ПОСТРОЕНИЕ И ИССЛЕДОВАНИЕ РЕШЕТОК ПОНЯТИЙ

Методические указания к лабораторной работе

1. ЦЕЛЬ РАБОТЫ

Целью работы является приобретение практических навыков работы с моделями знаний в виде решеток понятий.

2. КРАТКАЯ ТЕОРЕТИЧЕСКАЯ СПРАВКА

Решетки понятий применяются в направлении современного анализа текстовых данных, известном как Text Mining – *понимание текста*. Решетка понятий – это модель, отражающая структуру понятий, присущих тексту. Эти понятия строятся *формально* – как подмножества слов, связанных друг с другом отношением принадлежности «объект – атрибут», поэтому данный метод получил название *анализ формальных понятий*.

Анализ формальных понятий имеет строгое математическое обоснование. Он основан на теории решеток [1].

2.1. Упорядоченные множества и решетки

Упорядоченным множеством P называется непустое множество, на котором определено бинарное отношение $\leq$, удовлетворяющее для всех $x, y, z \in P$ следующим условиям:

1. Рефлексивность: $x \leq x$.

2. Антисимметричность. Если $x \leq y$ и $y \leq x$, то $x = y$.

3. Транзитивность. Если $x \leq y$ и $y \leq z$, то $x \leq z$.

Если $x \le y$ и $x \ne y$, то говорят, что x меньше y или y больше x, и пишут $x < y$ или $y > x$.

Примеры упорядоченных множеств:

1. Множество целых положительных чисел, а $x \le y$ означает, что x делит y.
2. Множество всех действительных функций $f(x)$ на отрезке $[a;b]$ и $f \le g$ означает, что $f(x) \le g(x)$ для $\forall x \in [a;b]$.

Цепью называется упорядоченное множество, на котором для любых x, y имеет место $x \le y$ или $y \le x$.

Используя отношение порядка, можно получить графическое представление любого конечного упорядоченного множества *P*. Изобразим каждый элемент множества *P* в виде небольшого кружка, располагая *x* выше *y*, если $x > y$. Соединим x и y отрезком. Полученная фигура называется *диаграммой* упорядоченного множества *P*.

Примеры диаграмм упорядоченного множества приведены на рис. 1:

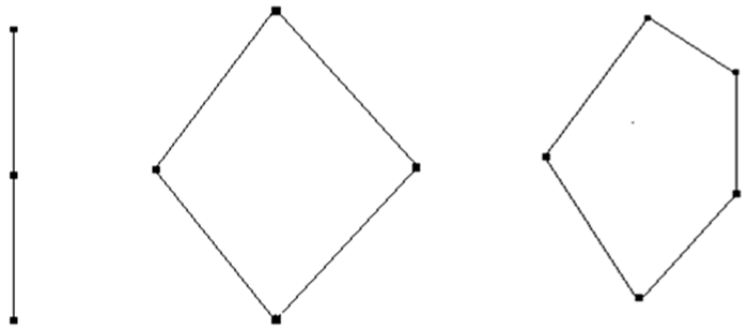

Рис. 1. Диаграммы упорядоченных множеств.

Верхней гранью подмножества *X* в упорядоченном множестве *P* называется элемент *a* из *P*, больший или равный всех *x* из *X*.

Точная верхняя грань подмножества *X* упорядоченного множества *P* – это такая его верхняя грань, которая меньше любой другой его верхней грани. Обозначается символом sup *X* и читается «*супремум X*».

Согласно аксиоме антисимметричности упорядоченного множества, если точная верхняя грань существует, то она единственна.

Понятия нижней грани и точной нижней грани (которая обозначается $\inf X$ и читается «*инфимум*») определяются двойственно. Также, согласно аксиоме антисимметричности упорядоченного множества, если точная нижняя грань X существует, то она единственна.

Решёткой $<L,\leq>$ называется упорядоченное множество L, в котором любые два элемента x и y имеют точную нижнюю грань, обозначаемую $x \wedge y$, и точную верхнюю грань, обозначаемую $x \vee y$ [1]

Примечание. Любая цепь является решёткой, т.к. $x \wedge y$ совпадает с меньшим, а $x \vee y$ с большим из элементов x, y.

Характерные примеры показаны на рис. 2.

Наибольший элемент, то есть элемент, больший или равный каждого элемента упорядоченного множества, он обозначен 1, а наименьший элемент, то есть меньший или равный каждого элемента упорядоченного множества, обозначен 0.

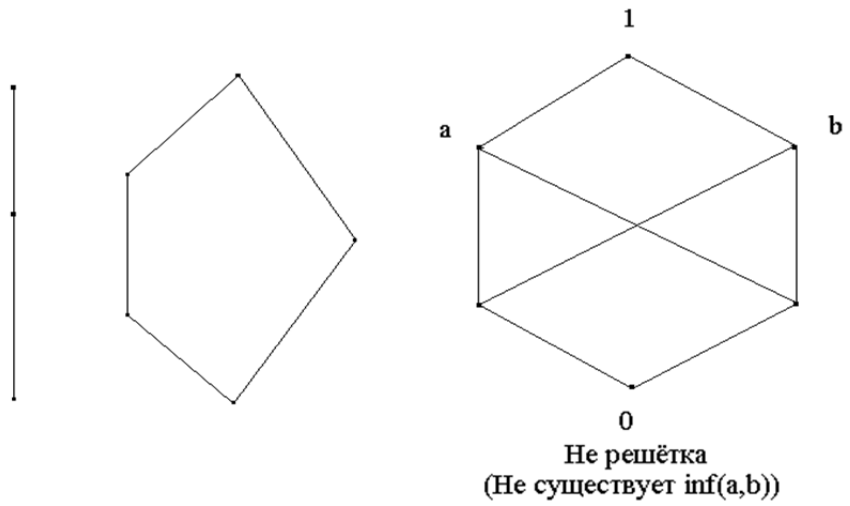

Рис. 2. Примеры диаграмм решеток

На решётке можно рассматривать две бинарные операции:

$a + b = a \vee b$ - сложение и

$a \cdot b = a \wedge b$ - произведение

Эти операции обладают следующими свойствами:

1. $a + a = a$, $a \cdot a + a$ идемпотентность;
2. $a + b = b + a$, $a \cdot b = b \cdot a$ коммутативность;

3. $(a+b)+c=a+(b+c)$, $(a\cdot b)\cdot c=a\cdot(b\cdot c)$ ассоциативность;

4. $a\cdot(a+b)=a$, $a+a\cdot b=a$ законы поглощения.

ТЕОРЕМА 1. Пусть L - множество с двумя бинарными операциями $+,\cdot$, обладающими свойствами (1) – (4). Тогда отношение $a\le b \Leftrightarrow a+b=b$ (или $a\cdot b=a$) является порядком на L, а возникающее упорядоченное множество оказывается решёткой, причём: $a+b=a\vee b$ и $a\cdot b=a\wedge b$.

2.2. Решетки понятий

Рассмотрим два множества: множество *объектов* G и принадлежащих им *атрибутов* M. Эти множества частично упорядочены некоторыми отношениями, которые мы обозначим $\sqsubseteq$ и $\Subset$, соответственно: $G=(G,\sqsubseteq)$, $M=(M,\Subset)$. На данных множествах определяется ***формальный контекст*** $\mathbf{K}=(G,M,I)$, в котором связь между объектами их атрибутами задается отношением $I\subseteq G\times M$, которое представляет собой набор кортежей $<g,m>\in I$.

Формальные понятия. Связи между объектами и атрибутами определяются следующим образом. Для подмножеств $A\subseteq G$ и $B\subseteq M$ объектов и атрибутов задаются отображения (функции) $A':A\to B$ и $B':B\to A$ со следующими свойствами: $A':=\{m\in M\,|\,\forall\, g\in A:<g,m>\in I\}$, $B':=\{g\in G|\forall\, m\in B:<g,m>\in I\}$. Пара множеств ($A$, B), таких, что $A'=B$, $B'=A$ называется ***формальным понятием*** контекста **K**.

Множества A и B замкнуты в силу композиции отображений: $A''=A$, $B''=B$. Множество A образует ***объем*** формального понятия (A, B), а множество B – его ***содержание***. Отношения частичного порядка $\sqsubseteq$, $\Subset$ на множествах G и M индуцируют отношение частичного порядка $\le$ на множестве понятий.

Если для понятий (A_1, B_1) и (A_2, B_2) $\mathrm{A}_1\subseteq\mathrm{A}_2$, что эквивалентно $\mathrm{B}_2\subseteq\mathrm{B}_1$, то ($A_1$, B_1) $\le$ (A_2, B_2). В этом случае логично считать понятие (A_1, B_1) менее общим, чем понятие (A_2, B_2). Формальный контекст имеет представление в виде матрицы инцидентности отношения I, в которой ненулевые элементы обозначают факт принадлежности атрибута $m\in M$ объекту $g\in G$.

На рис. 3 показан пример формального контекста на множествах

$G=$ {блок, устройство, котроллер, объект, сеть, форма, файл} и
$M=$ {генерация, отображение, поддержка, удаление, копирование},

соответствующих объектам и операциям с ними. В матрице контекста понятия (A, B) задаются максимальными по вложению подматрицами с ненулевыми элементами. Так понятием в контексте на рис. 3 будет пара

({объект, сеть, форма}, {удаление, копирование}). Понятия – подстроки и понятия – подстолбцы в матрице контекста также допустимы.

A	B	C	D	E	F
	генерация	отображение	поддержка	удаление	копирование
блок			X		
устройство		X	X		
котроллер					
объект	X	X		X	X
сеть				X	X
форма		X	X	X	X
файл				X	

Рис. 3. Пример формального контекста.

Множество G объектов контекста упорядочено естественным образом: блок более масштабный объект, чем файл. Если приоритет операций из множества атрибутов M не задан, то упорядоченность этого множества искусственная – его элементы просто пронумерованы.

ТЕОРЕМА 2. Частично упорядоченное по вложению объемов множество формальных понятий контекста **K** образует математический объект - *решетку*, которая называется *решетка понятий* [2].

На рисунке 4 изображена решетка понятий контекста на рис. 3. Для иллюстрации контекста и решетки понятий использовано программное средство [3].

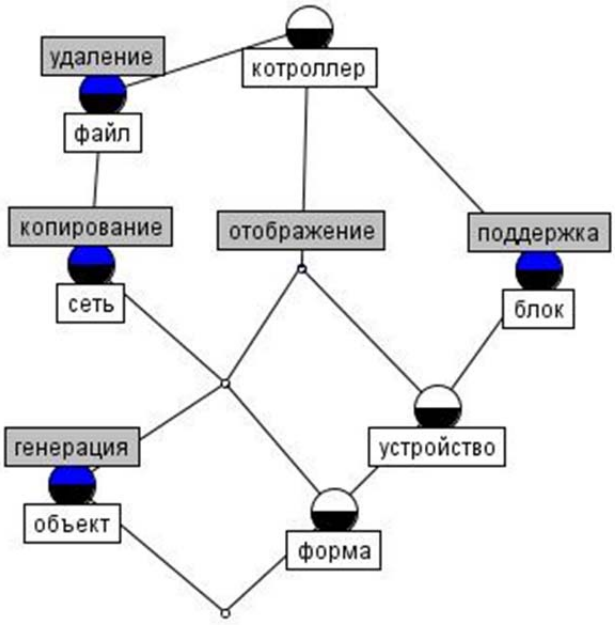

Рис. 4. Решетка понятий контекста на рис. 3.

Решетка понятий, построенная на формальном контексте, является инструментом представления и извлечения знаний из данных контекста. В роли знаний выступают понятия, организованные иерархично. При этом граф решетки понятий не является деревом, что характерно для графов многих концептуальных моделей, а имеет более общую структуру. Это позволяет представлять знания, выражающиеся понятиями, характеризующимися меньшей и большей общностью, меньшими и большими объемом и содержанием.

Инструментом извлечения знаний на решетках понятий являются методы Data Mining, использующие модели в виде:

- *импликаций*,
- *функциональных зависимостей*,
- *ассоциативных правил*.

Импликации $X \to Y$ на подмножествах признаков $X, Y \subseteq M$ имеют место, если $X^{'} \subseteq Y^{'}$, т.е. каждый объект, обладающий всеми признаками из множества *X*, также обладает всеми признаками из множества *Y*.

В решетке на рис. 4 имеем импликации:

копирование → удаление;
генерация → {отображение, удаление, копирование};
{отображение, удаление} → копирование;
{поддержка, удаление} → {отображение, копирование}.

На множестве импликаций решетки понятий строится система навигации в ней, позволяющая находить частные и общие понятия для заданного входа – узла решетки. В этом состоит большое преимущество решеток понятий как концептуальных моделей.

2.3. Построение решеток понятий на текстовых данных.

Исходные данные для построения решеток понятий содержит формальный контекст. Формальный контекст может быть построен на данных различной природы. В системах Text Mining, использующих решетки понятий, формальные контексты строятся на текстовых данных.

Построение формального контекста по тексту – это очень сложная задача. Например, слова, представленные в контексте на рис. 3, могут принадлежать тексту, в котором в произвольной форме рассказывается про котроллеры, блоки и т.д., и действия с ними. Этот текст может быть инструкцией по работе с некоторыми устройствами и тогда понимание такого текста – это проверка его корректности посредством выявления системы понятий. Построенная на таком тексте решетка понятий позволяет определить, как соотносятся между собой понятие, какое из них более общее, а какое менее общее.

Для построения формального контекста на текстовых данных необходимо выполнить следующее:

- определить подмножества слов, связанных отношениями «объект – атрибут»;
- в каждом из найденных подмножеств установить отношение частичного порядка.

Для определения слов, связанных отношениями «объект – атрибут», можно воспользоваться семантической моделью текста, известной как *концептуальный граф*.

Концептуальный граф признан в качестве одной из семантических моделей, применяемых для анализа текстов. Вместе с концептуальными решетками концептуальные графы относятся к *концептуальным структурам*, которые являются одним из формальных представлений знаний.

Концептуальный граф - это двудольный направленный граф, состоящий из двух типов узлов: *концептов* и *концептуальных отношений*.

На рис. 5 показан пример: концептуальный граф фразы «главный датчик мотора автомобиля».

Рис. 5. Концептуальный граф фразы «главный датчик мотора автомобиля»

На графе концепты изображены в виде прямоугольников, а концептуальные отношения – в виде эллипсов. Отношения «атрибут» и

«генитив» обозначают известные из лингвистики связи между словами – частями речи. Оба эти отношения являются отношениями принадлежности и могут быть непосредственно использованы для построения формального контекста.

Отношение частичного порядка на множестве слов текста может быть интерпретировано несколькими способами. Простейшим, естественным способом является учет порядка следования слов в тексте. Однако, не всегда данный способ отражает реальную, смысловую упорядоченность фрагментов текста. Для построения формального контекста допустимо принять порядок следования слов как способ упорядочивания слов, обозначающих объекты. Слова - атрибуты можно упорядочивать, применяя нумерацию.

3. ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ РАБОТЫ

В работе используется свободно распространяемое программное средство Concept Explorer [3], представляющее собой систему построения и анализа решеток понятий.

3.1. Краткая инструкция по применению Concept Explorer.

Система Concept Explorer работает на платформе Java. Это означает, что на компьютере, где запускается Concept Explorer, должна быть установлена виртуальная машина Java. Установить ее можно, используя ресурс [4].

После установки виртуальной машины Java Concept Explorer запускается инициализацией файла conexp.jar или файла conexp.bat.

Главное окно системы Concept Explorer показано на рис. 6. Система имеет интуитивно понятный интерфейс, элементы которого показаны на рисунке.

Система Concept Explorer имеет следующие функции:

- построение и редактирование контекстов;
- построение решеток понятий;
- построение базиса импликаций, допускаемых решеткой понятий;
- построение базиса ассоциативных правил, допускаемых решеткой понятий.

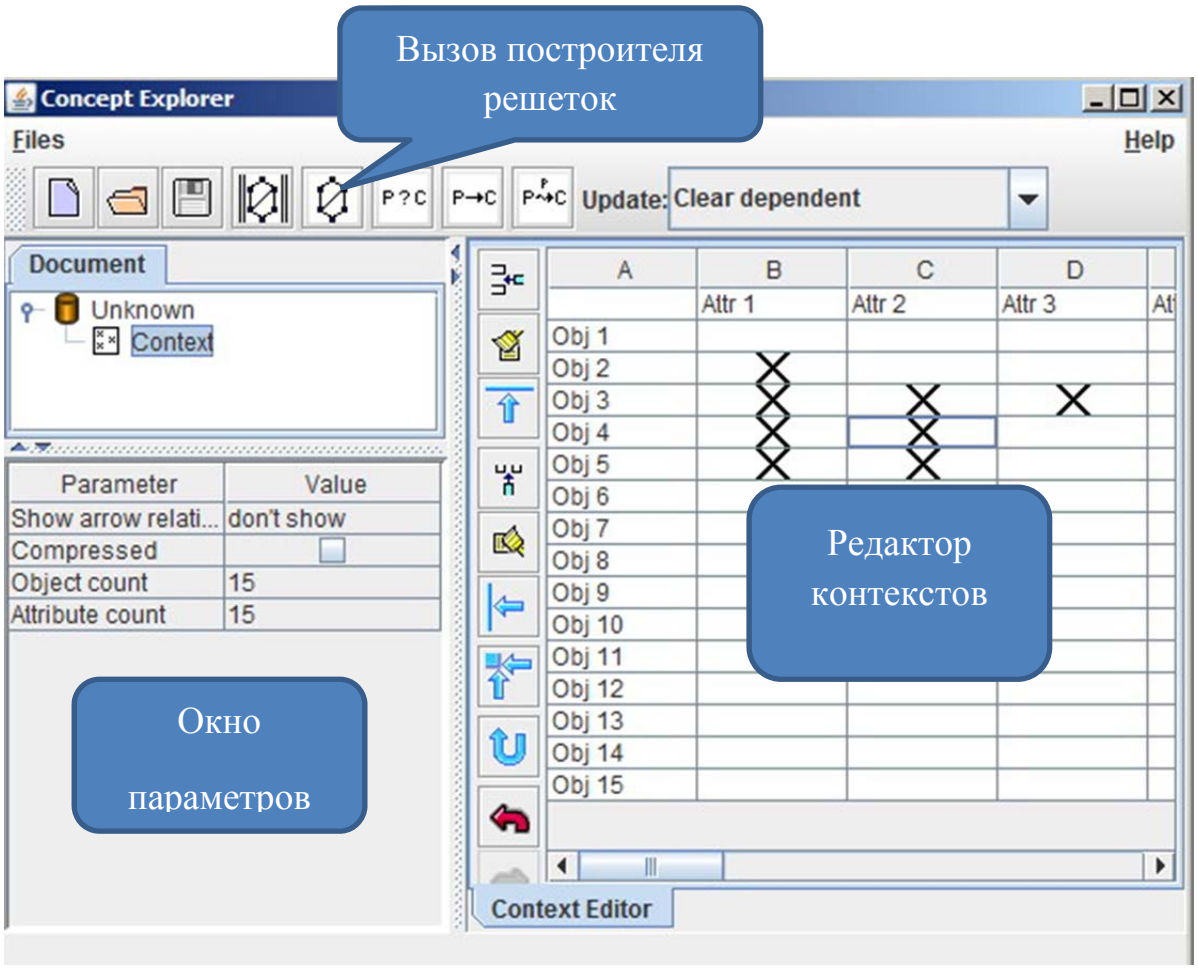

Рис. 6. Главное окно системы Concept Explorer.

Построение и редактирование контекстов выполняется в редакторе контекстов. Система стартует, предъявив пустой контекст.

Строки и столбцы матрицы контекста размечены как пронумерованные объекты (Obj*) и атрибуты (Attr*). Задание элементов контекста выполняется двойным кликированием клетки контекста, что приводит к появлению знака «X» в клетке. Этим устанавливается факт принадлежности атрибута Attr* объекту Obj*.

3.2. Применение концептуальных графов для построения формального контекста.

Для построения формального контекста можно использовать концептуальные графы, строящиеся для каждого предложения текста, как показано выше, в разделе 2.3.

Для построения концептуальных графов можно использовать программное средство прямого доступа по адресу: http://lis.tula.ru:8888/

4. ПОРЯДОК ВЫПОЛНЕНИЯ РАБОТЫ.

4.1. Построить формальный контекст для заданного текста.

4.2. Построить решетку понятий на формальном контексте.

4.3. Построить систему импликаций, допускаемых решеткой понятий.

4.4. Объяснить полученные импликации.

КОНТРОЛЬНЫЕ ВОПРОСЫ.

1. Доказать эквивалентность обоих определений решетки
2. Всегда ли решетка задает отношение включения? Приведите примеры.
3. Является ли понятием подматрица – строка (столбец) в матрице формального контекста?

СПИСОК ЛИТЕРАТУРЫ

1. Биркгоф Г. Теория решеток. М.: Наука, 1984. 284 с.
2. Ganter, Bernhard; Stumme, Gerd; Wille, Rudolf, eds., *Formal Concept Analysis: Foundations and Applications*, Lecture Notes in Artificial Intelligence, No. 3626, Springer-Verlag, 2005.
3. Электронный ресурс: http://conexp.sourceforge.net/download.html.
4. Электронный ресурс: https://www.java.com/ru/
5. Michael Bogatyrev and Alexey Kolosoff. Using Conceptual Graphs for Text Mining in Technical Support Services. Pattern Recognition and Machine Intelligence. - Lecture Notes in Computer Science, 2011, Volume 6744/2011, p.p. 466-471. Springer-Verlag. Heidelberg. 2011.