

УДК ????

Алгоритмы рисования линейных диаграмм

С. А. Евтушенко¹

Анализ формальных понятий - мощный метод анализа данных, который неоднократно успешно применялся на практике. Основным средством визуализации зависимостей в нем является линейная диаграмма решетки формальных понятий. В статье проводится обзор существующих методов рисования линейных диаграмм, а также сравниваются результаты, получаемые при применении наиболее распространенного метода разложения решетки на цепи и методов взаимодействия сил к различным видам диаграмм. Сравнимые алгоритмы реализованы в программной системе Concept Explorer.

Один рисунок стоит ста слов

Введение

Визуализация информации является важной частью интеллектуального анализа данных (ИАД). Большинство людей лучше всего воспринимают графическую информацию. Визуальное представление дает возможность наглядно представить информацию и часто позволяет с первого взгляда выявить закономерности, которые иначе можно найти лишь с помощью трудоемкого анализа.

При решении многих проблем возникает задача визуализации частично упорядоченных множеств, а также их частного случая – решеток. Особо актуальна данная проблема при использовании одного из наиболее мощных методов ИАД – анализа формальных понятий.

Анализ формальных понятий и линейные диаграммы

Анализ формальных понятий (Formal Concept Analysis, FCA) был предложен Wille в 1981 году [Wille, 1982] и активно развивается и сегодня. Основой данного метода является факт, что по любому бинарному отношению можно однозначно построить полную решетку.

FCA применяется для анализа качественной информации. Данные представляются с помощью формального контекста – таблицы, строкам

¹ Украина, 03056, Киев, пр Перемоги, 37, НТУУ “КПИ”, ФПМ, sye@mail.ru

которой соответствуют объекты, столбцам – атрибуты, и если объект обладает свойством, то на пересечении столбца и строки стоит 1.

Формальное понятие (концепт) состоит из объема и содержания. Содержание – это множество свойств, описывающих понятие. В объем входят все объекты из контекста, которые имеют все свойства из содержания. При этом содержание должно быть максимальным, т. е. включать все общие свойства объектов из объема понятия. Математически формальное понятие задается с помощью соответствий Галуа [Биркгоф, 1984].

Между понятиями устанавливается отношение предшествования: понятие $(A_1, B_1) \leq (A_2, B_2)$, если $A_1 \subseteq A_2$ и $B_1 \supseteq B_2$, где A_1, A_2 – объемы понятий 1 и 2, а B_1, B_2 – соответственно их содержания.

Множество всех понятий контекста образует полную решетку, которую называют *решеткой концептов*. Решетка концептов содержит всю информацию о взаимосвязях, существующих между атрибутами в контексте, по которому она была построена.

Традиционно для визуализации частично - упорядоченных множеств используются диаграммы Хассе. В FCA применяется их разновидность, в которой используется сокращенная пометка - каждый объект и 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. На рис. 1 представлены формальный контекст и линейная диаграмма решетки понятий с сокращенной пометкой.

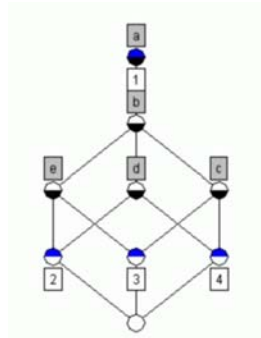
Требования, предъявляемые к изображениям линейных диаграмм

Хорошая линейная диаграмма должна быть «прозрачной, легко читаемой и облегчающей интерпретацию представленных данных» [Ganter et al., 1999]. Однако способы достижения этого зависят от целей интерпретации и от структуры решетки.

При рисовании линейной диаграммы обязательным является требование о том, что все подконцепты какого-либо концепта должны быть расположены ниже его.

	a	b	c	d	e
1	×				
2	×	×		×	×
3	×	×	×		×
4	×	×	×	×	

а)



б)

Рис. 1. а) формальный контекст б) линейная диаграмма решетки понятий

Желательным является выполнение следующих условий:

- ребра должны изображаться в виде прямых линий;
- две вершины не должны быть расположены в одной точке;
- количество пересечений между ребрами решетки должно быть как можно меньше;
- ребро не должно пересекать вершину, не являющуюся ее концом;
- должна быть наглядно представлена структура решетки;
- использование как можно меньшего количества различных направлений;
- максимизация количества параллельных линий.

Дополнительным требованием, которое можно предъявить к алгоритмам рисования линейных диаграмм, является симметричность изображения линейных диаграмм решеток, построенных по контексту и соответствующему ему транспонированному контексту (так как между этими диаграммами можно установить изоморфизм).

Рассмотрим теперь существующие подходы к изображению графов и линейных диаграмм.

Общие техники рисования графов

Большая группа методов рисования графов может быть представлена в обобщенном виде следующим образом:

1. рассчитать первоначальное положение графа;
2. пока возможно, производить оптимизацию выкладки графа;
3. произвести конечную подстройку положений отдельных вершин.

В рамках данной схеме могут быть описаны как методы для выкладки иерархических графов, так и методы «взаимодействия сил».

Традиционные методы рисования иерархических графов.

К данной группе методов принадлежат методы, основанные на схеме, впервые предложенной Sugiyama [Sugiyama et al., 1981]. Схема ориентирована в первую очередь на удовлетворение следующих эстетических критериев: минимизация пересечения между ребрами и сокращение длины ребер.

Алгоритм состоит из четырех этапов:

1. расчет начальных позиций вершин и преобразование графа в 2-слойный, у которого являются смежными вершины только двух соседних уровней за счет добавления фиктивных вершин;
2. минимизация пересечений с использованием барицентрической или медианной эвристики;
3. расчет координат вершин графа;
4. непосредственное рисование графа.

Методы взаимодействия сил

Широко распространенной группой методов для рисования графов общего вида являются методы, основанные на моделировании взаимодействия сил (*force-directed layouts*).

Первоначальная модель была предложена Р. Eades в [Eades, 1984].

В модели обычно задаются силы «отталкивания», действующие между всеми вершинами, и силы «притяжения», действующие между вершинами, соединенными ребрами. Также могут задаваться силы, действующие между ребрами и вершинами, с целью избежания пересечения ребер и вершин.

Алгоритмы «взаимодействия сил» задаются моделью сил, действующей между вершинами и ребрами графа, и процедурой оптимизации, используемой для нахождения положения, в котором достигается локальный минимум энергии.

Преимуществами данных методов являются выявление симметрий, существующих в графе, и легкая переносимость на случай трех и более измерений.

Существующие подходы к построению линейных диаграмм

Аддитивные линейные диаграммы

Данный метод был предложен Ganter и Wille [Ganter et al., 1999]. Для решетки L вводится «множество представления» X , и функция $rep: L \rightarrow \rho(X)$, которая ставит в соответствие каждому элементу решетки

некоторое подмножество множества X . При этом, если $x \leq y$, то $rep(x) \subseteq rep(y)$. Каждому элементу множества представления ставится в соответствие вектор в \mathbf{R}^2 , и позиция понятия вычисляется следующим образом: $pos p = n + \sum_{x \in rep(p)} vec(x)$, где n – вектор, который может быть выбран произвольно, для того, чтобы сдвинуть всю диаграмму.

Преимуществом представления в виде аддитивных линейных диаграмм является возможность легкой манипуляции изображением линейной диаграммы (изменение вектора, соответствующего одному элементу $x \in X$, приводит к изменению положения всех элементов p , для которых $x \in rep(p)$).

Стандартным методом построения аддитивных линейных диаграмм является метод разложения на цепи (chain-decomposition).

Метод разложения на цепи

Выделяя цепи в решетке понятий, можно определить минимальную сетку, в которую возможно погрузить решетку – если в решетке n независимых цепей, то погружение может быть осуществлено в n -мерную сетку, образованную произведением цепей.

Один из методов разложения решетки на цепи основан на теореме о том, что \vee размерность конечной решетки L (минимальное количество цепей, в произведение которых можно \vee вложить решетку) равна ширине множества всех \wedge несводимых элементов L .

В качестве множества представления используется множество всех несводимых признаков контекста (это \wedge несводимые элементы решетки) [Wille, 1989a].

Сначала находятся все «несводимые» (которые не могут быть получены как пересечение объемов других признаков) признаки $m \in M$. Затем вычисляется отношение частичного порядка на признаках $m_1 \leq m_2 \Leftrightarrow \{m_1\}' \subseteq \{m_2\}'$, и получается упорядоченное множество, изоморфное множеству \wedge несводимых элементов решетки L . Оно разлагается в минимальное количество цепей посредством версии алгоритма Форда-Фалкерсона.

После этого строится аддитивная линейная диаграмма. Также в данном алгоритме могут использоваться и другие множества представлений (например, несводимых объектов, которым соответствует множество \wedge несводимых элементов решетки).

Эвристики, используемые при построении линейных диаграмм

Правило параллелограмма. М. Skorsky предложил использовать для построения решеток правило параллелограмма: каждая четверка элементов a, b, c, d , таких, что $a\{b, a\{c, b\{d, c\{d,$, где $\{$ - отношение покрытия, изображается как параллелограмм, и показал, что локально-

дистрибутивная решетка может быть представлена таким образом [Skorsky, 1989].

Геометрическая эвристика. Наилучший известный способ построения концептуальных решеток на сегодня дает геометрическая эвристика, описанная в [Stumme et al., 1993], [Ganter et al., 1999]. Хотя геометрическое представление, используемое для построения линейной диаграммы, и может быть вычислено автоматически, насколько известно автору, программы, автоматически строящей линейную диаграмму по геометрическому представлению, не существует.

Гибридный метод

В [Cole, 2000] был предложен гибридный метод построения линейных диаграмм, основанный на использовании алгоритма удовлетворения ограничений (*constraint-satisfaction*) для перебора различных вариантов представления решеток и выбора наилучшего, который приводит к построению линейных диаграмм достаточно высокого качества.

Сравнение алгоритмов

Автором в рамках системы «CONCEPT EXPLORER» [Евтушенко, 2000] были реализованы следующие алгоритмы рисования решеток концептов:

1. Алгоритм, основанный на схеме Sugiyama, в котором применялась медианная эвристика для минимизации пересечений.
2. Алгоритм построения аддитивных линейных диаграмм на основании техники “chain-decomposition”.
3. Алгоритм, основанный на модели взаимодействия сил, на основании алгоритма, предложенного R. Freeze
4. Алгоритм, на основании модели взаимодействия сил для общих графов.

Для алгоритмов 3 и 4 (моделирования взаимодействия сил) использовались 3-х мерные варианты. В них применялись одинаковые формулы для сил «притяжения» и «отталкивания», однако по разному выбирались вершины, влияние которых учитывалось при расчете сил.

В оригинальной модели R. Freeze [Freeze, 1996] при расчете сил «притяжения» учитывалось влияние вершин, лежащих в фильтре текущего элемента, а при расчете сил «отталкивания» использовались несравнимые с текущим элементы решетки, имеющие ранг больший или равный текущему. При этом учитывалось влияние не всех элементов с рангом, равным текущему, а только элементов, получивших большие номера в результате топологической сортировки элементов решетки.

Ранг элемента равен высоте элемента плюс высота решетки - длина кратчайшего пути из единицы решетки до элемента.

Нами было внесено два изменения в модель.

1. У нас попеременно, в зависимости от итерации, при расчете сил «притяжения» учитывается влияние элементов, лежащих в фильтре либо идеале, и для силы «отталкивания» соответственно, элементы, несравнимые с текущим и имеющие ранг больший или равный ему, либо, соответственно, меньший либо равный ему.
2. Учитывается влияние всех элементов, имеющих одинаковый ранг с текущим.

В алгоритме 4 при расчете сил притяжения, действующих на вершину, учитывалось влияние вершин, непосредственно соединенных с текущей, а при расчете сил «отталкивания» - всех вершин решетки, лежащих не дальше определенного расстояния.

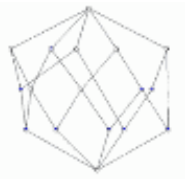


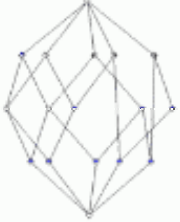



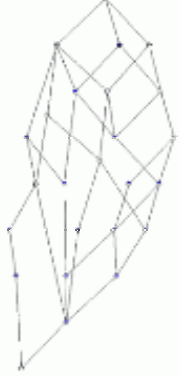
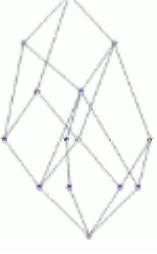
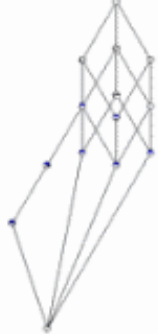


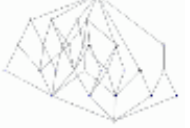

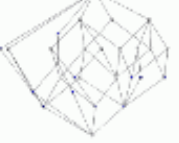
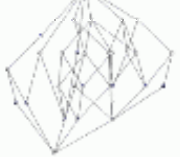
Тестовый набор решеток

Для сравнения были выбраны несколько решеток, которые являются типичными в теории решеток, а также решетки, которые были получены в результате исследований либо приводились в качестве примера в статьях по анализу формальных понятий.

В тестовый набор вошли: решетки B4 – булевская решетка, порождаемая сочетаниями 4 двuzначных атрибутов (1); FD3 – свободная дистрибутивная решетка, порождаемая 3 элементами (2); ID4 – интерординальная шкала для 4 интервалов (3); решетка треугольников из [Ganter et al., 1999] (4); решетка свойств цветов [Евтушенко, 2000] (5), решетка, описывающая взаимоотношения свойств конечных решеток [Wille, 1989b] (6); решетки, полученные в результате обработки репертуарных решеток (отношение студента к предметам [Таран и др., 2000] (7) и оценки большим знакомых [Burmeister, 1999] (8)). Результаты сравнений приведены в таблице 1.

Таблица 1

	Минимальные пересечения	Разложение на цепи	Схема Freeze	Общий метод взаимодействия сил
1				
2				
3				
4				

5				
6				
7				
8				

Выводы и направления дальнейших исследований

Наилучшие результаты при рисовании линейных диаграмм в проведенном сравнении продемонстрировали методы разложения на цепи и метод взаимодействия сил с использованием модели R. Freeze.

Метод разложения на цепи дает хорошие результаты для дистрибутивных решеток (см. решетки 1, 2). Он показал самое высокое быстродействие из сравниваемых методов. Однако при рисовании недистрибутивных решеток (см. решетки 3-8) выявляется его основной недостаток - нулевой элемент решетки может находиться на относительно большом удалении от других элементов решетки. Это связано с тем, что содержание нулевого элемента включает все атрибуты, и, следовательно, его координаты вычисляются как сумма векторов, приписанных всем атрибутам, а содержания других элементов включают меньшее количество атрибутов.

Метод построения линейной диаграммы с минимальным количеством пересечений приводит, в основном, к достаточно наглядным изображениям. При помощи этого метода было получено наиболее наглядное изображение решетки 8, которая вызвала наибольшие затруднения при рисовании. Однако ему присущ и ряд недостатков - длительное время построения диаграммы для больших решеток, возможны пересечения ребер и вершин, не принадлежащих этим ребрам (см. решетку 4), не выявляется структура решетки, симметричные части, и не удовлетворяется критерий максимизации параллельности линий.

Методы взаимодействия сил позволяют во многих случаях получить достаточно хорошие диаграммы. Они дают возможность выявить наличие симметричных структур. При изображении одной цепи вершины располагаются не на одной прямой (см. решетки 3, 4, 7), так как при расчете положения вершин для каждой вершины учитывается положение соседних вершин. Поэтому не выполняется критерий максимизации параллельности линий. Этим методам свойственно достаточно большое время работы и необходимость ручной подстройки параметров для получения наиболее наглядных изображений. Частично эти недостатки могут быть компенсированы за счет использования более эффективных процедур оптимизации и изменения моделей сил, используемых в алгоритме.

На основании проведенных экспериментов, наиболее перспективным представляется развитие гибридных методов. В них должна использоваться структурная информация о решетке, которую можно получить на основании анализа контекста (подобно методу разложения на цепи) и алгоритмы оптимизации, используемые в методе взаимодействия сил.

Направлениями дальнейшей работы является увеличение количества сравниваемых алгоритмов рисования диаграмм, разработка и реализация улучшенных модификаций методов с учетом результатов сравнений, а также расширение тестового набора решеток.

Список литературы

- [Биркгоф, 1984] Биркгоф Г. *Теория решеток*. - М.: Наука, 1984.
- [Евтушенко, 2000] Евтушенко С.А. Система анализа данных "CONCEPT EXPLORER"// КИИ-2000 / Труды конференции, Москва, изд. Физ. Мат. Литературы, 2000
- [Таран и др., 2000] Таран Т.А., Гатин А.Р., Ушинкина Е.С. Исследование когнитивного пространства личности // 4-я международная школа-семинар по искусственному интеллекту для студентов и аспирантов (Браславская школа - 2000) / Сб. научн. трудов. - Минск: БГУИР, 2000.
- [Burmeister, 1999] Burmeister P. Formal Concept Analysis with ConImp: Introduction to the Basic Features
- [Cole, 2000] Cole R. Automated Layout of Concept Lattice Using Layer Diagrams and Additive Diagrams.
- [Eades, 1984] Eades P. A heuristic for graph drawing //Congressus Numerantium, 42:, 1984
- [Freeze, 1996] Freeze R. LatDraw . <http://www.math.hawaii.edu/~ralph/LatDraw>
- [Ganter et al., 1999]Ganter B., Wille R. *Formal Concept Analysis: mathematical foundations*. Springer-Verlag, Berlin Heidelberg New York. - 1999
- [Skorsky, 1989] Skorsky M. How to draw concept lattices with parallelograms// *Klassifikation und ordnung* (R. Wille, eds) Frankfurt, INDEKS-Verlag, 1989
- [Stumme et al., 1993] Stumme G., Wille R. A geometrical heuristic for drawing concept lattice// R.Tamassia and I. G. Tollis (eds.) *Graph Drawing*, Berlin-Heidelberg-New York, Springer, 1993
- [Sugiyama et al., 1981] Sugiyama K., Tagawa S., Toda M. Methods for visual understanding of hierarchical systems //IEEE Transactions on Systems, Man and Cybernetics, SMC-11(2):109-125, February 1981.
- [Wille, 1982] Wille R. Restructuring Lattice Theory: an approach based on hierarchies of concept/*Ordered sets* (I. Rival, ed.) Reidel, Dordrecht-Boston, 1982.
- [Wille, 1989a] Wille R. Lattices in data analysis: how to draw them with computer// *Algorithms and Order* (I. Rival, ed.) Kluwer, Dordrecht-Boston. 1989
- [Wille, 1989b] Wille R. Knowledge acquisition by methods of formal concept analysis/ *Data Analysis, learning symbolic and numeric knowledge* (E.Diday, ed.) New York-Budapest, Nova Science Publisher, 1989