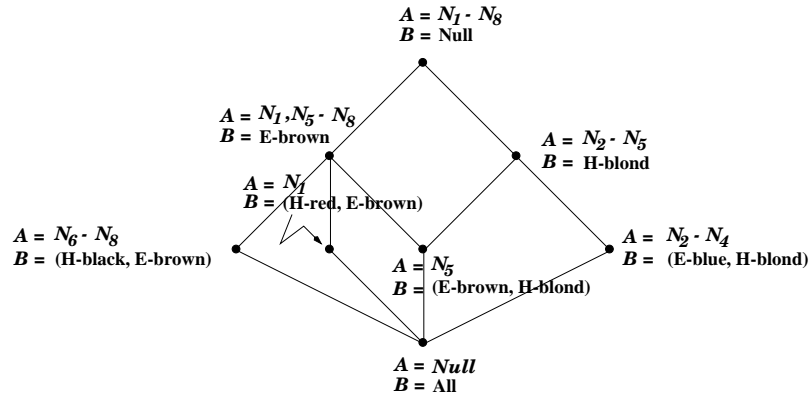


Figure 1: The Concept Lattice of Atomic Concepts

for data fusion). Standard lattice theory techniques, including sub-lattices, factor lattices, infimum-preserving and supremum-preserving embedding, direct product, horizontal sum, and scaling (to deal with multi-valued attributes) are being applied to these problems [?, ?, ?, ?]. In a lattice based application, it is conceivable that concept lattices grow to a fairly large size. However, a few researchers have proposed efficient techniques to alleviate the problem of size and render lattice theoretic approaches attractive to applications in information retrieval (IR) and in knowledge discovery in databases (KDD) [?, ?, ?, ?].