УДК ???.?

# СИСТЕМА АНАЛИЗА ДАННЫХ "CONCEPT EXPLORER"

С. А. Евтушенко[1]

В докладе описана программная система, предназначенная для интеллектуального анализа данных методом формального концептуального анализа (ФКА). В ней реализованы функции создания и редактирования контекстов, построения множества всех понятий и диаграмм концептуальных решеток. Решается задача выявления зависимостей между объектами и их свойствами, которые выражаются в виде логических импликаций. Разработан алгоритм, позволяющий находить базис импликаций.

В данной работе представлена система "Concept Explorer", реализующая метод формального концептуального анализа данных. Система реализована на языке программирования Java 2, с использованием библиотеки Swing для реализации графического интерфейса.

Формальный концептуальный анализ (ФКА) — логико-алгебраический метод анализа данных, предложенный в 1981 году Рудольфом Вилле[Wille, 1982]. В методе отражено философское понимание понятия как единицы мышления, определяемой своим объемом и содержанием. ФКА предназначен для исследования объектов, которые задаются имеющимися у них свойствами. Для установления связи между объектами и их свойствами служит формальный контекст.

Простой формальный контекст это тройка множество объектов, множество свойств, связь между объектами и свойствами.

**Формальным контекстом** называется тройка вида $K=(G, M, I)$, где $G$ и $M$ – множества, а $I$ – отношение на множестве $G$x$M$. $G$ представляет множество объектов, $M$ – множество свойств, а $gIm$ означает, что объект $g$ обладает свойством $m$.

Простой формальный контекст может быть задан бинарной матрицей $K$.

[1] *02056, Киев, пр. Перемоги, 37,НТУУ "КПИ", ФПМ; sye@mail.ru*

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | X | X | X | | | | |
| 2 | X | X | X | X | | | |
| 3 | | X | | X | X | X | |
| 4 | | | X | X | X | | |
| 5 | | | | X | X | | X |
| 6 | | | | | | X | X |

Рис. 1. Формальный контекст K = (G, M, I).

Одному понятию соответствует максимальная по вложению единичная подматрица матрицы *K*. Так, например, для заданного на рисунке 1 контекста одним из понятий является пара ({*1*, *2*},{*A*, *B*, *C*}).

Бинарное отношение I представимо в виде двудольного графа (рис. 2). Тогда одному понятию соответствует максимальный по вложению полный подграф двудольного графа.

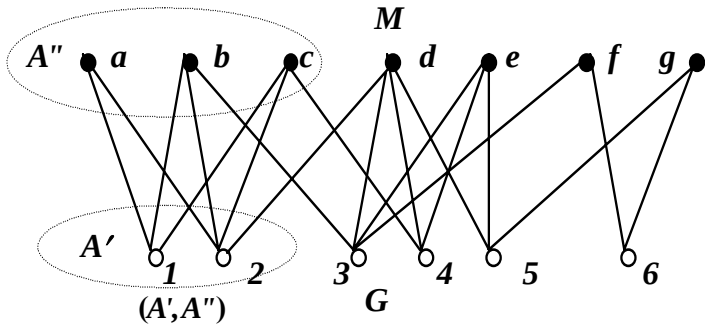



Рис. 2. Двудольный граф формального контекста K = (G, M, I).

В ФКА *понятия* принято называть *формальными концептами.* Множество объектов представляет объем (экстенсионал) понятия (концепта), а множество всех свойств, которыми они обладают, – его содержание (интенсионал).

Математически формальное понятие определяется с помощью соответствий Галуа.

**Соответствие Галуа** [Биркгоф 1984] – это пара отображений

$s: G \mid\rightarrow M$, $t: M \mid\rightarrow G$, такое, что если $A \subseteq G$ и $B \subseteq M$, то

$s(A) := \{m \in M \mid gIm$ для всех $g \in A\}$ ,

$t(B) := \{g \in G \mid gIm$ для всех $m \in B\}$ .

Таким образом, $A' \equiv s(A)$ есть множество тех свойств, которыми обладают все объекты подмножества $A \subseteq G$, а $B' \equiv t(B)$ есть множество тех объектов, которые обладают всеми свойствами из $B \subseteq M$. Тогда одно понятие - это пара $(A, B)$, где $A' = B$, $B' = A$. Поэтому понятие $(A, B)$ может обозначаться также как $(B', A') = (A'', A') = (B', B'')$.

Центральную роль в формальном концептуальном анализе играет следующая теорема [Wille, 1982]:

**Теорема.** Пусть $K = (G, M, I)$ есть формальный контекст, а $\underline{B}(G, M, I)$ есть множество всех концептов контекста $K$. Тогда $\underline{B}(G, M, I)$ является полной решеткой, в которой операции пересечения и объединения задаются следующим образом:

$$\bigwedge_{t \in T} (A_t, B_t) = \left( \bigcap_{t \in T} A_t, \left( \bigcup_{t \in T} B_t \right)'' \right)$$

$$\bigvee_{t \in T} (A_t, B_t) = \left( \left( \bigcup_{t \in T} A_t \right)'', \bigcap_{t \in T} B_t \right)$$

Решетку $\underline{B}(G, M, I)$ называют *концептуальной решеткой.*

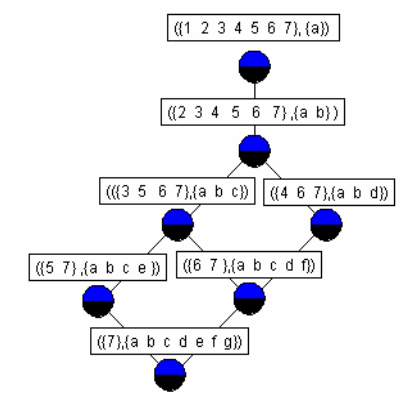
С помощью концептуальной решетки мы сразу же получаем классификацию объектов относительно заданных свойств. Концептуальные решетки позволяют "развернуть" и визуализировать структуру заданных данных, что дает возможность находить в них некоторые закономерности, регулярные структуры, исключения и т. п.

Решетки понятий можно изображать с помощью обычных диаграмм Хассе. Однако, при пометке каждого понятия его объемом и содержанием изображение не будет наглядным. Поэтому используется сокращенная пометка, в которой каждый объект и каждый атрибут изображаются на диаграмме всего один раз. Имя объекта приписывается пересечению всех понятий, в объемах которых содержится этот объект, а имя свойства приписывается объединению всех понятий, содержания которых включают это свойство. Таким образом, имя объекта приписывается наименьшему из понятий, в которых встречается данный объект, а имя свойства приписывается наибольшему из понятий, в которых присутствует это свойство. Такие диаграммы называются линейными диаграммами (рис. 3).

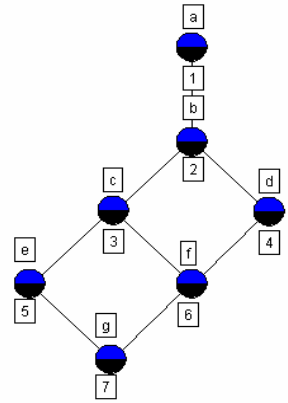
Из вышеизложенного можно сделать вывод, что в системе, реализующей метод формального концептуального анализа, должны быть реализованы функции создания и редактирования контекстов, построения множества всех понятий и линейных диаграмм концептуальных решеток.

| | a | b | c | d | e | f | g |
|---|---|---|---|---|---|---|---|
| 1 | X | | | | | | |
| 2 | X | X | | | | | |
| 3 | X | X | X | | | | |
| 4 | X | X | | X | | | |
| 5 | X | X | X | | X | | |
| 6 | X | X | X | X | | X | |
| 7 | X | X | X | X | X | X | X |

а)




б)




в)

Рис. 3. а) Формальный контекст,
б) Диаграмма Хассе соответствующей концептуальной решетки
в) Линейная диаграмма концептуальной решетки.

Система "Concept explorer" позволяет выполнять, среди прочих, и эти функции.

На рисунке 4 изображен внешний вид системы — главного окна системы, в котором расположен редактор контекста. В редакторе контекста, кроме сохранения и редактирования формальных контекстов, также можно выполнять ряд преобразований над контекстом — устранение повторяющихся объектов с одинаковыми свойствами, и повторяющихся атрибутов, получение по контексту редуцированного контекста — контекста, в котором содержатся только объекты и атрибуты, которые не могут быть получены как комбинация других объектов или атрибутов (т.е., остаются только независимые объекты и атрибуты). Ряд свойств концептуальной решетки может быть определен без ее

построения, с помощью только формального контекста, на основании отношения "стрелки". В системе имеется возможность визуализировать отношение "стрелки", что также можно заметить на рисунке 4.

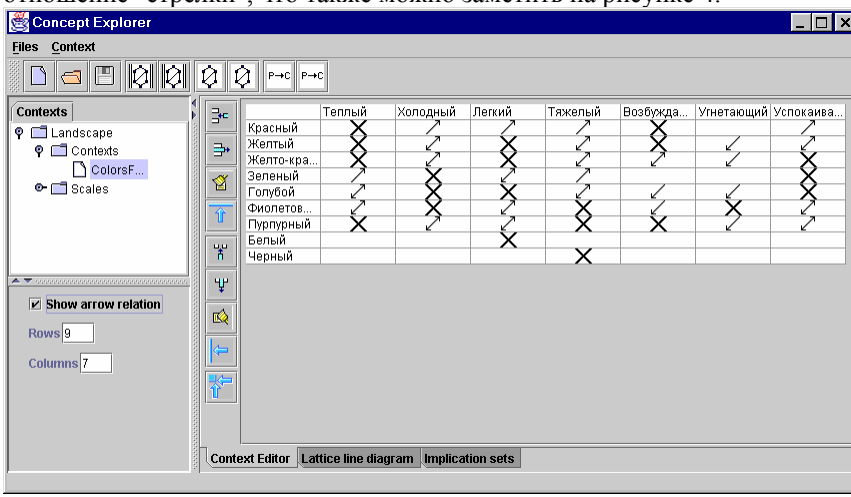



Рис. 4. Внешний вид системы "Concept Explorer"

Для построения концептуальных решеток и порождения множества всех понятий в системе реализованы оригинальные алгоритмы, разработанные автором.

На рисунке 5 изображено окно для работы с линейными диаграммами концептуальных решеток. В нем изображена концептуальная решетка, полученная на основании формального контекста, изображенного в на рисунке 4. Так как концептуальная решетка может быть достаточно сложной, то для ее визуализации применяется эвристический алгоритм, минимизирующий количество пересечений между ребрами решетки.

Кроме того, что ФКА является техникой для классификации и определения понятий по данным, решетку концептов можно использовать для выявления зависимостей между объектами и их свойствами.

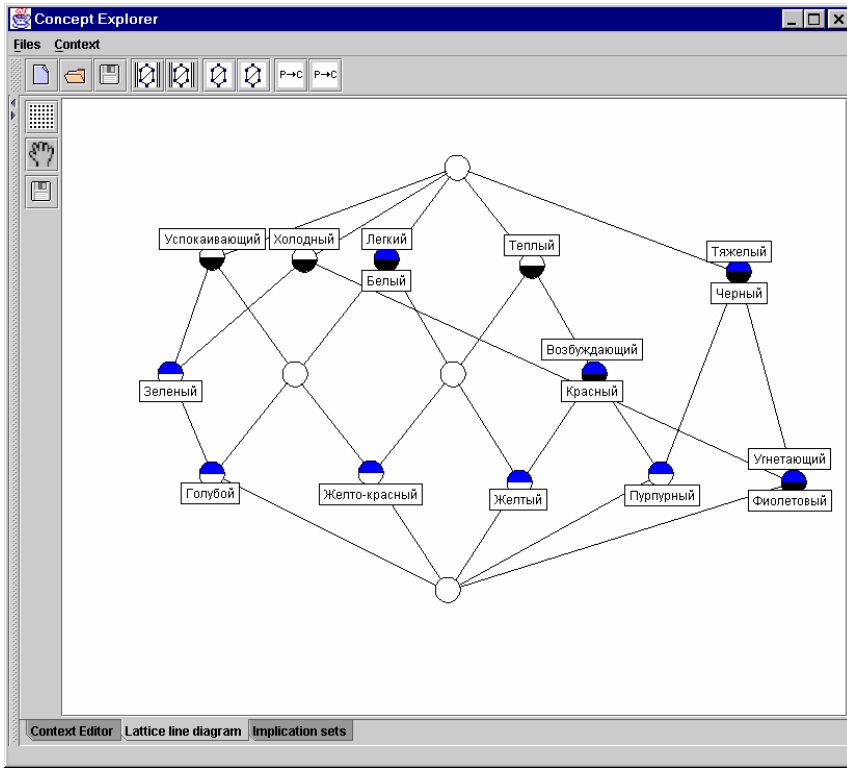



Рис. 5. Внешний вид программы — закладка для работы с линейными диаграммами

Понятие зависимости между атрибутами основано на следующей идее: если для всех объектов контекста, для которых справедливо некоторое свойства $P$, справедливо также некоторое свойство $C$, то является истинной импликация $P \rightarrow C$. Более точно, импликация $P \rightarrow C$ верна для контекста $K=(G, M, I)$, где $P \subseteq M$ и $C \subseteq M$, если для $g \in G$ выполняется следующее требование: если каждый атрибут из посылки $P$ применим к объекту g, то каждый атрибут из заключения $C$ также применим к $g$.

Даже для небольших контекстов множество всех импликаций может быть довольно большим. В связи с этим возникает проблема нахождения базиса всех импликаций контекста. V.Duquenne и J. L.Guigues установили , что базис для всех импликаций, верных в $(G, M, I)$ задается как $\{P \rightarrow P'' \mid P - псевдосодержание\}$, где *псевдосодержание* рекурсивно определяется следующим образом: множество атрибутов $P$ *псевдосодержание* в (G,M,I) , если $P \neq P''$ и $Q'' \subset P$ для *всех*

*псевдосодержаний Q,* таких, что где $Q \subset P$ [Guigues et al., 1986]. Данный базис можно также записать как { P→(P"\P) |P – *псевдосодержание*}

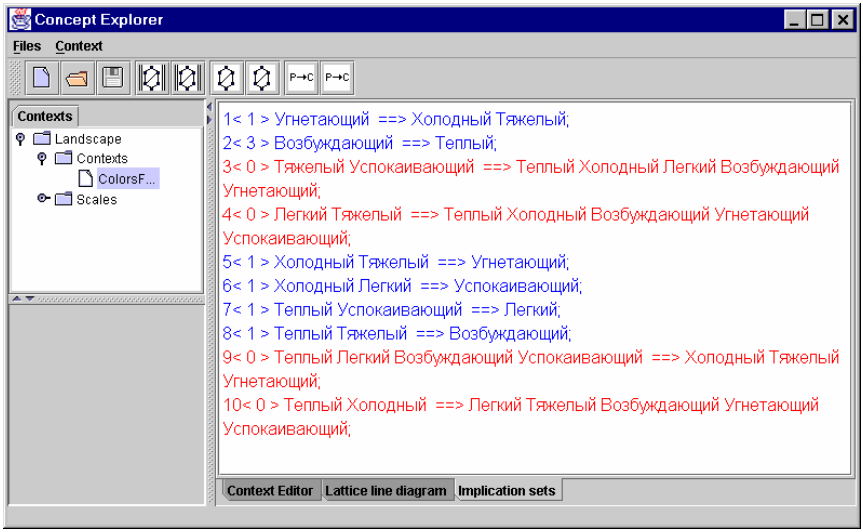



Рис. 6. Внешний вид программы — закладка для отображения базиса импликаций

Смысл базиса, полученного с помощью псевдосодержаний, можно пояснить следующим образом: из некоторой импликации вида P →P", где P≠P", может быть получена импликация, входящая в базис, если замыкание импликаций , следующих из посылки $P^{\bullet} \neq P"$, где $P^{\bullet}=P^{\circ} \cup P^{\circ\circ} \cup \ldots$, и $P^{\circ}=\{Q" \mid Q \subset P \ \& \ Q \rightarrow C\}$, т. е., после того, как к посылке были применены импликации, непосредственно из нее следующие, а к полученному результату были применены импликации, следующие из него и т. д., пока в результате мы не получим заключения.

На основании данного свойства разработан алгоритм, позволяющий находить базис импликаций, справедливых для контекста. На рисунке 6 можно видеть базис импликаций, соответствующий концептуальной решетке, изображенной на рисунке 5.

Работу алгоритма для построения базиса импликаций можно описать следующим образом:

1. Если существуют атрибуты, имеющиеся у всех объектов, то поместить в множество импликаций импликацию $\varnothing \to G'$
2. Множество атрибутов= $G'$
3. Пока не перечислены все множества, которые могут порождать импликации
   3.1. Сгенерировать следующее множество атрибутов P
   3.2. Если P≠P" , то вычислить его замыкание P* относительно ранее найденных импликаций
   3.3. Если P*≠P", то добавить P*→ P" в множество импликаций
4. После вычисления всех импликаций исключить из множества импликаций избыточные.

Для генерации множеств атрибутов используется процесс обхода концептуальной решетки поиском в глубину.

Для того, чтобы не генерировать избыточные множества атрибутов, используется следующее свойство – если мы переходим от одного понятия концептуальной решетки к понятию, лежащему ниже, то импликации могут быть образованы только за счет добавления атрибутов, имеющихся у объектов, лежащих в объеме предыдущего понятия, и не лежащих в объеме текущего понятия.

Программная система "Concept Explorer" может быть использования для исследования данных в области маркетинга, менеджмента, данных психологических и социологических исследований. В частности, система была использована для обработки данных психологического тестирования, которое проводилась на кафедре прикладной математики НТУУ “КПИ”.

## Список литературы

**[Wille, 1982]** Wille R. Restructuring Lattice Theory: an approach based on hierarchies of concept. / Ordered sets / editor I. Rival.— Reidel, Dordrecht-Boston , 1982.

**[Биркгоф 1984]**. Биркгоф Г. Теория решеток. - М.: Наука, 1984.

**[Guigues et al., 1986]** Guigues J.L., Duquenne V. Famillies minimales d’implications informatives resultant d’un tableau de donnees binaires // Math.Sci. Humaines 95.— 1986