Министерство образования и науки Украины

Национальный технический университет Украины
"Киевский политехнический институт"

А.А. Снарский, Д.В. Ландэ

МОДЕЛИРОВАНИЕ СЛОЖНЫХ СЕТЕЙ

Учебное пособие

Киев – 2015

УДК 004.5
ББК 22.18, 32.81, 60.54
С 95

Рекомендовано к печати ученым советом учебно-научного комплекса «Институт прикладного системного анализа» НТУУ «КПИ» (протокол № 8 от 30 сентября 2014 г.), ученым советом Института проблем регистрации информации НАН Украины (протокол № 6 от 20 января 2015 г.)

Рецензенты:
д.т.н., проф. А.Г. Додонов;
д.ф.-м.н., проф., член-корр. НАН Украины Б.И. Лев

С 95 **Снарский А.А., Ландэ Д.В. Моделирование сложных сетей: учебное пособие**. — К.: НТУУ «КПИ», 2015. — 212 с.

В учебном пособии рассматриваются базовые вопросы теории сложных сетей: характеристики, алгоритмы, модели, задачи поиска, ранжирования, а также приводятся сведения, необходимые для математического и компьютерного моделирования и анализа сложных сетей.

Теория сложных сетей – это комплексное научное направление, находящееся на стыке таких наук, как дискретная математика, теория графов, теория алгоритмов, нелинейная динамика, теория фазовых переходов, перколяции и др. Поэтому для успешного моделирования сложных сетей необходимы базовые сведения из всех этих областей, которые изложены в данном учебном пособии.

Издание предназначено для студентов и аспирантов высших учебных заведений, инженеров и научных сотрудников, работающих в областях системного анализа и прикладной математики.

УДК 004.5
ББК 22.18, 32.81, 60.54
ISBN _____
Тираж 300 экз.

Оглавление

Введение

Сейчас вряд ли можно возразить против того, что родилась новая парадигма, или более казенным языком, новое научное направление – теория сложных сетей. Вспомним историю 30-летней давности. Тогда, каждый, кто прочитал, пролистал или хотя бы просмотрел картинки и подписи к ним в книге Б. Мандельброта «Фрактальная геометрия природы» обнаруживал, что куда не глянь – одни фракталы. И деревья, и кусты, и профили гор и облаков, и знакомая еще со школы траектория броуновской частицы и т.д. и т.п. Термин фрактал стал модным, вошел (иногда и «не по делу») в терминологию статей и книг по самым разным областям науки – физики, химии, биологии, экономики, медицины, … И это при том, что первые фрактальные объекты изучались уже в XIX веке (кривая Пеано, множество Кантора и др.). Сейчас интерес к фракталам остыл, частота употребления термина фрактал в книгах, падает. Новая область, активно развивающаяся – это так называемые «сложные сети». Теперь, куда не глянешь, видишь сети – социальные сети, сети дружбы, соавторства в научных публикациях, сексуальных отношений, бизнес-связей, публикаций в СМИ, совместного употребления слов в текстах, метаболизмов (обменных процессов), кровеносных сосудов, транспорта, наконец, (см. рис.) Интернет и WWW. Огромное поле деятельности, много результатов, новые разделы в научных журналах и новые журналы.

Бенуа Мандельброт (1924 – 2010)

Мандельброт Б. Фрактальная геометрия природы = The Fractal Geometry of Nature. – М.: Институт компьютерн. исследований, 2002. – С. 656.

«Где вы работаете?» – «Я нуль-физик». Изумленно-восхищенный взгляд. «Слушайте, расскажите, пожалуйста, что такое – нуль-физика? Я никак не могу понять». – «Я тоже». *А. и Б. Стругацкие «Далекая Радуга».*

При первом знакомстве с книгами и обзорами по теории сложных сетей возникает вопрос. Не просто ли модное название эта теория сложных сетей? Экономическая составляющая модного термина хорошо известна, написал «нанофильтр» и получил мегагрант. Чем отличается сеть, пусть и сложная, от графа? А теория сложных сетей от теории графов – теории, начало которой положил еще Эйлер своей знаменитой задачей о Кенигсбергских мостах.

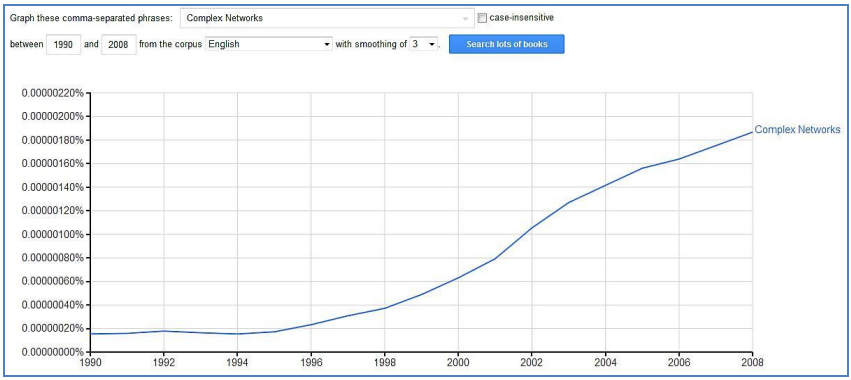


Рис. Динамика употребления словосочетания Complex Networks, полученная с помощью сервиса Google Ngram Viewer (https://books.google.com/ngrams)

Формально, любая сеть это граф. И, опять же, формально, теория, которая изучает свойства сложных сетей, должна была бы быть названа теорией графов. Приведем следующий контрпример. Классическая механика изучает, в том числе, движение материальных точек. Идеальный газ – это как раз движущиеся материальные точки. Однако свойства газа изучает другой раздел физики – статистическая механика.

У нее свои методы, термины, приемы. А «обычная» классическая механика с задачей описания газа справиться не может, слишком много частиц даже в небольшом «кусочке» газа. И пришлось Максвеллу, Больцману, Гиббсу и многим другим создать новую науку и ввести новые понятия, например, температуру и энтропию, которые не нужны и не вводятся принципиально в классической механике.

Аналогично и взаимоотношение теорий графов и сложных сетей. Как основой классической статистической физики является классическая механика, так основой теории сложных сетей является теория графов. При этом, теория графов может получать содержательные утверждения как правило для графов (сетей) небольшого размера или специальной структуры. В то время как теория сложных сетей имеет дело с большим количеством, как правило, случайным образом соединенных узлов. В связи с этим, с одной стороны, многие распространенные вопросы теории графов не представляют интереса для теории сложных сетей. Например, такой вопрос – планарен ли данный граф? Для случайного графа с большим числом узлов и связей ответ очевиден (и не представляет интереса) – нет. С другой стороны многие очень важные понятия в теории сложных сетей для графа с небольшим числом узлов и связей либо не представляют интереса, либо их просто трудно содержательно сформулировать. Например, такая важная характеристика в теории сложных сетей как функция распределения по степени узлов может быть подсчитана для любого, в т.ч. и малого количества узлов. Однако в силу своего вероятностного содержания это понятие оказывается полезным только в случае большого числа узлов и связей.

Dorogovtsev S.N., Mendes J.F.F. Evolution of Networks: From Biological Networks to the Internet and WWW. – Oxford, USA: Oxford University Press, 2003. – P. 280.

Mark Newman, Albert-Laszlo Barabasi, Duncan J. Watts. The Structure and Dynamics of Networks: (Princeton Studies in Complexity). – Princeton, USA: Princeton University Press, 2006. – P. 624.

Чем же занимается теория сложных сетей? Перечислим некоторые проблемы и задачи.

Во-первых, исследованием стандартных характеристик графов для сложных сетей разной природы – случайных графов, безмасштабных сетей, сетей малого мира и т.п.

Во-вторых, определением и изучением новых характеристик сложных сетей, таких, например, как средний

минимальный путь, посредничество, коэффициент кластеризации.

В-третьих, изучением различных «физических» процессов на сложных сетях – диффузии, эпидемических процессов, различных потоков (информации, электрического тока,…). В конце концов, знаменитый алгоритм PageRank рассматривает блуждание по связям (гиперссылкам) в сложной сети WWW.

В-четвертых, есть очень важное в прикладном отношении, направление – методы восстановления, защиты и уничтожения сетей. Такой вопрос - сколько узлов (связей) нужно «убить» для того, что бы, например, разрушился «гигантский кластер» или что бы значительно увеличился минимальный средний путь? Т.е., как говорят сисадмины, что бы «сеть легла». Сюда же примыкают и вопросы оптимизации сетей.

В-пятых, поиск неявных связей, тех, которые искусственно скрываются. Важное приложение этой задачи – поиск связей террористов. И, конечно, бизнес разведка.

Python-networkx package in Ubuntu (https:// launchpad. net/ubuntu/ +source/python-networkx)

Методы, которые используются для решения этих и многих других задач можно, условно, разделить на три типа. Методы теории графов, в значительной мере комбинаторные. Численное моделирование, в настоящее время хорошо развито, опробовано и приспособлено для целей исследователя сложных сетей. Например, для языка программирования Phyton разработан специальный пакет (python-networkx) – инструментарий для создания, манипулирования и изучения сложных сетей, позволяющий найти численно практически все возможные характеристики сложных сетей. Третий тип методов, который позволил установить основные закономерности в сложных сетях – методы теоретической физики. От теории среднего поля до ренорм-группы и диаграммной техники. Недаром значительное число ведущих исследователей сложных сетей физики теоретики.

О чем наша книга и для кого она предназначена? Сразу же надо сказать, что для тех, кто знаком с основными обзорами или книгами по теории сложных сетей наша книга не представляет интереса. Она предназначена, в первую очередь, тем кто только слышал термин «сложная сеть» или встречался с ним в статье по биологии, экономике и т.п. Заинтересовавшемуся читателю довольно сложно сразу же освоить большой обзор, терминологию, методы, основные задачи. Здесь и должна помочь предлагаемая книга. В первой ее части описаны, на простых примерах, основные характеристики и свойства нескольких, наиболее часто встречающихся сетей. Вторая часть книги предназначена тем читателям, которые бы хотели, в качестве примера, посмотреть «как работает» теория сложных сетей. Здесь приведены несколько задач (проблем) выбранных на вкус авторов.

И, наконец, выражаем искреннюю благодарность ученым, поддержавшим идею написания этой книги, нашим коллегам, имеющим непосредственное отношение ко многим результатам, изложенным здесь: М.И Женировскому, И.В. Безсуднову, А.Г. Додонову и С.М. Брайчевскому.

Часть I. Сложные сети

1. Основные понятия

1.0. Направление «сложных сетей»

Издавна среди жителей Кенигсберга была распространена такая загадка: как пройти по всем семи мостам (через реку Преголя, рис. 1.0.1), не проходя ни по одному из них дважды. Доказать или опровергнуть возможность существования такого маршрута никто не мог.

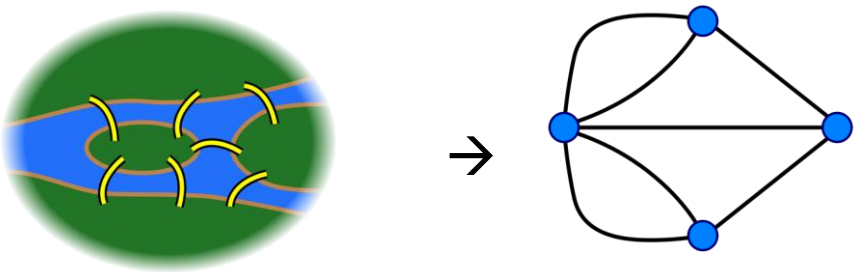


Рис. 1.0.1 – Схема мостов Кенигсберга

Леонард Эйлер (1707-1783)

Леонард Эйлер 13 марта 1736 г. в письме к итальянскому математику и инженеру Мариони написал, что нашел правило, пользуясь которым, легко определить, можно ли пройти по всем мостам, не проходя дважды ни по одному из них:

- Число нечетных вершин должно быть четно (нет такого графа, в котором нечетное число четных вершин)
- Если все вершины четные, то можно не отрывая карандаша от бумаги начертить граф. При этом можно начинать с любой вершины и завершить в той же вершине.

– Граф с более чем двумя нечетными вершинами невозможно начертить одним росчерком.

Граф Кенигсбергских мостов имеет пять нечетных вершин, то есть мосты нельзя обойти не проходя дважды по какому-нибудь.

P.S.: В 1905 году по приказу Кайзера Вильгельма был построен Императорский мост, который был впоследствии разрушен в ходе бомбардировки во время Второй мировой войны. В настоящее время в Калининграде семь мостов, и граф по-прежнему не имеет эйлерова пути.

Несмотря на то, что в рассмотрение теории сложных сетей попадают различные сети – электрические, транспортные, информационные, наибольший вклад в развитие этой теории внесли исследования социальных сетей.

Barnes, J.A. «Class and Committees in a Norwegian Island Parish», Human Relations 7:39-58.

Термин «социальная сеть» обозначает сосредоточение социальных объектов, которые можно рассматривать как сеть (или граф), узлы которой - объекты, а связи - социальные отношения. Этот термин был введен в 1954 году социологом из «Манчестерской школы» Дж. Барнсом (J. Barnes) в работе «Классы и сборы в норвежском островном приходе». Во второй половине XX столетия понятие «социальная сеть» стало популярным у западных исследователей. В теории социальных сетей получило развитие такое направление, как анализ социальных сетей (Social Network Analysis, SNA). Сегодня термин «социальная сеть» обозначает понятие, оказавшееся шире своего социального аспекта, оно включает, например, многие информационные сети, в том числе и WWW.

В теории сложных сетей выделяют три основных направления: исследование статистических свойств, которые характеризуют поведение сетей; создание модели сетей; предсказание поведения сетей при изменении структурных свойств. В прикладных исследованиях обычно применяют такие типичные для сетевого анализа характеристики, как размер сети, сетевая плотность, степень центральности и т.п.

При анализе сложных сетей как и в теории графов исследуются параметры отдельных узлов; параметры сети в целом; сетевые подструктуры.

Новая парадигма – «сложные сети» охватывает сети, обладающие свойствами:

1) большие размеры;
2) элементы случайности при формировании;
3) рост (изменение) во времени;
4) некоторые узлы могут образовывать компактные группы – ансамбли.

Сетевая плотность – соотношение наличествующих и возможных связей:

$$\Delta = \frac{2L}{n(n-1)},$$

где L – количество наблюдаемых связей, n – количество узлов в сети.

1.1. Характеристики сложных сетей

1.1.1. Параметры узлов сети

Для отдельных узлов выделяют следующие параметры:

- входная степень узла – количество ребер графа, которые входят в узел;
- выходная степень узла – количество ребер графа, которые выходят из узла;
- расстояние от одного узла до других;
- эксцентричность (eccentricity) – наибольшее из минимальных расстояний от данного узла до других;
- посредничество (betweenness), показывающее, сколько кратчайших путей проходит через данный узел;
- центральность – общее количество связей данного узла по отношению к другим.

1.1.2. Распределение степеней узлов

Сети в целом характеризуются такими параметрами, как количество узлов, количество связей, расстояние между узлами, среднее расстояние от одного узла до других, сетевая плотность, количество симметричных, транзитивных и

циклических триад, диаметр сети - наибольшее расстояние между узлами в сети и т.д.

Существует несколько актуальных задач исследования сложных сетей, среди которых можно выделить следующие основные:

- определение клик в сети. Клики – это подгруппы или кластеры, в которых узлы связаны между собой сильнее, чем с членами других клик;
- выделение компонент (частей сети), которые не связаны между собой, узлы которых связаны внутри этих компонент;
- нахождение блоков и перемычек. Узел называется перемычкой, если при его изъятии сеть распадается на несвязанные части;
- выделение группировок – групп эквивалентных узлов (которые имеют максимально похожие профили связей).

Классификация сетей по типам может быть произведена различными способами. Можно различать сети по тому, что представляют собой узлы и связи.

ER – сеть Эрдеша-Реньи, так называемый случайный граф

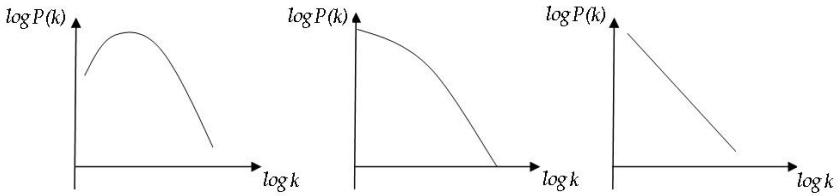
SF – сеть *scale free*, безмасштабная сеть

Для больших ($N>>1$) сетей со случайной структурой одной из самых важных характеристик является функция распределения $P(k)$ по степеням узлов. Наибольшая часть реальных CN похожи (близки) к следующим трем:

1. Случайная сеть или сеть Эрдеша-Реньи (*ER*) $P(k) \sim e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}$. Таким образом, в случае сети ER функция распределения есть функция Пуассона.
2. Сеть с экспоненциальным распределением $P(k) \sim e^{-k/\langle k \rangle}$.

3. Сеть со степенным распределением (Scale-Free) $P(k) = k^{-\gamma} / \varsigma(\gamma) \sim k^{-\gamma}$, где $\varsigma(\gamma)$ - дзета-функция Римана.

В двойном логарифмическом масштабе эти распределения имеют вид, представленный на рис. 1.1.2.



а б в

Рис. 1.1.2 – Плотности распределения $P(k)$ в двойном логарифмическом масштабе: а - распределение Пуассона (сеть *ER*); б – экспоненциальное распределение; в – степенное распределение

Сеть Эрдеша-Реньи можно построить, распределив случайным образом M связей между N узлами. Тогда с одной стороны $\langle k \rangle = 2M / N$, а с другой $\langle k \rangle = mN$, где m – вероятность соединения узлов. При $N \to \infty$ и $m \to 0$ распределение степени узлов является Пуассоновым.

К сетям со степенным распределением относятся сетями Барабаши-Альберта (BA), для построения которых используется специальная процедура, заключающаяся в том, что к изначально небольшому числу узлов N_0 постепенно добавляются новые узлы, связи от которых с большей вероятностью подсоединяются к тем узлам, у которых связей больше.

Альберт-Ласло Барабаши

Существуют процедуры построения сетей иного типа, когда к упорядоченной структуре добавляются случайные связи. Наиболее известный пример такой сети – так называемая сеть малого мира (Small World – SW).

Barabási A.L., Albert R. Emergence of scaling in random networks // *Science, 1999.* – **286** (5439): 509–512.

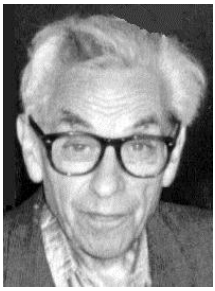
1.1.3. Кратчайший путь между узлами

Расстояние между узлами определяется как количество шагов, которые необходимо сделать, чтобы по существующим ребрам добраться от одного узла до другого. Естественно, узлы могут быть соединены прямо или опосредованно, через другие узлы.

Кратчайшим путем (SP, shortest path) между узлами назовем минимальное расстояние между ними. Для всей сети можно ввести понятие среднего кратчайшего пути, как среднее по всем парам узлов минимальное расстояния между ними:

$$l = \frac{2}{n(n+1)} \sum_{i \geq j} l_{ij}, \qquad (1.1.1)$$

где n – количество узлов, l_{ij} – кратчайшее расстояние между узлами i и j.



Пауль Эрдеш (1913-1996)

Erdős, P., Rényi A. On Random Graphs. // Publicationes Mathematicae, 1959. – № 6. – pp. 290-297.

Erdős P., Rényi A. On the evolution of random graphs // Publ. Math. Inst. Hungar. Acad. Sci., 1960. – № 5. – pp. 17-61. -1960.

П. Эрдешем (P. Erdős) и А. Реньи (A. Rényi) было показано, что средний кратчайший путь в случайном графе растет медленно – как логарифм от числа узлов.

С именем П. Эрдеша связаны не только исследования сложных сетей, но и популярное число Эрдеша, которое используется как один из критериев определения уровня

математиков в соответствующем социуме, базирующийся на так называемой сети соавторства. Известно, что Эрдеш написал около полутора тысяч статей, а также, что количество его соавторов превышало 500. Столь большое число соавторов и породило такое понятие, как число Эрдеша, которое определяется следующим образом: у самого Эрдеша это число равно нулю; у соавторов Эрдеша это число равно единице; соавторы людей с числом Эрдеша, равным единице, имеют число Эрдеша два; и т. д.

Таким образом, число Эрдеша это длина пути от некоторого автора до самого Эрдеша по совместным работам. Известен факт, что 90% математиков обладают числом Эрдеша не выше 8, что соответствует сетям «малых миров», речь о которых пойдет ниже.

Некоторые сети могут оказаться несвязными, т.е. найдутся узлы, расстояние между которыми окажется бесконечным. Соответственно, средний путь может оказаться также равным бесконечности. Для учета таких случаев вводится понятие среднего инверсного пути E между узлами, рассчитываемое по формуле:

$$E = \frac{2}{n(n-1)} \sum_{i>j} \frac{1}{l_{ij}}. \qquad (1.1.2)$$

Сети также характеризуются таким параметром как диаметр или максимальный кратчайший путь, равный максимальному значению из всех l_{ij}.

1.1.4. Коэффициент кластеризации

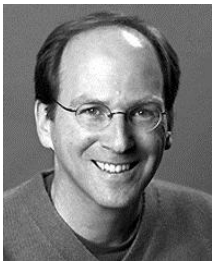
Д. Уаттс (D. Watts) и С. Строгатц (S. Strogatz) в 1998 году определили такой параметр сетей, как коэффициент кластеризации. Этот коэффициент характеризует тенденцию к образованию групп взаимосвязанных узлов, так называемых клик (Clique). Для конкретного узла коэффициент кластеризации показывает, сколько ближайших соседей данного узла являются также ближайшими соседями друг для друга.

Пусть из узла выходит k связей, которые соединяют его с k другими узлами, ближайшими соседями. Если предположить, что все ближайшие соседи соединены

непосредственно друг с другом, то количество связей между ними составляло бы $k(k-1)/2$. То есть это число, которое соответствует максимально возможному количеству связей, которыми могли бы соединяться ближайшие соседи выбранного узла.

Отношение реального количества связей, которые соединяют ближайших соседей данного узла i к максимально возможному (такому, при котором все ближайшие соседи данного узла были бы соединены непосредственно друг с другом) называется коэффициентом кластеризации узла C_i. Естественно, эта величина не превышает единицы.

Д.Уатте



С. Строгатц

Watts D.J., Strogatz S.H. Collective dynamics of “small-world” networks. // Nature, 1998. – **393**. – pp. 440-442.

Коэффициент кластеризации может определяться как для каждого узла, так и для всей сети:

$$C = \frac{1}{n}\sum_{i=1}^{n} C_i \,. \qquad (1.1.3)$$

В социальных сетях можно говорить о «структуре сообщества», когда существуют группы узлов, которые имеют высокую плотность связей между собой, при том, что плотность ребер между отдельными группами – низкая. Для больших социальных сетей наличие структуры сообществ оказалось неотъемлемым свойством. Традиционный метод для выявления структуры сообществ – кластерный анализ. Существуют десятки приемлемых для этого методов, которые базируются на разных мерах расстояний между узлами, взвешенных путевых индексах между узлами и т.п.

1.1.5. Посредничество

Посредничество (betweenness) – это параметр, показывающий, сколько кратчайших путей проходит через узел. Эта характеристика отражает роль данного узла в установлении связей в сети. Узлы с наибольшим посредничеством играют главную роль в установлении связей между другими узлами в сети. Посредничество b_m узла m определяется по формуле:

$$b_m = \sum_{i \neq j} \frac{B(i,m,j)}{B(i,j)}, \qquad (1.1.4)$$

где $B(i,j)$ – общее количество кратчайших путей между узлами i и j, $B(i,m,j)$ – количество кратчайших путей между узлами i и j, проходящих через узел m.

1.1.6. Эластичность сети

Свойство эластичности сетей относится к распределению расстояний между узлами при изъятии отдельных узлов (толерантность к атакам).

Эластичность сети зависит от ее связности, т.е. существовании путей между парами узлов. Если узел будет изъят из сети, типичная длина этих путей увеличится.

Albert R., Jeong H., Barabasi A. Error and attack tolerance of complex networks // Nature, 2000. – **406**. – pp. 378-382.

При исследовании атак на веб-серверы изучался эффект изъятии узлов сети, представляющей собой подмножество веб-пространства из 326000 страниц и около 1,5 млрд. гиперссылок. Для этой сети были определены параметры входного и выходного распределения степеней: $P(k) \sim k^{-\gamma}$, где $\gamma_{in} = 2{,}1$ и $\gamma_{out} = 2{,}45$.

Среднее расстояние между двумя узлами, как функция от количества изъятых узлов, почти не изменилось при случайном удалении узлов (высокая эластичность при атаке на сети со степенным распределением). Вместе с тем целенаправленное удаление узлов с наибольшим количеством

связей приводит к разрушению сети. Таким образом, веб-пространство является высокоэластичной сетью по отношению к случайному изъятию узла в сети, но высокочувствительной к намеренной атаке на узлы с высокими степенями связей с другими узлами.

1.1.7. Примеры вычисления харакетеристик сетей

Ниже будут перечислены основные характеристики сетей и одновременно они будут продемонстрированы на шести простых примерах. Для того чтобы «почувствовать» что именно описывают те или иные характеристики, читателю настоятельно рекомендуется получить «рукой» несколько приведенных в примерах чисел.

Рассматриваемые далее шесть примеров сетей будем обозначать номерами #1 – #6 (рис. 1.1.3).

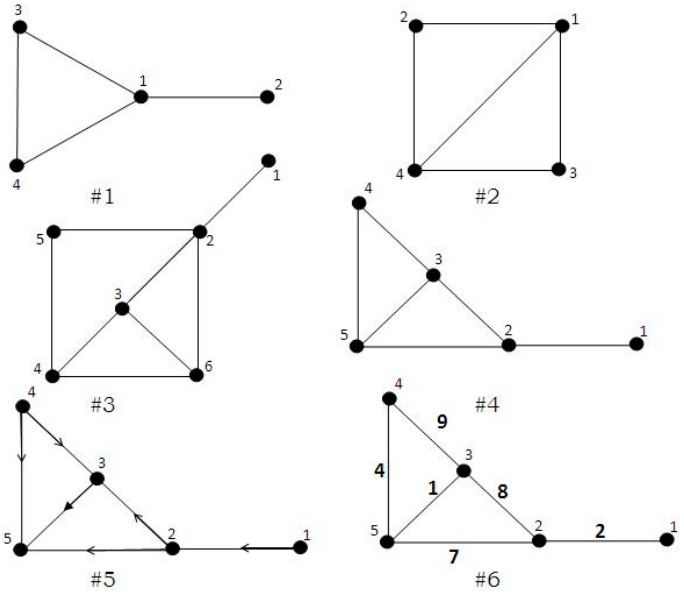


Рис. 1.1.3 – Примеры простейших сетей: сеть #5 – так называемый орграф (направленный граф), сеть #6 – сеть с весами связей (указаны цифры на связях)

Матрицы смежности сетей, изображенных на рис. 4 приведены ниже:

$$\boldsymbol{A}_1 = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix}, \quad \boldsymbol{A}_2 = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix},$$

Для сети #5 матрица смежности A_5 не симметрична.

Элементы матрицы смежности A_6 – это веса связей.

$$\boldsymbol{A}_3 = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 \end{pmatrix}, \quad \boldsymbol{A}_4 = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 \end{pmatrix},$$

$$\boldsymbol{A}_5 = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad \boldsymbol{A}_6 = \begin{pmatrix} 0 & 2 & 0 & 0 & 0 \\ 2 & 0 & 8 & 0 & 7 \\ 0 & 8 & 0 & 9 & 1 \\ 0 & 0 & 9 & 0 & 4 \\ 0 & 7 & 1 & 4 & 0 \end{pmatrix}.$$

Распределение узлов по их степеням $P(k)$ приведено на рис. 1.1.4, где также приведена такая характеристика сетей, как среднее число связей на один узел:

$$\langle k \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i = \sum_{i=1}^{N} k_i P(k_i), \qquad (1.1.5)$$

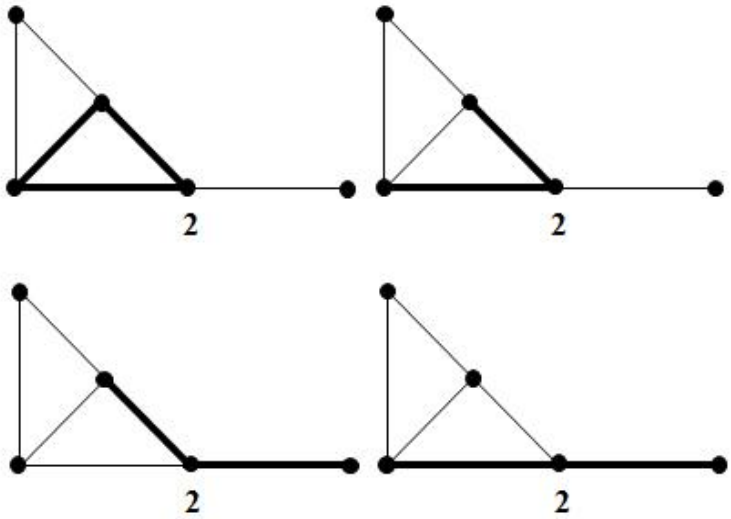
которая вычисляется как отношение всех связей в сети к числу узлов и, одновременно, как среднее дискретной переменной k_i с функцией распределения $P(k_i)$.

Важной характеристикой узла является его коэффициент кластеризации – C_i, характеризующий связность между собой соседей данного узла i. Коэффициент

кластеризации C_i может быть записан как отношение числа треугольников с вершиной i к числу вилок (две связи, выходящие из узла) с основанием в этом узле:

$$C_i = \frac{\textit{число треугольников с вершиной } i}{\textit{число вилок с основанием в } i}. \qquad (1.1.6)$$

Рассмотрим, например, узлы 2 и 3 сети #4. Для узла 2 жирными линиями обозначены вилки (их три) и треугольники (он один).



Как можно убедиться, для второго узла имеются три вилки и один треугольник, поэтому $C_2 = 1/3$.

Для третьего узла три вилки и два треугольника – $C_2 = 2/3$.

Еще одна сеть представлена на рис. 1.1.4.

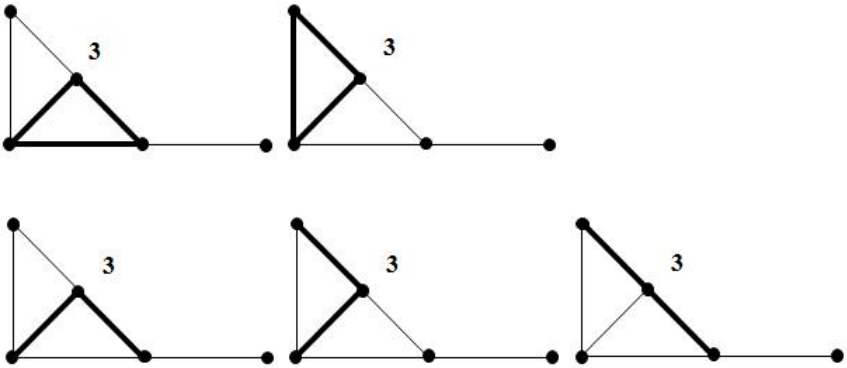


Рис. 1.1.4 – Варианты сети

Коэффициент кластеризации всей сети вычисляется по формуле:

$$C = \langle C_i \rangle = \frac{1}{N}\sum_{i=1}^{N} C_i. \qquad (1.1.7)$$

$$C_i = \frac{\sum \Delta_i}{\sum \vee_i}.$$

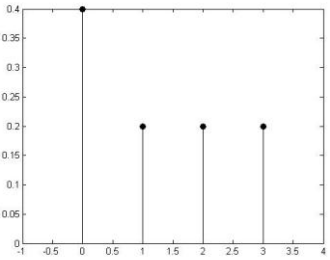
C – коэффициент кластеризации сети.

В табл. 1.1.1 приведены коэффициенты кластеризации для всех узлов сетей #1, #2, #3 и #4, а также коэффициент кластеризации всей сети C.

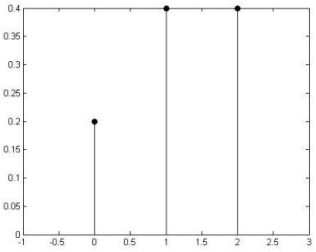
Как видно, наибольший коэффициент кластеризации имеет сеть #2, у нее, в среднем, наибольшее число соседей каждого узла соединены между собой. Наименьший коэффициент кластеризации имеет сеть #3.

Таблица 1.1.1 Коэффициенты кластеризации

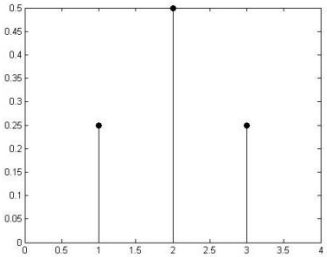
Сеть	C_1	C_2	C_3	C_4	C_5		C_6	C
# 1	1/3	0	1	1	1			7/12
# 2	2/3	1	1	1	2/3			10/12
# 3	0	1/6	2/3	2/3	1/3	0	2/3	11/36
# 4	0	1/3	2/3	2/3	1	2/3		8/15



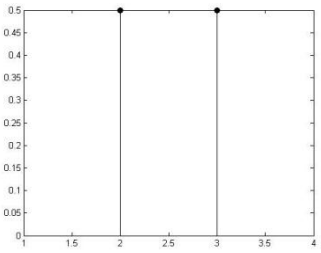
Сеть #1, $\langle k \rangle = 2$



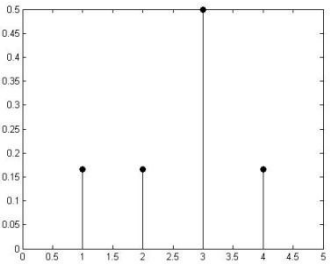
Сеть #2, $\langle k \rangle = 2{,}5$



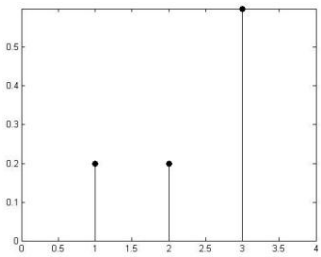
Сеть #3, $\langle k \rangle = 8/3 \approx 2.67$



Сеть #4, $\langle k \rangle = 12/5 = 2.4$



Сеть #5, входящие связи
$\langle k \rangle = 1.2$



Сеть #6, выходящие связи
$\langle k \rangle = 1.2$

Рис. 1.1.5 – Распределение узлов по степеням для сетей, изображенных на рис. 1.1.3

Коэффициент кластеризации узла можно вычислить и не прибегая к перечислению треугольников и вилок, непосредственно из матрицы смежности:

$$C_i = \frac{\sum_{j,m} A_{ij} A_{jm} A_{mi}}{k_i (k_i - 1)}, \qquad k_i = \sum_j A_{ij}, \tag{1.1.8}$$

где суммирование ведется по всем узлам.

На примерах простых сетей легко непосредственно убедиться, что определения (1.7) и (1.8) дают одни и те же значения C_i.

Кроме определения коэффициента кластеризации всей сети (1.1.7) и (1.1.8) в литературе встречается еще одно, близкое, но не тождественно определение, иногда называемое транзитивностью – T:

T – транзитивность, характери-стика, близкая к коэф-фициенту кластеризации.

$$T = \frac{\sum_{i=1}^{N} \left(\boldsymbol{A}^3\right)_{ii}}{\sum_{i=1}^{N} k_i (k_i - 1)} \equiv \frac{Tr\boldsymbol{A}^3}{\sum_{i=1}^{N} k_i (k_i - 1)}, \tag{1.1.9}$$

которое следующим образом выражается через количество треугольников и вилок во всей сети:

$$T = 3\frac{\textit{число треугольников в сети}}{\textit{число вилок в сети}}. \tag{1.1.10}$$

Для сравнения ниже приведены значения коэффициента кластеризации и транзитивности для сетей #1, #2, #3 и #4 – Табл. 1.1.2.

Еще одной важной характеристикой сетей является среднее минимальное расстояние между их узлами:

$$l = \left\langle l_{ij} \right\rangle = \frac{1}{N\left(N-1\right)} \sum_{i \neq j} l_{ij}, \qquad (1.1.11)$$

где l_{ij} - наименьшее число шагов от узла i к узлу j. При этом каждый шаг соответствует единичному расстоянию.

Таблица 1.1.2.
Коэффициенты кластеризации

Сеть	#1	#2	#3	#4
C	$\frac{7}{12} \approx 0.58$	$\frac{10}{12} \approx 0.83$	$\frac{11}{36} \approx 0.31$	$\frac{8}{15} \approx 0.53$
T	$\frac{3}{5} = 0.6$	$\frac{3}{4} = 0.75$	$\frac{3}{35} \approx 0.086$	$\frac{3}{5} = 0.6$

Аналогичное определение можно использовать и в нагруженных сетях, при этом каждый шаг может соответствовать как, расстоянию пропорциональному так и обратно пропорциональному весу:

$$\tilde{l} = \frac{1}{\left\langle 1/l_{ij} \right\rangle} = \left(\frac{1}{N\left(N-1\right)} \sum_{i \neq j} \frac{1}{l_{ij}} \right)^{-1}. \qquad (1.1.12)$$

В случае арифметического среднего (1.1.11) наибольший вклад будут давать наиболее длинные пути (из наиболее коротких), в случае же гармонического среднего – наиболее короткие. Если в сети имеются изолированные узлы, добраться до которых невозможно и для которых естественно считать $l_{ij} = \infty$, определение минимального среднего расстояния как арифметического среднего перестает быть информативным, поскольку даже при наличии одного такого изолированного узла $l = \left\langle l_{ij} \right\rangle = \infty$. В такой ситуации удобно пользоваться определением среднего

минимального расстояния как среднего гармонического – $\tilde{l}$. Иногда величину обратную $\tilde{l}$ обозначают $E = 1/\tilde{l}$ и называют действенность (efficiency).

Наименьшее число шагов l_{ij} от узла i к узлу j может быть записано в виде матрицы ***SP*** (short path). Численные значения этой матрицы для сетей #1 - #6 равны:

$$\mathbf{SP}(\#1) = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 2 & 2 \\ 1 & 2 & 0 & 1 \\ 1 & 2 & 1 & 0 \end{pmatrix}, \qquad \mathbf{SP}(\#2) = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 2 & 1 \\ 1 & 2 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix},$$

$$\mathbf{SP}(\#3) = \begin{pmatrix} 0 & 1 & 2 & 3 & 2 & 2 \\ 1 & 0 & 1 & 2 & 1 & 1 \\ 2 & 1 & 0 & 1 & 2 & 1 \\ 3 & 2 & 1 & 0 & 1 & 1 \\ 2 & 1 & 2 & 1 & 0 & 2 \\ 2 & 1 & 1 & 1 & 2 & 0 \end{pmatrix}, \quad \mathbf{SP}(\#4) = \begin{pmatrix} 0 & 1 & 2 & 3 & 2 \\ 1 & 0 & 1 & 2 & 1 \\ 2 & 1 & 0 & 1 & 1 \\ 3 & 2 & 1 & 0 & 1 \\ 2 & 1 & 1 & 1 & 0 \end{pmatrix},$$

$$\mathbf{SP}(\#5) = \begin{pmatrix} 0 & 1 & 2 & \infty & 2 \\ \infty & 0 & 1 & \infty & 1 \\ \infty & \infty & 0 & \infty & 1 \\ \infty & \infty & 1 & 0 & 1 \\ \infty & \infty & \infty & \infty & 0 \end{pmatrix}, \qquad \mathbf{SP}(\#6) = \begin{pmatrix} 0 & 2 & 10 & 13 & 9 \\ 2 & 0 & 8 & 11 & 7 \\ 10 & 8 & 0 & 5 & 1 \\ 13 & 11 & 5 & 0 & 4 \\ 9 & 7 & 1 & 4 & 0 \end{pmatrix}.$$

Отметим, что матрица $\mathbf{SP}(\#5)$ (для направленной сети) содержит бесконечные элементы, например $l_{21}(\#5) = \infty$ (сравните с $l_{21}(\#4) = 1$). Это, конечно, означает (см. рис. сети #5), что от узла 2 невозможно дойти до узла 1.

В сети с весами при подсчете l_{ij} каждый шаг по связи умножается на ее вес, так что суммируется не просто число шагов (связей), а их веса. Именно с этим связано то, что кратчайший путь в сети #6 между узлами 3 и 4 включает в себя два шага с весами 4 и 1, а не один шаг с весом 1.

Средние значения кратчайших путей (1.1.11) для рассматриваемых сетей равны:

$$\langle l(\#1)\rangle = \frac{4}{3} \approx 1.3, \qquad \langle l(\#2)\rangle = \frac{7}{6} \approx 1.16, \qquad \langle l(\#3)\rangle = \frac{23}{15} \approx 1.53,$$

$$\langle l(\#4)\rangle = \frac{3}{2} = 1.5, \qquad \langle l(\#5)\rangle = \infty, \qquad \langle l(\#6)\rangle = 7.$$

Гармоническое среднее:

$$\langle \tilde{l}(\#1)\rangle = 1.2, \qquad \langle \tilde{l}(\#2)\rangle \approx 1.09, \qquad \langle \tilde{l}(\#3)\rangle \approx 1.32,$$

$$\langle \tilde{l}(\#4)\rangle \approx 1.28, \qquad \langle \tilde{l}(\#5)\rangle \approx 2.86, \qquad \langle \tilde{l}(\#6)\rangle \approx 3.85.$$

Еще одной важной характеристикой сети является так называемое посредничество (betweenness). В таблице 1.1.3 приведены численные значения посредничества для всех узлов сетей #1 - #6.

Таблица. 1.1.3. Значения посредничества для сетей #1 - #6.

Сеть \ Узел	B_1	B_2	B_3	B_4	B_5	B_6
#1	4	0	0	0		
#2	1	0	0	1		
#3	0	10	4/3	2	4/3	4/3
#4	0	6	2	0	2	
#5	0	2	0	0	0	
#6	0	6	0	0	8	

Напомним, что при вычислении посредничества в сети с весом, кратчайший путь рассчитывается с учетом веса связей.

Кроме основной характеристики сети – распределения степени узлов по связям – $P(k)$, см. рис. 1.1.2 вводится также распределение концов связей – $P_{nn}(k)$:

$P_{nn}(k)$ – распределение концов связей.

$$P_{nn}(k) = \frac{kP(k)}{\langle k \rangle}. \qquad (1.1.13)$$

Это распределение называют также распределением степеней ближайших соседей (degree distribution of the nearest neighbor), откуда и берется обозначение $P_{nn}(k)$. На простом примере сети #1 легко убедиться, что (1.1.13) действительно дает распределение концов связей. Для сети #1 имеем восемь концов, т.е. вероятность попадания одного конца связи на узел равна $1/8$. У узла 1 одна связь, поэтому $P_{nn}(узел_1) = 1/8$. У каждого узлов 3 и 4 – $2/8$, т.е. $2 \cdot 2/8 = 1/2$, таким образом $P_{nn}(узел_3) = P_{nn}(узел_4) = 1/2$, $P_{nn}(узел_2) = 1/2$ и для узла 1 - $3 \cdot 1/8$, т.е. $P_{nn}(узел_3) = 3/8$.

С другой стороны эти же значения сразу получаются из (1.1.13), например, для первого узла

$$P_{nn}(узел_1) = \frac{1P(1)}{\langle k \rangle} = \frac{1 \cdot 1/4}{2} = \frac{1}{8},$$

и аналогично для других.

Качественно выражение для $P_{nn}(k)$ (1.1.13) можно пояснить так. Вероятность попадания случайно выбранной связи на узлы с k связями пропорционально числу всех связей таких узлов $P_{nn}(k) \sim NkP(k)$, где N - полное число

узлов в сети. С учетом того, что $\sum P_{nn}(k)=1$, нормировочная константа равна $1/\langle k\rangle$, откуда и следует (1.1.13).

Распределение $P_{nn}(k)$ для простых сетей совпадает с $P(k)$ только в исключительных случаях (когда для каждого узла выполняется равенство $k=\langle k\rangle$), например для бесконечной квадратной решетки.

Имея распределение концов связей $P_{nn}(k)$ можно ввести среднюю по этому распределению степень узлов:

$$\langle k\rangle_{nn}=\sum_{k=1}^{N}kP_{nn}(k)=\sum_{k=1}^{N}k\frac{kP(k)}{\langle k\rangle}=\frac{\langle k^2\rangle}{\langle k\rangle}. \quad (1.1.14)$$

Для некоррелированной сети знание распределения концов $P_{nn}(k)$ позволяет найти вероятность соединения узлов со степенями k и k' – $P(k,k')$. Т.к. вероятность того, что связь «воткнется» в узел со степенью k равна $P_{nn}(k)=kP(k)/\langle k\rangle$, вероятность того, что одновременно она «воткнется» и в связь со степенью k' равна произведению

$$P(k,k')=P_{nn}(k)P_{nn}(k')=\frac{kP(k)k'P(k')}{\langle k\rangle^2}. \quad (1.1.15)$$

Введем теперь понятие среднего числа первых, вторых и следующих соседей. Если узел имеет степень k, то у него k соседей, поэтому среднее число ближайших (первых) соседей z_1 равно, как и должно быть:

$$z_1=\sum k\cdot p_k=\langle k\rangle. \quad (1.1.16)$$

Первые соседи узла, в свою очередь, имеют своих соседей, которых логично назвать вторыми ближайшими к исходному узлу соседями. Если первый сосед имеет степень k, то у него $k-1$ соседей (одна из связей «ушла» к исходному узлу).

Число первых соседей со степенью k для всех узлов равно $N \cdot k \cdot p_k$, поэтому число всех вторых соседей равно:

$$z_2 = \frac{1}{N}\sum (k-1)k \cdot p_k = \sum \left(k^2 - k\right) p_k = \left\langle k^2 \right\rangle - \left\langle k \right\rangle. \quad (1.1.17)$$

Отношение среднего числа вторых соседей к первым:

$$B = \frac{z_2}{z_1} = \frac{\left\langle k^2 \right\rangle - \left\langle k \right\rangle}{\left\langle k \right\rangle} \qquad (1.1.18)$$

называют коэффициентом ветвления (branching coefficient), он характеризует степень «размножения» связей по мере удаления от узла.

Коэффициент ветвления можно получить и из таких соображений. Вероятность $P_{nn}(k)$ можно трактовать как вероятность $Q(k-1)$ того, что связь от некоторого выбранного узла A войдет в узел B с $k-1$ связями, так, что узел B будет иметь в сумме k связей:

$B-$ коэффициент ветвления.

$$Q(k-1) = P_{nn}(k) = \frac{kP(k)}{\left\langle k \right\rangle}. \qquad (1.1.19)$$

Тогда среднее число выходящих из узла B связей (без учета связи с узлом A) равно:

$$\sum_{k=0}^{N} kQ(k) = \sum_{k=0}^{N} \frac{k(k+1)P(k+1)}{\left\langle k \right\rangle} = \\ = \sum_{k=1}^{N} \frac{(k-1)kP(k)}{\left\langle k \right\rangle} = \frac{\left\langle k^2 \right\rangle - \left\langle k \right\rangle}{\left\langle k \right\rangle}. \qquad (1.1.20)$$

Последнее равенство справедливо при $N \gg 1$.

Зная коэффициент ветвления B (1.1.18) легко вычислить среднее число следующих (третьих, четвертых и т.д.) соседей:

$$z_m = Bz_{m-1} = \frac{z_2}{z_1} z_{m-1} = \frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle} z_{m-1}, \qquad (1.1.21)$$

$$z_m = \left(\frac{z_2}{z_1} \right)^{m-1} z_1. \qquad (1.1.22)$$

Выражение для среднего числа $m-$ных соседей позволяет получить ответ на очень важный вопрос – существует ли в сети так называемый гигантский кластер (Giant Connected Cluster или Giant Connected Component).

GC – Giant Cluster – гигантский кластер

Т.е. существует ли сети кластер (связанная между собой часть узлов), включающая в себя число узлов порядка всего числа узлов в сети. При вычислении все более дальних и дальних соседей ($m >> 1$) согласно (1.1.22) возможны два сценария – при $z_2 > z_1$, $z_m >> 1$ и при $z_2 < z_1$, $z_m << 1$.

Во втором случае (предполагается, что $N >> 1$) из (1.1.21) следует:

$$\sum_m z_m = z_1 \sum_m \left(\frac{z_2}{z_1} \right)^{m-1} \approx \frac{z_1^2}{z_1 - z_2}, \qquad (1.1.23)$$

а строгое равенство, конечно, имеет место в пределе $N \to \infty$.

Из (1.1.23) сразу же следует, что при $z_2 \geq z_1$, т.е. при

$$\langle k^2 \rangle - \langle k \rangle \geq \langle k \rangle, \qquad (1.1.24)$$

в сети существует бесконечный кластер.

Неравенство (1.1.24) называется критерий Моллоя-Рида, его часто записываю в таком виде

Критерий Моллоя-Рида

$$\langle k^2 \rangle - 2\langle k \rangle \geq 0. \qquad (1.1.25)$$

Таким образом, согласно (1.1.25) гигантский кластер существует, если число вторых соседей больше чем среднее число выходящих из узла связей.

Существование гигантского кластера означает также, что выбирая как угодно удаленные друг от друга узлы мы с ненулевой вероятностью сможем пройти по сети от одного узла к другому. Именно поэтому о тех значениях параметров, при которых $\langle k^2 \rangle$ достигает значения $2\langle k \rangle$, говорят как о пороге протекания. Значение порога протекания p_c при этом равна:

$$p_c = \frac{\langle k \rangle}{\langle k^2 \rangle - \langle k \rangle}. \qquad (1.1.26)$$

1.2. Модели артефактных сетей

1.2.1. Сети Эрдеша-Реньи

Сеть (граф) Эрдеша-Реньи (ER-сеть) это такая сеть, когда каждая пара узлов соединена с вероятностью p. В пределе большого числа узлов N функция распределения степеней узлов имеет вид:

$$P_k = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}. \qquad (1.2.1)$$

В пределе $N \to \infty$ значение $\langle k \rangle$ в построенной таким образом ER сети определено однозначно. В реальном случае для конечного значения числа узлов следует различать две модели ER сети – модель Гильберта (G_{np}-модель) и собственно модель Эрдеша-Реньи (G_{nm}). В модели G_{nm} фиксируется вероятность p. Для сети с конечным значением узлов N это означает, что $\langle k \rangle = pN$, при этом число связей M определено только в среднем $\langle M \rangle = pN(N-1)/2$. В модели Гильберта G_{np} задана не вероятность p, а число связей M, теперь вероятность p определена как $\langle M \rangle = pN(N-1)/2$, а средняя степень узла равна $\langle k \rangle = 2M/N$. Различие между этими моделями аналогично различию между каноническим и микроканоническим ансамблями в статистической физике. В микроканоническом ансамбле задана энергия системы, а в каноническом температура, при этом энергия флуктуирует вокруг среднего значения.

В пределе $N \to \infty$, $p \to 0$, pN – конечное число не равное нулю, оба определения сети Эрдеша-Реньи и Гильберта совпадают.

Сеть Эрдеша-Реньи является «хорошо соединенной» - среднее минимальное расстояние между узлами порядка $\ln N << N$ ($N >> 1$). Среднее минимальное расстояние между узлами легко оценить из следующих соображений. У каждого соседа есть еще в среднем по $\langle k \rangle$ соседей, к каждому из которых можно дойти за два шага. За l шагов можно в среднем дойти до $\langle k \rangle^l$ узлов.

Тогда для среднего минимального расстояния l между узлами сети из N узлов получаем:

Для сети *ER*:
$l \sim \ln N$

$$l = \frac{\ln N}{\ln \langle k \rangle}. \qquad (1.2.2)$$

Очень небольшое (логарифмически малое по сравнению с числом узлов) значение минимального расстояния, делает случайную ER-сеть так называемым «малым миром». Для каждого узла сети, состоящей, например, из $N = 10^9$ узлов (порядок числа людей на Земле) со средним числом связей $\langle k \rangle = 100$ (приблизительно столько людей мы лично знаем) минимальное среднее расстояние равно $l = \ln 10^9 / \ln 10^2 \approx 4.5$, т.е. не более пяти шагов.

Заметим, что для регулярной сети, например, для квадратной решетки это расстояние значительно больше $l \sim N^{1/2} >> \ln N$, для приведенного примера $l \sim \sqrt{10^9} \approx 30000 >> 4{,}5.$

Критическое значение p_c, при котором в сети ER рождается гигантский кластер сразу же находится из критерия Моллоя-Рида $p_c = \langle k \rangle / \left(\langle k^2 \rangle - \langle k \rangle \right)$. Так как $\langle k^2 \rangle = \langle k \rangle \left(\langle k \rangle + 1 \right)$, то для p_c получаем:

$$p_c = \frac{1}{\langle k \rangle}. \qquad (1.2.3)$$

1.2.2. Масштабно-инвариантные сети

Функция распределения степени узлов P_k для масштабно-инвариантной сети (Scale-Free – SF) имеет вид:

$$P_k = \frac{1}{k^\gamma}, \quad k \geq 0. \qquad (1.2.4)$$

Такое распределение хорошо известно в теории вероятности как распределение Парето. Обычно для реальних сетей показатель γ лежит в пределах от 2 до 3.

В том случае, когда k можно считать непрерывной переменной, обозначим ее – x $(x>0)$ необходимо учесть, что во всех реальных случаях существует минимальное значение этой переменной – $x_{\min}$. Нормировочная константа C для непрерывного случая определяется, при этом, из условия $\int_{x_{\min}}^{\infty} p(x)dx=1$ и равна

$$p(x)=Cx^{-\gamma}, \quad C=(\gamma-1)/x_{\min}^{\gamma-1}, \qquad (1.2.5)$$

откуда ясно, что распределение (1.31) имеет смысл только при $\gamma>1$.

Дополнительное ограничение $\gamma>2$ на показатель накладывает требование иметь конечное среднее значение – $\langle x\rangle$:

$$\langle x\rangle=\int_{x_{\min}}^{\infty} xp(x)dx=\frac{\gamma-1}{\gamma-2}x_{\min}, \quad \gamma>2, \qquad (1.2.6)$$

а для существования m-ного момента

$$\langle x^m\rangle=\frac{\gamma-1}{\gamma-1-m}x_{\min}^m,, \qquad (1.2.7)$$

требуется выполнения условия $\gamma>1+m$.

С распределением (1.2.5) связано так называемое правило $80/20$, в шуточном варианте – «20% людей выпивает 80% пива», и предполагается, что такого рода соотношение имеет место для многих других родов деятельности человека».

В самом деле, доля узлов $S(x)$ имеющих значение степени, больших x, составляет:

$$S(x)=\int_{x}^{\infty}p(x)dx=\left(\frac{x}{x_{\min}}\right)^{-\gamma+1}. \qquad (1.2.8)$$

Эти узлы в сумме содержат долю $W(x)$ от всех связей

$$W(x)=\frac{\int_{x}^{\infty}xp(x)dx}{\int_{x_{\min}}^{\infty}xp(x)dx}=\left(\frac{x}{x_{\min}}\right)^{-\gamma+2}. \qquad (1.2.9)$$

Из (1.2.8) и (1.2.9) сразу же следует

$$W=S^{\frac{\gamma-2}{\gamma-1}}=S^{1-\frac{1}{\gamma-1}}. \qquad (1.2.10)$$

На рис. 1.2.1 приведена зависимость W от S для различных значений γ. Очевидно, чем больше значение показателя γ, тем зависимость $W=W(S)$ ближе к линейной.

При показателе $\gamma=2.161$ для $S=0,2$ значение $W=0.8$ соответствует правилу Парето (80/20). Чем больше значение показателя γ, тем зависимость $W=W(S)$ ближе к линейной.

Поясним теперь, что означает термин «безмасштабная» по отношению к сети SF. Будем, для наглядности, считать, что степень узла (значение величины x) это богатство, которым обладает человек (данный узел). Тогда можно вычислить относительную долю $S(x_M)$ людей с богатством большим x_M, владеющих половиной всех денег.

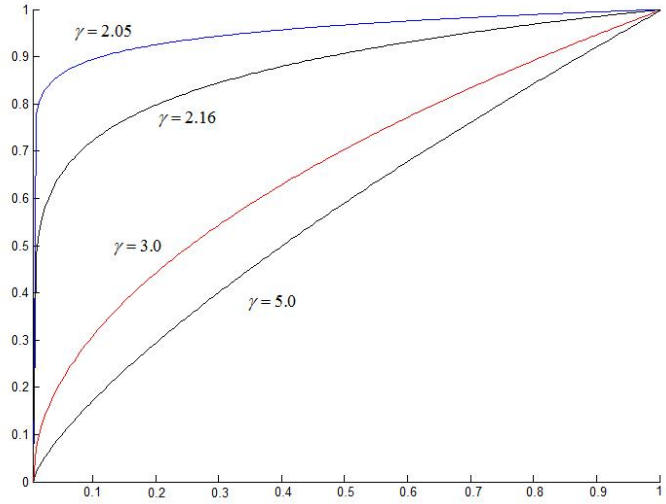


Рис. 1.2.1 – Зависимость $W = W(S)$

С одной стороны:

$$S(x_M) = \int_{x_M}^{\infty} p(x)dx = \left(\frac{x_M}{x_{\min}}\right)^{-\gamma+1}, \qquad (1.2.11)$$

с другой стороны, доля их богатства $W(x_M) = 1/2$, т.е.

$$W(x_M) = \frac{\int_{x_M}^{\infty} xp(x)dx}{\int_{x_{\min}}^{\infty} xp(x)d} = \left(\frac{x_M}{x_{\min}}\right)^{-\gamma+2} = \frac{1}{2}. \qquad (1.2.12)$$

Подставляя значения x_M из (1.2.11) в (1.2.12) находим

$$S(x_M) = 2^{-\frac{\gamma-1}{\gamma-2}}, \qquad S(x_M)\big|_{\gamma=2.161} = 6.7 \cdot 10^{-3} \qquad (1.2.13)$$

,

т.е. для значения показателя $\gamma = 2.161$ половиной всех денег владеет меньше 1% (0.67%) людей.

Назовем этих людей «богатыми». Безмасштабность означает, что среди «богатых» распределение на «богатых» и «бедных» в точности такое же. Т.е., как легко подсчитать, что 0.67% среди «богатых» владеют половиной всего «богатства» «богатых» (т.е. $1/4$ от всего богатства).

Для любого значения b для безмасштабного распределения

$$P(bx) = g(b)P(x), \qquad (1.2.14)$$

т.е. измените масштаба ($x \to bx$) приводит только к умножению распределения на константу.

Еще один, часто приводимый, пример для SF распределения связан с «компьютерной жизнью». Если в PC файлы размера 2 KB составляют четверть от файлов размером 1 KB, то число файлов размером 2 MB будет составлять четверть от числа файлов размером в 4 MB.

В дискретном варианте SF – распределение нормировано немного сложнее

$$1 = \sum_{k=1}^{\infty} P_k = C\sum_{k=1}^{\infty} \frac{1}{k^{\gamma}} = C\zeta(\gamma), \qquad (1.2.15)$$

где $\zeta(\gamma) = \sum_{k=1}^{\infty} k^{-\gamma}$ – ζ -функция Римана.

Таким образом, согласно (1.2.15):

$$P_k = \frac{k^{-\gamma}}{\zeta(\gamma)}$$

В том случае, когда обрезается «бедный хвост», т.е. не рассматриваются

Dorogovtsev, A.V. Goltsev, J.F.F. Mendes. Critical phenomena in complex networks, arXiv:0705.0010.

узлы со степенями меньшими $k_{\min}$, распределение p_k записывается так:

$$p_k = \frac{k^{-\gamma}}{\zeta(\gamma, k_{\min})}, \quad k \geq k_{\min}, \qquad (1.2.16)$$

а $\zeta(\gamma, k_{\min})$ равна

$$\zeta(\gamma, k_{\min}) = \sum_{k=k_{\min}}^{\infty} k^{-\gamma}, \qquad (1.2.17)$$

Сеть (граф) со SF распределением степени узлов может быть как случайной, так и детерминированной. Существует много «игрушечных» примеров детерминированных сетей со SF распределением, например, так называемые (u,v)-цветы и (u,v)-деревья. Построение (u,v)-цветка начинается с цепочки $w = u + v$ связей, после чего, на каждом следующем шаге, каждая связь заменяется на цепочку из двух частей u и v, как показано на рис. 5.

На рис. 1.2.2 показано несколько шагов построения (1,2), (1,3), и (2,2) цветов.

На n-ом шаге построения число связей в такой сети $M_n = (u+v)^n$.

В тоже время число узлов на n-м шаге – N_n подчиняется следующему итерационному соотношению

$$N_n = wN_{n-1} - w, \quad w = u + v, \qquad (1.2.18)$$

откуда

$$N_n = \left(\frac{w-2}{w-1}\right) w^n + \frac{w}{w-1}. \qquad (1.2.19)$$

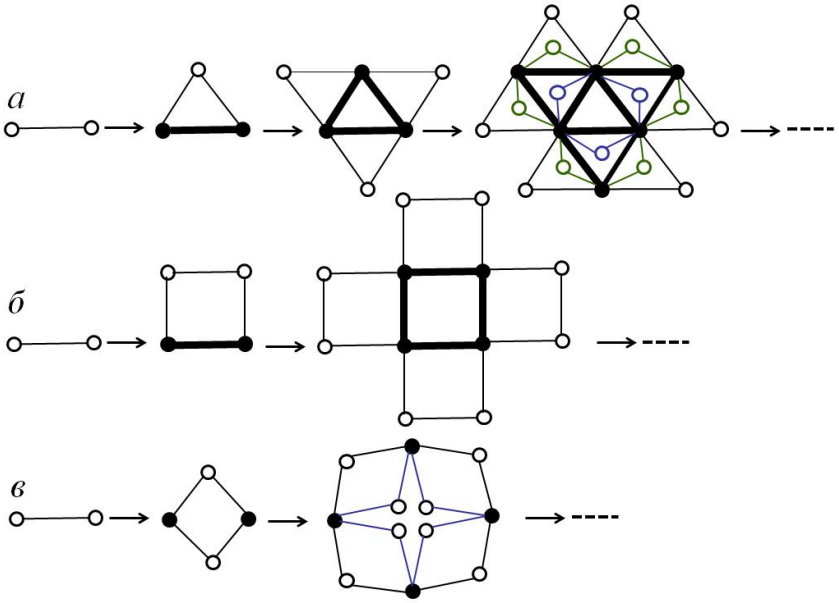


Рис. 1.2.2 – Примерs построения (u,v)–цветов: а – (1,2)-цветок; б – (1,3)-цветок; в – (2,2)-цветок:

○ – узлы, появляющиеся на данном шаге построения;

● – «старые узлы»; жирные линии – связи, появляющиеся на данном шаге построения.

Согласно правилу построения (u,v)-цветов, на n-ом шаге встречаются узлы только со степенями

$$k = 2^m, m = 1,2,...,n, \qquad (1.2.20)$$

обозначая N_n – число узлов на шаге n со степенью 2^m можно записать следующее итеративное соотношение:

$$N_n(m) = N_{n-1}(m-1) + (w-2)w^{n-1}\delta_{m,1} \qquad (1.2.21)$$

Откуда:

$$N_n(m) = \begin{cases} (w-2)w^{n-m}, & m < n, \\ w, & m = n. \end{cases} \qquad (1.2.22)$$

Так как при $n \gg 1$ $N_n(m) \sim p(k)$ из $N_n(m) \sim w^{-m}$ и $k = 2^m$ следует, что

$$p(k) \sim k^{-\gamma}, \qquad (1.2.23)$$

где

$$\gamma = 1 + \frac{\ln w}{\ln 2} = 1 + \frac{\ln(u+v)}{\ln 2} \qquad (1.2.24)$$

И, таким образом, (u,v)-цветы действительно является *SF*-сетями.

Для сети с $p_k \sim k^{-\gamma}$ имеются узлы с максимальным значением степени – $k_{\max}$. Значение $k_{\max}$, конечно, зависит от полного числа узлов – N. Эта зависимость степенная и следующим образом зависит от показателя γ:

$$k_{\max} \sim N^{-\frac{1}{\gamma-1}}. \qquad (1.2.25)$$

Коэффициент кластеризации для SF для случайного графа определяется из

$$C = \frac{\langle k \rangle}{N} \left(\frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle^2} \right)^2, \qquad (1.2.26)$$

где для $\gamma < 3$ имеем $\langle k^2 \rangle \sim k_{\max}^{3-\gamma}$ и, таким образом:

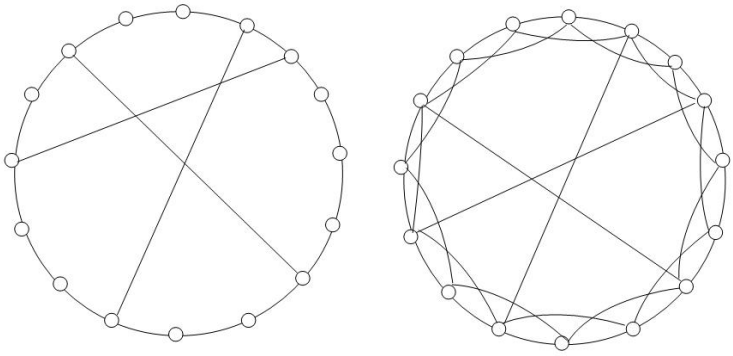
$$C \sim N^{-\beta}, \quad \beta = \frac{3\gamma - 7}{\gamma - 1}. \qquad (1.2.27)$$

1.2.3. Сети малого мира Ваттса – Строгатца

В 70-е года прошлого века американский психолог Милграм (Milgram) провел интересное исследование. Он задался вопросом, каково «расстояние» между двумя случайно выбранными людьми. Под расстоянием понимается

количество знакомств, необходимое для установления связи между данными людьми. Милграм поступил следующим образом – поскольку он жил в Бостоне, то был выбран далекий от Бостона город – Небраска, и случайно выбранным людям были розданы конверты, которые нужно было передать в Бостон. Конверты можно было передавать только через своих знакомых и родственников. Милграм получил весьма неожиданный результат: в среднем каждый конверт прошел через шесть человек. Так и родилась теория «шести рукопожатий». Т.е. каждый человек связан с любым другим цепью не больше шести личных знакомств. В этом смысле о нашем мире говорят как о малом мире – “small world”.

Модель перехода от большого (регулярного) мира к малому предложили Ваттс (Watts) и Строгац (Strogatz). Эта модель представляет собой одномерную регулярную решетку, состоящую из N узлов, где каждый узел соединен только со своими k ближайшими соседями и наложены периодические граничные условия, т.е. решетку свернули в кольцо, см. рис. 1.2.3. После чего каждую связь с вероятностью $\phi \ll 1$ перебрасывали на другой случайно выбранный узел. Правда при такой процедуре есть вероятность появления изолированных узлов.



а б

Рис. 1.2.3 – Пример малого мира с тремя перебросами ($N=16$): а – каждый узел соединен со своими ближайшими соседями ($k=2$), б – каждый узел соединен с четырьмя соседями ($k=4$)

Формально расстояние до них от любого узла будет бесконечным. Во избежание этого, Ньюман (Newman) и Ваттс предложили связи не перебрасывать, а просто добавлять. Остановимся на этом варианте модели подробнее.

Среднее расстояние между концами добавленных связей есть: $N/(2\cdot\phi\cdot N)=1/(2\cdot\phi)$. Для удобства опустим двойку в знаменателе и определим ξ как:

$$\xi=\frac{1}{\phi}. \qquad (1.2.28)$$

Для $k\geq 2$ естественное обобщение дает:

$$\xi=\frac{1}{k\cdot\phi}. \qquad (1.2.29)$$

Так как существует только один характерный размер системы ξ, то и безразмерное отношение среднего расстояния между узлами графа к числу всех узлов графа l/N может зависеть только от безразмерной величины N/ξ. Т.е. можно написать:

$$l=N\cdot f\left(\frac{N}{\xi}\right), \qquad (1.2.30)$$

где $f(x)$ – скейлинговая функция со следующими асимптотиками:

$$f(x)\sim\begin{cases} cons, & x\ll 1 \\ \dfrac{\log(x)}{x} & x\gg 1 \end{cases}. \qquad (1.2.31)$$

Как уже упоминалось выше, существует много способов определения корреляционного радиуса. Предположим что

$\xi \sim \phi^{-\tau}$. Покажем с помощью ренорм-группового преобразования (для $k = 2$) что $\tau = 1$. Итак пусть имеем:

$$l = N \cdot f\left(N \cdot \phi^{\tau}\right). \qquad (1.2.32)$$

Выполним ренорм-групповое преобразование, показанное на рис. 1.2.4, а именно: объединим в пары соседние узлы в графе, при этом в новом графе узлы соединены добавленной связью, если в исходном графе такая связь была хотя бы у одной из пар.

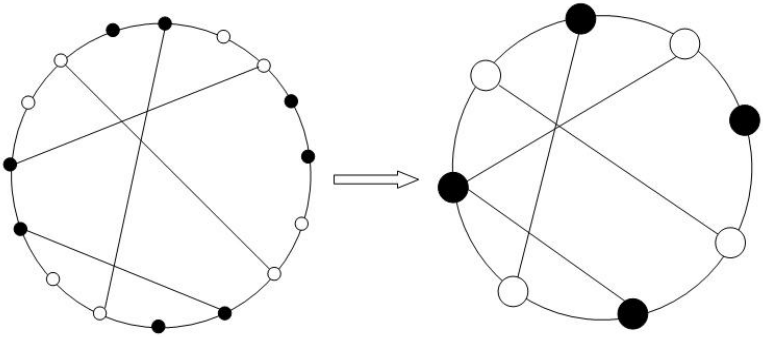


Рис. 1.2.4 – Ренорм-групповое преобразование. Два соседних черных узла объединяются в один большой черный узел в новом графе и аналогично для белых узлов

При таком преобразовании с очевидностью можно записать, что:

$$N' = \frac{1}{2}N, \quad \phi' = 2 \cdot \phi, \qquad (1.2.33)$$

где штрихованные величины относятся к правому графу на рис. 1.2.4.

Также понятно, что и среднее минимальное расстояние в новом графе l' будет отличаться в два раза.

$$l' = \frac{1}{2}l. \qquad (1.2.34)$$

Подставляя (1.2.33) и (1.2.34) в (1.2.32) получаем:

$$\tau = \frac{\log\left(N/N'\right)}{\log\left(\phi'/\phi\right)} = 1. \qquad (1.2.35)$$

Для $k > 2$ можно провести аналогичное преобразование, только теперь необходимо группировать не по два узла, а по $k/2$ узлов. Результат естественно остается тем же – $\tau = 1$.

Выше описанную модель малого мира можно обобщить на большие размерности. Так, например, в двумерном случае это может быть регулярная квадратная решетка с дополнительными связями, как показано на рис. 1.2.5. Здесь далее под k понимается степень узла, т.е. число ближайших соседей. Но нужно иметь в виду, что в литературе при описании малого мира под k часто понимают количество соседей в одном направлении.

Тогда вместо (1.2.29) будем иметь:

$$\xi = \frac{1}{\left(k \cdot \phi \cdot d\right)^{1/d}}, \qquad (1.2.36)$$

где – d размерность малого мира. И выражение (1.2.30) примет вид:

$$l = \frac{N}{k} \cdot f\left(\left(\phi \cdot k\right)^{1/d} \cdot N\right). \qquad (1.2.37)$$

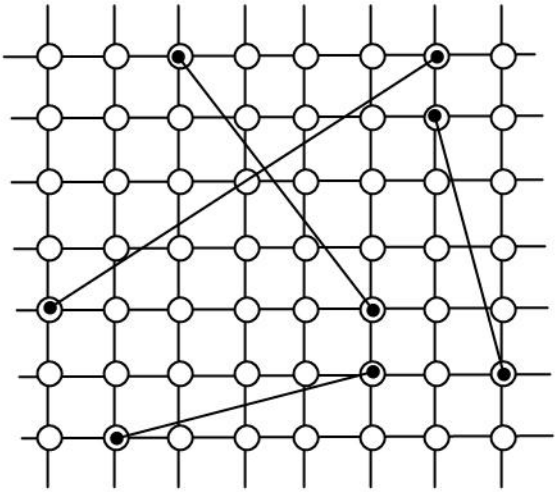


Рис. 1.2.5 – Пример двумерного ($d=2$) малого мира, $k=4$.

Для одномерного малого мира можно найти в явном виде кластерность C и средне минимальное расстояние l для малого мира, приведем конечные выражения:

$$c(\phi)=\frac{3\cdot(k-2)}{4\cdot(k-1)}\cdot(1-\phi)^3 \qquad (1.2.38)$$

и

$$l(\phi)=\frac{1}{\phi\cdot k\cdot\sqrt{1+\frac{2}{N\cdot\phi}}}\cdot \operatorname{arth}\left(\frac{1}{\sqrt{1+\frac{2}{N\cdot\phi}}}\right) \qquad (1.2.39)$$

.

Как и должно быть при записи выражения (1.2.39) в форме (1.2.38), функция $f(x)$ имеет следующие асимптотики:

$$f(x) \sim \begin{cases} \dfrac{1}{4}, & x \ll 1 \\ \dfrac{\log(2 \cdot x)}{4 \cdot x} & x \gg 1 \end{cases} \qquad (1.2.41)$$

.

На рис. 1.2.6. приведены графики нормированных зависимостей кластерности $\overline{C}$ и среднего расстояния $\overline{l}$ от концентрации перебросов p.

Подробнее о скейлинге в малом мире можно прочитать здесь:

M. E. J. Newman and D. J. Watts. Scaling and percolation in the small-world network model // Phys. Rev. E 60, 7332–7342 (1999) (arXiv: 9904419v2).

О ренорм-групповом преобразовании можно прочитать здесь: *Newman M. E. J., Watts D. J.* Renormalization group analysis of the small-world network mode // Phys. Lett. A 263, 341–346 (1999).

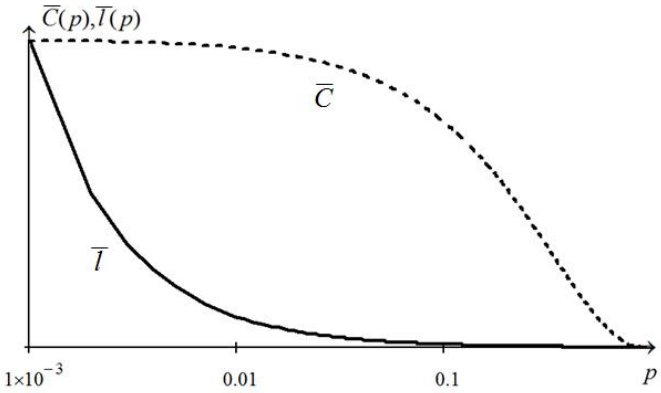


Рис. 1.2.6 – Пунктирная линия – нормированная кластерность. Сплошная линия – нормированное средне минимальное расстояние. Нормировка происходит на регулярный граф (без перебросов) – $\overline{C}(0)=1,\ \overline{l}(0)=1$

Для обычного регулярного графа (например, сетки) характерно большое среднее минимальное расстояние и

большая (близкая к единице) кластерность. А для полностью случайного графа обе эти величины падают. Поэтому обычно большое l ассоциируется с большим C и наоборот малое l с малым C. Тут же мы видим (см. рис. 1.2.6) что есть большая область значений, в которой относительно большое C и малое l. Это и является характерным признаком малого мира.

Подробный вывод формулы (1.2.41) основанный на приближении теории среднего поля описан здесь: *"Mean-field solution of the small-world network model" - M. E. J. Newman, C. Moore and D. J. Watts, Phys. Rev. Lett. 84, 3201–3204 (2000) (arXiv: 9909165v2).*

1.2.4. Перколяционные сети

Коротко остановимся на еще одном типе сетей – перколяционных. В простейшем варианте перколяционная сеть строится из регулярной, например, квадратной решетки путем вырывания (разрушения) случайным образом выбранных связей. На рис. 1.2.7 оставшиеся связи обозначены жирной линией, вырванные тонкой. Пусть при создании перколяционной сети с вероятностью 1 – p вырываются связи (узлы), тогда она будет будет состоять из $p \cdot N$ (N – целое число связей (узлов)) так называемых черных связей (узлов).

Для каждой такой сети (квадратной, треугольной, шестиугольной, кубической, …) существует такое значение p_c (порог протекания), что при $p > p_c$ можно пройти по черным связям через всю сеть, а при $p < p_c$ нет.

На рис. 1.2.8 показана перколяционная сеть большого размера для двух случаев *а* – ниже порога протекания $(p < p_c)$ и *б* - выше $(p > p_c)$.

Для разных решеток – треугольной, квадратной, кубической и т.п. свой порог протекания p_c, численное значение которого можно (за редким исключением) только путем численного моделирования. В Таблице 1.2.1 приведены значения порога протекания p_c для разных решеток.

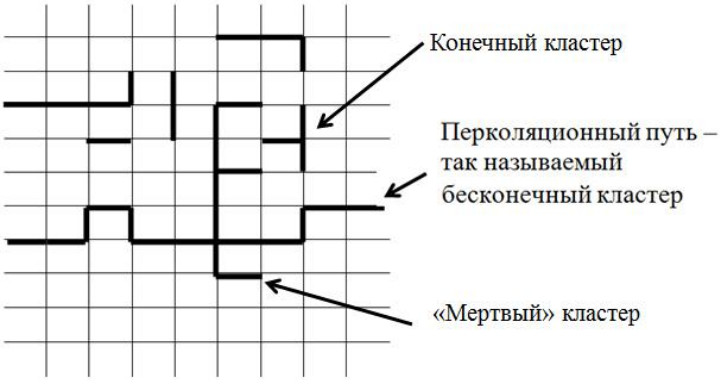


Рис. 1.2.7 – Квадратная решетка со случайно вырванными связями (тонкие линии)

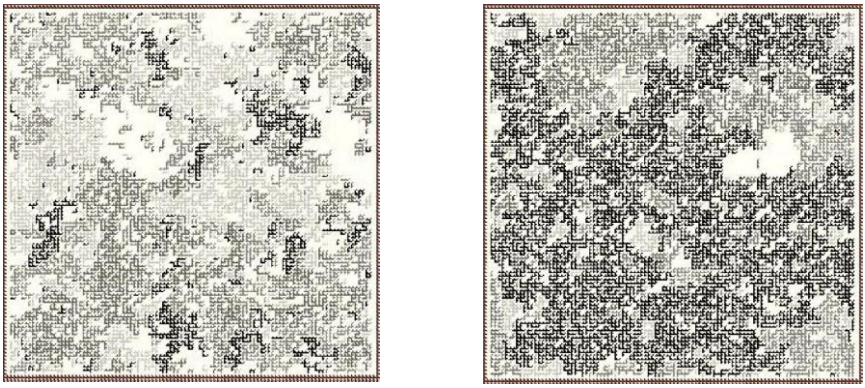


Рис. 1.2.8 – Перколяционная сеть большого размера. Слева, случай, когда концентрация связей меньше пороговой, справа – больше

Подчеркнуты значения p_c, которые определены точно. Первый столбец – задача узлов, второй – задача связей.

При вычислении p_c размер решетки выбирается достаточно большим (в теории – бесконечным), так что

значение p_c перестает зависеть от полного числа узлов – N или связей.

Главное в перколяции – это образование так называемого бесконечного кластера, позволяющего пройти по связям через всю сеть (говорят из бесконечности в бесконечность, подразумевая в пределе бесконечный размер сети). Число связей в бесконечном кластере порядка всех связей сети.

Таблица 1.2.1. Численное значение порога протекания p_c

Решетка	Задача узлов	Задача связей
Шестиугольная	0.69	0.65
Квадратная	0.59	$\frac{1}{2}$
Треугольная	$\frac{1}{2}$	0.35
Кубическая	0.31	0.25
$d=4$ гиперкубическая	0.197	0.16
$d=5$ гиперкубическая	0.14	0.12
$d=6$ гиперкубическая	0.11	0.09

В этом отношении – бесконечный кластер аналогичен Giant Cluster в сложных сетях, например в сети ER. Существует как общее так и отличное в свойствах бесконечного кластера в теории протекания и в Giant Cluster в теории сложных сетей.

Обозначая для краткости Giant Cluster как GC и бесконечный кластер как PC отметим следующее.

Для перколяции на решетке Келли $p_c = 1/(z-1)$, где z – координатное число, для GC ER сети координатное число из N связей равно $N-1$, а для $N >> 1$ $p_c = 1/N$. Таким образом при увеличении N p_c уменьшается, что аналогично увеличению пространственной размерности перколяционной

сети. В этом случае теория сложных сетей (ER) в пределе $N \to \infty$ аналогично перколяции в бесконечно мерном пространстве.

Как в задаче о GC так и в задаче о PC при $p < p_c$ вероятность появления GC и PC равна нулю.

Выше порога протекания, при $p > p_c$ размер GC равен $\left(f\left(p_c N\right) - f\left(pN\right)\right)N$, где f – экспоненциально убывающая функция с $f(1) = 1$, в то время как размер PC равен $\left(p - p_c\right)N$.

Еще одно отличие заключается в структуре, GC - это деревья, в то время как PC имеет фрактальную структуру.

Более подробно перколяционные сети описаны во второй части данного учебного пособия.

1.3. Примеры реальных сетей

Изучение значительного числа сложных артефактных (искусстыенно созданных) сетей, некоторые из которых описываются здесь, было инициировано желанием понять и описать многочисленные реальные сети – от сетей коммуникации до экологических сетей. Вот некоторые из них:

1. World Wide Web: Количество веб-сайтов – 919,533,715 (по состоянию на март 2014 года по данным службы Netcraft, рис. 1.3.1), охватывающих свыше триллиона ($\approx 10^{12}$) веб-страниц. Однонаправленные связи между отдельными веб-страницами реализуются в виде гиперссылок.
2. Internet – «Физическая сеть» (рис. 1.3.2, 1.3.3);
3. Протеиновые сети (рис. 1.3.4);
4. Сеть метаболизма;
5. Экологические сети;
6. Сеть телефонных звонков;
7. Сеть связей террористов;
8. Сеть цитирования (ациклическая);
9. Лингвистическая (сеть связанных слов);
10. Нейронные сети.

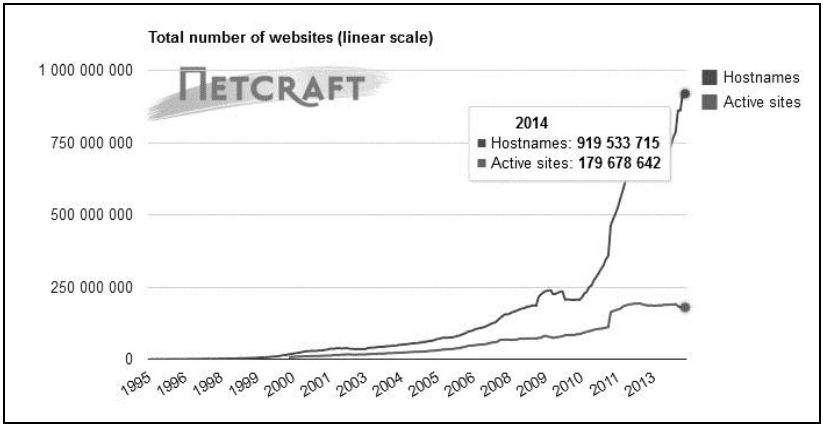


Рис. 1.3.1 – Динамика развития сети World Wide Web по данным службы Netcraft (http://netcraft.com)

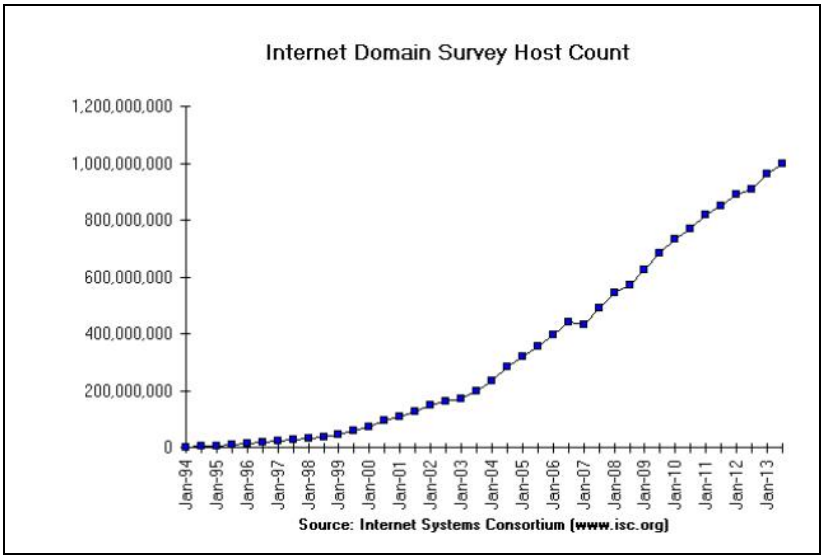


Рис. 1.3.2 – Динамика роста количества Интернет-доменов по информации службы isc.org

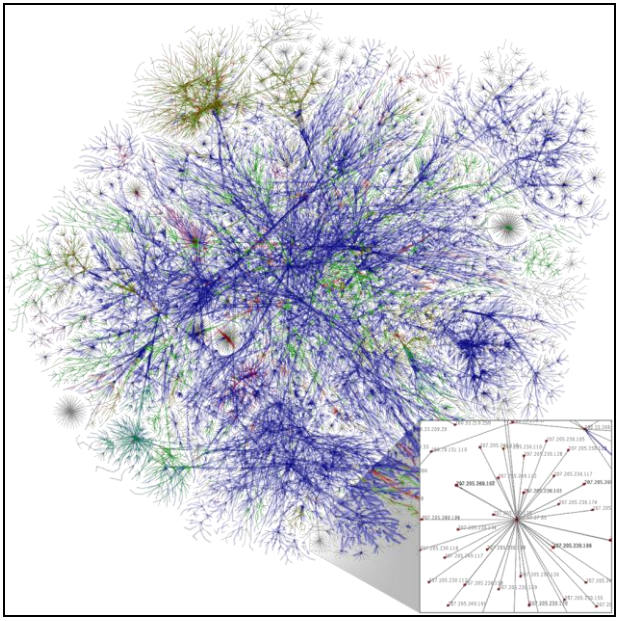


Рис. 1.3.3 – Карта связей Интернет-серверов как сложная сеть (по данным wikipedia.org)

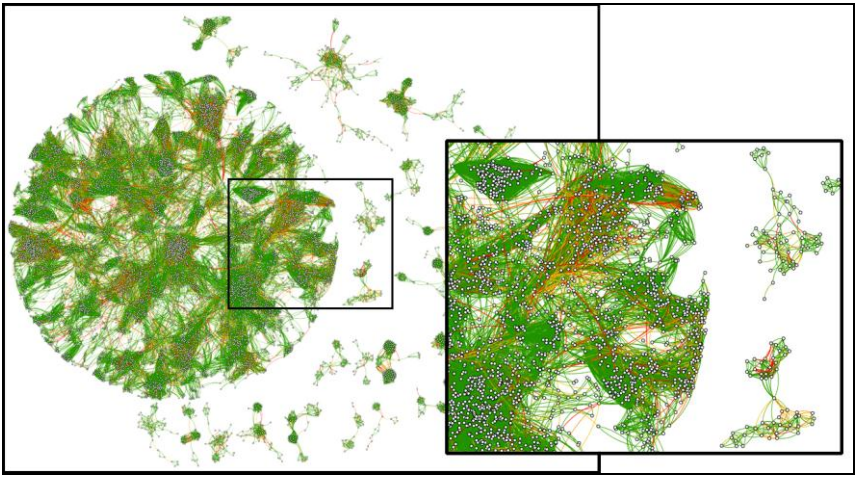
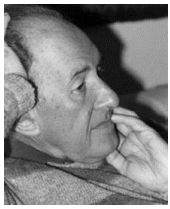


Рис. 1.3.4 – Протеиновая сеть (проект Protein Structure Initiative, сайт www.helixscript.com)

2. Задачи поиска в сетях

2.1. Векторно-пространственная модель поиска

Большинство известных информационно-поисковых систем базируется на использовании векторно-пространственной модели описания данных (Vector Space Model), предложенной Г. Солтоном в 1975 г. и примененной им в системе SMART. Данная модель является классической алгебраической. В рамках этой модели документ описывается вектором в евклидовом пространстве, в котором каждому терму, использующемуся в документе, ставится в соответствие его весовое значение, которое определяется на основе статистической информации о его появлении как в отдельном документе, так и во всем документальном массиве. Описание запроса, соответствующего необходимой пользователю тематике, также представляет собой вектор в том же евклидовом пространстве термов. Для оценки близости запроса и документа используется скалярное произведение соответствующих векторов запроса и документа.



Герхард Солтон (*1927-1995*)

Ландэ Д.В., Снарский А.А., Безсуднов И.В. Интернетика: Навигация в сложных сетях: модели и алгоритмы. – М.: Либроком (Editorial URSS), 2009. – 264 с.

В рамках этой модели каждому терму t_i в документе d_j соответствует некоторый неотрицательный вес w_{ij}.

В этой модели запросу q, который представляет собой также множество термов, не соединенных между собой никакими логическими операторами, также соответствует вектор весовых значений w_{iq}.

Таким образом, каждый документ и запрос могут быть представлены в виде $n-$мерного вектора, где n – общее количество термов в словаре модели. В соответствии с рассматриваемой моделью, близость документа d_j к запросу q, которые, как и в предыдущих моделях, рассматриваются как информационные векторы $\boldsymbol{d}_j = (w_{1j},\ w_{2j},...,w_{nj})$ и $\boldsymbol{q} = (w_{1q},\ w_{2q},...,w_{nq})$, оценивается как их скалярное произведение. При этом вес отдельных термов можно вычислять разными способами. Один из возможных простейших подходов – использовать как вес терма w_{ij} в документе нормализованную частоту $freq_{ij}$ его встречаемости в данном документе, то есть:

$$w_{ij} = tf_{ij} = freq_{ij} / \max_{1 \le i \le n}(freq_{ij}). \qquad (2.1.1)$$

Однако этот подход не учитывает, насколько часто данный терм используется во всем массиве документов, так называемую, дискриминационную силу терма. Поэтому в случае, когда доступна статистика использования термов во всем документальном массиве, более эффективно следующее правило вычисления веса:

$$w_{ij} = tf_{ij} \cdot \log \frac{N}{n_i}, \qquad (2.1.2)$$

где $n_i -$ число документов, в которых используется терм t_j, а $N -$ общее количество документов в массиве.

Следует отметить, что приведенная выше формула многократно уточнялась с целью наиболее точного соответствия выдаваемых системами документов запросам пользователей. В 1988 году Солтоном был предложен такой вариант для вычисления веса терма t_i из запроса:

$$w_{iq} = \left(0.5 + \frac{freq_{iq}}{\max\limits_{1 \le l \le n} freq_{lq}}\right) \cdot \log \frac{N}{n_i}, \qquad (2.1.3)$$

где $freq_{iq}$ - частота терма t_i из запроса в тексте документа, состоящего из n термов.

Обычно весовые значения w_{ij} нормируются, что позволяет рассматривать документ как ортонормированный вектор. Такой метод взвешивания термов имеет стандартное обозначение - $TF \cdot IDF$, где TF указывает на частоту появления терма в документе (term frequency), а IDF – на величину, обратную числу документов массива, содержащих данный терм (inverse document frequency).

Когда возникает задача определения тематической близости двух документов или документа и запроса, в этой модели используется простое скалярное произведение $sim(\boldsymbol{d}_1, \boldsymbol{d}_2)$, двух соответствующих векторов весовых значений $\left(w_{i1},\ ...,\ w_{n1}\right)$ и $\left(w_{i2},\ ...,\ w_{n2}\right)$, которое, очевидно, соответствует косинусу угла между векторами - образами документов d_1 и d_2. Очевидно, $sim(\boldsymbol{d}_1, \boldsymbol{d}_2)$ принадлежит диапазону [0, 1]. Чем больше величина $sim(d_1, d_2)$ – тем более близки документы d_1 и d_2. Для любого документа d имеем $sim(\boldsymbol{d}, \boldsymbol{d}) = 1$. Аналогично мерой близости документа d_j и запроса q является величина:

$$sim(d_j, q) = \frac{\boldsymbol{d}_j \cdot \boldsymbol{q}}{|\boldsymbol{d}_j| \cdot |\boldsymbol{q}|} = \frac{\sum_{i=1}^{n} w_{ij} w_{iq}}{\sqrt{\sum_{i=1}^{n} w_{ij}^2}\ \sqrt{\sum_{i=1}^{n} w_{iq}^2}}. \qquad (2.1.4)$$

Векторно-пространственная модель представления данных обеспечивает системам, построенным на ее основе, такие возможности:

- обработку запросов без ограничений их длины;
- простоту реализации режима поиска подобных документов (каждый документ может расматриваться как запрос);
- сохранение результатов поиска с возможностью выполнения уточняющего поиска.

Вместе с тем в векторно-пространственной модели не предусмотрена реализация запросов, реализующих логические операции, что существенно ограничивает ее применимость. Кроме того, являясь методологической основой других, в том числе, сетевых моделей поиска, классическая векторно-пространственная модель ориентированна на поиск массивах информации, не обладающих явно выраженной сетевой структурой.

2.2. Модели поиска в пиринговых сетях

Модели поиск в сетевой среде рассмотрим на примере поиска в так называемых пиринговых сетях. Пиринговые сети (Peer-to-peer, P2P – равный с равным) - это компьютерные сети, основанные на равноправии участников. В таких сетях отсутствуют выделенные серверы, а каждый узел (peer) является как клиентом, так и сервером. Впервые фраза «peer-to-peer» была использована в 1984 году Парбауэллом Йохнухуйтсманом (Parbawell Yohnuhuitsman) при разработке архитектуры Advanced Peer to Peer Networking фирмы IBM.

Гуркин Ю.Н. Семенов Ю.А. Файлообмен-ные сети P2P: основные принципы, протоколы, безопасность // “Сети и системы связи” - № 11. - Стр. 62. - 2006.

P2P – это сетевой протокол, обеспечивающий возможность создания и функционирования сети равноправных узлов, их взаимодействия. Во многих случаях P2P являются наложенными сетями, использующими существующие транспортные протоколы стека TCP/IP - TCP или UDP. Следует отметить, что на практике пиринговые сети состоят из узлов, каждый из которых взаимодействует лишь с некоторым подмножеством других узлов сети (из-за ограниченности ресурсов). Для реализации протокола P2P используются клиентские программы, обеспечивающие функциональность как отдельных узлов, так и всей пиринговой сети.

Процедуры поиска в пиринговых сетях подразумевают учет их разнообразной топологии, зачастую децентрализованной. Сегодня не существует унифицированных подходов к организации поисковых

процедур, поэтому применяются самые разнообразные методики. Именно благодаря пиринговым сетям были разработаны многие методы поиска в сетевой среде, подробное описание которых будет рассматриваться ниже.

Достаточно часто пиринговые сети дополняются выделенными серверами, несущими организационные функции, например авторизацию. В частности, известны библиотечные пиринговые сети, в которых используются выделенные серверы, играющие роль центров авторизации, хеширования и репликации библиографических данных.

Централизованная архитектура «клиент/сервер» подразумевает, что сеть зависит от центральных узлов (серверов), обеспечивающих подключенные к сети терминалы (клиенты) необходимыми сервисами. В этой архитектуре ключевая роль отводится серверам, которые определяют сеть независимо от наличия клиентов. Несмотря на то, что все узлы в P2P имеют одинаковый статус, реальные возможности их могут существенно отличаться.

Очевидно, что рост количества клиентов сети типа «клиент/сервер» приводит к росту нагрузок на серверную часть, в результате чего она может оказаться перегруженной.

Децентрализованная пиринговая сеть, напротив, становится более производительной при увеличении количества узлов, подключенных к ней. Действительно, каждый узел добавляет в сеть P2P свои ресурсы (дисковое пространство и вычислительные возможности), в результате суммарные ресурсы сети увеличиваются.

По сравнению с клиент/серверной архитектурой P2P обладает такими преимуществами, как самоорганизованность, отказоустойчивость при потере связи с узлами сети (высокая живучесть), возможность разделения ресурсов без привязки к конкретным адресам, увеличение скорости копирования информации за счет использования сразу нескольких источников, широкая полоса пропускания, гибкая балансировка нагрузки.

Помимо названных выше преимуществ пиринговых сетей, им присущ также ряд недостатков.

Первая группа недостатков связана со сложностью управления такими сетями по сравнению с клиент-

серверными системами. Приходится тратить значительные усилия на поддержку стабильного уровня их производительности, резервное копирование данных, антивирусную защиту, защиту от информационного шума и других злонамеренных действий пользователей.

Большая проблема – это легитимность контента, передаваемого в таких P2P-сетях. Неудовлетворительное решение этой проблемы привело уже к скандальному закрытию многих таких сетей (например, Napster в июле 2001 года). Есть и другие, проблемы, имеющие социальную природу. Так в системе Gnutella, например, 70% пользователей не добавляют вообще никаких файлов в сеть. Более половины ресурсов в этой сети предоставляется одним процентом пользователей, т.е. сеть эволюционирует в направлении клиент-серверной архитектуры.

Еще одна проблема P2P-сетей связана с качеством и достоверностью предоставляемого контента. Серьезной проблемой является фальсификация файлов и распространение фальшивых ресурсов. Защита распределенной сети от хакерских атак, ботнетов, вирусов и «троянских коней» является очень сложной задачей. Зачастую информация с данными об участниках P2P-сетей хранится в открытом виде, доступном для перехвата. Еще одной проблемой является возможность фальсификации ID узлов.

Главная задача информационного поиска в пиринговых сетях – быстрое и эффективное нахождение наиболее релевантных откликов на запрос, передаваемый от узла ко всей сети. При этом, естественно, поиск реализуется без участия центрального сервера, т.е. децентрализовано. В частности, при такой организации поиска актуальна задача получения

Kalogeraki V., Gunopulos D., Zeinalipour-Yazti D. A Local Search Mechanism for Peer-to-Peer Networks // Proc. of CIKM'02, McLean VA, USA, 2002.

Zeinalipour-Yazti D. Information Retrieval in Peer-to-Peer Systems. // M.Sc Thesis, Dept. of Computer Science, University of California - Riverside, June 2003.

Zeinalipour-Yazti D., Kalogeraki V., Gunopulos D. Information Retrieval in Peer-to-Peer Networks // IEEE CiSE Magazine, Special Issue on Web Engineering, 2004. – pp. 1-13.

качественного результата при общем уменьшении сетевого трафика.

Рассмотрим подробно такие методы (алгоритмы) децентрализованного поиска:

– алгоритм поиска ресурсов по ключам;

– метод широкого первичного поиска (Breadth First Search);

– метод случайного широкого первичного поиска (Random Breadth First Search);

– интеллектуальный поисковый механизм (Intelligent Search Mechanism);

– метод «большинства результатов по прошлой эвристике» (>RES);

– алгоритм «случайных блужданий» (Random Walkers Algorithm).

Алгоритм поиска ресурсов по ключам

В большинстве пиринговых сетей, ориентированных на обмен файлами, используется два вида сущностей, которым приписываются соответствующие идентификаторы (ID): узлы и ресурсы, характеризующиеся ключами (Key), т.е. сеть может быть представлена двумерной матрицей размерности MN, где M – количество узлов, N – количество ресурсов. В данном случае задача поиска сводится к нахождению ID узла, на котором хранится ключ ресурса. На рис. 2.2.1 представлен процесс поиска ресурса, запускаемый с узла с ID 0.

В данном случае с узла с ID 0 запускается поиск ресурса с ключем 14. Запрос проходит определенный маршрут и достигает узла, на котором находится ключ 14. Далее узел с ID 14 пересылает узлу с ID 0, адреса всех узлов, обладающих ресурсом, соответствующим ключу 14.

Рассмотрим алгоритмы поиска в пиринговых сетях, ограничившись основными методами поиска по ключевым словам.

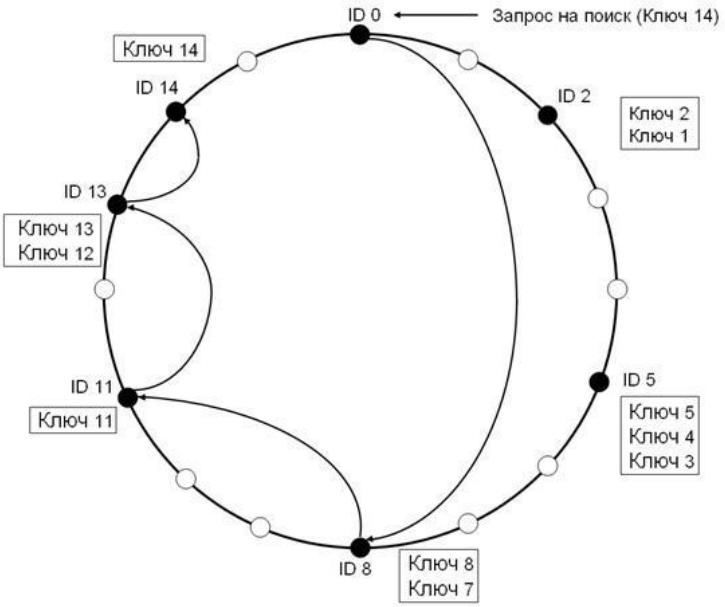


Рис. 2.2.1 – Модель поиска ресурса по ключу (в черных узлах содержатся документы с ключами – в белых – не содержатся)

Метод широкого первичного поиска

Метод широкого первичного поиска (Breadth First Search, BFS) широко используется в реальных файлообменных сетях P2P, таких как, например, Gnutella (www.gnutella.com). Метод BFS (рис. 9) в сети P2P размерности N реализуется следующим образом. Узел q генерирует запрос, который адресуется ко всем соседям (ближайшим по некоторым критериям узлам). Когда узел p получает запрос, выполняется поиск в его локальном индексе. Если некоторый узел r принимает запрос (Query) и обрабатывает его, то он генерирует сообщение-отклик (QueryHit), чтобы возвратить результат. Сообщение QueryHit включает информацию о релевантных документах, которая доставляется по сети запрашивающему узлу (рис. 2.2.2).

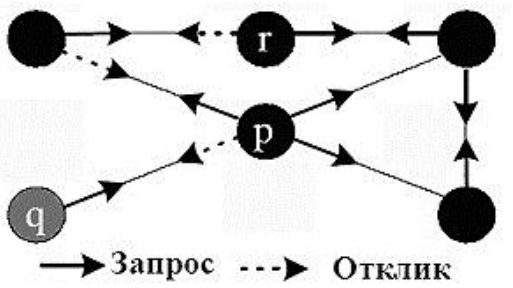


Рис. 2.2.2 – Метод BFS

Когда узел *q* получает QueryHits от более чем одного узла, он может загрузить файл с наиболее доступного ресурса. Сообщения QueryHit возвращаются тем же путем, что и первичный запрос. В BFS каждый запрос вызывает чрезмерную нагрузку сети, так как он передается по всем связям (в том числе и узлам с высоким временами ожидания). Поэтому узел с низкой пропускной способностью может стать узким местом. Один метод, позволяющий избежать перегрузки всей сети сообщениями заключается в приписывании каждому запросу с параметром времени жизни (Time-to-level, TTL). Параметр TTL определяет максимальное число переходов, по которым можно пересылать запрос. При типичном поиске начальное значение для TTL составляет обычно 5-7, которое уменьшается каждый раз, когда запрос пересылается. Когда TTL становится равным 0, сообщение больше не передается. BFS гарантирует высокий уровень качества совпадений за счет большого числа сообщений.

Метод случайного широкого первичного поиска

Метод случайного широкого первичного поиска (Random Breadth First Search, RBFS) был предложен как улучшение «наивного» подхода BFS. В методе RBFS (рис. 2.2.3) узел *q* пересылает поисковое предписание только части узлов сети, выбранной в случайном порядке. Какая именно часть узлов – это параметр метода RBFS.

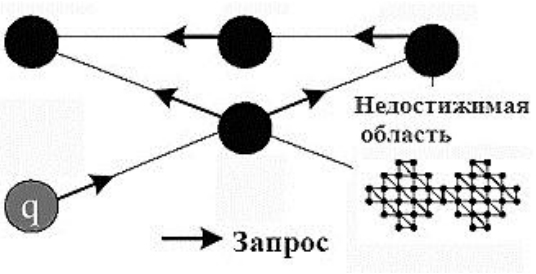


Рис. 2.2.3 – Метод RBFS

Преимущество RBFS заключается в том, что не требуется глобальной информации о состоянии контента сети; узел может получать локальные решения так быстро, как это потребуется. С другой стороны, этот метод вероятностный. Поэтому некоторые большие сегменты сети могут оказаться недостижимыми.

Интеллектуальный поисковый механизм

Интеллектуальный поисковый механизм (Intelligent Search Mechanism, ISM) - новый метод поиска в сетях P2P (рис. 2.2.4). Улучшение скорости и эффективности поиска информации с помощью данного метода достигается за счет минимизации затрат на связи, то есть на число сообщений, передающихся между узлами, и минимизация количества узлов, которые опрашиваются для каждого поискового запроса. Чтобы достичь этого, для каждого запроса оцениваются лишь те узлы, которые наиболее соответствуют данному запросу.

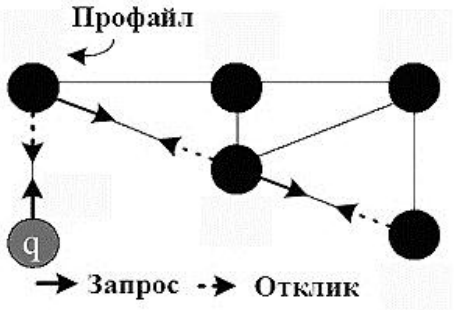


Рис. 2.2.4 – Метод ISM

Интеллектуальный поисковый механизм состоит из двух компонент – профайла (profile) и способа его ранжирования, так называемого ранга релевантности. Каждый узел сети строит информационный профайл для каждого из соседних узлов. Профайл содержит последние ответы каждого из узлов. С помощью ранга релевантности осуществляется ранжирование профайлов узлов для выбора тех соседних, которые будут давать наиболее релевантные документы по запросу.

Механизм профайлов используется для того, чтобы сохранять последние запросы, а также количественные характеристике результатов поиска.

При реализации модели ISM используется единый стек запросов размером $O(TN)$, в котором сохраняется по T запросов для N узлов. Как только стек заполняется, происходит замена «последнего наименее используемого» (Least Recently Used, LRU) для сохранения последних запросов. Функция «ранг релевантности» (Relevance Rank, RR) используется узлом P_l, чтобы выполнять оперативную классификацию его соседей для определения тех, которые следует опрашивать первыми по запросу q. Для вычисления ранга релевантности каждого узла P_i, P_l сравнивает q со всеми запросами в структуре профайла, для которого известен список ответов на предыдущие запросы, и вычисляется $RR(P_i, q)$:

$$RR(P_i, q) = \sum_{j \in Q} Sim(q_j, q)^{\alpha} \cdot S(P_i, q_j). \qquad (2.2.1)$$

В этой формуле Q - множество запросов, на которые был ответ у узла P_i, $S(P_i, q_j)$ – число результатов, возвращаемых узлом P_i по запросу q_j, а метрика Sim рассчитывается по правилу косинуса, аналогично рассмотренному в векторно-пространственной модели поиска:

$$Qsim(\boldsymbol{q}_j, \boldsymbol{q}) = \frac{\boldsymbol{q}_j \cdot \boldsymbol{q}}{|\boldsymbol{q}_j||\boldsymbol{q}|}. \qquad (2.2.2)$$

RR обеспечивает более высокий ранг узла, который возвращает больше результатов. Кроме того, используется параметр α, который позволяет увеличивать вес запросов, наиболее подобных исходному.

В случае, когда α большое, запросы с большим подобием $Qsim(q_j, q)$ доминируют в приведенной выше формуле. Рассмотрим ситуацию, когда узел P_1 соответствует запросам q_1 и q_2 значениями подобия для запроса q: $Qsim(q_1, q) = 0.5$ и $Qsim(q_2, q) = 0.1$, а узел P_2 соответствуют запросам q_3 и q_4 значениями $Qsim(q_3, q) = 0.4$ и $Qsim(q_4, q) = 0.3$. Если выбрать $\alpha = 10,$ то $Qsim(q_1, q)$ доминирует, так как $0.5^{10} + 0.1^{10} > 0.4^{10} + 0.3^{10}$.

Однако для $\alpha = 1$ все запросы весят одинаково, и P_2 дает более высокую релевантность. При $\alpha = 0$ учитывается только количество результатов, возвращенных каждым узлом.

Метод ISM эффективно работает в сетях, где узлы содержат некоторые специализированные сведения. В частности, исследование сети Gnutella показывает, что качество поиска очень зависит от «окружения» узла, с которого задается запрос. Еще одна проблема в методе ISM состоит в том, что поисковые сообщения могут зацикливаться, поэтому не в состоянии достичь некоторых частей сети. Чтобы разрешить эту проблему в был предложен такой подход. Выбиралось маленькое случайное подмножество узлов (в эксперименте дополнительно выбирался один случайный узел) и оно добавилось к набору релевантных узлов для каждого запроса. В результате механизм ISM стал охватывать большую часть сети.

Методы «большинства результатов по прошлой эвристике»

В рамках метода «большинства результатов по прошлой эвристике» (>RES) (рис. 2.2.5) каждый узел пересылал запрос подмножеству своих узлов, образованному на основании некоторой обобщенной статистики.

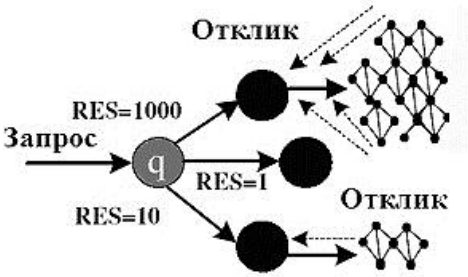


Рис. 2.2.5 – Метод >RES

Запрос в методе >RES является удовлетворительным, если выдается Z или больше результатов (Z – некоторая постоянная). В методе >RES узел q пересылает запросы к k узлам, выдавшим наибольшие результаты для последних m запросов. В их экспериментах k изменялось от 1 до 10 и таким путем метод >RES варьировался от BFS до подхода глубинного первичного поиска (Depth-first-search). Метод >RES подобен методу ISM, который рассматривался, но использует более простую информацию об узлах. Его главный недостаток по сравнению с ISM – отсутствие анализа параметров узлов, содержание которых связано с запросом. Поэтому метод >RES характеризуется скорее как количественный, а не качественный подход. Из опыта известно, что >RES хорош тем, что он маршрутизирует запросы в большие сегменты сети (которые возможно, также содержат более релевантные ответы). Он также захватывает соседей, которые менее перегружены, начиная с тех, которые обычно возвращают больше результатов.

Метод «случайных блужданий»

Ключевая идея алгоритма «случайных блужданий» (Random Walkers Algorithm, RWA) заключается в том, что каждый узел случайным образом пересылает сообщение с запросом, именуемое «посылкой» одному из своих узлов. Чтобы сократить время, необходимое для получения результатов, идея одной «посылки» расширена до «k-посылок», где k- число независимых посылок, последовательно запущенных с поискового узла. Ожидается, что «k-посылок» после T шагов достигнет тех же

результатов, что и одна посылка за kT шагов. Этот алгоритм напоминает метод RBFS, но в RBFS каждый узел пересылает сообщение запроса части из его соседей. К тому же, в RBFS предполагается экспоненциальное увеличение пересылаемых сообщений, а в методе случайных блужданий – линейное. Оба метода – и RBFS, и RWA не используют никаких явных правил для того чтобы адресовать поисковый запрос к наиболее релевантному содержанию.

Еще одной методикой, подобной RWA, является «адаптивный вероятностный поиск» (Adaptive Probabilistic Search, APS). В APS каждый узел развертывает на своих ресурсах локальный индекс, содержащий значения условных вероятностей для каждого соседа, который может быть выбран для следующего перехода для будущего запроса. Главное отличие от RWA в данном случае - это то, что в APS узел использует обратную связь от предыдущих поисков (в виде условных вероятностей) вместо полностью случайных переходов. Поэтому метод APS часто дает лучшие результаты, чем RWA.

Другой алгоритм, разработанный в Калифорнийском университете, построенный на методе случайных блужданий, использует принцип порога перколяции связей, то есть порога протекания или просачивания связей между тесно связанными узлами в сети. На этапе перколяции связей запрос попадает на один из базовых серверов, которые соединены друг с другом мощными каналами связи. Оказалось, что полноценный процесс поиска можно проводить «локально», то есть при опрашивании только соседних серверов. При таком подходе каждый запрос генерирует относительно небольшой трафик.

Существует много областей, где успешно применяется P2P-технология, например, параллельное программирование, кэширование данных, резервное копирование данных.

Благодаря таким характеристикам, как живучесть, отказоустойчивость, способность к саморазвитию, пиринговые сети находят все большее применение в системах управления производствами и организациями (например, P2P-технология сегодня применяется в Государственном Департаменте США). В данном случае

возможный выход из строя части узлов или серверов не существенно влияют на управляемость всей системы. Система доменных имен (DNS) в сети Интернет также фактически является сетью обмена данными, построенной по принципу P2P.

Реализацией технологии P2P является также популярная в настоящее время система распределенных вычислений GRID. Еще одним примером распределенных вычислений может служить проект distributed.net, участники которого занимаются легальным взломом криптографических шифров, чтобы проверить их надежность.

2.3. Ранговые характеристики

Ранжирование - процесс, при котором поисковая система выстраивает результаты поиска в определенном порядке по принципу наибольшего соответствия конкретному запросу. Таким образом, представление результатов поиска зависит от алгоритма ранжирования, который используется в поисковой системе.

В результате поиска пользователь может получить большой список релевантных документов. Сортировку этого списка таким образом, чтобы наиболее важные для пользователя документы были в его начале, в технологиях информационного поиска принято называть ранжированием откликов информационно-поисковых систем.

Ранжирование результатов поиска по уровню релевантности возможно не для всех моделей поиска (например, невозможно для булевой модели).

Перспективный подход к ранжированию - использование многопрофильных шкал, сформированных на основе метаданных, сетевых свойств, данных о пользователях.

Например, реализация сюжетных цепочек в тематических информационных массивах и их взвешивание рассматриваются как один из алгоритмов ранжирования.

Ранжирование текстовых и гипертекстовых документов имеет существенные отличия. Ранжирование текстовых документов может осуществляться по уровню релевантности и другим параметрам, в том числе экстрагируемым из текстов.

Ранжирование гипертекстовых документов возможно также по свойствам, обуславливаемым сетевой структурой, гиперссылками.

В Интернет для определения авторитетности веб-страницы как источника информации или посредника используется анализ топологии сети, образованной документами и соответствующими гиперссылками. Два алгоритма ранжирования веб-страниц, основанных на связях, HITS (hyperlink induced topic search) и PageRank, были разработаны в 1996 году в IBM Дж. Клейнбергом [94] и в Стенфордском Университете С. Брином и Л. Пейджем [74].

Kleinberg J.M. Authoritative sources in a hyperlink environment. // In Processing of ACM-SIAM Symposium on Discrete Algorithms, 1998, 46(5):604-632.

Brin S., Page L. The Anatomy of a Large-Scale Hypertextual Web Search Engine. WWW7, - 1998.

Оба алгоритма предназначены для решения "проблемы избыточности", свойственной широким запросам, увеличения точности результатов поиска на основе методов анализа сложных сетей.

2.3.1.Алгоритм HITS

Алгоритм HITS (Hyperlink Induced Topic Search), предложенный Дж. Клейнбергом, обеспечивает выбор из информационного массива лучших «авторов» (первоисточников, на которые введут ссылки) и «посредников» (документов, от которых идут ссылки цитирования). Понятно, что страница является хорошим посредником, если она содержит ссылки на ценные первоисточники, и наоборот, страница является хорошим автором, если она упоминается хорошими посредниками.

Дж. Клейнберг

Для каждого документа d_j

($j = 1,...,N$) рекурсивно вычисляется его значимость как автора $a(d_j)$ и посредника $h(d_j)$ по формулам:

$$a(d_j) = \sum_{i \to j}^{N} h(d_i), \;\; h(d_j) = \sum_{j \to i}^{N} a(d_i). \qquad (2.3.1)$$

Если ввести понятие матрицы инциденций $\boldsymbol{A}$, элемент которой a_{ij} равен единице, когда документ d_i содержит ссылку на документ d_j, и нулю в противном случае, то алгоритм HITS обеспечивает выбор наиболее авторитетных документов (авторов или посредников). Эти документы соответствуют собственным векторам матриц $\boldsymbol{A}\boldsymbol{A}^T$ и $\boldsymbol{A}^T\boldsymbol{A}$ с наибольшими модулями собственных значений (здесь $\boldsymbol{A}^T$ - транспонированная матрица $\boldsymbol{A}$).

Алгоритм вычисления рангов HITS приводит к росту рангов документов при увеличении количества и степени связанности документов соответствующего сообщества. В этом случае в результаты выдачи информационно-поисковой системы, использующей алгоритм HITS, могут попасть в большом количестве документы по темам, отличным от информационной потребности пользователя, т.е. часть выдаваемых результатов может отклониться от доминирующей тематики, происходит, так называемый, сдвиг тематики (topic drift).

Для решения этой проблемы как альтернатива стандартному алгоритму HITS был предложен алгоритм PHITS. В рамках этого алгоритма предполагается: D – множество цитирующих документов, C – множество ссылок, Z – множество классов (близких по какому-нибудь критерию документов), на которые разделяются документы. Предполагается также, что событие $d \in D$ (то, что выбранный для рассмотрения наугад документ является документом d) происходит с вероятностью $P(d)$.

Условные вероятности $P(c \mid z)$ (вероятность того, что ссылка из класса z – это c) и $P(z \mid d)$ (вероятность того, что выбранный документ d относится к классу z) используются

для описания зависимостей между наличием ссылки $c \in C$, фактором $z \in Z$ и документом $d \in D$.

Оценивается функция правдоподобия:

$$L(D,C) = \prod_{c \in C, d \in D} P(d,c) = \\ = \prod_{c \in C, d \in D} P(d)P(c \mid d), \qquad (2.3.2)$$

где

$$P(c \mid d) = \sum_{z \in Z} P(c \mid z)P(z \mid d). \qquad (2.3.3)$$

Цель алгоритма PHITS состоит в том, чтобы подобрать $P(z)$, $P(c \mid z)$, $P(d \mid z)$, чтобы максимизировать $L(D,C)$.

После этого:

$P(c \mid z)$ – ранги авторов;

$P(d \mid z)$ – ранги посредников.

Функция правдоподобия в этом случае показывает, насколько правдоподобно наличие ссылок при выбранных документах.

Для вычисления рангов необходимо задать количество факторов в Z, и тогда $P(c \mid z)$ будет характеризовать качество страницы как автора в контексте тематики z. К недостаткам метода надо отнести то, что итеративный процесс чаще всего останавливается не на абсолютном, а на локальном максимуме функции правдоподобия L. Вместе с там в ситуациях, когда в множестве найденных веб-страниц нет явного доминирования тематики запроса, PHITS превосходит алгоритм HITS.

2.3.2. Алгоритм PageRank

Алгоритм PageRank был придуман основателями компании Google для ранжирования веб-страниц. Он получил

название по фамилии одного из его изобретателей Лари Пейджа (Larry Page).

Главную идею этого алгоритма можно описать следующими словами: значимость (ранг) страницы тем выше, чем больше ссылок на нее с других значимых страниц. Т.е. PageRank рассчитывает вероятность того, что человек, случайно переходящий по ссылкам, доберется до некоторой страницы. Чем больше ссылок ведет на данную страницу с других популярных страниц, тем выше вероятность, что экспериментатор чисто случайно наткнется на нее. В нашей терминологии страница это узел сети, а линк на страницу это направленная связь.

Алгоритм PageRank близок по идеологии к литературному индексу цитирования, который рассчитывается для произвольного документа с учетом количества ссылок от других документов на данный документ, но при этом в PageRank как и в HITS, в отличие от литературного индекса цитирования, не все ссылки считаются равнозначными.

Авторы PageRank

Лэрри Пейдж

Сергей Брин

Принцип рассчета ранга веб-страницы PageRank следующий: рассматривается модель – процесс, при котором некоторый пользователь Интернет открывает случайную веб-страницу, с которой переходит по случайно выбранной гиперссылке. Потом он перемещается на другую веб-страницу и снова активизирует случайную гиперссылку и т. д., постоянно переходя от страницы к странице, никогда не возвращаясь. Иногда ему такое блуждание надоедает, и он снова переходит на случайную веб-страницу - не по ссылке, а набрав вручную некоторый URL. В этом случае вероятность того, что блуждающий в Сети пользователь перейдет на некоторую определенную веб-страницу – это ее ранг. Очевидно, PageRank веб-страницы тем выше, чем больше

других страниц ссылается на нее, и чем эти страницы популярнее.

Пусть есть n страниц $D = \{d_1, ..., d_n\}$, которые ссылаются на данный документ (веб-страницу A) , а $C(A)$ – общее число ссылок с веб-страницы A на другие документы. В соответствии с приведенной моделью поведения пользователя определяется некоторое фиксированное значение δ (damping factor) как вероятность того, что пользователь, пересматривая какую-нибудь веб-страницу из множества D, перейдет на страницу A по ссылке, а не набирая ее URL в явном виде. В рамках модели вероятность продолжения этим пользователем веб-серфинга по сети из N веб-страниц без использования гиперссылок, путем ручного ввода адреса (URL) со случайной страницы составит $1-\delta$ (альтернатива перехода по гиперссылкам). Индекс PageRank $PR(A)$ для страницы A рассматривается как вероятность того, что пользователь окажется в некоторый случайный момент времени на этой странице:

$$PR(A) = (1-\delta)/N + \delta \sum_{i=1}^{n} \frac{PR(d_i)}{C(d_i)}. \qquad (2.3.4)$$

По этой формуле индекс страницы легко подсчитывается простым итерационным алгоритмом. На практике применяется до 30 шагов итерации для достижения устойчивых результатов.

Несмотря на различия HITS и PageRank, в этих алгоритмах общее то, что авторитетность (вес) узла зависит от веса других узлов, а уровень "посредника" зависит от того, насколько авторитетны узлы, на которые он ссылается.

Расчет авторитетности отдельных документов сегодня широко используется в таких приложениях, как определение порядка сканирования документов в сети роботами поисковых система, ранжирование результатов поиска, формирование тематических обзоров и т.п.

Проиллюстрируем вышесказанное на примере. Рассмотрим небольшую часть сети, состоящую из узлов A с

рангом $r_A = 0.5$, B с рангом $r_B = 0.3$ и определим ранг узла C – см. рис 2.3.1.

Каждый из узлов сети A и B ссылается на узел C и для них ранг уже известен. Узел А при этом ссылается еще на три узла, а узел В еще на два узла. Чтобы вычислить ранг узла C ранги каждого узла, ссылающегося на C, делятся на общее количество ссылок для этого узла, после чего полученные значения складываются.

$$r_C = r_A \cdot \frac{1}{4} + r_B \cdot \frac{1}{3} = 0.225. \qquad (2.3.5)$$

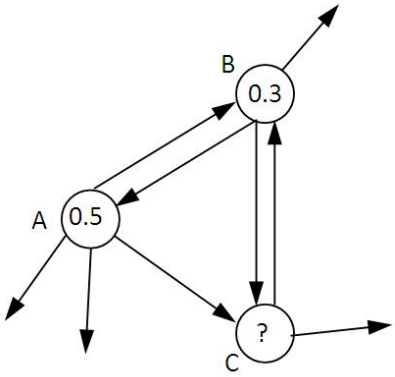


Рис. 2.3.1 – Фрагмент сети

Каждый из узлов сети A и B ссылается на узел C и для них ранг уже известен. Узел А при этом ссылается еще на три узла, а узел В еще на два узла. Чтобы вычислить ранг узла C ранги каждого узла, ссылающегося на C, делятся на общее количество ссылок для этого узла, после чего полученные значения складываются.

$$r_C = r_A \cdot \frac{1}{4} + r_B \cdot \frac{1}{3} = 0.225. \qquad (2.3.6)$$

В данном примере для всех узлов, ссылающихся на C, уже вычислен ранг. Но невозможно вычислить ранг узла, пока неизвестны ранги ссылающихся на него узлов, а эти ранги можно вычислить, только зная ранги узлов, которые

ссылаются на них. Так как же вычислить значение ранга для множества узлов, ранги которых еще неизвестны? Решение состоит в том, чтобы присвоить всем узлам произвольный начальный ранг и провести несколько итераций.

В общем виде можно записать следующую формулу для $(n+1)$-ого шага итерации:

$$r_i^{(n+1)} = \sum_{j \in E(i)} \left(r_j^{(n)} \cdot \frac{1}{kout_j} \right), \qquad (2.3.7)$$

где суммирование по $j \in E(i)$ означает суммирование по всем узлам, которые имеют ссылку на i-ый узел, а $kout_j$ – число исходящих связей для узла j. В матричном виде (2) записывается следующим образом:

$$\mathbf{r}^{(n+1)T} = \mathbf{r}^{(n)T} \cdot \mathbf{H}, \qquad (2.3.8)$$

где $H_{ij} = A_{ij} \Big/ \sum_j A_{ij}$ – нормализованная матрица инцидентности и $\mathbf{A}$ – матрица инцидентности сети.

Для итерационного процесса (2.3.8) естественно возникают следующие вопросы. Сходится ли данный процесс? Какими свойствами должна обладать матрица $\mathbf{H}$, чтобы процесс гарантировано сходился? Зависит ли конечный вектор $\boldsymbol{r}^{(\infty)}$ от начальных условий? И так далее. Этими же вопросами задались и авторы алгоритма. Рассмотрим два простых примера.

Первый:

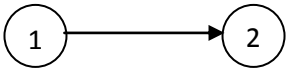


Рис. 2.3.2 – Пример 1

После выполнения итераций (2.3.7) получаем:

	Итерация 0	Итерация 1	Итерация 2	Итерация 3
Ранг узла 1	1	0	0	0
Ранг узла 2	0	1	0	0

Результат явно не соответствует ожиданиям. Рассмотрим второй, похожий пример.

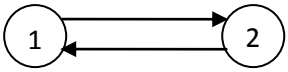


Рис. 2.3.3 – Пример 2

Здесь получаем следующую таблицу:

	Итерация 0	Итерация 1	Итерация 2	Итерация 3
Ранг узла 1	1	0	1	0
Ранг узла 2	0	1	0	1

Из приведенной таблицы видно, что сходимости процесса не наблюдается. Можно привести еще много контрпримеров, показывающих ограниченность формулы (2.3.8). *С* другой стороны, можно увидеть что (2.3.8) напоминает “power method” для вычисления собственного вектора матрицы **H** (соответствующего собственному значению – единица) примененному к Марковским цепям с матрицей вероятностей переходов $\mathbf{P} = \mathbf{H}$. В данном случае речь идет о «левом» собственном векторе. А поскольку теория Марковских цепей очень хорошо изучена, то можно сразу ответить на вопрос каким свойствам должна удовлетворять

матрица вероятностей переходов $\mathbf{P}$, чтобы процесс сходился, не зависел от начальных условий и т.д.

Матрица вероятностей переходов должна быть стохастической, неприводимой и непериодической. Первое условие удовлетворяется путем перехода к матрице $\mathbf{S}$.

$$\mathbf{S} = \mathbf{H} + \mathbf{a} \cdot \left(\frac{1}{n} \cdot \mathbf{e}^T \right), \qquad (2.3.9)$$

где $a_i = 1$ если из i-ого узла не выходит ни одна связь (так называемый “dangling node”) и равен нулю в другом случае, $\mathbf{e}$ - вектор состоящей из n единиц, n – число узлов в сети. И чтобы удовлетворить оставшимся двум условиям запишем матрицу $\mathbf{G}$:

$$\begin{aligned} \mathbf{G} &= \alpha \cdot \mathbf{S} + (1-\alpha) \cdot \frac{1}{n} \cdot \left(\mathbf{e} \cdot \mathbf{e}^T \right) = \\ &= \alpha \cdot \mathbf{H} + \left(\alpha \cdot \mathbf{a} + (1-\alpha) \cdot \mathbf{e} \right) \cdot \frac{1}{n} \cdot \mathbf{e}^T, \end{aligned} \qquad (2.3.10)$$

где α – так называемы коэффициент затухания. Он означает, что пользователь продолжит переходить по ссылкам, имеющимся на текущей странице, с вероятностью α, которое находится в интервале от нуля до единицы. Обычно принимают $\alpha = 0.85$. Итак, наша задача сводится к вычислению собственного левого вектора гугловской матрицы $\mathbf{G}$.

Рассмотрим первый контрпример. Для него:

$$\mathbf{H} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad \mathbf{a} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \mathbf{e} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}. \qquad (2.3.11)$$

Из (2.3.10) получаем матрицу $\mathbf{G}$ (для определенности положим $\alpha = 0.85$).

$$\mathbf{G} = \begin{pmatrix} 0.075 & 0.925 \\ 0.5 & 0.5 \end{pmatrix}. \qquad (2.3.12)$$

Находим левые собственные значения матрицы $\mathbf{G}$.

$$\mathbf{r}^T = \begin{pmatrix} 0.351 & 0.649 \end{pmatrix}. \qquad (2.3.13)$$

Видно, что второй узел более значимый, нежели первый, что полностью соответствует интуитивному представлению. Т.к. первый узел ссылается на второй (см. рис. 2.3.2). Совершенно аналогично для второго контрпримера получаем следующие ранги:

$$\mathbf{r}^T = \begin{pmatrix} 05 & 0.5 \end{pmatrix}. \qquad (2.3.14)$$

Как и должно было быть (см рис. 2.3.3), мы получили равнозначные значения для рангов.

Рассмотрим более сложный пример сети, изображенный на рис. 2.3.4.

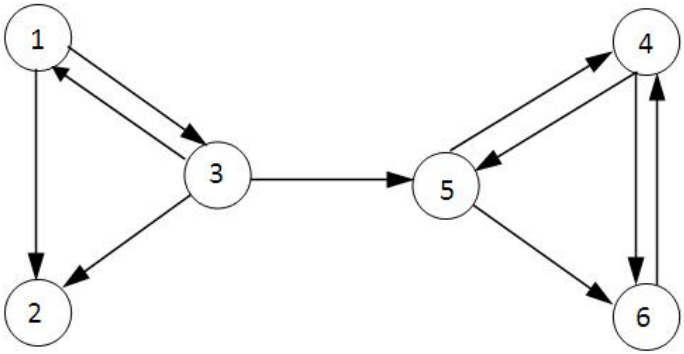


Рис. 2.3.4 – Пример сети

Запишем матрицу **H** и вектор **a** для этой сети.

$$\mathbf{H} = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}. \qquad (2.3.15)$$

$$\mathbf{a}^T = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}. \qquad (2.3.16)$$

Найдем матрицу $\mathbf{G}$, взяв $\alpha = 0.85$:

$$\mathbf{G} = \begin{pmatrix} 0.025 & 0.45 & 0.45 & 0.025 & 0.025 & 0.025 \\ 0.167 & 0.167 & 0.167 & 0.167 & 0.167 & 0.167 \\ 0.308 & 0.308 & 0.025 & 0.025 & 0.308 & 0.025 \\ 0.025 & 0.025 & 0.025 & 0.025 & 0.45 & 0.45 \\ 0.025 & 0.025 & 0.025 & 0.45 & 0.025 & 0.45 \\ 0.025 & 0.025 & 0.025 & 0.875 & 0.025 & 0.025 \end{pmatrix}. \qquad (2.3.17)$$

Находим левый собственный вектор (для собственного значения - единица):

$$\mathbf{r}^T = \begin{pmatrix} 0.052 & 0.074 & 0.057 & 0.349 & 0.2 & 0.269 \end{pmatrix}. \qquad (2.3.18)$$

Для наглядности перенумеруем узлы в соответствии с их рангами – первый номер присвоим узлу с самым большим рангом – см. рис. 2.3.5.

Можно отметить, что без вычислений (интуитивно) даже такую несложную сеть практически невозможно ранжировать.

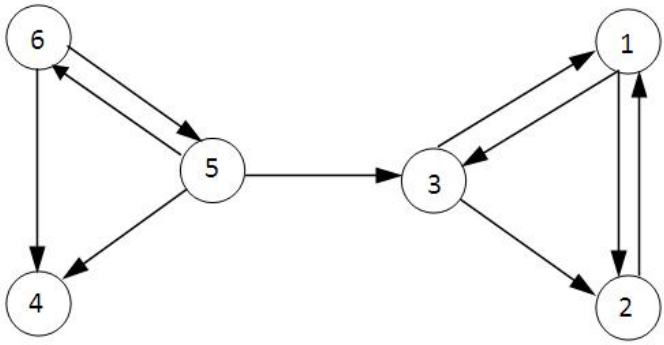


Рис. 2.3.5 – Сеть с ранжированными узлами

2.3.3. Алгоритм Salsa

Алгоритм ранжирования Salsa (Stochastic Approach for Link-Structure Analysis - Стохастический Алгоритм Анализа Структуры Связей) был предложен Ш. Мораном (Sh. Moran) и Р. Лемпелем (R. Lempel) как некоторый симбиоз алгоритмов PageRank и HITS, позволяющий сократить последствия образования так называемых TKC (Tightly-Knit Community) – тесно связанных сообществ». Этот эффект заключается в наличии в выдаче посковой системы множества документов, тесно связанных между собой, тематика которых несколько отличается от информационной потребности пользователя, т.е. наблюдается эффект «сдвига» тематики (topic draft) при отображении результатов, часть из которых может отклоняться от доминирующей тематики.

Lempel R. and Moran S. The stochastic approach for link-structure analysis (SALSA) and the TKC effect // In Proc. of the 9th International WWW Conference, Amsterdam, The Netherlands, 2000. - pp. 387–401.

Как и в методе PageRank в случае Salsa предполагается модель случайного блуждания пользователя по сети (веб-графу), однако предполагается наличие двухстороннего «серфинга». В соответствии с алгоритмом Salsa заданное заранее количество раз выполняется простая двухшаговая процедура:

1. Из произвольного узла v, пользователь случайным образом возвращается к узлу u, – случайно выбранному из множества узлов, ссылающихся на v. Выбор v делается случайно (при этом узлы v и u принадлежат сети).

2. Из узла u наугад осуществляется переход к узлу w, если существует связь (u, w).

Веб-граф G (рис. 2.3.6 *а*) может быть преобразован в двудольный ненаправленный граф G_{bip}, (рис. 2.3.6 *б*) и определен как совокупность $G_{bip} = (V_h, V_a, E)$, где h обозначает посредников, V_h - совокупность узлов-посредников (тех, из которых исходят ссылки), a – авторов, V_a - совокупность узлов-авторов (тех, на которые ведут ссылки).

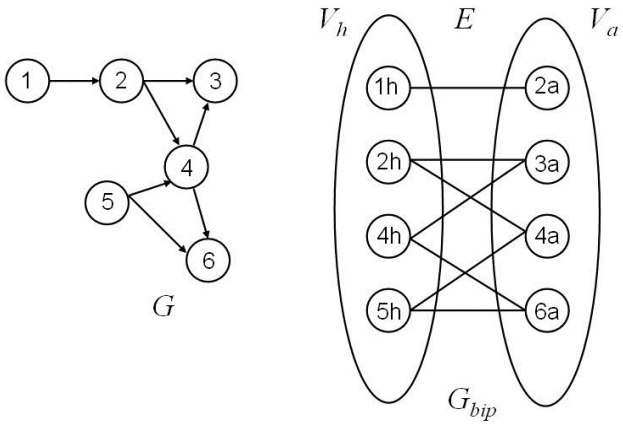


Рис. 2.3.6 – Salsa: конструкция двудольного графа

Необходимо отметить, что одни и те же узлы могут быть одновременно и авторами и посредниками.

Каждый неизолированный узел $s \in G$ представлен в G_{bip} одним или двумя узлами s_h и s_a. В этом двудольном

графе Salsa реализует два разных случайных перехода. При каждом переходе возможно «посещение» узлов только из одной из двух долей графа G_{bip}.

Каждый путь длины два в G_{bip} представляет собой обход одной гиперсвязи (при прохождении от доли посредников к доле авторов в G_{bip}), и отход вдоль гиперсвязи (при прохождении в обратном направлении). Это движение в обратном направлении напоминает танец сальса, который ассоциируется с названием данного алгоритма.

Так как посредники и авторы, относящиеся к теме t должны быть явно выражены в G_{bip} (доступны из многих узлов благодаря прямым ссылкам или коротким путям), предполагается что авторы из V_a и посредники из V_h, относящиеся к теме t, будут наиболее часто посещаемыми при случайных «блужданиях» пользователей.

В алгоритме Salsa исследуются две различных цепи Маркова, которые ассоциируются с этими случайными блужданиями: цепь на стороне авторов G_{bip} (цепь авторов), и цепь на стороне посредников G_{bip}.

Такой подход позволяет ввести две стохастические матрицы перехода цепей Маркова, которые определяются следующим образом: строится матрица инциденций W ориентированного графа G. Обозначим как W_r матрицу, полученную делением каждого ненулевого элемента W на сумму значений соответствующей строки, а через W_c - матрицу, полученную делением каждого ненулевого элемента W на сумму элементов в соответствующем столбце. Тогда, матрица H, соответствующая посредникам будет состоять из ненулевых строк и столбцов $W_r W_c^T$, а матрица авторов A, соответственно, будет состоять из ненулевых строк и столбцов $W_c^T W_r$. В рамках алгоритма Salsa игнорируются строки и столбцы матриц A и H, которые состоят полностью из нулей, так как по определению, все узлы G_{bip} имеют не менее одной связи. В результате матрицы A и H

используются для вычисления рангов тем же путем, что и в алгоритме HITS.

Сходящиеся в процессе итерационного процесса вероятность перехода к узлу v как к автору, имеет очень простую форму:

$$\pi_v = c_1 \cdot InDegree(v), \qquad (2.3.19)$$

а вероятность возврата к узлу u как к посреднику:

$$\pi_u = c_2 \cdot OutDegree(u), \qquad (2.3.20)$$

где c_1 и c_2 - некоторые константы, а $InDegree$ и $OutDegree$ - это количество исходящих и входящих ссылок, соответственно.

Р. Лемпель и С. Моран продемонстрировали, что алгоритм Salsa менее чувствителен к эффекту тесно связанных сообществ, чем HITS, но при условиях, что вручную в документах удаляются ссылки, не относящиеся к исследуемой теме. Это требование на практике ведет к большим издержкам, в результате чего авторам пока не известно случаев использования этого алгоритма ранжирования в реально работающих системах.

Многие из реальных сложных сетей являются контентными, т.е. такими, в узлах которых хранятся текстовые документы (веб-страницы, сообщения блогов, официальные документы и т.п.). Поэтому остановимся детально на методах группирования, а именно на задачах их классификации и кластеризации. Эти методы позволяют объединять различные узлы сети в категории, что существенно влияет на сокращение количества различных содержательных объектов в сетях. При дальнейшем изложении будем называть все объекты документами, а параметры объектов – терминами, что не ограничивает общность последующих выводов.

2.4. Классификация

2.4.1. Формальное описание классификации

Пусть $D = \{d_1, ..., d_{|D|}\}$ - множество объектов (узлов сети или, например, их содержательных элементов – документов),

$C = \{c_1, ..., c_{|C|}\}$ – множество категорий, Ф – целевая функция, которая по паре $< d_i, c_j >$ определяет, относится ли документ d_i к категории c_j (1 или True) или нет (0 или False). Задача классификации состоит в построении функции Ф', максимально близкой к Ф.

Методы машинного обучения, которые применяются для классификации, предусматривают наличие коллекции заранее классифицированных экспертами объектов, т.е. таких, для которых уже точно известно значение целевой функции. Для того чтобы после построения классификатора можно было оценить его эффективность, эта коллекция разбивается на две части, не обязательно равного размера:

1. Учебная (training-and-validation, TV) коллекция. Классификатор Ф' строится на основе характеристик этих объектов.

2. Тестовая (test, Te) коллекция. На ней проверяется качество классификации. Объекты из Te не должны использоваться в процессе построения классификатора.

Рассматриваемая классификация называется четкой бинарной, т.е. подразумевается, что существуют только две категории, которые не пересекаются. К такой классификации сводится много задач, например, классификация по множеству категорий $C = \{c_1, ..., c_{|C|}\}$ разбивается на $|C|$ бинарных классификаций по множествам $\{c_i, \overline{c_i}\}$.

Часто используется ранжирование, при котором множество значений целевой функции - это отрезок [0, 1]. Объект при ранжировании может относиться не только к одной, а сразу к нескольким категориям с разной степенью принадлежности, т.е. категории могут пересекаться между собой.

2.4.2. Ранжирование и четкая классификация

Предположим, что для каждой категории c_i построена

фунция CSV_i. Рассмотрим задачу, заключающуюся в том, чтобы от функции ранжирования перейти к точной классификации. Наиболее простой способ - для каждой категории c_i выбрать предельное значение (порог) τ_i. Если $CSV_i(d) > \tau_i$, то документ d соответствует категории c_i. Другой подход: для каждого документа d выбирать k ближайших категорий, т.е. k категорий, на которых $CSV_i(d)$ принимают наибольшие значения.

Выбирать пороговое значение можно несколькими способами:

- Пропорциональный метод. Учебная коллекция разбивается на две части. Для каждой категории c_i на одной части учебной коллекции вычисляется, какая часть документов ей принадлежит. Пороговые значения выбирается так, чтобы на другой части учебной коллекции количество оставшихся документов, отнесенных c_i, было таким же.
- Метод k ближайших категорий. Каждый документ d_i считается принадлежащим к k ближайшим категориям и соответственно этому выбирается пороговое значение.

CSV (*Categorization Status Value* – статус классификации) – функция, отображающая множество документов *D* на отрезок [0; 1], которая задает степень принадлежности документа категории.

2.4.3. Мера близости объекта и категории

В этом методе правилом классификатора является скалярное произведение. Пусть каждой категории C_i соответствует вектор $\boldsymbol{C_i} = (c_{i1}, ..., c_{iN})$, где N - размерность пространства термов. В качестве правила классификатора используется формула:

$$CSV_i(d) = \boldsymbol{d} \cdot \boldsymbol{C}_i = \sum_{j=1}^{N} c_{ij} d_j. \qquad (2.4.1)$$

Нормализация проводится обычно таким образом,

чтобы итоговая формула для $CSV_i(d)$ – это нормированное скалярное произведение - косинус угла между вектором категории c_i и вектором из весовых значений термов, входящих в документ d – $\boldsymbol{d} = (d_1, \ldots, d_N)$:

$$CSV_i(d) = \frac{\boldsymbol{d} \cdot \boldsymbol{C_i}}{|\boldsymbol{d}| \cdot |\boldsymbol{C_i}|}. \qquad (2.4.2)$$

Координаты вектора C_i определяются в ходе обучения, которое проводится по каждой категории независимо от других.

2.4.4. Метод Rocchio

Профайл (profile) – прототип документа, категории или массива документов, чаще всего совокупность взвешенных термов

Некоторые классификаторы используют так называемый профайл для определения категории. Профайл – это список взвешенных термов, присутствие или отсутствие которых позволяет наиболее точно отличать конкретную категорию от других категорий. К таким методам классификации относится и метод Rocchio, который относится к линейным классификаторам, в которых каждый документ представляется в виде вектора весовых значений термов. Профайл категории i будем рассматривать как вектор $\boldsymbol{C_i} = (c_{1i}, \ldots, c_{Ni})$ (N – количество термов в словаре), значения элементов которого c_{ki} в рамках метода Rocchio рассчитывается по формуле:

$$c_{ki} = \frac{\alpha}{|POS_i|} \cdot \sum_{d_j \in POS_i} w_{kj} - \frac{\beta}{|NEG_i|} \cdot \sum_{d_j \in NEG_i} w_{kj}, \qquad (2.4.3)$$

где w_{kj} - это вес терма t_k в документе d_j (рассчитанный, например, по принципу *TF IDF*), $POS_i = \{d_j \mid \Phi(d_j, c_i) = 1\}$ и $NEG_i = \{d_j \mid \Phi(d_j, c_i) = 0\}$. В этой формуле, α и β – контрольные параметры, которые характеризуют значимость

положительных и отрицательных примеров. Например, если $\alpha = 1$ и $\beta = 0$, C_i будет центром масс всех документов, относящихся к соответствующей категории.

Функция $CSV_i(d)$ определяется либо как величина обратная расстоянию от вектора из весовых значений термов, входящих в документ d, до профайла категории i – C_i, либо как скалярное произведение этих векторов. Метод Rocchio дает удовлетворительные результаты когда документы из одной категории близки друг к другу по расстоянию.

2.4.5. Метод линейной регрессии

Регрессионный анализ используется, когда признаки категорий могут быть выражены количественно в виде некоторой комбинации векторов весовых значений термов, входящих в документы из учебной коллекции. Полученная комбинация может использоваться для определения категории, к которой будет относиться новый документ. В простейшем случае для решения этой задачи используются стандартные статистические методы, такие как линейная регрессия.

Метод регрессии является вариантом линейной классификации, обучаемой сразу на всей коллекции. При применении регрессионного анализа к классификации текстов рассматривается множество термов (F) и множество категорий (C). В этом случае учебной коллекции документов ставится в соответствие две матрицы:

- матрица документов $\boldsymbol{D}$ в учебной коллекции, в которой каждая строка – это документ, а столбец – терм, количество строк N – количество документов в учебной коллекции;
- матрица ответов $\boldsymbol{O} = \left\| o_{i,j} \right\|$, в которой строка i соответствует документу ($i = 1, ..., N$), столбец j – категории ($j = 1, ..., K$), а $o_{i,j}$ - значению $CSV_j(d_i)$.

Метод регрессии базируется на алгоритме нахождения

матрицы правил $\boldsymbol{M}$, которая минимизирует значение нормы матрицы $\|\boldsymbol{MD} - \boldsymbol{O}\|_F$, то есть:

$$M = \arg\min_{M} \|\boldsymbol{MD} - \boldsymbol{O}\|_F . \qquad (2.4.4)$$

Напомним, что в линейной алгебре под нормой матрицы понимается функция, которая ставит в соответствие матрице числовую характеристику. Норма матрицы отражает порядок величины матричных элементов. В данном случае рекомендуется использовать норму Фробениуса $\|\cdot\|_F$, равную корню квадратному из суммы квадратов всех элементов соответствующей матрицы:

$$\|A\|_F = \sqrt{\sum_{i,j} a_{ij}^2}. \qquad (2.4.5)$$

Элемент m_{ij} искомой матрицы $\boldsymbol{M}$ будет отражать степень принадлежности i–го терма j–й категорий.

2.4.6. ДНФ-классификатор

ДНФ-классификатор состоит из множества правил, условия которых задаются некоторой ДНФ-формулой (ДНФ - дизъюнктивная нормальная форма), представляющей собой дизъюнкцию нескольких выражений, элементы которых соединены некоторым количеством (возможно, нулевым) конъюнкций. В этом случае документ относится к категории, если он удовлетворяет этой формуле, т.е. удовлетворяет хотя бы одному члену дизъюнкции.

На начальной стадии для каждой категории c_i, которая состоит из документов $\{d_1^i, \ldots, d_{|c_i|}^i\}$, определяется следующая формула:

ЕСЛИ $(x = d_1^i)$ ИЛИ $(x = d_2^i)$ ИЛИ . . . ИЛИ $(x = d_{|c_i|}^i)$, ТО c_i.

Классификатор, основанный на таком множестве формул, абсолютно правильно работает на учебной коллекции, но, во-первых, он не может работать на других документах, во-вторых, пользоваться таким

классификатором неудобно ввиду большого количества правил. В реально работающих ДНФ-классификаторах происходит переход от документов к множествам термов, которые определяются на основании анализа содержания документов, принадлежащих той или иной категории. Кроме того, проводится ряд упрощений, связанных с объединением или удалением некоторых условий. Приведем небольшой пример:

ЕСЛИ ((кофе & эспрессо) ИЛИ

(кофе & молоко) ИЛИ

(чай & стакан & лимон) ИЛИ

(кофе & чашка & ¬зерна))

ТО Напиток

ИНАЧЕ ¬Напиток

Подобные действия улучшают показатель полноты классификации, но при этом может существенно пострадать точность даже на учебной коллекции.

2.4.7. Байесовская логистическая регрессия

В модели байесовской логистической регрессии рассматривается условная вероятность принадлежности документа D классу C: $p(C|D)$.

Предполагается, что документ определяется термами, входящими в него, т.е. в рамках данной модели документ – это вектор: $D=(w_1, ..., w_N)$, где w_i - вес терма i, а N – размер словаря.

Модель байесовской логистической регрессии задается формулой:

$$p(C|D)=\varphi(\beta\cdot D)=\varphi(\sum_{i=1}^{N}\beta_i\cdot w_i), \qquad (2.4.6)$$

где $C\in\{0,1\}$, $\beta=\{\beta_1, ..., \beta_N\}$ - вектор параметров модели, а φ – логистическая функция, в качестве которой рекомендуется использовать:

$$\varphi(x) = \frac{1}{1+\exp(-x)}. \qquad (2.4.7)$$

Основная идея подхода состоит в том, чтобы использовать предшествующее распределение вектора параметров β, в котором каждое конкретное значение β_i с большой вероятностью может принимать значение, близкое к 0. При реальных расчетах принимаются гипотезы о Гауссовском или Лапласовом распределении значений β_i, а также то, что все величины β_i взаимно независимы.

2.4.8. Наивная байесовская модель

Рассматривается условная вероятность принадлежности объекта классу C при том, что он обладает признаками $F_1, \ldots, F_n$:

$$p(C \mid F_1, \ldots, F_n). \qquad (2.4.8)$$

В соответствии с теоремой Байеса:

$$p(C \mid F_1, \ldots, F_n) = \frac{p(C)\,p(F_1, \ldots, F_n \mid C)}{p(F_1, \ldots, F_n)}. \qquad (2.4.9)$$

По определению условной вероятности:

$$\begin{aligned} p(C \mid F_1, \ldots, F_n) &= p(C)\,p(F_1, \ldots, F_n \mid C) = \\ &= p(c)\,p(F_1 \mid C)\,p(F_2, \ldots, F_n \mid C, F_1) = \\ &= p(c)\,p(F_1 \mid C)\,p(F_2 \mid C)\,p(F_3, \ldots, F_n \mid C, F_1, F_2). \end{aligned} \qquad (2.4.10)$$

В соответствии с «наивным» байесовским подходом предполагается, что события F_i, F_j независимы для любых $i \neq j$:

$$p(F_i \mid C, F_j) = p(F_i \mid C). \qquad (2.4.11)$$

Соответственно:

$$p(C \mid F_1,\ ...,\ F_n) =$$
$$= p(C)p(F_1 \mid C)p(F_2 \mid C)\cdot ...\cdot (F_n \mid C) = \qquad (2.4.12)$$
$$= p(C)\prod_{i=1}^{n} p(F_i \mid C).$$

Перейдем к классификации документов. В случае бинарной классификации «наивная» байесовкая вероятность принадлежности документа классу определяется по формуле:

$$p(D \mid C) = \prod_i p(w_i \mid C). \qquad (2.4.13)$$

В соответствии с теоремой Байеса:

$$p(C \mid D) = \frac{p(C)}{p(D)} p(D \mid C). \qquad (2.4.14)$$

Допустим, классификация происходит только по двум классам - C и $\bar{C}$. Тогда в соответствии с формулой Байеса имеем:

$$p(C \mid D) = \frac{p(C)}{p(D)} \prod_i p(w_i \mid C); \qquad (2.4.15)$$

$$p(\bar{C} \mid D) = \frac{p(\bar{C})}{p(D)} \prod_i p(w_i \mid \bar{C}). \qquad (2.4.16)$$

В качестве критерия принадлежности документа к категории рассматривается следующее отношение вероятностей:

$$\frac{p(C \mid D)}{p(\bar{C} \mid D)} = \frac{p(C)}{p(\bar{C})} \prod_i \frac{p(w_i \mid C)}{p(w_i \mid \bar{C})}. \qquad (2.4.17)$$

На практике используется логарифм отношения вероятностей:

$$\ln \frac{p(C \mid D)}{p(\bar{C} \mid D)} = \ln \frac{p(C)}{p(\bar{C})} + \sum_i \ln \frac{p(w_i \mid C)}{p(w_i \mid \bar{C})}. \qquad (2.4.18)$$

Если выполняется неравенство $\ln\frac{p(C\mid D)}{p(\overline{C}\mid D)}>0,$ то считается, что документ D относится к категории C.

2.4.9. Метод опорных векторов

В.Н. Вапник

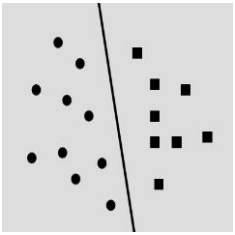
SVM – Support Vector Mashine

Vapnik V.N. Statistical Learning Theory. – NY: John Wiley, 1998.

Метод опорных векторов (Support Vector Mashine, SVM), предложенный В.Н. Вапником, относится к группе граничных методов классификации. Он определяет принадлежность объектов к классам с помощью границ областей.

Рассматривается бинарная классификация, т.е. только по двум категориям C и $\overline{C}$ (принимается во внимание то, что этот подход может быть расширен на любое конечное количество категорий). Кроме того предполагается, что каждый объект классификации является вектором в N-мерном пространстве. Каждая координата вектора - это некоторый признак, количественно тем больший, чем больше этот признак выражен в данном объекте.

Предполагается, что существует учебная коллекция – это множество векторов $\{x_1, ..., x_n\}\in R^N$ и чисел $\{y_1, ..., y_n\}\in\{-1, 1\}$. Число y_i равно 1 в случае принадлежности соответствующего вектора x_i категории c, и -1 – в противном случае.



Разделение классов прямой

Как было показано выше, линейный классификатор - это один из простейших способов решения задачи классификации. В этом случае ищется прямая (гиперплоскость в N-мерном пространстве), отделяющая все точки одного класса от точек другого класса. Если удается

найти такую прямую, то задача классификации сводится к определению взаимного расположения точки и прямой: если новая точка лежит с одной стороны прямой (гиперплоскости), то она принадлежит классу c, если с другой – классу $\overline{c}$.

Формализуем эту классификацию: необходимо найти вектор $\boldsymbol{w}$ такой, что для некоторого предельного значения b и новой точки x_i выполняется:

$$y_i = \begin{cases} +1, \ \textit{если } w \cdot x_i \geq b, \\ \\ -1, \ \textit{если } w \cdot x_i < b, \end{cases} \qquad (2.4.19)$$

где $\boldsymbol{w} \cdot \boldsymbol{x}_i$ - скалярное произведение векторов w и $\boldsymbol{x}_i$:

$$\boldsymbol{w} \cdot \boldsymbol{x}_i = \sum_{j=1}^{N} w_j x_{i,j}. \qquad (2.4.20)$$

Уравнение $\boldsymbol{w} \cdot \boldsymbol{x}_i = b$ описывает гиперплоскость, которая разделяет классы. То есть, если скалярное произведение вектора w на x_i не меньше значения b, то новая точка принадлежит первому классу, если меньше – ко второму. Известно, что вектор $\boldsymbol{w}$ перпендикулярен искомой разделяющей прямой, а значение b зависит от кратчайшего расстояния между разделяющей прямой и началом координат. Очевидно, если существует одна разделяющая прямая, то она не единственная. Возникает вопрос, какая из прямых разделяет классы лучше всего?

Метод SVM базируется на таком постулате: наилучшая разделяющая прямая – это та, которая максимально далеко отстоит от ближайших до нее точек обоих классов. То есть задача SVM состоит в том, чтобы найти такие вектор w и число b, чтобы для некоторого $\varepsilon > 0$ (половина ширины разделяющей поверхности) выполнялось:

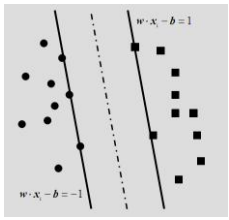
$$\begin{cases} \boldsymbol{w} \cdot \boldsymbol{x}_i \geq b + \varepsilon \ \Rightarrow \ y_i = +1, \\ \\ \boldsymbol{w} \cdot \boldsymbol{x}_i \leq b - \varepsilon \ \Rightarrow \ y_i = -1. \end{cases} \qquad (2.4.21)$$

Умножим после этого обе части неравенства на $1/\varepsilon$ и, не ограничивая общности, выберем ε равным единице. Таким

образом, для всех векторов x_i из учебной коллекции будет справедливо:

$$\begin{cases} w \cdot x_i - b \geq +1, \text{ если } y_i = +1, \\ w \cdot x_i - b \leq -1, \text{ если } y_i = -1. \end{cases} \quad (2.4.22)$$

Условие $-1 < w \cdot x_i - b < 1$ задает полосу, которая разделяет классы. Границами полосы являются две параллельные гиперплоскости с направляющим вектором w. Точки, ближайшие к разделяющей гиперплоскости, расположены точно на границах полосы.



Разделяющая полоса

Чем шире полоса, тем увереннее можно классифицировать документы, соответственно, в методе SVM предполагается, что самая широкая полоса является наилучшей.

Сформулируем условия задачи оптимальной разделяющей полосы, определяемой неравенством: $y_i(w \cdot x_i - b) \geq 1$ (так переписывается система уравнений, исходя из того, что $y_i \in \{-1, 1\}$). Ни одна из точек обучающей выборки не может лежать внутри этой разделяющей полосы. При этих ограничениях x_i и y_i - постоянные, как элементы учебной коллекции, а w и b - переменные.

Из геометрических соображений известно, что ширина разделяющей полосы равна $2/\|w\|$. Поэтому необходимо найти такие значения w и b, чтобы выполнялись приведенные линейные ограничения, и при этом как можно меньше была норма вектора w, то есть необходимо минимизировать:

$$\|w\|^2 = w \cdot w. \quad (2.4.23)$$

Это известная задача квадратичной оптимизации при линейных ограничениях.

Если предположить, что на учебных документах возможно были допущены ошибки экспертами при классификации, то необходимо ввести набор дополнительных переменных $\xi_i \geq 0$, характеризующих величину ошибок на объектах $\{x_1, \ldots, x_n\}$. Это позволяет смягчить ограничения:

$$y_i(w \cdot x_i - b) \geq 1 - \xi_i. \qquad (2.4.24)$$

Предполагается, что если $\xi_i = 0$, то на документе x_i ошибки нет. Если $\xi_i > 1$, то на документе x_i допускается ошибка. Если $0 < \xi_i < 1$, то объект попадает внутрь разделяющей полосы, но относится алгоритмом к своему классу.

Задачу поиска оптимальной разделяющей полосы можно в этом случае переформулировать следующим образом: при определенных ограничениях минимизировать сумму:

$$\|w\|^2 + C\sum_i \xi_i. \qquad (2.4.25)$$

Коэффициент C - это параметр настройки метода, который позволяет регулировать соотношение между максимизацией ширины разделяющей полосы и минимизацией суммарной ошибки. Приведенная задача осталась задачей квадратичного программирования, которую можно переписать в следующем виде:

$$\begin{cases} \dfrac{\|w\|^2}{2} + C\sum_i \xi_i \to \min; \\ y_i(w \cdot x_i - b) + \xi_i \geq 1, \quad i = 1, \ldots, n. \end{cases} \qquad (2.4.26)$$

По известной теореме Куна-Такера такая задача эквивалентна двойственной задаче поиска седловой точки функции Лагранжа:

$$\begin{cases} \frac{1}{2} w \cdot w + C \sum_i \xi_i - \\ \quad - \sum_i \lambda_i (\xi_i + y_i (w \cdot x_i - b) - 1) \to \min\limits_{w,b} \max\limits_{\lambda}; \\ \xi_i \geq 0, \quad \lambda_i \geq 0, \quad i = 1, \ ..., \ n. \end{cases} \qquad (2.4.27)$$

Необходимым условием метода Лагранжа является равенство нулю производных лагранжиана по переменным w и b, откуда получаем:

$$w = \sum_{i=1} \lambda_i y_i x_i, \qquad (2.4.28)$$

т.е. искомый вектор – это линейная комбинация учебных векторов, для которых $\lambda_i \neq 0$. Если $\lambda_i > 0$, то документ обучающей коллекции называется опорным вектором.

Таким образом, уравнение разделяющей плоскости имеет вид:

$$\sum_{i=1} \lambda_i y_i x_i \cdot x - b = 0 . \qquad (2.4.29)$$

Приравняв производную лагранжиана по b нулю, получим:

$$\sum_{i=1} \lambda_i y_i = 0. \qquad (2.4.30)$$

Подставляя последнее выражение и выражение для w в лагранжиан получим эквивалентную задачу квадратичного программирования, содержащую только двойственные переменные:

$$\begin{cases} \sum_i \lambda_i - \frac{1}{2}\sum_{i,j} \lambda_i \lambda_j y_i y_j (x_i \cdot x_j) \to \min\limits_{\lambda}; \\ \sum_{i=1} \lambda_i y_i = 0; \\ C \geq \lambda_i \geq 0, \quad i = 1, \ ..., \ n. \end{cases} \qquad (2.4.31)$$

При этом очень важно, что целевая функция зависит не от конкретных значений x_i, а от скалярных произведений между ними.

Следует отметить, что целевая функция является выпуклой, поэтому любой ее локальный минимум является глобальным.

Метод классификации разделяющей полосой имеет два недостатка:

- при поиске разделяющей полосы важное значение имеют только пограничные точки;
- во многих случаях найти оптимальную разделяющую полосу невозможно.

Для улучшения метода применяется идея расширенного пространства, для чего:

1. Выбирается отображение $\phi(x)$ векторов $\boldsymbol{x}$ в новое, расширенное пространство.
2. Автоматически применяется новая функция скалярного произведения, которая применяется при решении задачи квадратичного программирования, так называемую функцию ядра (kernel function): $K(x, y) = \phi(x) \cdot \phi(y)$. На практике обычно выбирают не отображение $\phi(x)$, а сразу функцию $K(x, y)$, которая могла бы быть скалярным произведением при некотором отображении $\phi(x)$. Функция ядра - главный параметр настраивания машины опорных векторов.
3. Находим разделяющую гиперплоскость в новом

пространстве: с помощью функции $K(x,y)$ устанавливается новая матрица коэффициентов для задачи оптимизации. При этом вместо $x_i \cdot x_j$ подставляются значение $K(x_i, x_j)$, и решается новая задача оптимизации.

4. Найдя $\boldsymbol{w}$ и b, получаем поверхность, которая классифицирует, $w \cdot \phi(x) - b$ в новом, расширенном пространстве.

Приведем более строгое определение функции ядра. Функция $K: X \times X \to R$ называется ядром, если она представима в виде $K(x,y) = \phi(x) \cdot \phi(y)$ при некотором отображении $\phi: X \to H$, где H – пространство со скалярным произведением.

Ядром может быть не всякая функция, однако класс допустимых ядер достаточно широк. Известна теорема Мерсера о том, что функция $K(x,y)$ является ядром тогда и только тогда, когда она симметрична и неотрицательно определена, т.е. когда $\int\limits_X \int\limits_X K(x,y)g(x)g(y)dxdy \geq 0$ для любой функции $g: X \to R$.

Рассмотрим некоторые свойства функций ядра, которые позволяют строить их в практических задачах:

1. Любое скалярное произведение является ядром.
2. Тождественная единица ($K(x,y) = 1$) является ядром.
3. Произведение ядер является ядром.
4. Для любой функции $\phi: X \to R$ произведение $K(x,y) = \phi(x) \cdot \phi(y)$ является ядром.
5. Линейная комбинация ядер с неотрицательными коэффициентами является ядром.
6. Композиция произвольной функции и ядра $K(x,y) = K_0\left(\phi(x), \phi(y)\right)$ является ядром.
7. Если $s: X \times X \to R$ симметричная интегрируемая

функция, то $K(x,y)=\int_X s(x,z)s(y,z)dz$ является ядром.

8. Предел локально-равномерно сходящейся последовательности ядер является ядром.

9. Композиция произвольного ядра K_0 и произвольной функции $f: R \to R$, представимой в виде сходящегося степенного ряда с неотрицательными коэффициентами $K(x,y)=f\left(K_0(x,y)\right)$ является ядром. В частности, ядрами являются функции $f(z)=\exp(z)$ и $f(z)=1/(1-z)$ от ядра.

Например, в системе классификации новостного контента с применением известного пакета LibSVM (http://www.csie.ntu.edu.tw/~cjlin/libsvm) в качестве функции ядра рекомендуется использовать радиальную базисную функцию:

$$K(x,y)=\exp(-\gamma\|x-y\|^2), \qquad (2.4.32)$$

где γ - настраиваемый параметр.

Рассмотрим наглядный пример перехода к расширенному пространству, изображенный на рис. 2.4.1. По всей видимости, круглые и квадратные фигуры не разделяются линейной полосой. Если же «изогнуть» пространство, перейдя к третьему измерению, то эти фигуры можно разделить плоскостью, которая отрезает часть поверхности с квадратными точками. Таким образом, выгнув пространство с помощью отображения $\phi(x)$, можно найти разделяющую гиперплоскость.

Метод SVM обладает такими преимуществами:

- на тестах с документальными массивами превосходит другие методы;
- при выборах разных ядер позволяет эмулировать другие подходы. Например, большой класс нейронных сетей можно представить с помощью SVM с определенными ядрами;
- итоговое правило выбирается не с помощью

некоторых эвристик, а путем оптимизации некоторой целевой функции.

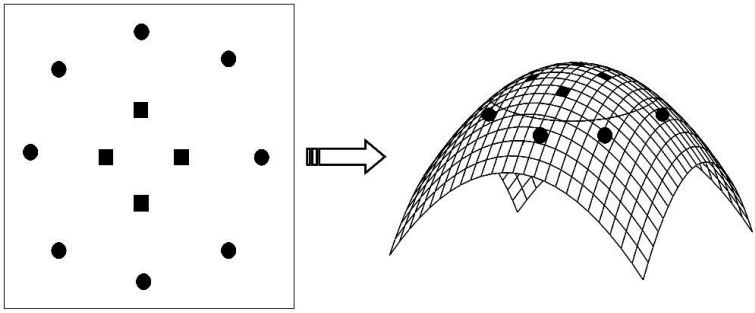


Рис. 2.4.1 – Пример перехода к расширенному пространству

К недостаткам можно отнести:

- мало параметров для настройки: после того как фиксируется ядро, единственным параметром, который варьируется, остается коэффициент ошибки C;
- нет четких критериев выбора ядра;
- довольно медленное обучение системы классификации.

2.5. Кластеризация

Все рассмотренные выше классические модели информационного поиска имеют общий недостаток, связанный с большими размерностями. Для обеспечения эффективной работы необходимо группирование как термов, так и тематически подобных документов. Только в этом случае может быть обеспечена обработка современных информационных массивов в режиме реального времени. В данном случае на помощь приходят два основных приема - классификация и кластеризация. Классификация - это отнесение каждого документа к определенному классу с заранее известными признаками, полученными на этапе обучения системы. Число классов при классификации строго ограничено.

Кластеризация - разбиение множества документов на кластеры - подмножества, смысловые параметры которых

заранее неизвестны. Количество кластеров может быть произвольным или фиксированным. Если классификация допускает приписывание документам определенных, известных заранее признаков, то кластеризация более сложный процесс, который допускает не только приписывание документам некоторых признаков, но и выявление самых этих признаков - классов.

Классификация и кластеризация представляют собой два уровня человеческого участия в процессе группирования документов. Механизм классификации обычно обучается на отобранных документах только после того, как заканчивается стадия обучения путем автоматического выявления классов (кластеров).

Задачей кластеризации документов является автоматическое выявление групп семантически подобных документов [24]. Однако, в отличие от классификации, тематическая ориентация этих групп не известна заранее. Цель всех методов кластеризации массивов документов состоит в том, чтобы подобие документов, которые попадают в кластер, было максимальным. Поэтому методы кластерного анализа базируются на таких определениях кластера, как множества документов, значение семантической близости между любыми двумя элементами которых не меньше определенного порога или значение близости между любым документом множества и центром кластера также не меньше определенного порога.

При использовании численных методов кластерного анализа определения близости используются такие основные метрики:

Евклидово расстояние:

$$D(x_i, x_j) = \sqrt{\sum_{k=1}^{N} (x_{ik} - x_{jk})^2}, \qquad (2.5.1)$$

которое является частным случаем метрики Минковского при $p = 2$:

$$D_p(x_i, x_j) = \left(\sum_{k=1}^{N} (x_{ik} - x_{jk})^p \right)^{\frac{1}{p}}. \qquad (2.5.2)$$

Для группирования документов, представленных в виде векторов весовых значений входящих в них термов, часто используется метрика, базирующаяся на скалярном произведении весовых векторов:

$$Sim(x_i,\ x_j) = \hat{x}_i \cdot \hat{x}_j = \sum_{k=1}^{N} \hat{x}_{ik} \cdot \hat{x}_{jk}, \qquad (2.5.3)$$

где x_i, x_j - документы x_{ik} - элемент матрицы весовых значений термов, входящих в x_i, $(i = 1, ..., N)$, $x_i^{'}$ - нормализированный вектор $\hat{x}_i = x_i / |x_i|$.

Начальным пространством признаков обычно выбирается пространство термов, которое образуется в результате анализа большого массива документов. Для проведения такого анализа используются разные подходы - весовой, вероятностный, семантический и т. д.

В области информационного поиска кластерный анализ чаще всего применяется для решения двух задач - группирования документов в базах данных (информационных массивах) и группирования результатов поиска.

Для статических документальных массивов методы кластерного анализа в настоящее время получили большое развитие и популярность [99, 121, 31]. Вместе с тем открытым остается вопрос применения этих методов к динамично изменяемым информационным потокам, которым присущи, кроме динамики, еще и большие объемы [49, 50].

Методы кластерного анализа находят широкое применение в процедурах ранжирования откликов информационно-поисковых систем, при построении персонализированных папок поиска, персональных поисковых интерфейсов пользователей информационно-поисковых систем.

2.5.1. Метод *k*-means

Итеративный алгоритм кластерного анализа *k*-means (*k*-средних) группировки документов по фиксированному количеству кластеров заключается в следующем: случайным

образом выбирается k векторов, которые определяются как центроиды (наиболее типичные представители) кластеров. Затем k кластеров $\{C_1,\ C_2,\ ...,\ C_k\}$ наполняются – для каждого из векторов, которые остались, некоторым образом определяется близость к центроиду соответствующего кластера. Близость может определяться разными способами, в частности, как нормированное скалярное произведение:

$$Sim(x,\ c^j) = \frac{\sum_{k=1}^{N} x_k c_k^j}{\left|x\right|\left|c^j\right|}, \qquad (2.5.4)$$

где x – документ, c^j $(j = 1,\ ...,\ k)$ – центроид кластера C_j, N – размерность пространства термов.

После этого вектор приписывается к тому кластеру, к которому он наиболее близок. Далее векторы группируются и перенумеровываются соответственно принадлежности к кластерам. Потом для каждого из новых кластеров заново определяется центроид $c^i = (c_1^i,\ ...,\ c_N^i)$ – вектор, наиболее близкий ко всем векторам из данного кластера, координаты которого определяются, например, следующим образом:

$$c_k^i = \frac{1}{|C_i|}\sum_{x \in C_i} x_k. \qquad (2.5.5)$$

После этого снова осуществляется процесс наполнения кластеров, затем вычисление новых центроидов и т.д., пока процесс формирования кластеров не стабилизируется (или если уменьшение суммы расстояния от каждого элемента до центра его кластера меньше некоторого заданного порогового значения).

Алгоритм k-means максимизирует функцию качества кластеризации Q:

$$Q(C_1,\ ...,\ C_k) = \sum_{i=1}^{k}\sum_{x \in C_i} Sim(x, c^j). \qquad (2.5.6)$$

В отличие от метода LSI, k-means может использоваться для группирования динамических информационных потоков

благодаря своей вычислительной простоте – $O(kn)$, где n – количество объектов группирования (документов). Недостатком метода является то, что каждый документ может попасть всего лишь в один кластер.

2.5.2. Иерархическое группирование-объединение

Иерархическое группирование-объединение (Hierarchical Agglomerative Clustering, HAC) начинается с того, что каждому объекту в соответствие ставится отдельный кластер, а затем происходит объединение кластеров, которые наиболее близки друг к другу, в соответствии с выбранным критерием. Алгоритм завершается, когда все объекты объединяются в единый кластер. История объединений образует бинарное дерево иерархии кластеров.

Разновидности алгоритма HAC различаются выбором критериев близости (подобия) между кластерами. Например, близость между двумя кластерами может вычисляться как максимальная близость между объектами из этих кластеров.

Иерархическая кластеризация очень часто применяется при социологическом анализе, в биологии, экономике и т.д. Главным образом там, где заранее неизвестно количество кластеров.

Для иерархической кластеризации необходимо каким-либо образом определить расстояние между узлами нашего графа (сети). Т.е. нам необходимо получить количественную оценку близости узлов, аналогичную расстоянию в обычном евклидовом пространстве. Рассмотрим два наиболее используемых определения. Первое это Евклидово расстояние (Euclidean distance), определяется следующим образом:

$$x_{i,j} = \sqrt{\sum_{k \neq i,j}^{N} \left(A_{ik} - A_{jk} \right)^2}, \qquad (2.5.7)$$

здесь N – число узлов в сети. Евклидово расстояние точно равно нулю для полностью структурно - эквивалентных узлов и увеличивается для узлов, которые не имеют общих соседей.

Второе определение базируется на корреляции Пирсона между строками (столбцами) матрицы инцидентности.

$$x_{i,j} = \frac{\frac{1}{N} \cdot \sum_{k=1}^{N} (A_{ik} - \mu_i) \cdot (A_{ik} - \mu_i)}{\sigma_i \cdot \sigma_j}, \qquad (2.5.8)$$

где

$$\mu_i = \frac{1}{N} \cdot \sum_{j}^{N} A_{ij}, \quad \sigma_i^2 = \frac{1}{N} \cdot \sum_{j}^{N} \left(A_{ij} - \mu_i\right)^2. \qquad (2.5.9)$$

Здесь структурно эквивалентные узлы те, которые имеют большой коэффициент корреляции.

После того как мы любым способом определили матрицу расстояний можно приступать непосредственно к иерархической кластеризации. Остановимся подробнее на агломеративном алгоритме. Вначале опишем идею алгоритма. Пусть у нас есть точки $x_1, x_2 \ldots x_N$ и матрица относительных расстояний x_{ij}. На первом шаге каждую точку считаем отдельным кластером. Затем ближайшие (в смысле расстояния) точки объединяем и считаем одним кластером, и так далее пока не задействуются все точки. На выходе мы получим дерево (дендограмму). При подсчете расстояний между кластерами наиболее часто используют один из двух следующих алгоритмов. Single-link алгоритм считает минимум из возможных расстояний между парами узлов, находящихся в кластере, Complete-link алгоритм считает максимум из этих расстояний.

Приведем пошаговый пример для Single-link алгоритма. Пусть есть граф, показанный на рис. 2.5.1.

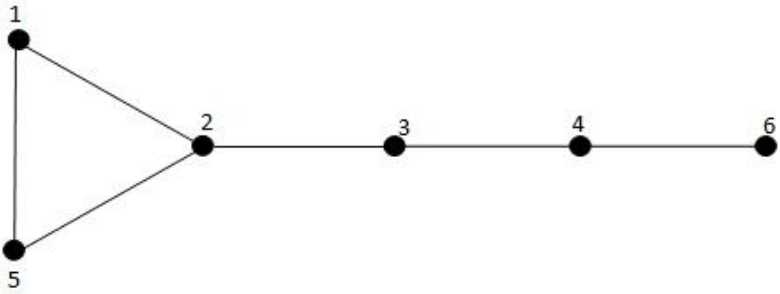


Рис. 2.5.1 – Пример графа

Записав матрицу Евклидовых расстояний (1), видим, что минимальное расстояние равное нулю соответствует узлам 1 и 5. Поэтому на первом шаге объединяем эти узлы в один кластер.

$$x = \begin{array}{c} \\ (1) \\ (2) \\ (3) \\ (4) \\ (5) \\ (6) \end{array} \begin{array}{cccccc} (1) & (2) & (3) & (4) & (5) & (6) \\ 0 & 1 & 1.414 & 2 & 0 & 1.732 \\ 1 & 0 & 1.732 & 1.732 & 1 & 2 \\ 1.414 & 1.732 & 0 & 1.414 & 1.414 & 1 \\ 2 & 1.732 & 1.414 & 0 & 2 & 1 \\ 0 & 1 & 1.414 & 2 & 0 & 1.732 \\ 1.732 & 2 & 1 & 1 & 1.732 & 0 \end{array}$$

На втором шаге снова записываем матрицу расстояний, используя алгоритм Single-link. Следующее расстояние после нулевого расстояния равно единице. И на этом шаге мы получаем два кластера, которые суммарно включают все узлы сети.

Если разрезать дендограмму вдоль пунктирной линии 1 – то мы получим два кластера содержащие узлы (1, 5, 2) и (3, 4, 6) соответственно. Что соответствует и интуитивному разбитию на кластера, см. рис. 1. Если резать вдоль второй пунктирной линии, то получаем три кластера (1, 5), (2) и (3, 4, 6).

	$\binom{1}{5}$	(2)	(3)	(4)	(6)
$\binom{1}{5}$	0	1	1.414	2	1.732
(2)	1	0	1.732	1.732	2
(3)	1.414	1.732	0	1.414	1
(4)	2	1.732	1.414	0	1
(6)	1.727	2	1	1	0

На рис. 2.5.2 показана окончательная дендограмма сети.

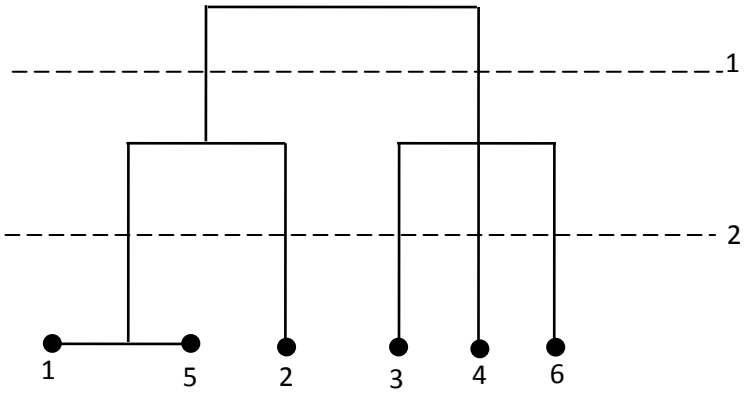


Рис. 2.5.2 – Использован алгоритм Single-link

Если при иерархической кластеризации использовать алгоритм Complete-link, то получим немного другую дендограмму, см рис. 2.5.3.

Отличие в алгоритмах проявится на втором шаге построения. При использовании алгоритма Single-link в один кластер к узлам (4, 6) добавляется узел 3, поскольку он связан с узлом 6 единичным расстоянием, таким же, как и внутрекластеное расстояние. А при использовании Complete-link алгоритма, узел 3 присоединяется на следующем шаге

т.к. здесь необходимо смотреть по максимальному расстоянию, т.е. по расстоянию от узла 3 до узла 4, которое больше внутрикластерного расстояния между узлами 4 и 6.

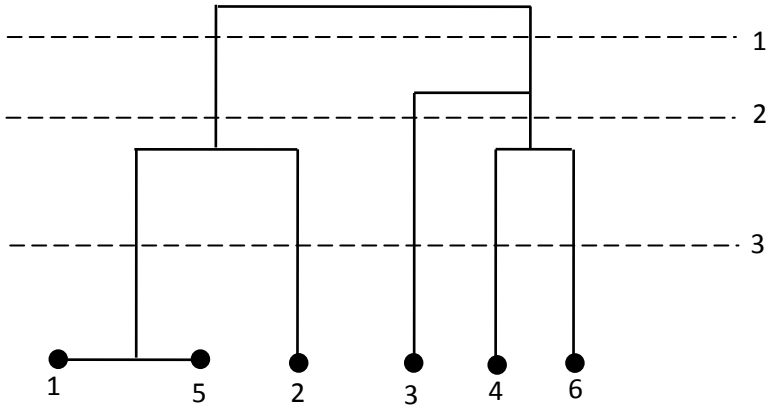


Рис. 2.5.3 – Использован алгоритм Complete-link

2.5.3. Латентно-семантический анализ

Метод кластерного анализа LSA/LSI (латентного семантического анализа/ индексирования) [базируется на сингулярном разложении матриц (SVD, Singular Value Decomposition). Пусть массиву документов $D=\{d_j \mid j=1,\ ...,\ n\}$ ставится в соответствие матрица A, строки которой соответствуют документам, а столбцы – весовым значениям термов (размер словаря термов – m). Сингулярным разложением матрицы A ранга r размерности $m\times n$ называется ее разложение вида $A=USV^T$, где U и V – ортогональные матрицы размерности $m\times r$ и $r\times n$, соответственно, а S – диагональная матрица, элементы которой $s_{ij}=0$, если $i\neq j$, а диагональные элементы $s_{ii}\geq 0$. Диагональные элементы матрицы S называют сингулярными значениями матрицы A.

Ортогональные матрицы U и V обладают таким свойством:

$$UU^T = V^T V = I. \qquad (2.5.9)$$

Доказано, что приведенное выше разбиение матрицы A обладает той особенностью, что если в матрице S оставить только k наибольших сингулярных значений (обозначим такую матрицу как S_k), а в матрицах U и V – только соответствующие этим значениям колонки (соответственно, матрицы U_k, V_k), то матрица $A_k = U_k \cdot S_k \cdot V_k^T$ будет наилучшей по Фробениусу аппроксимацией исходной матрицы A матрицей с рангом, не превышающим k. Напомним, что нормой матрицы X размерности $M \cdot N$ по Фробениусу является выражение:

$$\|X\|_F = \sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N} x_{ij}^2}. \qquad (2.5.10)$$

Таким образом, для матриц A и A_k доказано, что:

$$A_k = \underset{X:rank(X)=k}{\arg\min} \|A - X\|_F. \qquad (2.5.11)$$

В соответствии с методом LSA в рассмотрение берутся не все, а лишь k наибольших сингулярных значений матрицы A, и каждому такому значению ставится в соответствие один кластер.

A_k определяет k-мерное факторное пространство, на которое проецируются, как документы (с помощью матрицы V), так и термины (с помощью матрицы U). В полученном факторном пространстве документы и термины группируются в области, имеющие некоторый общий скрытый смысл. Т.е. получаемые области и представляют собой кластеры.

Выбор наилучшей размерности k для LSA – это проблема отдельных исследований. В идеале, k должно быть достаточно велико для отображения всей реально существующей структуры данных, но в то же время достаточно мало чтобы не учитывать шума - случайных зависимостей.

В практике информационного поиска особое значение отводится матрицам U_k и V_k^T. Строки матрицы U_k рассматриваются как образы термов в $k-$мерном вещественном пространстве. Аналогично, столбцы матрицы V_k^T рассматриваются как образы документов в том же $k-$мерном пространстве. Иными словами, эти векторы задают искомое представление термов и документов в $k-$мерном пространстве скрытых факторов.

Существуют также методы инкрементного обновления всех значений, используемых в LSA. При пополнении новым документом d (например, новым результатом поиска по запросу) информационного массива, для которого уже проведено сингулярное разложение, можно не вычислять разложение заново. Достаточно аппроксимировать его, вычисляя образ нового документа на основе ранее вычисленных образов термов и весов факторов. Пусть $d-$ вектор весов термов нового документа (новый столбец матрицы A), тогда его образ можно вычислить по формуле: $d' = S_k^{-1} U_k^T d.$

Если q – вектор запроса пользователя размерности m, $i-$й элемент которого равен 1, если терм с номером i входит в запрос, и 0 – в противном случае, то образ запроса q в пространстве латентных факторов будет иметь вид: $q' = q^T U_k S_k^{-1}.$

В этом случае мера близости запроса q и документа d оценивается величиной скалярного произведения векторов q' и $V_k^T\{d\}$ (здесь $V_k^T\{d\}$ обозначает $d-$й столбец матрицы V_k^T).

При информационном поиске, в результате того, что отбрасываются наименее значимые сингулярные значения, формируется пространство ортогональных факторов, играющих роль обобщенных термов. В результате происходит «сближение» документов из близких по содержанию предметных областей, частично решаются проблемы синонимии и омонимии термов.

Метод LSA широко применяется при ранжировании выдачи информационно-поисковых систем, основанных на цитировании. Это алгоритм HITS (Hyperlink Induced Topic Search) – один из двух самых популярных сегодня в области информационного поиска. Если вспомнить понятие матрицы инциденций A (п. 2.8), то алгоритм HITS обеспечивает выбор наиболее авторитетных документов (авторов – a_p или посредников – h_p), которые соответствуют собственным векторам матриц AA^T и A^TA с наибольшими модулями собственных значений. Покажем, что алгоритм HITS эквивалентен LSA. Действительно, пусть, в соответствии с сингулярным разложением: $A = USV^T$, S – квадратная диагональная матрица. Тогда $A^TA = USV^TVSU^T = USISU^T = US^2U^T$, где S^2 – диагональная матрица с элементами s_{ii}^2. Очевидно, что как и при LSA, собственные векторы, которые соответствуют наибольшим сингулярным значениям AA^T и/или A^TA, будут соответствовать статистически наиболее важным авторам и/или посредникам.

Наряду с тем, что метод LSA не нуждается в предварительной настройке на специфический набор документов, качественно выявляет латентные факторы, к его недостаткам можно отнести невысокую производительность (скорость вычисления SVD соответствует порядку $O(N^2 \cdot k)$, где $N = |D| + |T|$, D – множество документов, T – множество термов, k – размерность пространства факторов) и то, что он не предусматривает возможность пересечения кластеров, что противоречит практике. Кроме того, ввиду своей вычислительной трудоемкости метод LSA применяется только для относительно небольших матриц.

Часть II. Алгоритмы, методы, феномены

3. Некоторые методы и приемы

Вторая часть пособия представляет собой в значительной мере методические заметки для лектора. Введение этого материала в курс лекций зависит как от предпочтений лектора, так и от подготовленности аудитории.

Часть материала необходима для более глубокого понимания основной части (например, материал, связанный с теорией протекания или теорией фазовых переходов второго рода). Другая часть материала – то, что как считается «знает каждый образованный человек», но как правило (на удивление) не читается в стандартных курсах. В основном, это, так называемые, приемы, позволяющие получить, пусть и приближенно и не всегда математически строго, результат. «Чистых» математиков такие методы не интересуют, а если они и излагаются, то на таком уровне «глубокой» теории, что о конкретном применении и речи не идет.

В качестве примера таких приемов можно назвать метод малого параметра (без которого не обходится ни один прикладник), или, например, исключительно полезный метод аппроксимант Паде.

Даже в том случае, когда задачу удается решить точно, строгое решение может быть настолько громоздким, что удобнее пользоваться приближенным решением. Здесь мы приведем пример из книги Гринберга, который наглядно показывает, что решение полученное «в лоб», стандартным методом разделения переменных, может давать совершенно «несъедобное» решение.

3.0. Простая краевая задача

1. Пример одной простой краевой задачи математической физики, или о том, как надо и как не надо решать задачи.

Граничная задача первого рода или задача Дирихле (внутренняя) для прямоугольника

$$\frac{\partial^2 \varphi}{\partial x^2} + \frac{\partial^2 \varphi}{\partial y^2} = 0. \qquad (3.0.1)$$

Стандартный метод решения – разделение переменных:

$$\varphi(x, y) = X(x)Y(y) \qquad (3.0.2)$$

для аналитической реализации которого потребовалось пять страниц текста.

Гринберг Г.А. Избранные вопросы математической теории электрических и магнитных явлений. – М.; Л.: Академия наук СССР, 1948.

Тихонов А.Н., Самарский А.А. Уравнения математической физики (5-е изд.). – М.: Наука, 1977. – 735 с.

В классических учебниках (например Тихонова и Самарского) практически *нигде* не учат приемам, а учат общим методам.

Г.А. Гринбергу, которому приходилось заниматься расчетом конкретных приборов, необходимо было получать конкретное решение, а не «теорему о его существовании в данном классе функций». Г.А. Гринбергом был разработан метод, позволяющий получать решения с хорошо сходящимися рядами как для самого решения, так и ее производных. Этот метод принципиально отличается от метода разделения переменных – в упомянутой книге этот метод четко сформулирован – подчеркивается, что решение полученное таким образом не является суммой частных решений предложенного уравнения, а дает разложение решения в ряд по некоторым собственным функциям уравнений.

Рассмотрим совсем простой пример, приведенный в книге Гринберга, который даже не требует применения указанного выше метода.

Стандартное решение задачи Дирихле базируется на разделении переменных и приводит к выражению:

$$\varphi = u^{(1)} + u^{(2)}, \qquad (3.0.2)$$

где $u^{(1)}(x, y)$ и $u^{(2)}(x, y)$ являются суммами произведений:

$$u^{(1)} = \frac{2}{b}\sum_{k=1}^{\infty} \frac{sh\frac{k\pi(a-x)}{b}\int_0^b F_1(y)\sin\frac{k\pi y}{b}dy + sh\frac{k\pi x}{b}\int_0^b F_2(y)\sin\frac{k\pi y}{b}dy}{sh\frac{k\pi a}{b}}\sin\frac{k\pi y}{b},$$

$$u^{(2)} = \frac{2}{a}\sum_{k=1}^{\infty} \frac{sh\frac{k\pi(b-y)}{a}\int_0^a \psi_1(x)\sin\frac{k\pi x}{a}dx + sh\frac{k\pi x}{a}\int_0^a \psi_2(x)\sin\frac{k\pi x}{a}dx}{sh\frac{k\pi b}{a}}\sin\frac{k\pi x}{a},$$

(3.0.3)

где граничные условия представлены в общем виде:

$$\begin{aligned} &\varphi|_{x=0} = F_1(y), \varphi|_{x=a} = F_2(y), \\ &\varphi|_{y=0} = \psi_1(x), \varphi|_{y=b} = \psi_2(x). \end{aligned} \qquad (3.0.4)$$

Теперь нас интересует конкретный случай, когда

$$\begin{aligned} &F_1(y) = 0, \quad \psi_1(x) = 0, \\ &F_2(y) = ay, \quad \psi_2(x) = bx, \end{aligned} \qquad (3.0.5)$$

т. е.:

$$\begin{aligned} &\varphi|_{x=0} = 0, \varphi|_{x=a} = ay, \\ &\varphi|_{y=0} = 0, \varphi|_{y=b} = bx. \end{aligned} \qquad (3.0.6)$$

Тогда, согласно вышеизложенному, подставляя эти значения в ряды для $u^{(1)}$ и $u^{(2)}$ находим:

$$\varphi(x,y) = \frac{2ab}{\pi}\left\{\sum_{k=1}^{\infty}(-1)^{k+1}\frac{sh\frac{k\pi x}{b}}{ksh\frac{k\pi a}{b}}\sin\frac{k\pi y}{b} + \sum_{k=1}^{\infty}(-1)^{k+1}\frac{sh\frac{k\pi y}{a}}{ksh\frac{k\pi b}{a}}\sin\frac{k\pi x}{a}\right\}.$$

(3.0.7)

Это весьма сложное выражение, представляющее собой две бесконечные суммы, по виду которого невозможно понять, как ведет себя решение. Легко непосредственно убедиться, что уравнение удовлетворяет условию:

$$\varphi(x,y)=x\cdot y, \qquad (3.0.8)$$

а так как решение единственно, то это и есть решение, т.е. вместо сложных рядов можно использовать лишь простое выражение.

Так в чем же заключаетсяьв рассмотренном случае «прием»? Его суть можно сфрмулировать таким образом: Прежде чем решать уравнение мат. физики, попробуй сделать подстановку для упрощения:

$$\varphi(x,y)=<\text{некоторая функция}>+u(x,y).$$

3.1. Малый параметр

Метод малого параметра применяется во многих областях науки и техники, например, теоретической физике, теории детерминированного хаоса, трудно сказать, где он не применяется. В книге Блехмана, Мышкиса и Пановко отмечено, что «метод возмущений является в прикладной математике одним из самых распространенных».

Блехман И.И., Мышкис А.Д., Пановко Я.Г. Механика и прикладная математика. Логика и особенности приложений математики. – М.: Мир, 1983.

Найфэ А. Введение в методы возмущений. – М.: Мир, 1984.

Мышкис А.Д. Элементы теории математических моделей. – М.: Наука, 1994.

Методу возмущений посвящено большое число монографий, например, монография Найфэ, в котрой приведено множество примеров. Здесь мы рассмотрим только один простой пример – одну из наиболее известных и полезных задач – задачу о ангармоническом осцилляторе. Нулевое приближение этой задачи – одномерный гармонический осциллятор («шарик на пружинке»), описывается линейным уравнением

$$m\frac{d^2x}{dt^2}=-kx, \qquad (3.1.1)$$

где m – масса, k – жесткость пружины, x – отклонение от положения равновесия, $F=-kx$ –закон Гука.

Это уравнение удобно записать в виде:

$$\frac{d^2x}{dt^2} + \omega^2 x = 0\,, \qquad (3.1.2)$$

где $\omega = \sqrt{k/m}$ – циклическая частота.

В таком виде это уравнение описывает огромный класс явлений, не только «шарик на пружинке», но и маятник (в этом случае x – угол отклонения), колебания в LC-контуре (x – заряд конденсатора) и многое другое.

Решение уравнения (2), как легко убедиться простой подстановкой

$$x(t) = x_0 \sin(\omega t + \varphi_0)\,, \qquad (3.1.3)$$

где φ_0 – начальная фаза, а x_0 – амлитуда.

Возможны ситуации, когда линейного приближения недостаточно и закон Гука длжен быть заменен на более сложный, нелинейный

$$F(x) = -kx + \gamma x^2 + \ldots \qquad (3.1.4)$$

где предполагается, что каждый следующий член разложения много меньше предыдущего, т.е.

$$\gamma x_0^2 << k|x_0|\,, \ldots \qquad (3.1.5)$$

Вообще говоря, (3.1.4) это следствие разложения в ряд Тейлора потенциальной энергии $U(x)$, напомним, что сила $F = -gradU$. Колеблющаяся частица находится в потенциальной яме и в простейшем случае, когда яма параболическая $U(x) \sim x^2$ имеет место закон Гука.

В более общем случае, учитывая только два первых слагаемых в разложении (3.1.4) вместо уравнения гармонического осциллятора (3.1.1) получаем нелинейное уравнение

$$m\frac{d^2x}{dt^2} + kx = \gamma x^2\,, \qquad (3.1.6)$$

Решение которого в аналитическом виде не представляется возможным.

Для дальнейшего удобно записать (3.1.6) в безразмерном виде

$$\frac{d^2\psi}{d\tau^2}+\psi=\varepsilon\psi^2, \qquad (3.1.7)$$

где $\psi = x/x_0$ – безразмерная координата, а τ – время, и $\varepsilon = \gamma x_0/\omega$ – безразмерный параметр, который в принятом приближении ($\gamma x_0^2 << k|x_0|$) и является малым параметром $\varepsilon \ll 1$.

Приближенное решение уравнения (3.1.7) будем искать в виде разложения по малому параметру

$$\psi(\tau)=\psi_0(\tau)+\varepsilon\psi_1(\tau)+... \qquad (3.1.8)$$

Подставляя (5) в (4) получаем

$$\frac{d^2\psi_0}{d\tau^2}+\varepsilon\frac{d^2\psi_1}{d\tau^2}+\psi_0+\varepsilon\psi_1=\varepsilon{\psi_0}^2+\varepsilon^2\psi_0\psi_1+... \qquad (3.1.9)$$

В нулевом приближении по малому параметру, т.е. оставляя в (3.1.9) только слагаемые с ε^0 и отбрасывая все слагаемые с ε^1, ε^2 и т.д. имеем

$$\frac{d^2\psi_0}{d\tau^2}+\psi_0=0, \qquad (3.1.10)$$

Линейное уравнение (уравнение гармонического осциллятора), решение которого можно записать в виде

$$\psi_0(\tau)=\sin\tau \ . \qquad (3.1.11)$$

В первом приближении по малому параметру ε в (3.1.9) откидываются слагаемые с ε^2 и выше и получаем

$$\frac{d^2\psi_1}{d\tau^2}+\psi_1={\psi_0}^2. \qquad (3.1.12)$$

Это уравнение линейно по ψ_1, а функция ψ_0 найдена ранее (3.1.11). Решение уравнения (3.1.12) ($\sin^2\tau=(1-\cos 2\tau)/2$) имеет вид

$$\psi_1(\tau)=\frac{1}{2}+\frac{1}{6}\cos 2\tau \ . \qquad (3.1.13)$$

Ограничиваясь первым приближением по ε для (3.1.9) имеем:

$$\psi(\tau)=\psi_0+\varepsilon\psi_1=\sin\tau+\frac{1}{2}\varepsilon+\frac{1}{6}\varepsilon\cos 2\tau. \qquad (3.1.14)$$

Или возвращаясь от безразмерного вида к первоначальным обозначениям

$$x(t)=x_0\sin(\omega t)+\frac{\gamma x_0^2}{2\omega}+\frac{\gamma x_0^2}{6\omega}\cos(2\omega t). \qquad (3.1.15)$$

Таким образом, решение нелинейного уравнения методом малого параметра свелось к последовательному решению линейных уравнений.

Уже первое приближение по малому параметру (3.1.15) позволяет сделать два нетривиальных вывода влияния ангармонизма:

1. кроме колебаний с частотой ω, которые имеют место в линейном случае, появляются колебания с удвоенной частотой
2. средняя точка колебаний, которая в гармоническом осцилляторе была выбрана равной нулю, теперь смещена (см. второе слагаемое в (3.14.15)).

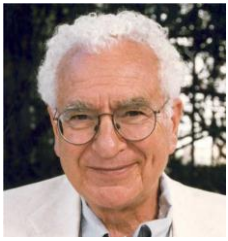
3.2. Асимптотические ряды и разложения

Приходится удивляться, что приближенное решение уравнений с использованием метода разложения по малому параметру не входит в стандартные курсы математического анализа.

Малый параметр обычно обозначают буквой ε/ До каких значений $(\varepsilon \ll 1)$ можно считать малым, – наиболее острый вопрос.

Пример: два ряда $n-$е члены которых пропорциональны: 1) $\frac{1000^n}{n!}$; 2) $\frac{n!}{1000^n}$. Формально 1-й ряд сходится и быстро, его миллионный член всего 1/999999; 2-й ряд расходится.

Пуанкаре: «астрономы наоборот (получив приближенное решение в виде



Мюррей Гелл-Манн, лауреат Нобелевской премии: «На деле всякий теоретик в своей собственной работе полагает какие-то параметры малыми, а затем нападает на других, поступающих так же, обвиняя их в неестественности».

таких рядов, с *конечным* числом членов) будут считать 1-й ряд расходящимся, т.к. первые 1000 членов ряда возрастают; а 2-й – сходящимся, т.к. первые 1000 его членов убывают.

Для того, чтобы найти сумму ряда $\sum_n a_n$, необходимо около 140 слагаемых, причем, прежде чем a_n станут достаточно маленькими и перестанут изменять сумму, их значения вырастут до величины $\sim 10^{42}$ (см. рис. 3.2.1).

Рассмотрим теперь два ряда, для $x \ll 1$

$$S_1(x,N) = \sum_k a_k x^k = \sum_{k=1}^{N} \frac{100^k}{k!} x^k \qquad (3.2.1)$$

и

$$S_2(x,N) = \sum_k b_k x^k = \sum_{k=1}^{N} \frac{k!}{100^k} x^k \qquad (3.2.2)$$

Первый ряд $S_1(x,N)$ – разложение по малому параметру $x \ll 1$, его сумма тем точнее задает функцию, чем больше N. Ряд при $N \to \infty$ сходится.

Второй ряд при $N \to \infty$ расходится, и как нас учат в стандартных курсах математического анализа пользоваться им не имеет смысла.

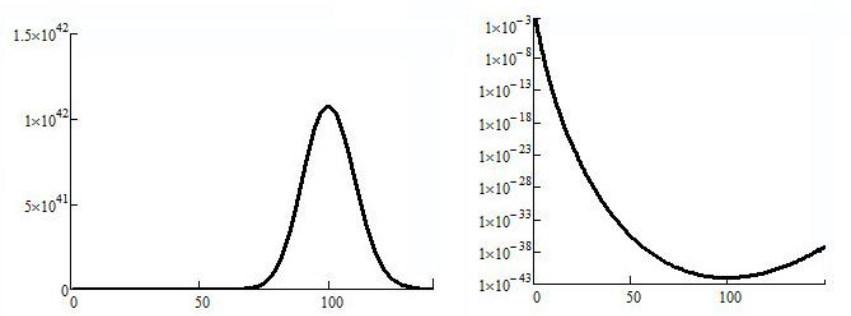


Рис. 3.2.1 – Поведение a_n и b_n при росте n

На практике происходит все наоборот. Не смотря на сходимость ряда $S_1(x,N)$ пользоваться им исключительно неудобно, число членов ряда должно быть (см. рис. 3.2.2) не

менее 20. Что же касается второго ряда $S_2(x,N)$, то при $x \ll 1$ уже при малых значениях N сумма перестает расти.

Ниже приводятся примеры асимптотических рядов для двух функций.

Функция $f(x)$ задана интегралом:

$$f(x) = \int_0^\infty \frac{xe^{-\xi}}{x+\xi} d\xi. \qquad (3.2.3)$$

Необходимо получить поведение $f(x)$ при $x \gg 1$.

Разлагаем $x/(x+\xi)$ в ряд по степеням ξ:

$$\frac{x}{x+\xi} = \sum_{n=0}^{\infty} \frac{(-1)^n \xi^n}{x^n}. \qquad (3.2.4)$$

Жюль Анри Пуанкаре (1854 - 1912)

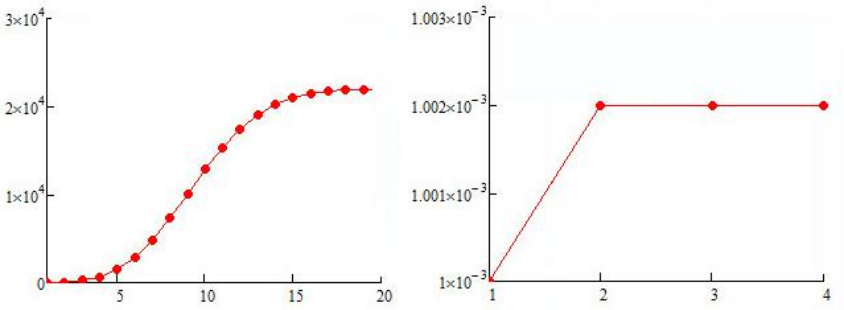


Рис. 3.2.2 – Зависимость суммы радов $S_1(x,N)$ и $S_2(x,N)$ от N

Подставляя это разложение в интеграл находим:

$$f(x)=\sum_{n=0}^{\infty}\frac{(-1)^n n!}{x^n} \qquad (3.2.5)$$

Этот ряд, конечно, расходится т.е. эта сумма равна бесконечности. Тем не менее, его можно использовать. Для этого представим это сумму в виде двух слагаемых – усеченного ряда и остатка:

$$f(x)=\sum_{n=0}^{N}\frac{(-1)^n n!}{x^n}+R_N(x) \qquad (3.2.6)$$

Можно показать, что

$$\left|R_N(x)\right|<\frac{N!}{x^N} \qquad (3.2.7)$$

т.е. эта замена $f(x)$ усеченным на N-ом члене рядом дает ошибку, не превышающую первого отброшенного члена, а именно $N+1$-го. Таким образом, при фиксированном N и $x\gg 1$ $(x\to\infty)$ ошибка $R_N\to 0$ (как угодно мала) и $f(x)$ можно представить в виде асимптотического ряда

$$f(x)\sim\sum_{n=0}^{N}\frac{(-1)^n n!}{x^n} \qquad (3.2.8)$$

В рассмотренном примере малым параметром является $\varepsilon=1/x$, что дает, например, для $\varepsilon=0.1$ на десятом шаге ошибку $R_{10}(x=1/\varepsilon=10)\approx 4\cdot 10^{-4}$, что вполне приемлемо для многих приближенных расчетов. Если же малый параметр еще в десять раз меньше $\varepsilon=0.01$, то погрешность становится ничтожной $R_{10}(x=1/\varepsilon=100)\approx 4\cdot 10^{-14}$. Теперь можно кратко сформулировать различие между разложением по малому параметру и асимптотическим разложением. В случае разложения по малому параметру предельный переход в сумме от $n=0$ до N имеет вид:

$\varepsilon\ll 1$ - конечно, $N\to\infty$, (3.2.9)

а для асимптотического ряда

Андрианов И. В., Баранцев Р. Г., Маневич Л. И. Асимптотическая математика и синергетика: путь к целостной простоте. – М.: Эдиториал УРСС, 2004.

Андрианов И. В., Маневич Л. И. Асимптотология: идеи, методы, результаты. – М.: Аслан, 1994.

$\varepsilon \to 0$, N - конечно. (3.2.10)

Приведем второй пример, разложениие функции Бесселя нулевого порядка $J_0(x)$ при $x \gg 1$ (малый параметр $\varepsilon = 1/x$). Асимптотическое разложение $J_0(x)$ имеет вид

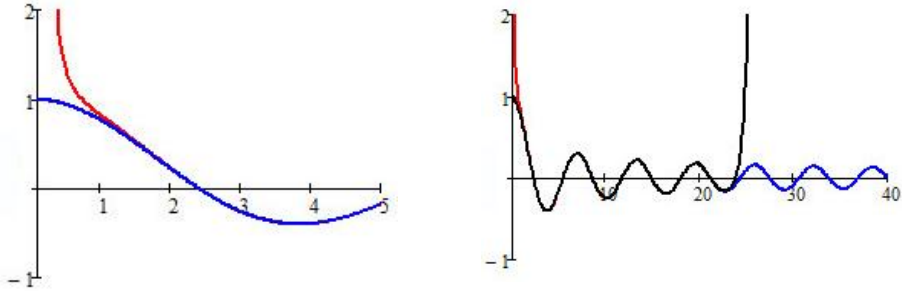
$$J_0(x) = \sqrt{\frac{2}{\pi x}}\left[u(x)\cos\left(x-\frac{\pi}{4}\right)+v(x)\sin\left(x-\frac{\pi}{4}\right)\right], \quad (3.2.11)$$

где

$$u(x) = 1 - \frac{1^2 \cdot 3^2}{4^2 \cdot 2^2 \cdot 2! x^2} + \frac{1^2 \cdot 3^2 \cdot 5^2 \cdot 7^2}{4^4 \cdot 2^4 \cdot 4! x^4} + \ldots \quad (3.2.12)$$

$$v(x) = \frac{1}{4 \cdot 2x} - \frac{1^2 \cdot 3^2 \cdot 5^2}{4^3 \cdot 2^3 \cdot 3! x^3} + \ldots \quad (3.2.13)$$

Ряды $u(x)$ и $v(x)$ расходятся, однако как видно из рис. 3.2.3 уже первые несколько членов ряда прекрасно описывают $J_0(x)$ при $x > 1$, т.е. при малом параметре порядка единицы.



а б

Рис. 3.2.3. Функция Бесселя нулевого порядка: а – $J_0(x)$ на промежутке $(0 \div 5)$, 1 – точное значение, 2 – асимптотическое разложение; б – на промежутке $(1 \div 40)$, сплошная линия – точное значение $J_0(x)$ и асимптотическое разложение

3.3. Паде-аппроксиманты и разложение по малому параметру

Паде-аппроксимацией функции $f(x)$, разложенной в степенной ряд

$$f(x)=\sum a_n x^n , \qquad (3.3.1)$$

называется дробно-степенная функция

$$R(x)=\frac{\sum_{n=0}^{N} p_n x^n}{\sum_{m=0}^{M} q_m x^m}, \qquad (3.3.2)$$

такая, что при разложении в степенной ряд совпадает с разложением $f(x)$ с точностью до коэффициентов при x^{N+M}.

Г. Стенли «Фазовые переходы и критические явления» М.: Мир, 1973.

За этим сухим определением скрывается идея, которая, вместе с тем, что очень проста, в ряде случаев дает возможность получить «фантастические результаты».

В качестве примера рассмотрим приближение функции $tg\ x$ степенным рядом – рис. Это разложение, естественно не работает вблизи $x=\frac{\pi}{2}$, где $tg\ x$ расходится, поскольку полином не может расходиться (принимать бесконечное значение) при конечном значении аргумента. При малых значениях x разложение $tg\ x$ по степеням x неплохо описывает $tg\ x$, однако чем ближе x к $\pi/2$ тем больше расхождение между функцией и разложением. Как видно из рис. увеличение числа слагаемых в разложении практически не улучшает ситуации вблизи $\pi/2$. Так на рис. 3.3.1. нижняя линия – разложение $tg\ x$ до 9-ой степени x:

$$f_9(x)=x+\frac{x^3}{3}+\frac{2}{15}x^5+\frac{17}{315}x^7+\frac{1382}{155925}x^9, \qquad (3.3.3)$$

вторая снизу до 15-ой степени, и наконец третья снизу до 29-ой степени, так, что последнее слагаемое этого разложения равно

$$\frac{689005380505609448}{263505041412702261046875}\cdot x^{29} \qquad (3.3.4)$$

Но даже разложение такое (которое вряд ли кто-то делал бы вручную) слабо помогает.

Посмотрим теперь, как с этой задачей справится метод аппроксимант Паде. Выберем его в виде:

$$R(x)=\frac{ax}{1+bx^2+cx^4}, \qquad (3.3.5)$$

т. е. в виде отношения полиномов не выше четвертой степени.

Разложение $R(x)$ в степенной ряд по x имеет вид:

$$R(x)\approx ax-abx^3-a\left(c-b^2\right)x^5, \qquad (3.3.6)$$

Сравнивая это разложение с разложением $tg\ x$

$$tg\ x\approx x+\frac{x^3}{3}+\frac{2}{15}x^5+..., \qquad (3.3.7)$$

и приравнивая коэффициенты при одинаковых степенях переменной получаем систему уравнений для коэффициентов a, b и c:

$$a=1,\quad ab=\frac{1}{3},\quad a\left(c-b^2\right)=\frac{2}{15}, \qquad (3.3.8)$$

откуда $a=1$, $b=-1/3$, $c=-0.022$, и, следовательно, аппроксиманта Паде для функции $tg\ x$ имеет вид:

$$R(x)=\frac{x}{1-\frac{1}{3}x^2-0.022\ x^3}, \qquad (3.3.9)$$

см. третью снизу кривую на рис. 7, самая верхняя кривая это функция $tg\ x$.

Сравнение между собой $tg\,x$, $R(x)$, $f_{15}(x)$ и $f_9(x)$ показывает, что аппроксимант Паде всего четвертой степени много лучше описывает поведение $tg\,x$ во всем диапазоне от 0 до $\pi/2$. И что очень важно описывает (пусть и не очень точно) расходимость вблизи $\pi/2$.

Второй пример связан с реальной физической задачей – определения вязкости суспензии, представляющей собой жидкость с вязкостью μ_0 в которой находятся жесткие частицы с концентрацией p.

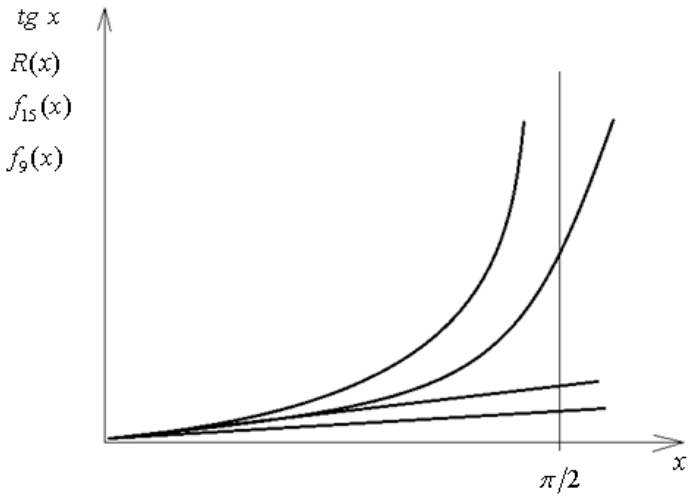


Рис. 3.3.1. Сравнение функции $tg\,x$, аппроксиматы Паде и двух разложений в ряд $f_{15}(x)$ (до 15-ой степени x) и $f_9(x)$ (до 9-ой). Графики функций расположены сверху вниз согласно обозначениям на вертикальной оси

Чем больше концентрация частиц, тем больше вязкость суспензии – μ_e. Первым такую задачу для $p \ll 1$ решил А. Эйнштейн в 1905 году:

$$\mu_e = \mu_0(1+\frac{5}{2}p), \quad p \ll 1, \qquad (3.3.10)$$

получив приближенное решение с точностью до первого порядка малой величины концентрации частиц.

Только через 67 лет, в 1972 г. Г. Бэтчелор и Дж. Грин (после весьма сложных вычислений) сумели получить следующее приближение

$$\mu_e = \mu_0\left(1+\frac{5}{2}\mathrm{p}+\left(5,2\pm 0,3\right)\mathrm{p}^2\right). \qquad (3.3.11)$$

Построим теперь аппроксимант Паде для этих двух приближений.

$$R_1(p) = \mu_0\frac{1}{1+ap} \approx \mu_0(1-ap), \qquad (3.3.12)$$

откуда, согласно приближению А. Эйнштейна $a = 5/2$, и, следовательно:

$$\mu_e \approx \mu_0\frac{1}{1-\frac{5}{2}p} \qquad (3.3.13)$$

откуда сразу же следует принципиально новый факт – при приближении p к определенному значению, в данном случае к $p_c = 2/5$ эффективная вязкость μ_c расходится. Такая расходимость действительно имеет место в реальности, эксперимент показывает, что при определенной концентрации жестких частиц жидкость перестанет течь.

Заметим, так же, что если разложить аппроксимант Паде (3.3.13) $\mu_e = \mu_0 / \left(1-\frac{5}{2}p\right)$ в ряд до p^2:

$$\mu_e \approx \mu_0\frac{1}{1-\frac{5}{2}p} \approx \mu_0\left(1+\frac{5}{2}p+\frac{25}{4}p^2\right), \qquad (3.3.14)$$

то получится неплохое второе приближение, где около p^2 стоит $25/4 \approx 6,25$ вместо $5,2\pm 0,3$, полученного Бэтчелором и Грином.

Для уточнения p_c – того значения, при котором эффективная вязкость становится очень большой можно построить аппроксимант Паде используя второе

приближение, полученное Бэтчелором и Грином, предположив, что коэффициент при p^2 равен точно 5 (рис. 3.3.2).

$$R_2 = \mu_0 \frac{1+Ap}{1-Bp} \approx \approx \mu_0 \left(1+\left(A+B\right)p + B\left(A+B\right)p^2\right), \qquad (3.3.15)$$

откуда

$$A+B=\frac{5}{2}, \; B\left(A+B\right)=5, \qquad (3.3.16)$$

и следовательно

$$\mu_e \approx \mu_0 \frac{1+p/2}{1-2p}, \qquad (3.3.17)$$

что для p_c даст $p_c = 1/2$.

Это значение очень хорошо согласуется с экспериментальным значением.

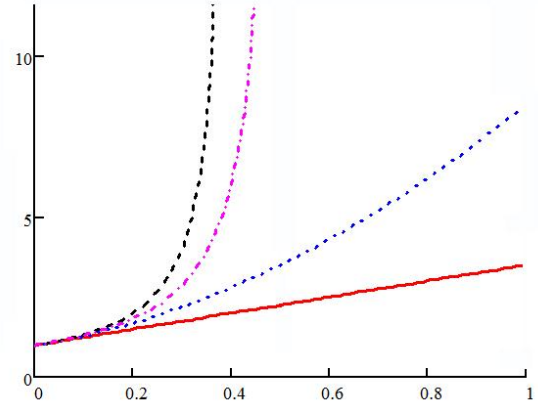


Рис. 3.2.2 – Эффективная вязкость. Нижняя линия – приближение Эйнштейна, выше нее – выражение Бэтчелора и Грина, еще выше – апроксиманта Падэ приближения Эйнштейна, самая верхняя – аппроксиманта Паде приближения Бэтчелора-Грина

3.4. Вероятностные распределения

Наиболее частыми (как обычно считается), универсальными законами распределения случайных величин, встречаемыми в различных естественнонаучных исследованиях, является нормальный закон – распределение Гаусса и так называемое логнормальное распределение (рис. 3.4.1):

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}, \qquad (3.4.1)$$

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{\ln x^2}{2\sigma^2}}, \quad x > 0 \qquad (3.4.2)$$

Частая встречаемость нормального закона объясняется тем, что когда распределение случайной величины связано с суммой независимых процессов, распределение приближается к нормальному. Именно это утверждение и является содержанием центральной предельной теоремы теории вероятности. Заметим, что часто в конкретных исследованиях гауссово распределение случайной величины принимается в силу привычки или удобства.

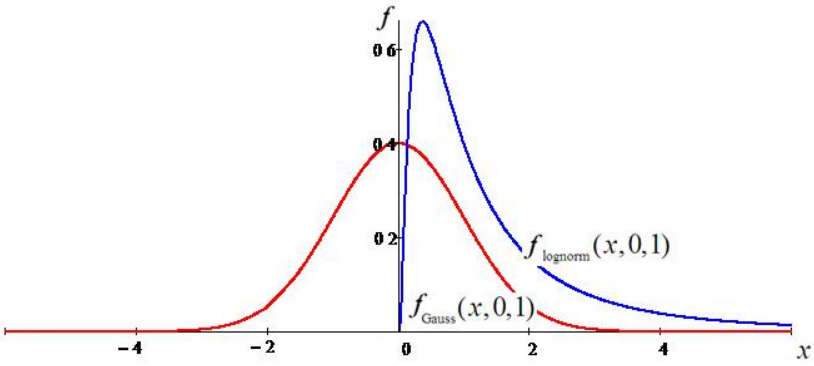


Рис. 3.4.1 – Графики нормального и логнормального распределения. Среднее значение для нормального распределения выбрано равным нулю

Б. Мандельброт был одним из первых, кто обратил пристальное внимание на то, что не менее универсальным, часто встречаемым законом распределения случайной величины является степенное (часто говорят гиперболическое) распределение с плотностью вероятности:

$$f\ x = \frac{B}{x^{\beta}}, \qquad (3.4.3)$$

или

$$P\ X \geq x = \frac{A}{x^{\alpha}}, \quad x > 0, \quad \alpha = \beta - 1, \qquad (3.4.4)$$

где $P\ X \geq x$ - вероятность того, что $X \geq x$, а A и α – некоторые положительные константы, параметры распределения.

Соответственно,

$$P\ X \geq x = \int_{x}^{\infty} \frac{B}{x^{\beta}} dx, \qquad (3.4.5)$$

$$\frac{dP\ x}{dx} = -f\ x\ .$$

Следует отметить, что приведенное выше распределение рассматривалось Мандельбротом как уточнение закона Ципфа и его часто называют распределением Ципфа-Мандельброта. При этом оказалось, что α – близкая к единице величина, которая может изменяться в зависимости от свойств текста и языка.

Справедливости ради, надо отметить, что степенные функции распределения рассматривались еще Коши. Как наглядный пример распределения Коши можно привести модель стрельбы из вращающегося с постоянной угловой скоростью ω в горизонтальной плоскости пулемета (рис. 3.4.3).

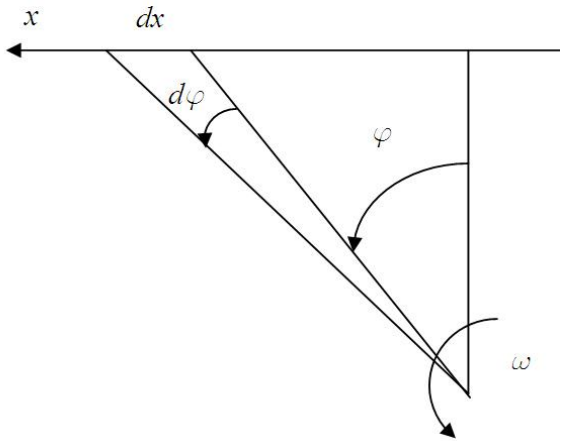


Рис. 3.4.2. – Пример распределения Коши

Если, производя одиночные выстрелы, нажимать на курок равновероятно при любом его положении, то функция распределения выстрелов по углу – φ будет величиной постоянной: $F(\varphi) = const$. С другой стороны, вероятность попадания в бесконечно малый участок dx бесконечной плоской мишени равна $f(x)\,dx = F(\varphi)\,d\varphi$. Откуда, с учетом $x = a \cdot tg\,\varphi$, после элементарных преобразований находим распределение Коши:

$$f(x) = \frac{1}{\pi}\frac{a}{a^2 + x^2}, \quad -\infty < x < \infty. \qquad (3.4.6)$$

Так как для этой функции среднее $\langle x^\alpha \rangle$ от x^α ($\langle x^\alpha \rangle = \int_{-\infty}^{\infty} x^\alpha f(x)dx$) не определено для $\alpha \geq 1$, то ни математическое ожидание (т.е. среднее от x^α), ни дисперсия $\langle x^2 \rangle$, ни моменты старших порядков этого распределения не определены. В этом случае говорят, что математическое ожидание не определено, а дисперсия бесконечна.

Напомним, частный случай степенного распределения – гиперболическое распределение A/x названо в честь В. Парето, а дискретный закон распределения с ранжированной переменной был назван в честь Дж. Ципфа, который сформулировал его для описания частоты употребления слов.

3.5. Скейлинг. Однородные функции

3.5.1. Однородная функция одной переменной

Однородную функцию можно определить так:

$$f(\lambda x) = g(\lambda) f(x). \qquad (3.5.1)$$

Пример:

$$f(x) = \alpha x^3. \qquad (3.5.2)$$

В этом случае

$$f(\lambda x) = \alpha(\lambda x)^3 = \alpha\lambda^3 x^3 = \lambda^3 \alpha x^3 = \lambda^3 f(x). \qquad (3.5.3)$$

Какова «выгода» от однородной функции?

Оказывается, огромная!

Например, пусть мы знаем, что $f(x)$ – однородная функция. Тогда если мы знаем ее значение в одной единственной точке, то можем узнать в любой другой:

Пусть знаем в x_0 – $f(x_0)$, а хотим в x. Обозначим $x = \lambda x_0$.

$$f(x) = f(\lambda x_0) = g(\lambda) f(x_0), \qquad (3.5.4)$$

где $\lambda = x/x_0$.

Если $f(x)$ непрерывна и дифференцируема, то можно показать, что из определения $\boxed{f(\lambda x) = g(\lambda) f(x)}$ следует $\boxed{g(\lambda) = \lambda^p}$ (с точностью до сомножителя).

И таким образом, однородная непрерывная и дифференцируемая функция – это степенная функция.

Существуют ли другие (однородные), не степенные функции, для которых $f(\lambda x) = g(\lambda) f(x)$?

Да, существуют, но при этом приходится отказаться от гладкости.

Например, δ-функция Дирака:

$$\delta(x) = \begin{cases} 0, & x \neq 0 \\ \neq 0, & x = 0 \end{cases} \quad ; \quad \int_{-\infty}^{+\infty} \delta(x) dx = 1. \qquad (3.5.5)$$

3.5.2. Однородные функции нескольких переменных

Для однородной функции нескольких переменных имеет место:

$$f(\lambda x_1, \lambda x_2, ..., \lambda x_n) = g(\lambda) f(x_1, x_2, ..., x_n), \qquad (3.5.6)$$

в частности, для функции двух переменных

$$f(\lambda x, \lambda y) = g(\lambda) f(x, y), \; g(\lambda) = \lambda^p. \qquad (3.5.7)$$

Можно расширить понятие однородной функции нескольких переменных, введя обобщенную функцию нескольких переменных. Например, для двух переменных обобщенная однородная функция

$$f\left(\lambda^a x, \lambda^b y\right) = \lambda^p f\left(x, y\right), \qquad (3.5.8)$$

что означает, что каждая переменная имеет свою масштабную константу – λ^a для x и λ^b для y.

Снова возникает вопрос, какова же «выгода» от того, что функция, например двух переменных – однородна? Оказывается, обобщение однородной функции на несколько переменных приводит к намного более важным вопросам, чем для функции одной переменной.

В самом деле, однородную функцию двух переменных

можно свести к функции одной переменной. В самом деле, пусть $f(\lambda x, \lambda y) = \lambda^p f(x, y)$, тогда выбираем $\lambda = 1/y$, откуда:

$$\lambda^p f(x,y) \Leftarrow f(\lambda x, \lambda y) \Rightarrow f\left(\frac{x}{y}, 1\right);$$

$$\lambda^p = y^{-p} \ . \qquad (3.5.9)$$

Следовательно,

$$y^{-p} f(x,y) = f\left(\frac{x}{y}, 1\right). \qquad (3.5.10)$$

Обозначим $f\left(\frac{x}{y}, 1\right) = F\left(x/y\right)$, тогда окончательно:

$$f(x,y) = y^p F\left(x/y\right). \qquad (3.5.11)$$

Таким образом функция двух переменных $f(x,y)$ выражается через функцию одной $F\left(x/y\right)$.

Рассмотрим несколько примеров.

Формула Планка теплового излучения:

$$u(\omega, T) = \frac{\hbar\omega^3}{\pi^2 c^3} \frac{1}{e^{\frac{\hbar\omega}{KT}} - 1}, \qquad (3.5.12)$$

где $\hbar$ - постоянная Планка, c - скорость света, k – постоянная Больцмана, ω – циклическая частота, T – абсолютная температура.

Нет надобности объяснять программистам, как важно, если *двумерный* массив можно заменить на *одномерный*.

Функция $F(z)$ называется скейлинговой (или автомо-дельной) функцией.

Выберем более удобные обозначения

$$f(x,y)=\frac{x^3}{e^{\frac{x}{y}}-1}, \qquad (3.5.13)$$

где $x=\hbar\omega$, $y=kT$, $f(x,y)=\pi^2\hbar^2c^3u(\omega,T)$.

Функция $f(x,y)$ является функцией двух переменных x и y и для ее задания необходим двумерный массив, например, набор зависимостей $f(x,y)$ как функции от x при разных значениях y рис 3.5.1.

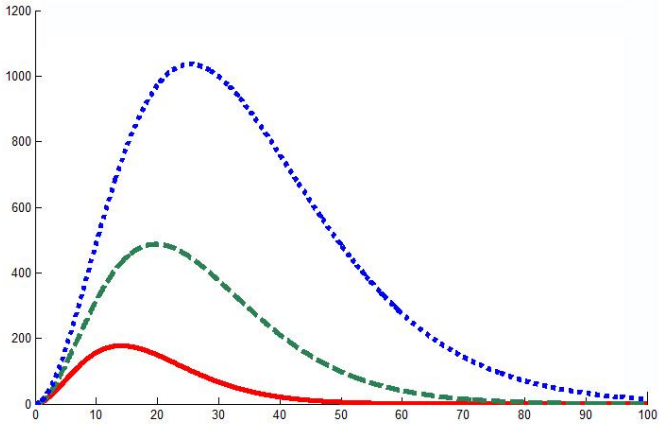


Рис. 3.5.1. Зависимости функции $f(x,y)$ от x при разных значениях y: 5 (сплошная линия), 7 (штриховая линия), 9 (пунктирная линия)

На рис. 3.5.2 приведена зависимость функции $f(x,y)$ от y при разных x.

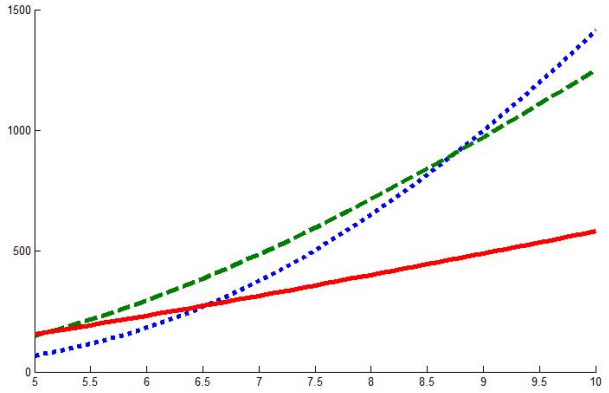


Рис. 3.5.2. Зависимости функции $f(x,y)$ от y при разных значениях x: 10 (сплошная линия), 20 (штриховая линия), 30 (пунктирная линия)

Двумерный массив – $f(x,y)$ можно изобразить так же как поверхность – рис 3.5.3.

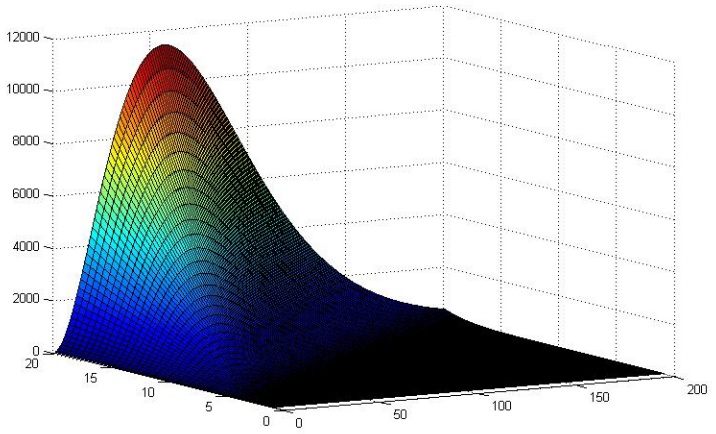


Рис. 3.5.3. Поверхность, задаваемая функцией $f(x,y)$

Если бы стояла задача описать функцию $f(x,y)$ экспериментально, то согласно рис. или рис. необходимо

было бы измерить ее значение для разных x и y, некоторые значения которых могли быть сложны для измерения (например очень высокие температуры).

До того как Планк написал свою формулу для теплового излучения, был известен закон Вина, согласно которому

$$u(\omega,T)=\omega^3 F\left(\frac{\omega}{T}\right), \qquad (3.5.14)$$

или в используемых нами обозначениях

$$f(x,y)=x^3 F\left(\frac{x}{y}\right). \qquad (3.5.15)$$

Заметим, что это типичное скейлинговое соотношение, что резко упрощает задачу экспериментального определения $f(x,y)$. В самом деле, вводя функцию

$$\varphi(x,y)=\frac{f(x,y)}{x^3}. \qquad (3.5.16)$$

Находим, что она является функцией одной переменной, и что для ее определения достаточно одномерного массива (рис. 3.5.4).

Еще один пример из так называемой экономической физики.

В экономике важную роль играет спрос.

Чернавский Д.С., Старков Н.И., Малков С.Ю., Коссе Ю.В., Щербаков А.В. Об эконофизике и ее месте в современной теоретической экономике, УФН, 2011, т. 181, № 7, с. 15

Баренблатт Г.И. Подобие, автомодельность, промежуточная асимптотика / 2-е изд., перераб. и доп. – Ленинград: Гидрометеоиздат, 1982. – 256 с.

Функция спроса на данный товар (или группу сходных товаров) зависит от:

1) количества товара *Q*, потребляемого в единицу времени;

2) наличия у потребителя денег *U* (или доходов *D*);

3) цены товара – *p*

$$Q=Q(U,p),$$

где U и p – условны и измеряются в различных единицах (\$, €, ¥, …), поэтому представляется правдоподобным, что $Q(U,p)$ – однородная функция этих переменных.

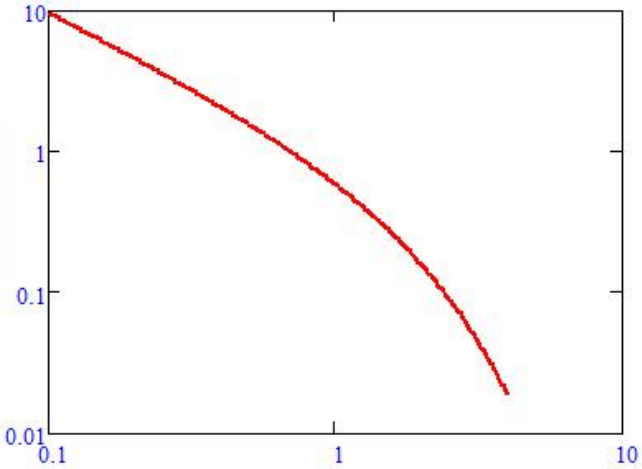


Рис. 3.5.4 – $\varphi(x,y)$ как функция одной переменной $z = x / y$

Действительно, если записать $Q(U,p) = F\left(U/p\right)$, то как оказывается эмпирические данные неплохо укладываются на одну кривую.

Такой подход позволяет сформулировать задачу и получить аналитическое представление функции спроса и выявить очень важный параметр $r_{\min}$.

С понятием скейлинга тесно связаны понятия автомодельность (скейлинг), подобие, промежуточная асимптотика.

При обработке опытных данных автомодельность приводила к тому, что, казалось бы, беспорядочное в обычных координатах облако опытных точек ложилось на единую кривую или поверхность.

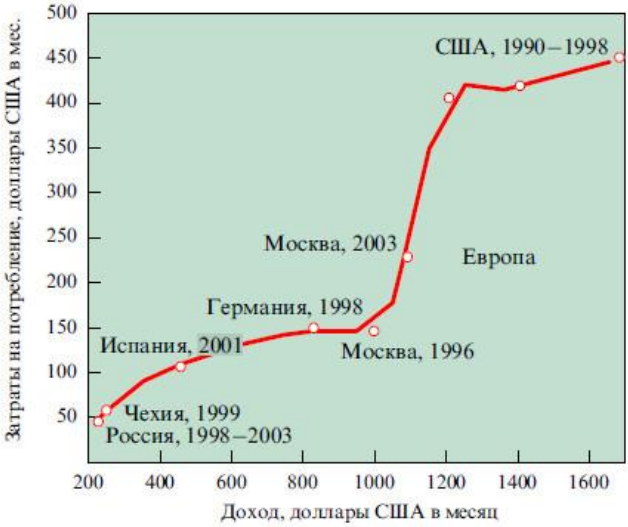


Рис. 3.5.5 – $r = \dfrac{U}{p}$ – покупательная способность

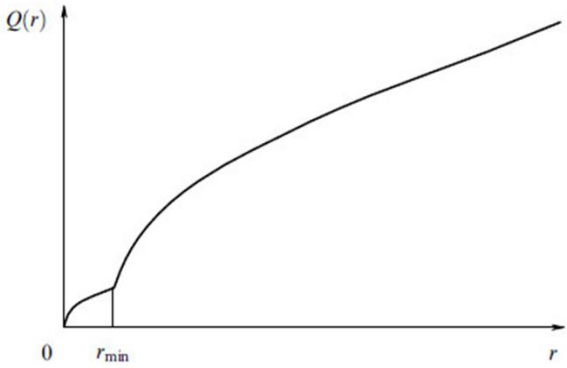


Рис. 6. $r < r_{\min}$ – потребитель не покупает товаров долговременного пользования (машины, телевизоры, дачи, элитарные товары), все идет на пищу, одежду, ЖКХ

3.6. Производящие функции

Производящей функцией $f(x)$ произвольной бесконечной последовательности $a_0, a_1, a_2, ..., a_k, ...$ называется выражение

$$f(x) = \sum_{k=0}^{\infty} a_k x^k, \qquad (3.6.1)$$

Ландо С.К. Лекции о производящих функциях. – М.: МЦНМО, 2007.

Wilf H.S. Generatingfunctionology. – Academic Press, 1994.

при этом переменная x является формальной и сама сумма ряда смысла не имеет.

Такое утверждение весьма неожиданно (может поставить в тупик – что же это за функция, переменная, которой не имеет смысла). Ряд может расходиться и единственное, что определено – это значение в точке $x = 0$, т. е. $f(0) = a_0$.

Не смотря на такие удивительные утверждения, производящая функция это мощный инструмент (метод), позволяющий простым элегантным способом решать весьма сложные задачи. Здесь будет рассмотрен метод производящих функций для решения рекуррентных соотношений, что позволяет получить много важных результатов в теории сложных сетей.

Производящая функция превращает дискретный набор – последовательность a_k $(k = 0, 1, ...)$ в непрерывную функцию, что позволяет «включить» математический аппарат, работающий с непрерывными объектами – анализ непрерывных функций (решение уравнений, дифференцирования, интегрирования и т. д.), после чего можно опять вернуться в «дискретный» мир.

Постановка задачи о рекуррентных соотношениях выглядит так: пусть есть соотношение

$$x_{k+1} = f(x_k). \qquad (3.6.2)$$

необходимо найти явную зависимость

$$x_k = F(k). \qquad (3.6.3)$$

Возможны и более общие постановки, например, когда задано рекуррентное соотношение

$$x_{k+2} = f\left(x_k, x_{k+1}\right) \qquad (3.6.4)$$

Рассмотрим несколько примеров.

Первый пример, тривиальный, для простой демонстрации метода. Пусть дано итерационное соотношение $a_{k+1} = a_k + 1,\ k = 1,\ 2,\ ...,\ a_1 = 1$ и необходимо найти зависимость a_k от k. Вводим производящую функцию:

$$G(x) = \sum_{k=1}^{\infty} a_k x^k, \qquad (3.6.4)$$

умножаем ее на x и проводим ряд элементарных преобразований:

$$\begin{aligned} xG(x) &= x\sum_{k=1}^{\infty} a_k x^k = \sum_{k=1}^{\infty} a_k x^{k+1} = \left(a_k = a_{k+1} - 1\right) = \\ &= \sum_{k=1}^{\infty} a_{k+1} x^{k+1} - \sum_{k=1}^{\infty} x^{k+1} = \sum_{n=2}^{\infty} a_n x^n - \sum_{n=2}^{\infty} x^n = \\ &= -a_1 x + \left(a_1 x + \sum_{n=2}^{\infty} a_n x^n\right) - \left(1 + x + \sum_{n=2}^{\infty} x^n\right) + 1 + x. \end{aligned} \qquad (3.6.5)$$

Первое слагаемое в скобках это $G(x)$, второе $\sum_{n=0}^{\infty} x^n = 1/(1-x)$, и, учитывая, что $a_1 = 1$, находим

$$xG(x) = -x + G(x) - \frac{1}{1-x} + 1 + x, \qquad (3.6.6)$$

откуда следует выражение для производящей функции

$$G(x) = \frac{x}{(1-x)^2}. \qquad (3.6.7)$$

Разлагая плученное выражение $G(x)$ в ряд Тейлора находим:

$$G(x)=\frac{x}{(1-x)^2}=x\left(1+2x+3x^2+...\right)=\sum_{k=1}^{\infty}k\cdot x^k. \quad (3.6.8)$$

Поскольку $G(x)=\sum_{k=1}^{\infty}a_k x^k$, для a_k получим

$$a_k=k, \quad (3.6.9)$$

и, таким образом, искомая зависимость найдена.

Конечно, в рассмотренном простом примере, из рекуррентного соотношения $a_{k+1}=a_k+1$ можно сразу догадаться, что при $a_1=1$ речь идет о последовательности 1, 2, 3, ..., т. е., что $a_k=k$.

Однако второй и третий примеры демонстрируют случаи, когда догадка мало вероятна, и тогда метод производящей функции оказывается практически единственным выходом.

Пример второй, стандартный пример демонстрации метода производящей функции – числа Фибоначчи. Фибоначчи рассматривал идеализированную популяцию кроликов, когда изначально есть новорожденная пара, которая со второго месяца после своего рождения начинает спариваться и каждый месяц производить новую пару, которая ведет себя точно так же. Вопрос: если кролики не умирают, то сколько их будет через n месяцев.

Процесс описывается итерационным соотношением

$$F_{n+1}=F_n+F_{n-1}, \quad n\geq 1,\ F_0=0,\ F_1=1, \quad (3.6.10)$$

где F_n - число новорожденных пар в n-й месяц.

Вводим производящую функцию:

$$G(x)=\sum_{n=0}^{\infty}F_n x^n. \quad (3.6.11)$$

Умножим итерационное соотношение на x^n и просуммируем:

$$\sum_{n=1}^{\infty} F_{n+1} x^n = \sum_{n=1}^{\infty} F_n x^n + \sum_{n=1}^{\infty} F_{n-1} x^n, \qquad (3.6.12)$$

заметим, что суммирование возможно начинать только с $n=1$, т. к. при $n=0$ в каждой сумме справа появляется неопределенное слагаемое F_{-1}.

Для левой части выражения (3.6.12) имеем:

$$\begin{aligned} &\sum_{n=1}^{\infty} F_{n+1} x^n = \frac{1}{x} \sum_{n=1}^{\infty} x^{n+1} F_{n+1} = \\ &= \frac{1}{x}\left(1 + xF_1 + x^2 F_2 + \ldots - 1 - F_1 x\right) = \\ &= \frac{1}{x}\left(\sum_{n=0}^{\infty} F_1 x^n - (1+x)\right) = \\ &= \frac{1}{x}\left(G(x) - (1+x)\right), \end{aligned} \qquad (3.6.13)$$

где было учтено, что $F_1 = 1$.

Для суммы $\sum_{n=1}^{\infty} F_n x^n$ легко получить:

$$\sum_{n=1}^{\infty} F_n x^n = \sum_{n=0}^{\infty} F_n x^n - 1 = G(x) - 1. \qquad (3.6.14)$$

И, наконец, для последней суммы имеем:

$$\begin{aligned} &\sum_{n=1}^{\infty} F_{n-1} x^n = x \sum_{n=1}^{\infty} F_{n-1} x^{n-1} = \\ &= x \sum_{n=0}^{\infty} F_n x^n = xG(x). \end{aligned} \qquad (3.6.15)$$

Таким образом, итерационное соотношение может быть записано как:

$$\frac{G(x)}{x}-\frac{1+x}{x}=G(x)(1-x)-1, \qquad (3.6.16)$$

что дает следующее выражение для производящей функции:

$$G(\mathrm{x})=\frac{x}{1-x-x^2}. \qquad (3.6.17)$$

Раскладывая $G(x)$ на простые дроби, находим:

$$\frac{x}{1-x-x^2}=\frac{1}{a-b}\left(\frac{1}{1-ax}-\frac{1}{1-bx}\right)= \\ =\frac{1}{a-b}\left(\sum_{k=0}^{\infty}a^k x^k-\sum_{k=0}^{\infty}b^k x^k\right), \qquad (3.6.18)$$

где $b=\frac{1-\sqrt{5}}{2}$, а $a=\frac{1+\sqrt{5}}{2}$ - золотое сечение.

Таким образом, для $G(x)$ находим:

$$G(x)=\sum_{k=0}^{\infty}\frac{a^n-b^n}{\sqrt{5}}x^n= \\ =\sum_{k=0}^{\infty}\frac{1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^n-\left(\frac{1-\sqrt{5}}{2}\right)^n\right]x^n, \qquad (3.6.19)$$

что для искомого F_n дает:

$$F_n=\frac{1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^n-\left(\frac{1-\sqrt{5}}{2}\right)^n\right]. \qquad (3.6.20)$$

Последний пример относится к исследованию детерминированных SF-сетей, а именно (u,v)-цветам. Ранее было без вывода указано, что из итерационного соотношения

$$N_n=wN_{n-1}-w,\ n=1,\ 2,\ \ldots,\ N_1=w, \qquad (3.6.20)$$

следует, что

$$N_n = \frac{w-2}{w-1} \cdot w^n + \frac{w}{w-1}. \qquad (3.6.21)$$

Покажем, как можно получить $N_n = f(n)$ методом производящей функции

$$G(x) = \sum_{n=2}^{\infty} N_n w^n \ . \qquad (3.6.22)$$

Проделывая очевидные преобразования

$$\begin{aligned} G(x) &= N_1 w^1 + \sum_{n=2}^{\infty} N_n x^n = wx + \sum_{n=2}^{\infty} \left(wN_{n-1} - w \right) x^n = \\ &= wx + wx \sum_{n=1}^{\infty} \left(wN_n - w \right) x^n - w \left(\sum_{n=0}^{\infty} x^n - (1+x) \right) = \\ &= wx + wxG(x) - w \left(\frac{1}{1-x} - (1+x) \right), \end{aligned} \qquad (3.6.23)$$

для $G(x)$ получаем следующее выражение

$$G(x) = w \cdot \frac{x - 2x^2}{(1-wx)(1-x)}, \qquad (3.6.24)$$

Поскольку

$$\frac{1}{(1-wx)(1-x)} = \frac{1}{1-w} \left(\frac{1}{1-x} - \frac{w}{1-wx} \right), \qquad (3.6.25)$$

Производящую функцию можно записать через сумму дробей

$$G(\mathrm{x}) = \frac{w}{1-w} \left(\frac{x}{1-x} - \frac{2x^2}{1-x} - \frac{wx}{1-wx} + \frac{2x^2}{1-wx} \right), \qquad (3.6.26)$$

каждая из которых может быть, в свою очередь, записана в виде суммы:

$$\frac{x}{1-x} = x \sum_{n=0}^{\infty} x^n = \sum_{n=0}^{\infty} x^{n+1} = \sum_{n=1}^{\infty} x^n, \qquad (3.6.27)$$

$$\frac{\alpha x}{1-wx}=\sum_{n=1}^{\infty}w^n x^n, \qquad (3.6.28)$$

$$\frac{x^2}{1-x}=x^2\sum_{n=0}^{\infty}x^n=x^2\left(1+x+x^2+...\right)= \\ =x^2+x^3+...+x-x=\sum_{n=1}^{\infty}x^n-x, \qquad (3.6.29)$$

$$\frac{wx^2}{1-wx}=\frac{1}{w}\sum_{n=1}^{\infty}w^n x^n-x, \qquad (3.6.30)$$

подставляя которые в выражение для $G(x)$ (свободные «x» появляющиеся на месте второго и четвертого слагаемого сокращаются) находим:

$$G(x)=\frac{w}{1-w}\sum_{n=1}^{\infty}\left(\frac{2-w}{w}w^k-1\right)x^n, \qquad (3.6.31)$$

откуда и следует выражение для N_n.

Приведенные примеры показывают, что задача получения явной зависимости от номера шага из итерационного соотношения чем-то похожа на шифровку-дешифровку – разгадать сложно, зашифровать просто. Действительно имея явное выражение, например $N_n=\left(\frac{w-2}{w-1}\right)w^n+\frac{w}{w-1}$ очень просто убедится, что оно удовлетворяет итерационному соотношению $N_n=wN_{n-1}-w$. Обратная задача намного сложнее и здесь метод производящих функций приносит огромную пользу. Тем не менее, при использовании этого метода, нет четкого алгоритма. Что надо делать – умножать на x, x^2, ..., делить, разбивать ряд на части и т. п. В каждой конкретной задачи свое действие. Однако, в общем (не полностью конкретном) описании действий присутствует несколько ясных,

последовательных шагов, первый – переписать (преобразовать) производящую функцию, с учетом заданного итерационного соотношения так, что бы получить уравнение для производящий функции относительно введенной формальной переменной. После решения этого уравнения полученную функцию от x необходимо разложить в ряд, выписав в явном математическом виде коэффициенты разложения при степенях формальной переменной, что и дает решение.

В некоторых случаях такой путь довольно громоздок. Здесь на помощь могут прийти производящие функции другого типа. Если до сих пор под производящей функцией понимался степенной ряд

$$G(x)=\sum a_n x^n, \qquad (3.6.32)$$

то как оказывается, иногда удобнее воспользоваться производящей функцией Дирихле, основанной на дзета функции Римана $\zeta(x)=1+\frac{1}{2^x}+\frac{1}{3^x}+...$

$$G(x)=\frac{a_1}{1^x}+\frac{a}{2^x}+...=\sum_{n=1}^{\infty}\frac{a_n}{n^x}, \qquad (3.6.33)$$

или на экспоненциальной производящей функции

$$G(x)=\sum_n a_n\frac{x^n}{n!}. \qquad (3.6.34)$$

В заключении заметим, что существуют так же производящие функции нескольких переменных.

3.7. Дельта-функция Дирака

Дельта функция (δ-функция) была введена английским физиком П. А. М. Дираком «по необходимости», когда он создавал математический аппарат квантовой механики. Математики некоторое время «не признавали» ее, после чего создали теорию обобщенных функций, частным случаем которых является δ-функция.

Согласно (наивному) определению δ-функция равна нулю всюду кроме одной точки, но при этом площадь охватываемая этой функцией равна единице:

$$\delta(x)=\begin{cases}0, & x\neq 0\\ \infty, & x=0\end{cases}, \qquad \int_{-\infty}^{+\infty}\delta(x)dx=1. \qquad (3.7.1)$$

Эти противоречивые требования не могут быть удовлетворены функцией «обычного» типа.

Зельдович Я.Б. Высшая математика для начинающих физиков и техников. –М.: Наука, 1982.

На самом деле, как дифференциал *dx* не есть число (равное нулю), а словосочетание «бесконечно-малая величина» трудно понимаемо качественно, правильно понимать *dx* не как число, а как предел (процесс), так же δ-функцию правильно понимать как предел (процесс). На рис. 3.7.1 и 3.7.2 изображено несколько функций (зависящих от параметра), пределом которых и является δ-функция. Таких функций бесконечно много – каждый может выбрать свою.

δ-функция обладает многими полезными свойствами, являясь, в частности, континуальным аналогом символ Кронекера δ_{ik}

$$\int_{-\infty}^{+\infty} f(x)\delta(x-x_0)dx = f(x_0). \qquad (3.7.2)$$

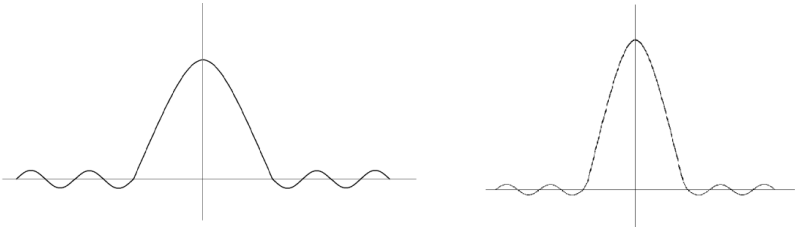
сравните с

$$\sum_i f_i\delta_{ik} = f_k\,. \qquad (3.7.3)$$

Еще одно удивительное соотношение указывает как можно продифференцировать проинтегрировав:

$$\int_{-\infty}^{+\infty} f(x)\delta'(x-x_0)dx = f'(x_0), \qquad (3.7.4)$$

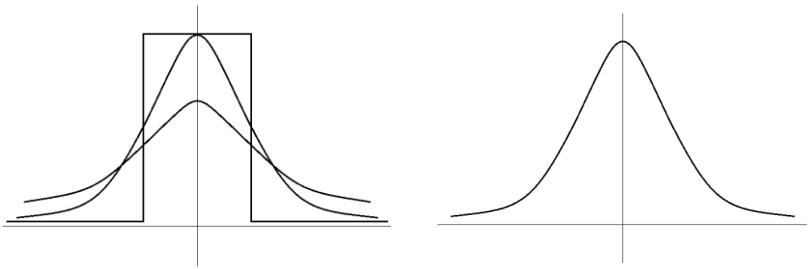
где δ' – производная δ-функции.



а) б)

Рис. 3.7.1 – Два последовательных приближения к δ-функции Дирака. Изображена функция $R(x,\omega)=\frac{1}{\pi}\frac{\sin(\omega x)}{x}$,

а) $\omega=10$, б) $\omega=20$



а) б)

Рис. 3.7.2 – Две функции, которые в пределе $a\to\infty$ дают δ-функции:

а) $R(x,a)=\frac{a}{2(1+a^2x^2)^{3/2}}$, б) $R(x,a)=\frac{a}{\pi(1+a^2x^2)}$.

Наконец заметим, что интервал от δ-функции:

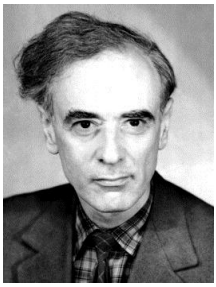
$$\int_{-\infty}^{x}\delta(x)dx=\theta(x), \qquad (3.7.5)$$

где $\theta(x)$ – функция Хевисайда,

$$\theta(x)=\begin{cases}0,\ x<0\\1,\ x>0\end{cases}, \qquad (3.7.6)$$

ступенька, с разрывом в точке $x=0$.

3.8. Фазовые переходы



Лев Давидович Ландау (1908-1968)

Для того чтобы говорить о фазовых переходах, необходимо определить, что такое фазы. Понятие фаз встречается во множестве явлений, поэтому вместо того, чтобы давать общее определение (чем оно более общее, тем оно, как и положено, более абстрактное и ненаглядное), приведем несколько примеров.

Вначале пример их физики. Для обычной, наиболее часто встречаемой в нашей жизни жидкости – воды известно три фазы: жидкая, твердая (лед) и газообразная (пар). Каждая из них характеризуется своими значениями параметров. Существенно то, что при изменении внешних условий одна фаза (лед) переходит в другую (жидкость). Еще один любимый объект теоретиков – ферромагнетик (железо, никель и множество других чистых металлов и сплавов). При низких температурах (для никеля ниже $T_c=360^0C$) образец из никеля является ферромагнетиком, при снятии внешнего магнитного поля он остается намагниченным, т.е. может использоваться как постоянный магнит. При температуре выше T_c это свойство теряется, при выключении внешнего магнитного поля он переходит в парамагнитное состояние и не является постоянным магнитом. При изменении температуры происходит переход – фазовый переход – из одной фазы в другую.

Приведем еще один геометрический пример из теории перколяции. Случайно вырезая из сетки связи, в конце концов, когда концентрация оставшихся связей – p станет меньше некоторого значения p_c, по решетке уже нельзя

будет пройти «из одного конца в другой». Таким образом, сетка из состояния протекания - фаза «протекания», перейдет в состояние фазы «непротекания».

Из этих примеров ясно, что для каждой из рассмотренных систем существует так называемый параметр порядка, определяющий в какой из фаз находится система. В ферромагнетизме параметр порядка – намагниченность в нулевом внешнем поле, в теории перколяции – связность сетки, или, например, ее проводимость или плотность бесконечного кластера.

Фазовые переходы бывают разного рода. Фазовый переходы I рода - это такой переход, когда в системе может одновременно существовать несколько фаз. Например, при температуре $0^\circ C$ лед плавает в воде. Если система находится в термодинамическом равновесии (нет подвода и отвода тепла), то лед не тает и не нарастает. Для фазовых переходов II рода существование одновременно нескольких фаз невозможно. Кусочек никеля либо находится в парамагнитном состоянии, либо в ферромагнитном. Сетка со случайно вырезанными связями либо связна, либо нет.

Решающим в создании теории фазовых переходов II рода, начало которой положил Л.Д. Ландау, было введение параметра порядка (будем обозначать его η) как отличительного признака фазы системы. В одной из фаз, например, парамагнитной, $\eta = 0$, а в другой, ферромагнитной, $\eta \neq 0$. Для магнитных явлений параметр порядка η – это намагниченность системы.

Для описания фазовых переходов вводится некоторая функция параметров, определяющих состояние системы – $G(\eta, T, ...)$. В физических системах это энергия Гиббса. В каждом явлении (перколяция, сеть «малых миров» и т.д.) эта функция определятся «самостоятельно». Главное свойство этой функции, первое предположение Л.Д. Ландау – в состоянии равновесия эта функция принимает минимальное значение:

$$\frac{\partial G}{\partial \eta}=0,\ \frac{\partial^2 G}{\partial \eta^2}>0. \qquad (3.8.1)$$

В физических системах говорят о термодинамическом равновесии, в теории сложных цепей можно говорить об устойчивости. Заметим, что условие минимальности определяется варьированием параметра порядка.

Второе предположение Л.Д. Ландау – при фазовом превращении $\eta=0$. Согласно этому предположению, функцию $G(\eta,T,...)$ вблизи точки фазового перехода можно разложить в ряд по степеням параметра порядка η:

$$G\ \eta,T\ =G_0\ T\ +A\ T\ \eta^2+B\eta^4+..., \qquad (3.8.2)$$

где $\eta=0$ в одной фазе (парамагнитной, если речь идет о магнетизме и несвязной, если о сетке) и $\eta\neq 0$ в другой (ферромагнитной или связной).

Из условия $\partial G/\partial\eta=0$ находим:

$$2A\eta+B\eta^3=0, \qquad (3.8.3)$$

что дает нам два решения $\eta=0$ и $\eta^2=-A/2B\neq 0.$

Для $T>T_c$ должно иметь место решение $\eta=0$, а для $T<T_c$ решение $\eta\neq 0$. Этому можно удовлетворить, если для случая $T>T_c$ и $\eta=0$ выбрать $A>0$. В этом случае второго корня не существует. А для случая $T<T_c$ должно иметь место второе решение, т.е. должно выполняться $A<0$. Таким образом:

$A>0$ при $T>T_c$, $A<0$ при $T<T_c$,

Второе предположение Ландау требует выполнения $A(T_c)=0$. Простейший вид функции $A(T)$, удовлетворяющий этим требованиям, есть

$$A=\alpha\ T-T_c\ . \qquad (3.8.4)$$

Тогда

$$\eta^2 = -A/2B \sim \alpha\ T - T_c\ , \qquad (3.8.5)$$

откуда

$\eta \sim \sqrt{T_c - T}$, или $\eta \sim (T_c - T)^{\beta}$, $T < T_c$, где – $\beta = 1/2$ – так называемый критический индекс, а функция $G(\eta, T)$ принимает вид:

$$G\ \eta, T\ = G_0\ T\ + \alpha\ T - T_c\ \eta^2 + B\eta^4 + \ldots \quad (3.8.6)$$

На рис. 3.8.1 изображена зависимость $G(\eta, T)$ для $T > T_c$ и $T < T_c$.

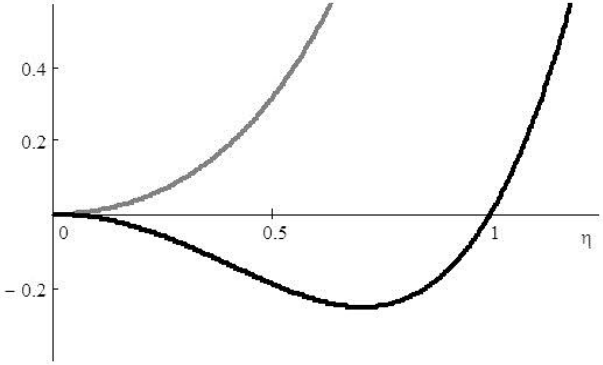


Рис. 3.8.1 – Графики функции параметров $G(\eta, T)$ для $T > T_c$ и $T < T_c$

Качественная зависимость параметров $G(\eta, T)$ от параметра порядка η изображена на рис. 3.8.1 $\left(G_0 = 0\right)$. Зависимость параметра порядка η от температуры представлена на рис. 3.8.2.

Постон Т., Стюарт И. Теория катастроф и ее приложения. – М.: Мир, 1980.

Гилмор Р. Прикладная теория катастроф. – М., Мир, 1984.

В более продвинутой теории учтено, что при $T > T_c$ параметр порядка η, хоть и очень маленький, но не равен точно нулю.

Переход системы из состояния с $\eta = 0$ при $T > T_c$ в состояние с $\eta \geq 0$ при уменьшении T и достижении значений $T \leq T_c$ можно понимать, как потерю устойчивости положения $\eta = 0$ при $T \leq T_c$. Недавно появилась математическая теория со звучным названием «Теория Катастроф» описывающая с единой точки зрения множество различных явлений. С точки зрения теории катастроф – фазовый переход второго рода это «катастрофа сборки».

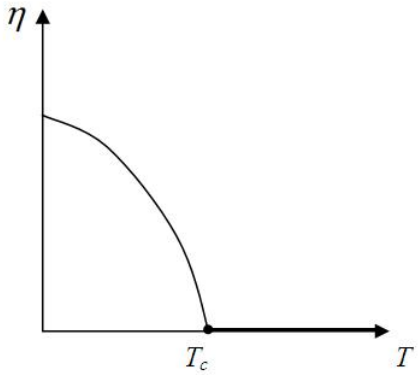


Рис. 3.8.2 – Зависимость параметра порядка η от температуры: при $T < T_c$ и вблизи T_c параметр порядка η ведет себя как степенная функция, а при $T > T_c$ $\eta = 0$

3.9. Клеточные автоматы

Дж. Фон. Нейман (1903-1957)

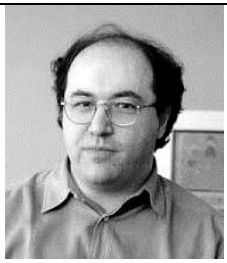
Концепция клеточных автоматов была впервые предложена больше полстолетия тому назад Дж. Фон Нейманом (J. Von Neumann) и развита С. Вольфрамом (S. Wolfram) в фундаментальной монографии “Новый вид науки”.

Клеточные автоматы являются полезными дискретными моделями для исследования динамических систем. Дискретность модели, а точнее, возможность представить модель в дискретной форме, может

считаться важным преимуществом, поскольку открывает широкие возможности использования компьютерных технологий.

Продолжительное время клеточные автоматы воспринимались как забавная игра, которая не имеет практической ценности. Но в последние годы, в связи с развитием компьютерных технологий, они начинают быстро входить в арсенал инструментальных средств, которые используются на практике в различных областях науки и техники.

Нейман Дж. Теория самово–спроизводящихся автоматов. – М.: Мир, 1971. – 382 с.



С. Вольфрам

Wolfram S. A New Kind of Science. – Champaign, IL: Wolfram Media Inc., 2002.

Wolfram S. ed. Theory and Applications of Cellular Automats. – Singapore: World Scientific. 1986.

Тоффоли Т., Марголус Н. Машины клеточных автоматов. – М.: Мир, 1991.

Клеточный автомат представляет собой дискретную динамическую систему, совокупность одинаковых клеток, одинаковым образом соединенных между собой. Все клетки образуют сеть (решетку) клеточных автоматов. Состояние каждой клетки определяются состоянием клеток, входящих в ее локальную окрестность и называемых ближайшими соседями. Окрестностью клетки с номером j – $O(j)$ называется множество его ближайших соседей. Состояние j-й клетки в момент времени $t+1$ определяется некоторым правилом F, которое можно выразить, например, языком булевой алгебры:

$$y_j(t+1) = F\left(y_j, O(j), t\right). \qquad (3.9.1)$$

Во многих задачах считается, что сама клетка относится к своим ближайшим соседям, т.е. $y_j \in O(j)$, в этом случае формула упрощается: $y_j(t+1) = F\left(O(j), t\right)$. Клеточные автоматы удовлетворяют таким правилам:

- изменение значений всех клеток происходит одновременно (единица измерения - такт);

-сеть клеточных автоматов однородная, т.е. правила изменения состояний для всех клеток одинаковые;

-на клетку могут повлиять лишь клетки из ее локальной окрестности;

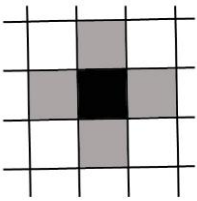
-множество состояний клетки конечное.

Клеточные автоматы могут иметь любую размерность, однако чаще всего рассматривают одномерные и двумерные системы клеточных автоматов.

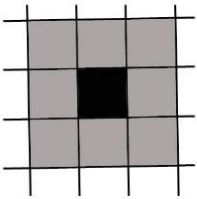
В случае двумерной решетки, элементами которой являются квадраты каждую клетку удобно задавать двумя индексами – $y_{i,j}$. Ближайшими соседями, входящими в окрестность элемента $y_{i,j}$, являются клетки, расположенные вверх-вниз и влево-вправо от него (так называемая окрестность фон Неймана: $O^N(y_{i,j}) = \left(y_{i-1,j}, y_{i,j-1}, y_{i,j}, y_{i,j+1}, y_{i+1,j}\right)$, можно добавить и диагональные элементы – окрестность Мура (G. Moore): $O^M(y_{i,j}) =$

$= \left(y_{i-1,j-1}, y_{i-1,j}, y_{i-1,j+1}, y_{i,j-1}, y_{i,j}, y_{i,j+1}, y_{i+1,j-1}, y_{i+1,j}, y_{i+1,j+1}\right)$.

Окрестность Неймана:



Окрестность Мура



В модели Мура каждая клетка имеет восемь соседей. Для устранения краевых эффектов решетка топологически «сворачивается в тор», т.е. первая строка считается продолжением последней, а последняя – предшествующей первой – так называемые граничные условия.

Это позволяет определять общее соотношение значения клетки на шаге $t+1$ по сравнению с шагом t:

$$y_{i,j}(t+1) = (y_{i-1,j-1}(t), y_{i-1,j}(t), y_{i-1,j+1}(t), y_{i,j-1}(t), y_{i,j}(t), y_{i,j+1}(t), y_{i+1,j-1}(t), y_{i+1,j}(t), y_{i+1,j+1}(t)). \quad (3.9.2)$$

Рассмотрим один из примеров использования клеточных автоматов – модель распространения инноваций и ее обобщение – модель распространения новостей. Модель диффузии (распространения) инноваций функционирует по

следующим правилам: каждый индивид, который способен принять инновацию, соответствует одной квадратной клетке на двумерной плоскости.

Bhargava S.C., Kumar A., Mukherjee A. A stochastic cellular automata model of innovation diffusion // Technological forecasting and social change, 1993. - Vol. 44. - № 1. - pp. 87-97.

При этом: 1) каждая клетка может находиться в двух состояниях: 1 - новинка принята; 0 - новинка не принята. 2) автомат, восприняв инновацию один раз, запоминает ее навсегда (состояние 1, которое не может быть измененным). 3) автомат одобряет решение относительно принятии новинки, ориентируясь на мнение восьми ближайших соседей, т.е. если в окрестности данной клетки (используется окрестность Мура) есть m приверженцев новинки, p – вероятность ее принятия, (если $pm > R$ (R – фиксированное значение – порог), то клетка принимает инновацию (значение 1). Клеточное моделирование позволяет строить значительно более реалистические модели рынка инноваций, чем традиционные подходы.

В модели диффузии информации предполагалось, что клетка может быть в одном из трех состояний: 1 - «свежая новость» (клетка окрашивается в черный цвет); 2 - новость, которая устарела, но сохраненная в виде сведений (серая клетка); 3 - клетка не имеет информации, переданной новостным сообщением (клетка белая, информация не дошла или уже забыта). В модели приняты такие правила распространения сообщений:

- сначала все поле состоит из белых клеток за исключением одной, черной, которая первой «приняла» новость (рис. 3.9.1 а);

- белая клетка может перекрашиваться только в черный цвет или оставаться белой (она может получать новость или оставаться «в неведении»);

- белая клетка перекрашивается, если выполняется условие, аналогичное модели диффузии инноваций: $pm > 1$ (это условие модифицируется для $m \le 2$: $1.5 \cdot pm > 1$);

-если клетка черная, а вокруг нее исключительно черные и серые, то она перекрашивается в серые цвета (новость устаревает, но сохраняется как сведения);

-если клетка серая, а вокруг нее исключительно серые и черные, то она перекрашивается в белый цвет (происходит старение новости при ее общеизвестности).

Описанная система клеточных автоматов качественно отражает процесс распространения сообщений среди отдельных информационных источников. Выяснилось, что состояние системы клеточных автоматов полностью стабилизируется за ограниченное количество ходов, т.е. процесс эволюции оказался сходящимся. Пример работы модели приведен на рис. 3.9.1.

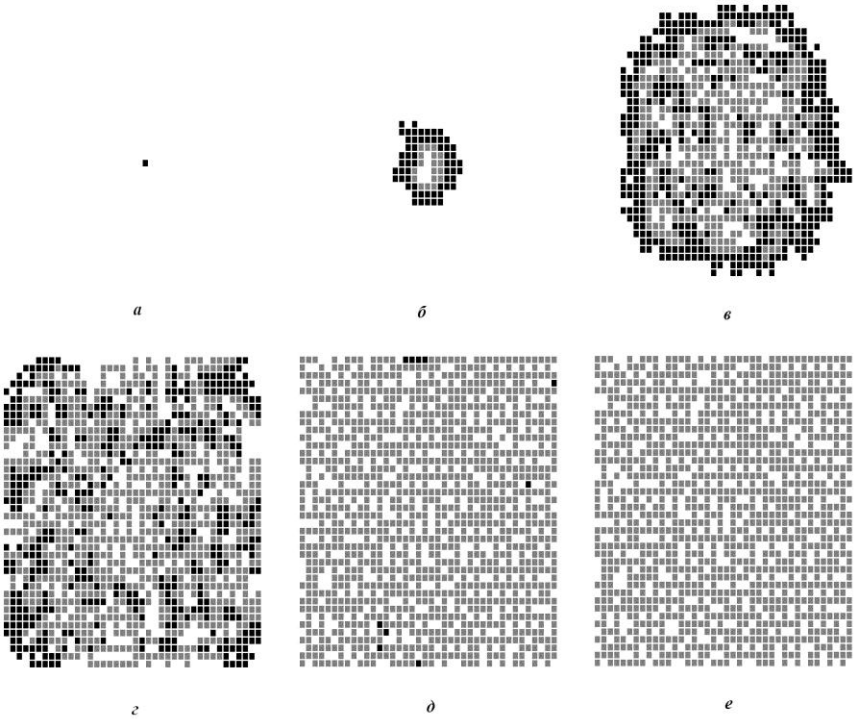


Рис. 3.9.1 – Процесс эволюции системы клеточных автоматов «диффузии новостей»: а - исходное состояние; б-д - промежуточные состояния; е - конечное состояние

Типичные зависимости количества клеток, которые находятся в разных состояниях в зависимости от шага итерации приведены на рис. 3.9.2. При этом, очевидно, суммарное количество клеток, которые находятся во всех трех состояниях на каждом шагу итерации постоянно и равно количеству клеток, а при стабилизации системы клеточных автоматов соотношения серых, белых и черных клеток приблизительно составляет: 3:1:0.

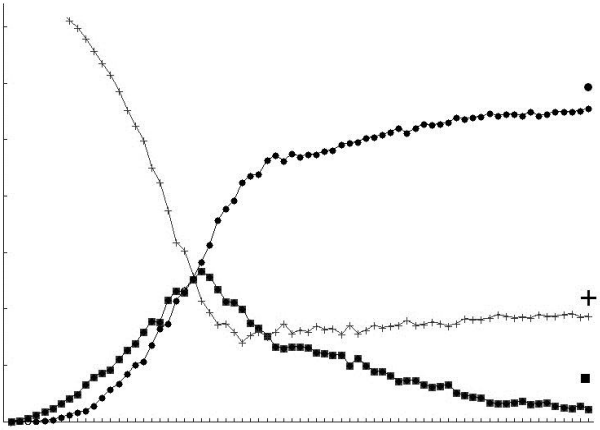


Рис. 3.9.2 – Количество клеток каждого из цветов в зависимости от шага эволюции: белые клетки - (+); серые клетки - (•); черные клетки – (■)

Детальный анализ полученных зависимостей позволил провести аналогии данной модели «диффузии информации» с некоторыми аналитическими соображениями. Результаты моделирования дают основания предположить, что эволюция серых клеток описывается некоторой непрерывной функцией:

$$x_g = f\left(t, \tau_g, \gamma_g\right), \qquad (3.9.3)$$

где t – время (шаг эволюции), τ_g – сдвиг по времени, обеспечивающий получение необходимого фрагмента

аналитической функции, γ_g – параметр крутизны данной функции.

Соответственно, динамика белых клеток x_w (количество клеток в момент t) может моделироваться «перевернутой» функцией x_g с аналогичными параметрами:

$$x_w = 1 - f\left(t, \tau_w, \gamma_w\right). \qquad (3.9.4)$$

Поскольку, как было сказано выше, всегда выполняется условие баланса, т.е. общее количество клеток в любой момент времени всегда постоянно, то условие нормирования можно записать следующим образом:

$$x_g + x_w + x_b = 1, \qquad (3.9.5)$$

где x_b – количество черных клеток в момент времени t.

Таким образом, получаем:

$$x_b = 1 - x_g - x_w = f\left(t, \tau_w, \gamma_w\right) - f\left(t, \tau_g, \gamma_g\right). \qquad (3.9.6)$$

Вид представленной на рис. 55 зависимости позволяет предположить, что в качестве функции $f\left(t, \tau, \gamma\right)$ может быть выбрано следующее выражение (логистическая функция):

$$f\left(t, \tau, \gamma\right) = \frac{C}{1 + e^{\gamma(t-\tau)}}, \qquad (3.9.7)$$

где C – некоторая нормирующая константа.

На рис. 3.9.3 приведены графики зависимости x_g, x_w, x_b от шага эволюции системы клеточных автоматов, полученные в результате аналитического моделирования.

Следует отметить, что зависимость диффузии новостей, полученная в результате моделирования, хорошо согласуется с «жизненным» поведением тематических информационных потоков на интернет-источниках (веб-сайтах), а на локальных временных промежутках – с традиционными моделями.

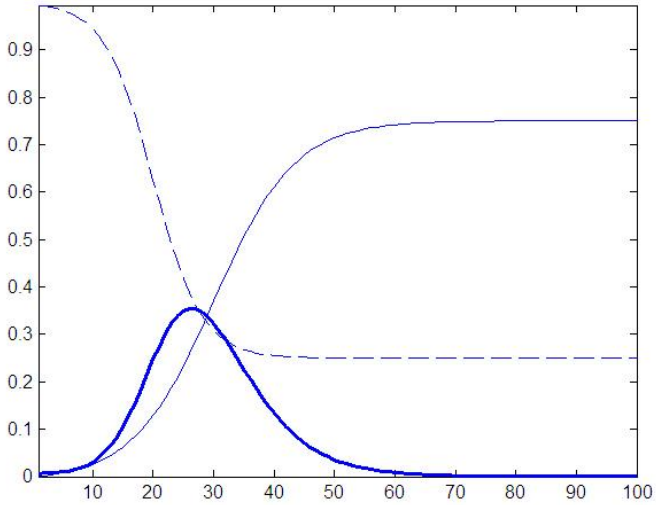


Рис. 3.9.3 – Непрерывные зависимости, полученные в результате аналитического моделирования, в зависимости от шага эволюции: сплошная линия – серые (x_g); пунктирная линия – белые (x_w); сплошная жирная линия – черные (x_b)

3.10. Самоорганизованная критичность

Пер Бак

(1948 – 2002)

Под системой, порождающей информацию, чаще всего предполагают реальную социальную или экономическую систему, ожидать от которой простой предсказуемой информации или единообразного поведения не приходится. В реальной системе информационное событие можно рассматривать в каком-то смысле как катастрофу, поскольку оно неожиданно. Если в большинстве случаев предсказание отдельного события кажется невозможным, то поведение системы в целом, ее отклик на воздействие или возмущение частично предсказуемо и является объектом научного исследования.

Термин «самоорганизация» (self-organizing), связанный с общей теорией систем, был введен В. Ашби (V. Ashby) в 1947 году и воспринят новой тогда кибернетикой, ее создателями Н. Винером (N. Winner), Г. Форстером (G. Forster) и др. В настоящее время это понятие чаще всего ассоциируется с именем П. Бака (P. Bak). В 1987-1988 году П. Бак, Ч. Танг (C. Tang) и К. Визенфельд (K. Wiesenfeld) в своих работах [70, 71] впервые детально описали клеточный автомат, приводивший систему к статистически одному и тому же «критическому» состоянию, названному ими состоянием самоорганизованной критичности. Типичная стратегия физики заключается в уменьшении количества степеней свободы в исследуемой задаче, например, в теории среднего поля, где окружение воздействует на оставшуюся степень свободы системы как некоторое среднее поле, оставляя для исследования только одну переменную.

Bak P., Tang C., Wiesenfeld K. Self-organized criticality: An explanation of 1/f-noise // Phys. Rev. Lett. 1987. –**59.** - pp. 381-384.

Bak P., Tang C., Wiesenfeld K. Self-organized criticality // Phys. Rev. A., 1988. – **38**. – № 1. – pp. 364-374.

Самой наглядной моделью, демонстрирующей самоорганизованную критичность, является куча песка, знакомая каждому с детства. Если песок сухой, то никакой кулич из него не построить, все тут же осыпается. В детские годы мало кто задумывался о том, как это происходит. Какой бы высоты куча не была, угол наклона конуса осыпания оставался неизменным, Это еще раз в эксперименте доказали в эксперименте выросшие дети, в Чикагском университете под руководством Х. Ягера (H. Yager) они экспериментировали с самой настоящей кучей песка. Состояние этой кучи можно назвать критическим, поскольку приложив минимальное возмущение, бросив сверху одну песчинку, поверхность кучи выйдет из равновесия, вниз сойдет лавина. А после ее схода останется снова куча песка поменьше, новые падающие песчинки достроят кучу до того же критического наклона, а новая брошенная песчинка вновь вызовет лавину. Куча всегда находится в критическом состоянии – малые возмущения вызывают реакцию,

непредсказуемую по размеру, и всегда самоорганизуется – сохраняет угол наклона поверхности (рис. 3.10.1).

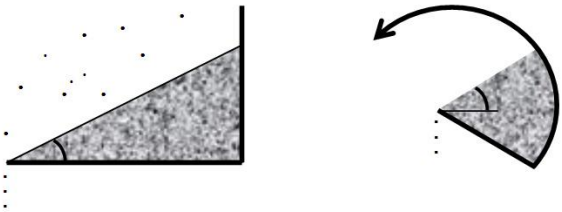


Рис. 3.10.1 – Коробка и вращающийся барабан с песком с одинаковым углом наклона боковой плоскости

При моделировании самоорганизованной критичности исследуется статистика схода лавин, когда одна брошенная песчинка вызывает лавину из других, лежащих на поверхности.

Рассмотрим дискретную систему, аналогичную куче песка в одномерном случае. Пусть $h(x)$ – высота кучи в точке x $(x=1,\ 2,\ ...,\ N)$. Кучу удобно изображать в двух видах – исходном (рис. 58 *а*) как функцию высоты от координат $h=h\ x$ и в виде приращений $z\ x=h\ x-h\ x+1$, которые показывают отличие высот в соседних точках – рис. 58 *б*. Левые части рис. 57 показывают начальное состояние кучи.

Введем правило 1: если разница высот в точке x больше некоторого критического значения $h\ x>h_c$, то лишние песчинки скатываются на соседние точки. Выбирая критическое значение $h_c=3$ правило 1 можно записать следующим образом:

$$\left.\begin{array}{l} z(x)\to z(x)-2, \\ z(x\pm1)\to z(x\pm1)+1 \end{array}\right|_{z(x)>2} \qquad (3.10.1)$$

Первое соотношение означает, что высота кучи в точке X уменьшается на две песчинки, второе, что в соседних точках (левой и правой) высота увеличится на одну песчинку.

На границах кучи будут выполняться граничные условия 1:

$$z(1)=0,$$
$$z(N)\to z(N)-1,$$
$$z(N-1)\to z(N-1)+1 \quad \Big|_{z(N)>2} \qquad (3.10.2)$$

Первое из условий 1 можно назвать «закрытым», поскольку наружу системы частица при этом никогда не выйдет, в противоположность «открытым» условиям на другой стороне, когда частица скатывается наружу и падает.

На рис. 3.10.2 слева изображено начальное состояние кучи – высоты $h\ x$ и приращения $z\ x$, справа – после осыпания. Так, например, при $x=6$ приращение до осыпания было равно $z\ 6=3>2$. После осыпания, две песчинки с позиции $x=6$, переходят по одной налево $x=5$ и направо $x=7$ – рис. 3.10.2 слева.

С одной стороны правило *1*, это дискретное нелинейное уравнение диффузии а, с другой стороны, это клеточный автомат, в котором состояние ячейки x в момент времени $t+1$ определяется состоянием/соседних ячеек в предыдущий момент времени t. Графически действие правила 1 можно представить так, как это сделано на рис. 58.

Очевидно, что у одномерной кучи, когда она осыпается согласно правилу 1 и условиям 1 из неравновесного состояния с $z\ x>>2$, есть одно критическое состояние $z\ x=2$ для любого x. В одномерном случае любое стабильное состояние является в определенном смысле критическим, поскольку любое малое возмущение будет приводить к тому, что оно пройдет по всей системе, а любое уменьшение наклона до $z\ x<2$ в любой точке остановит его. Это очень похоже на другие одномерные критические явления, такие как перколяция. Также следует отметить, что у такого состояния в одномерной модели нет пространственной структуры.

Аналогично одномерному П. Бак предложил правило для двумерной кучи (правило 2 и условия 2). В такой системе сохраняется действие правил 1 для каждого из направлений x и y, критическое значение традиционно выбирается равным 3:

$$\left.\begin{aligned} &z(x,y) \to z(x,y) - 4, \\ &z(x,y \pm 1) \to z(x,y \pm 1) + 1, \\ &z(x \pm 1,y) \to z(x \pm 1,y) + 1 \end{aligned}\right|_{z(x,y)>3} \qquad \begin{matrix} (3.10.3) \\ \text{(Правило 2)} \end{matrix}$$

$$z(0,y) = z(x,0) = z(N+1,y) = z(x,N+1) = 0 \qquad \begin{matrix} (3.10.4) \\ \text{(Условия 2)} \end{matrix}$$

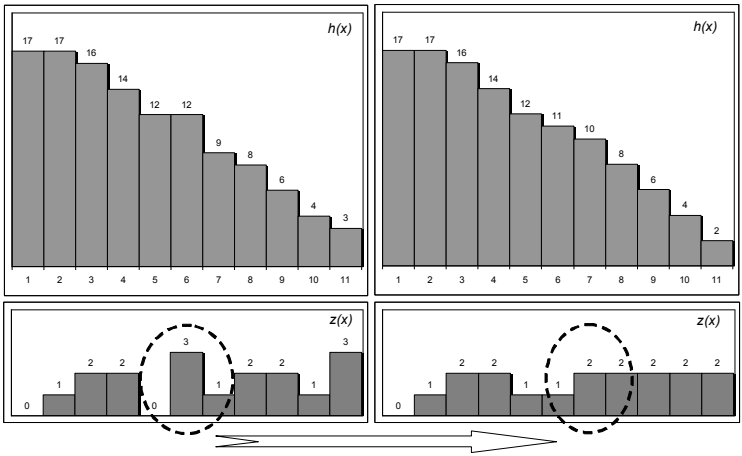


Рис. 3.10.4 – Одномерная куча до и после осыпания. Осыпался столбец 6 и крайний правый столбец 11 с «открытыми» граничными условиями

Указан вариант «закрытых» граничных условий 2 по всем направлениям. Конечно возможна любая комбинация «открытых» и «закрытых» условий. Условия 2 также могут быть модифицированы для «настоящей» кучи, насыпанной в углу, скажем, обувной коробки. Решетки, на которых строились самоорганизующиеся критические системы также разнообразны, например, проводились эксперименты на квадратных решетках в больших размерностях, на гексагональных решетках были даже получены точные аналитические результаты. Аналогичные клеточные автоматы возможно построить и на нерегулярных решетках.

Соответствие между величиной $z(x,y)$ и наклоном кучи не такое однозначное, как в 1D, поскольку теперь значение наклона $z(x,y)$ представляет собой средний наклон по диагонали системы и при осыпании частицы начнут двигаться в обоих направлениях x и y. В двумерном

случае уже нельзя говорить о том, что из неустойчивого состояния с $z(x,y) \gg 4$ система перейдет в одно и то же состояние, поскольку неустойчивость будет распространяться по обоим направлениям взаимозависимо. Конечное состояние системы будет существенно зависеть от начальных условий, но свойства этого получившегося состояния, например, наклон, будут всегда одинаковыми.

Получить систему в состоянии самоорганизованной критичности можно двумя различными способами. Либо осыпанием системы из случайного состояния с $z(x,y) \gg 4$ до равновесного состояния, либо насыпая на ровную поверхность $z\ x,y = 0$ песчинки одну за другой в случайно выбранных точках и выполняя процедуру согласно правилу 2 тогда, когда это будет необходимо. Определить момент когда система достигает критического уровня можно по тому, средний наклон кучи перестанет изменяться. Эксперимент показывает, что свойства систем полученных обоими способами не отличаются друг от друга.

Результаты, приведенные ниже, получены на квадратной решетке в двумерном случае согласно определенным ранее правилам и условиям. Сначала случайным образом выбирались $z\ x,y \gg 4$, после чего проводилась «релаксация» кучи и она осыпалась до устойчивого состояния. На рис. 59 представлено одно из таких стабильных состояний на двумерной решетке 500 х 500, где цвета от черного до белого соответствуют значениям $z\ x,y$ от 0 до 3.

Если в одной из наиболее неустойчивых точек системы (в нашем случае $z\ x,y = 3$), запустить процесс по *правилу 2* с *условиями 2*, положив $z\ x,y = 4$, т.е. добавить одну песчинку, то система начнет осыпаться, покатится лавина песка. Для каждой такой точки x, y системы, область, затронутая осыпанием, будет различна. На рис. 3.10.4 представлены несколько таких лавин, полученных осыпанием. Исходной послужила система, изображенная на рис. 3.10.3. Центры лавин отмечены на фоне белых снежных лавин черными точками, а пары цифр в скобках обозначают время осыпания и размер лавины.

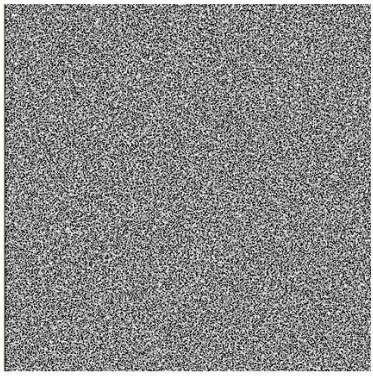


Рис. 3.10.3 – Стабильное состояние

Определим $D(s)$ – функцию распределения размеров возникающих лавин. Чтобы получить эту функцию, в каждой точке системы, где $z\ x,y = 3$ положим $z\ x,y = 4$ и запустим сход лавины, определим ее площадь – S, для получения достаточного количества лавин обработаем таким образом несколько систем в самоорганизованном состоянии, получая их из случайного начального состояния с $z\ x,y \gg 4$. На рис. 61 *а*) представлен вид зависимости $D(s)$, полученный обработкой набора систем размера 500 х 500. Распределение размеров лавин подчиняется степенному закону:

$$D(s) \sim s^{-\tau}, \quad \tau_{2D} \approx 1{,}1. \qquad (3.10.5)$$

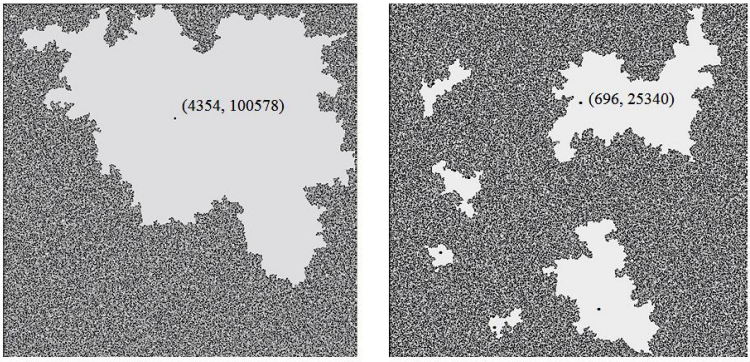
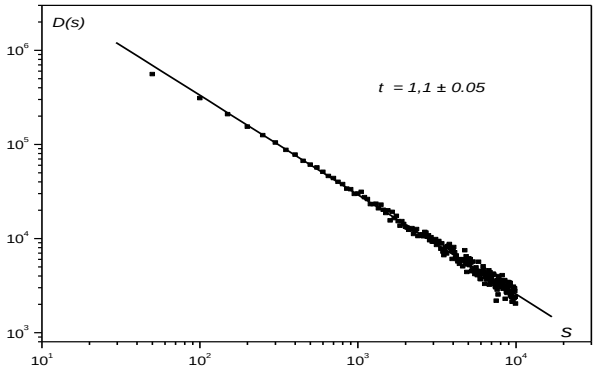


Рис. 3.10.4 – Лавины, полученные осыпанием

Аналогично возможно исследовать и временные характеристики этого процесса, введя $D(t)$ – функцию распределения времен осыпания этих лавин. В общем случае площадь S лавины больше, чем время ее осыпания t, поскольку в один момент осыпаются несколько песчинок. Распределение развития лавин также подчиняется степенному закону:

$$D(t) \sim t^{-\alpha}, \ \ \alpha_{2D} \approx 0,43. \qquad (3.10.6)$$

а



б

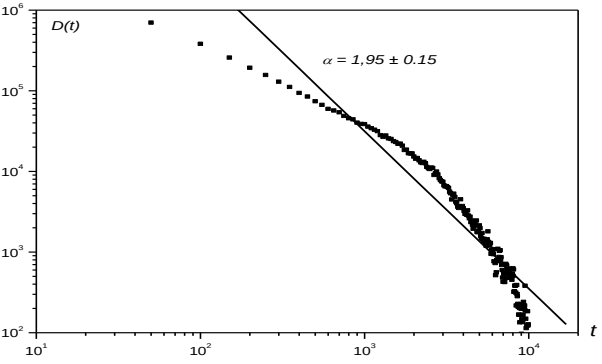


Рис. 3.10.5 – Распределения (а) размеров лавин $D(s)$ и (б) времен их осыпаний $D(t)$

Безусловно, индексы τ и α связаны как друг с другом, так и с другими индексами, характеризующими самморганизованное критическое состояние.

Размер «лавины» новостей, возникающих в информационных потоках при появлении новой темы порой кажется непредсказуемым, однако вполне поддается моделированию. Например, степенные распределения количества тематических публикаций, вполне соответствуют приведенным выше распределениям размеров рассматриваемых лавин «песка». В последнее десятилетие моделирование информационных процессов с помощью методов клеточных автоматов и теории самоорганизованной критичности получили большое распространение.

3.11. Перколяция

Хорошо изученными и важными в практическом отношении являются перколяционные сети.

Тарасевич Ю.Ю. Перколяция: теория, приложения, практика. – М.: УРСС, 2002.

Рассмотрим одну из наиболее простых постановок перколяционной задачи. Дана квадратная сетка (бесконечная), каждая связь которой имеет сопротивление r_1, такую связь для удобства будем называть черной. Случайным образом черные (проводящие) связи разрываются. Можно говорить о том, что при этом черные связи заменяются на белые с сопротивлением $r_2 = \infty$. Задача – необходимо найти такую концентрацию черных связей p_c при и выше которой существует связная часть черных связей, по которой можно добраться из одной бесконечности в другую, без перепрыгивания через белую связь. Такая, связная часть, простирающаяся на бесконечность называется бесконечным кластером. Конечно, реальная сетка всегда конечна, поэтому, в данном случае, предполагается, что ее размер намного больше так называемой корреляционной длины, в этом случае разные реализации структуры черных связей, получаемы при случайном вырезании имеют одни и те же свойства. Вероятность же специальных вырожденных распределений черных связей пренебрежимо мала. Здесь имеет место аналогия со случайным распределением молекул

газа в некотором объеме. Вероятность того, что в одной половине объема соберется весь газ, настолько маловероятна, что никогда не принимается во внимание. В тоже время, если молекул всего две, то эта вероятность равна $1/4$.

Кроме геометрической постановки задачи перколяции – возникновение бесконечного черного кластера, можно предложить и физическую постановку задачи, например, о протекании тока по черным связям. Черные связи проводят ток, белые нет. Необходимо определить сопротивление (проводимость) сетки в целом.

При $p > p_c$ проводимость всей сетки в целом – G не равна нулю (ток находит свой путь от одного контакта на «бесконечности» к другому по черному бесконечному кластеру). При $p < p_c$ $G = 0$, и, соответственно сопротивление всей сетки $R = 1/G = \infty$.

Снарский А.А., Безсуднов И.В., Севрюков В.А. Процессы переноса в макроскопи-чески неупоря-доченных средах. – М.: УРСС, 2007. – 304 с.

Конечно, в теории протекания не обязательно рассматривать именно двумерную квадратную сетку. Возможны любые размерности и типы сеток (однородных в среднем). Кроме того, можно говорить не только о задаче связей, а и задаче узлов, когда все связи проводящие, а проводящие (черные) узлы случайным образом с данной вероятностью вырезаны.

Как оказалось, задача протекания, появившаяся при формулировке прикладной инженерной задачи о протекании газа или жидкости через пористый фильтр, является одним из наиболее простых и наглядных примеров теории фазовых переходов второго рода и критических явлений. Так, многие характеристики, описываемые геометрические и физические свойства вблизи порога протекания p_c ведут себя универсальным образом, описываются критическими индексами, численное значение которых не зависит от вида сетки.

Рассмотрим некоторые геометрические характеристики перколяционной сетки. Таких геометрических характеристик много – среднее число кластеров размера s, распределение кластеров по размерам, средний размер кластера, мощность

бесконечного кластера, характеристики различных частей бесконечного кластера скелета, скелетона, мертвых концов,…) и т.п. Здесь мы рассмотрим только некоторые характеристики. Первая из них это $n_s\ p$ – распределение конечных кластеров по размерам, т.е. число кластеров из s узлов (связей) приходящихся на один узел (связь) решетки. Вторая характеристика удачно подходящая на роль параметра порядка $P\ p$ - мощность бесконечного кластера, вероятность того, что произвольный узел (связь) принадлежит бесконечному кластеру. Мощность бесконечного кластера – $P\ p$ выражается через $n_s\ p$, для этого достаточно учесть, что вероятность попасть на черный узел – p есть сумма вероятности попасть на бесконечный кластер $P\ p$ или на любой конечный $\sum_s sn_s\ p$:

$$P\ p + \sum_s sn_s\ p = p, \qquad (3.11.1)$$

откуда:

$$P\ p = p - \sum_s sn_s\ p\ . \qquad (3.11.2)$$

Вблизи порога протекания p_c мощность бесконечного кластера ведет себя аналогично параметру порядка в теории фазовых переходов второго рода:

$$P\ p\ \sim\ p - p_c\ ^{\beta}, \quad p - p_c\ / p_c\ . \qquad (3.11.3)$$

Роль температуры – T и критической температуры – T_c в фазовых переходах теперь играют концентрация – p хорошо проводящих связей/узлов (черная фаза) и порог протекания - p_c.

Близость к порогу протекания будем обозначать:

$$\tau = \frac{p - p_c}{p_c}. \qquad (3.11.4)$$

Рассмотренную аналогию между теорией фазовых переходов и теорией протекания можно углубить, введя в теорию протекания аналог безразмерного магнитного поля – h. В рассматриваемых здесь геометрических характеристиках перколяционных систем это делается довольно искусным и искусственным образом, вводится так называемый демон Кастеляйна-Фортуина – черный узел вне решетки, связанный с каждым черным узлом с вероятностью $1 - \exp(-h)$. Аналог свободной энергии в теории фазовых переходов может быть записан как

$$G(\tau, h) = \sum_s n_s e^{-hs}, \qquad (3.11.5)$$

где $e^{-hs} = (e^{-h})^s$ – доля конечных кластеров из s узлов, у которых ни один из узлов не связан с демоном Кастеляйна-Фортуина.

Параметр порядка в теории фазовых переходов может быть найден из G как производная по полю h, при $h = 0$

$$\left.\frac{\partial G}{\partial h}\right|_{h=0} = -\eta, \qquad (3.11.6)$$

и из (3.11.5) находим:

$$\left.\frac{\partial G}{\partial h}\right|_{h=0} = -\sum_s s n_s e^{-hs}\Big|_{h=0} = -\sum_s s n_s, \qquad (3.11.7)$$

что дает главную (сингулярную) часть $P_\infty(p)$ (3.11.2):

$$P_\infty(p) = p - \sum_s s n_s = p - \left.\frac{\partial G}{\partial h}\right|_{h=0}. \qquad (3.11.8)$$

При нулевом поле $h=0$, параметр порядка $P\ p$ ниже порога протекания $p<p_c$ равен нулю, что полностью аналогично ситуации в теории фазовых переходов – при $T>T_c$ ферромагнитное состояние (намагниченность $m\neq 0$) переходит в парамагнитное ($m=0$). При $h\neq 0$ и при $T>T_c$ имеется ненулевой параметр порядка (3.11.5) «обязанный» внешнему магнитному полю – h и поэтому пропорциональный ему. Легко увидеть, что введение демона Кастеляйна-Фортуина также оставляет параметр порядка теории протекания – $P\ p$ не равным нулю ниже порога протекания, т.е. и при $p<p_c$ существует бесконечный кластер, величина которого пропорциональна h. В самом деле, при $p<p_c$ формально нет бесконечного черного кластера, но при $h\neq 0$ ($h<<1$) каждый черный узел связан с другим через демона Кастеляйна-Фортуина с вероятностью

$$1-e^{-h}\approx 1-1+h=h, \qquad (3.11.9)$$

пропорциональной полю.

Что и означает существование бесконечного кластера ($P\ p\neq 0$) пропорционального h

$$P\ p<p_c\ \sim h. \qquad (3.11.10)$$

Обратимся теперь к физическим характеристикам в теории протекания, позволяющим намного более наглядным образом пояснить основные закономерности фазовых переходов. Будем теперь считать, что черные связи в перколяционной сети имеют сопротивление r_1, а разорванные (белые) – $r_2=\infty$. При размерах сетки $L>>\xi$ намного больше так называемого корреляционного радиуса ξ влияние конкретного случайного распределение черных и белых связей (случайная реализация структуры) становится несущественной и хорошо определенной величиной является полное сопротивление – R. Для того, чтобы абстрагироваться от данного конкретного размера сетки (L) удобно перейти от

сопротивления всего образца (решетки) к удельной эффективной проводимости – σ_e:

$$R = \frac{1}{\sigma_e}\frac{L}{L^{d-2}}, \qquad (3.11.11)$$

где $d = 2,3...$ размерность сетки.

По определению на размерах порядка и более корреляционного радиуса все свойства сетки в целом (в данном случае удельная эффективная проводимость) одинаковы, соответственно должны быть одинаковы и основные, главные элементы их протекательной структуры.

В качестве параметра порядка, описывающего фазовый переход, удобно ввести величину пропорциональную эффективной проводимости – σ_e. Как и для параметра порядка P p эффективная проводимость σ_e при $p > p_c$ не равна нулю, а ниже порога при $r_2 = \infty$ равна σ_e $p < p_c$ $= 0$. Такое поведение легко пояснить – выше порога протекания ток может протечь от одного контакта к другому (которые могут быть формально разнесены и на бесконечное расстояние) проходя по черным – проводящим, связям.

Это и означает, что существует бесконечный черный кластер. При $p < p_c$ существуют только конечные кластеры, с размером меньше корреляционной длины, они изолированы друг от друга т.к. $r_2 = \infty$ и поэтому ток через сетку пройти не может. Т.о. при $r_2 = \infty$ эффективная проводимость ноль σ_e $p < p_c$ $= 0$, а если r_2 большое, но конечное, т.е. $h = r_1 / r_2 << 1$, но не ноль, то ток сможет протечь от одного конечного кластера к другому. При этом, конечно же проводимость сетки будет пропорциональна h σ_e $p < p_c$ $\sim h$ и при $h \to 0$ эффективная проводимость σ_e $p < p_c$ $\to 0$. Таким образом нет необходимости вводить демона Кастеляйна-Фортуина, его роль играют белые сопротивления с большим, но конечным сопротивлением. Роль внешнего поля теперь играет отношение $h = r_1 / r_2 \ll 1$.

Критическим поведением вблизи порога протекания обладает не только плотность бесконечного кластера $P(p)$, но и множество других важных характеристик перколяционной сети, например, корреляционная длина, которая расходится при приближении к p_c:

$$\xi \sim (p - p_c)^{-\nu} , \qquad (3.11.12)$$

где ν - критический индекс корреляционной длины.

Критическим образом ведет себя и удельная эффективная проводимость p_e. Вблизи порога протекания выше $(p > p_c)$ и ниже $(p < p_c)$ имеет место:

$$\rho_e = \begin{cases} \rho_1 \tau^{-t}, & p > p_c, \\ \rho_1 h \tau^q \equiv \rho_2 \tau^q, & p < p_c. \end{cases} \qquad (3.11.13)$$

Аналогия между фазовым переходом второго рода и перколяционным переходом проявляется здесь в том, что если критическая температура – T_c и пороги протекания – p_c, для каждого материала, или, соответственно, решетки имеет свое численное значение, то критические индексы являются универсальными, зависящими только от размерности задачи, но не зависящими от типа решетки.

Рассмотрим вопрос применения ренорм-группового метода для вычисления критических индексов.

Вблизи порога протекания структура связных частей перколяционной сети (бесконечный кластер, при $p > p_c$ и «решеточные звери» при $p < p_c$) имеют фрактальную структуру, т.е. являются статистически самоподобными, Таким образом, переходя от одного масштаба к другому и требуя масштабной инвариантности можно получить приближенное значение критических индексов. Ниже мы рассмотрим несколько примеров использования метода ренорм-группы для вычисления порогов протекания и критического индекса корреляционной длинны ν.

Пример 1. Порог протекания треугольной решетки, задача узлов

Для удобства будем рассуждать в терминах протекания тока – проводящий узел (черный) проводит ток, не проводящий (белый) не проводит, все связи проводящие. На рис 3.11.1 *а* – треугольная решетка с обозначением проводящих (черных) и непроводящих (белых) узлов, с размером треугольной ячейки равным b; б - ренормированная решетка, треугольные ячейки (обозначены серым) представляют теперь новые узлы, связи между которыми обозначены жирными черными линиями. Новые узлы образуют новую (ренормированную) треугольную решетку с размером ячейки $b' = \sqrt{3} \cdot b$.

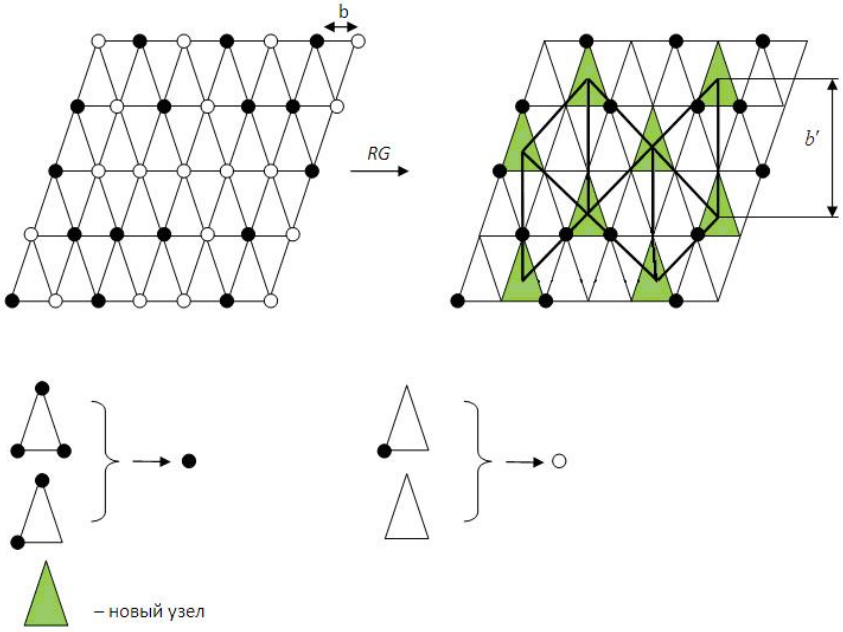


Рис. 3.11.1 – Схематическое изображение кластеров черной (хорошо проводящей фазы)

Правила преобразования черных узлов следующие - серый треугольник решетки переходит в черный узел ренормированной, если у него 2 или 3 узла черные, в противном случае он переходит в белый, непроводящий узел.

Вероятность черного узла в треугольной решетке равна p, поэтому вероятность встретить в новой ренормированной решетке проводящий, черный, узел равна $p^3+3p^2(1-p)$, где первое слагаемое «обязано» своим происхождением серому треугольнику с тремя черными узлами, а второе – с двумя. При этом, т.к. расположение черных и белых узлов во втором случае возможно тремя разными способами, второе слагаемое имеет сомножитель 3. Таким образом, вероятность встретить черный узел в ренормированной решетке p' равна:

$$p' = p^3 + 3\cdot p^2(1-p). \qquad (3.11.13)$$

Сеть, находящаяся на пороге протекания при ренорм-групповом преобразовании остается на пороге:

$$p'(p_c) = p_c, \qquad (3.11.14)$$

т.е. p_c – является неподвижной точкой преобразования. Тогда из уравнения, связывающего между собой p' и p получаем:

$$p_c = p_c^3 + 3\cdot p_c^2(1-p_c) \qquad (3.11.15)$$

Это уравнения имеет три решения $p_c = 1$, $p_c = 0$, $p_c = 1/2$. Первые два из них тривиальны – полностью «белая» или «черная» решетка остается таковой. Третье решение

$$p_c = \frac{1}{2}, \qquad (3.11.16)$$

и есть искомый порог протекания треугольной решетки для задачи узлов.

В данном примере, в отличие от нескольких других, все складывается настолько удачно, что полученное ренорм-групповым методом выражение для порога протекания совпадает с точным значением.

Покажем теперь, как можно выразить критический индекс корреляционной длины – ν.

Пусть в некоторой решетке $\xi = \xi_0 \left| p - p_c \right|^{-\nu}$, тогда в ренормированной $\xi' = \xi_0 \left| p' - p_c \right|^{-\nu}$ и $\xi' = \xi / a$, где $a = b' / b$. Таким образом:

$$a \left| p' - p_c \right|^{-\nu} = \left| p - p_c \right|^{-\nu} , \qquad (3.11.17)$$

откуда для критического индекса находим:

$$\frac{1}{\nu} = \frac{\ln \dfrac{p' - p_c}{p - p_c}}{\ln a} \equiv \frac{\ln \lambda}{\ln a}, \qquad \lambda = \frac{p' - p_c}{p - p_c} . \qquad (3.11.18)$$

Устремляя концентрацию к пороговой p_c для λ можно записать:

$$\lambda = \lim_{p \to p_c} \frac{p' - p_c}{p - p_c} = \left. \frac{dp'}{dp} \right|_{p = p' = p_c} , \qquad (3.11.19)$$

и для критического индекса получается простое выражение

$$\frac{1}{\nu} = \frac{\ln \left(dp' / dp \big|_{p_c} \right)}{\ln a} . \qquad (3.11.20)$$

Ранее, для треугольной решетки было получено $p' = p^3 + 3 \cdot p^2 (1 - p)$ и $p_c = 1/2$, откуда:

$$\lambda = \frac{\ln a}{\ln b} = \frac{1}{2} \frac{\ln 3}{\ln (3/2)} \approx 1.355 \qquad (3.11.21)$$

Точное значение (которое возможно получить для данной решетки) $\nu = 4/3 \approx 1.33$, и т.о. применение метода ренорм-группы дает очень хорошее приближение.

Пример 2. Квадратная решетка, задача связей

Одна из трудностей применения ренорм-группового метода является определение ренормализуемой ячейки и требований протекания к ней. В треугольной решетке этой трудности не было, ячейка выбиралась треугольной и протекание определялось наличием двух и более черных узлов. В случае квадратной решетки требуется более аккуратное рассуждение. Под перколирующим кластером мы понимаем как кластер, соединяющий «верх и низ», так и кластер, соединяющий «лево и право». Поэтому протекательное состояние для ячейки выберем, например, как протекание «слева направо». Тогда в качестве ячейки квадратной решетки можно выбрать конфигурацию, изображенную на рис. 3.11.2.

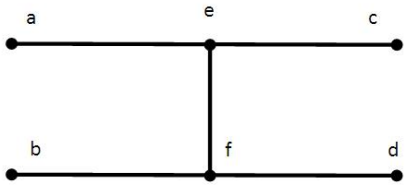


Рис. 3.11.2 – Ячейка квадратной решетки

Ячейка квадратной решетки, для исследование протекания «слева-направо». a и b – левые, c и d – правые контакты. Каждая из связей - ae, ec,… является проводящей с вероятностью p.

При ренормализации ячейка превращается в одну связь:

Соответственно $a = b'/b = 2$.

Ниже изображено, с указанием вероятности протекания, все проводящие «слева направо» конфигурации

ячейки, пунктиром указаны разорванные (выкинутые) связи, см. рис. 3.11.3.

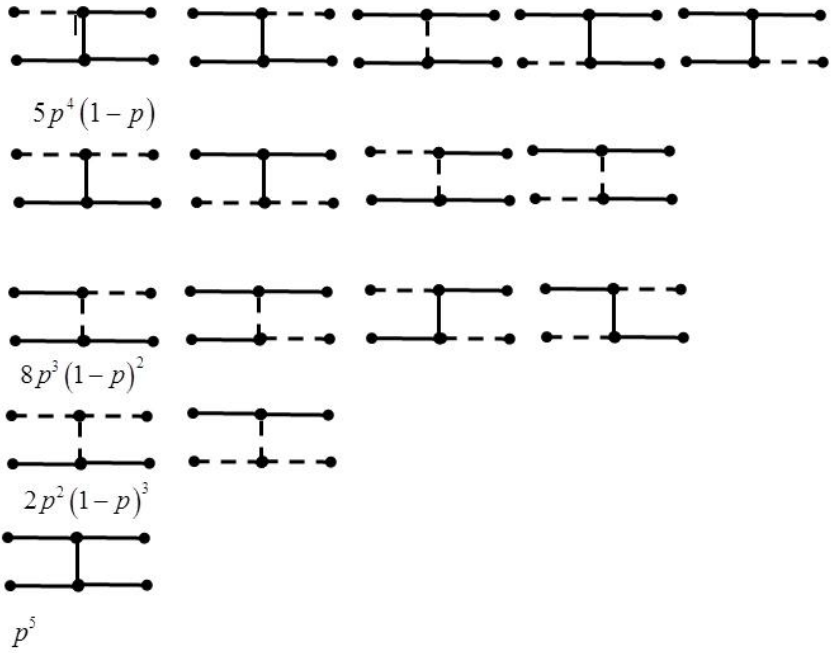


Рис. 3.11.3 – Проводящие конфигурации

Таким образом, вероятность получить протекательную ренормированную конфигурацию равна:

$$p' = p^5 + 5p^4(1-p) + 8p^3(1-p)^2 + 2p^2(1-p)^3 . \qquad (3.11.22)$$

Подставляя в правую и левую части $p' = p = p_c$ легко убедится, что решение полученного уравнения дает:

$$p_c = \frac{1}{2} . \qquad (3.11.23)$$

Сразу же можно вычислить и критический индекс корреляционной длинны:

$$\nu = \frac{\ln b}{\ln \lambda}, \quad \lambda = \left.\frac{dp'}{dp}\right|_{p=p_c}, \qquad (3.11.24)$$

беря производную от $p' = p'(p)$, по p и подставляя $p = p_c = 1/2$. Находим:

$$\left.\frac{dp'}{dp}\right|_{p=p_c} = \left(10p^4 - 20p^3 + 6p^2 + 4p\right)\Big|_{p=p_c=1/2} = \frac{13}{8}. \qquad (3.11.25)$$

Таким образом:

$$\nu = \frac{\ln 2}{\ln\left(13/8\right)} \approx 1.43 \qquad (3.11.26)$$

При точном значении $\nu = 4/3$.

Далеко не всегда метод ренормгруппы дает такие хорошие результаты. Для уточнения необходимо брать большие размеры ренормирующей ячейки, например, вместо ячейки с $b = 2$ брать ячейку с $b = 5$.

На первый взгляд такое уточнение не представляет значительного усложнения. Однако, на самом деле, это не просто значительное, а принципиально значительное усложнение, так как вместо 32-х для случая $b = 2$ комбинаций из которых 20 протекательных, в случае $b = 5$ число комбинаций равно $2^{25} \approx 3 \cdot 10^7$, из которых для начала необходимо выбрать протекательные, а затем еще и найти вероятность их появления.

3.12. Корреляционный и фрактальный анализ

3.12.1. Понятие фрактала

Термин фрактал был предложен Бенуа Мандельбротом (B. Mandelbrot) в 1975 году для обозначения нерегулярных самоподобных математических структур. Основное определение фрактала, данное Мандельбротом, звучало так: «Фракталом называется структура, которая состоит из частей, которые в каком-то смысле подобны целому». Следует признать, что это определение, ввиду своей нестрогости, не всегда верно. Можно привести много примеров самоподобных объектов, не являющихся фракталами, например, картина сходящихя к горизонту железнодорожных путей.

Мандельброт Б. Фрактальная геометрия природы. – М.: Институт компьютерных исследований, 2002. – 656 с.

Мандельброт Б. Фракталы, случай и финансы. – М.: Регулярная и хаотическая динамика, 2004. – 256 с.

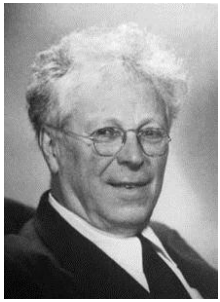
В самом простом случае небольшая часть фрактала содержит информацию обо всем фрактале. Строгое определение самоподобных множеств было дано Дж. Хатчинсоном (J. Hutchinson) в 1981 году. Он назвал множество самоподобным, если оно состоит из нескольких компонент, подобных всему этому множеству, т.е. компонент получаемых афинными преобразованиями - поворотом, сжатием и отражением исходного множества. Заметим, что такое строгое определение малопродуктивно. Во многих случаях исследуются подмножества, в которых компонента подобия всему множеству только приближенная, да и в самом множестве Мандельброта компоненты только похожи (а иногда и не похожи) на все множество.

Однако самоподобие – это хотя и необходимое, но далеко не достаточное свойство фракталов. Ведь нельзя же, в самом деле, считать фракталом точку, или плоскость, расчерченную клетками. Главная особенность фрактальных объектов состоит в том, что для их описания недостаточно «стандартной» топологической размерности d_τ, которая, как

известно, для линии равна 1 ($d_\tau = 1$ - линия одномерный объект), для поверхности $d_\tau = 2$, и т.д. Фракталам характерна геометрическая «изрезанность».

Поэтому используется специальное понятие фрактальной размерности, введенное Ф. Хаусдорфом (F. Hausdorf) и А.С. Безиковичем. Применительно к идеальным объектам классической евклидовой геометрии (линии, плоскости…) она давала те же численные значения, что и топологическая размерность, однако новая размерность обладала более тонкой чувствительностью ко всякого рода несовершенствам реальных объектов, позволяя различать и индивидуализировать то, что прежде было безлико и неразличимо. Размерность Хаусдорфа-Безиковича как раз и позволяет измерять степень «изрезанности». Размерность фрактальных объектов не является целым числом, характерным для привычных геометрических объектов. Вместе с тем, в большинстве случаев фракталы напоминают объекты, плотно занимающие реальное пространство, но не использующие его полностью.

Феликс Хаусдорф (1868-1942)



Абрам Самойлович Безикович (1891-1970)

Пусть есть множество G в евклидовом пространстве размерности d_τ. Это множество покрывается кубиками размерности d_τ, при этом длина ребра любого кубика не превышает некоторого значения δ, т.е. $\delta_i < \delta$.

Вводится зависящая от некоторого параметра d и δ сумма по всем элементам покрытия:

$$l_d(\delta) = \sum_i \delta_i^d \cdot \qquad (3.12.1)$$

Определим нижнюю грань данной суммы:

$$L_d(\delta) = \inf \sum_{i,\delta_i<\delta} \delta_i^d. \qquad (3.12.2)$$

При уменьшении максимальной длины δ, если параметр d будет достаточно велик, очевидно, будет выполняться:

$$\lim_{\delta\to 0} L_d(\delta) \to 0. \qquad (3.12.3)$$

При некотором достаточно малом значении параметра d будет выполняться:

$$\lim_{\delta\to 0} L_d(\delta) \to \infty. \qquad (3.12.4)$$

Промежуточное, критическое значение d_x, для которого выполняется:

$$\lim_{\delta\to 0} L_d(\delta) = \begin{cases} 0, & d > d_x, \\ \infty, & d < d_x, \end{cases} \qquad (3.12.5)$$

и называется размерностью Хаусдорфа-Безиковича (или фрактальной размерностью). Для простых геометрических объектов размерность Хаусдорфа-Безиковича совпадает с топологической (для отрезка d_x = 1, для квадрата d_x = 2, для куба d_x = 3 и т.д.)

Несмотря на то, что размерность Хаусдорфа-Безиковича с теоретической точки зрения определена безупречно, для реальных фрактальных объектов расчет этой размерности является весьма затруднительным. Поэтому вводится несколько упрощенный показатель – емкостная размерность d_c.

Гринченко В.Т. , Мацыпура В.Т. , Снарский А. А. Введение в нелинейную динамику. Хаос и фракталы. – К.: ЛКИ, 2007. – 264 с.

При определении этой размерности используются кубы (квадраты, отрезки …) с гранями одинакового размера. В этом случае, естественно, справедливо:

$$L_d(\delta) = N(\delta)\delta^{d_c}, \qquad (3.12.6)$$

где $N(\delta)$ – количество кубиков, покрывающего область G. Путем логарифмирования и перехода к пределу при уменьшении грани кубика ($\delta \to 0$) получаем:

$$d_c = -\lim_{\delta \to 0} \frac{\log N(\delta)}{\log \delta}, \qquad (3.12.7)$$

если этот предел существует. Следует отметить, что в большинстве численных методов определения фрактальной размерности используется именно d_c, при этом необходимо учитывать, что всегда справедливо условие: $d_x \le d_c$. Для регулярных самоподобных фракталов емкостная размерность и размерность Хаусдорфа-Безиковича совпадают, поэтому терминологически их часто не различают и говорят просто о фрактальной размерности объекта.

При проведении практических вычислений фрактальной размерности для реальных объектов используют следующий методический прием. Пусть на некотором этапе покрытия фрактала пришлось использовать $N(\delta)$ кубиков с гранями размера δ, а на другом – $N(\delta')$ элементов с гранями размера δ'. Ввиду предполагаемой степенной зависимости справедливо:

$$N(\delta) \sim \frac{1}{\delta^{d_i}}, \quad N(\delta') \sim \frac{1}{\delta'^{d_i}}, \qquad (3.12.8)$$

откуда значение d_c может оцениваться как:

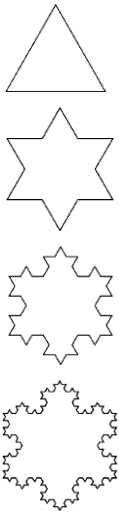
$$d_c = -\frac{\log\left(N(\delta)/N(\delta')\right)}{\log\left(\delta/\delta'\right)}. \qquad (3.12.9)$$

3.12.2. Абстрактные фракталы

Рассмотрим принципы формирования нескольких абстрактных фрактальных объектов, которые обладают выраженными свойствами самоподобия.

Нильс Фабиан Хельге фон Кох (1870-1924)

Построение фрактального множества, снежинки Хельге фон Коха (H. Von Koch) , начинается с правильного треугольника, длина стороны которого равна 1. Сторона треугольника считается базовым звеном. Далее, на любом шаге итерации каждое звено заменяется на образующий элемент – ломаную, которая состоит по краям исходного звена из отрезков длиной 1/3 от длины этого звена, между которыми размещаются две стороны правильного треугольника со стороной также в 1/3 длины звена. Все отрезки - стороны полученной ломаной считаются базовыми звеньями для следующей итерации. Кривая, получаемая в результате n-ой итерации при любом конечном n, называется предфракталом, и лишь при n, стремящимся к бесконечности, кривая Коха становится фракталом. Полученное в результате итерационного процесса фрактальное множество представляет собой линию бесконечной длины, которая ограничивает конечную площадь. Действительно, при каждом шаге число сторон результирующего многоугольника увеличивается в 4 раза, а длина каждой стороны уменьшается только в 3 раза, т.е. длина многоугольника на n-ой итерации равна $3\times\left(4/3\right)^n$ и стремится к



Первые 4 поколения снежинки Коха

бесконечности с ростом n. Площадь под кривой, если принять площадь первого образующего треугольника за единицу, равна:

$$S = 1 + \frac{1}{3}\sum_{k=0}^{\infty}\left(\frac{4}{9}\right)^k = \frac{8}{5} = 1{,}6. \qquad (3.12.10)$$

Подсчитаем фрактальную размерность снежинки Коха. Пусть длина стороны исходного треугольника равна единице. В данном случае роль кубиков, покрывающих рассматриваемую фигуру играют отрезки прямой. Тогда на нулевом шаге имеем: $\delta = 1,\ N(\delta) = 3$. Для второго шага справедливо: $\delta' = 1/3,\ N(\delta') = 12$. Этих данных достаточно для оценки фрактальной размерности:

$$d_c = -\frac{\log\left(N(\delta)/N(\delta')\right)}{\log\left(\delta/\delta'\right)} = -\frac{\log\left(3/12\right)}{\log\left(3\right)} = \frac{\log 4}{\log 3} \sim 1{,}26 \qquad (3.12.11)$$

Вацлав Франциск Серпинский

(1882-1969)

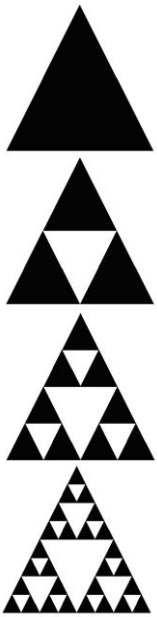
Самоподобный фрактал, предложенный в 1915 г. В. Серпинским (V. Serpinski), формируется по следующим правилам. Исходным множеством, соответствующим нулевому шагу, является равносторонний треугольник. Затем он разбивается на четыре области путем соединения серидины сторон исходного треугольника отрезками прямых. Затем удаляется внутренность центральной области исходного треугольника – малый внутренний «перевернутый треугольник». Затем, на следующем шаге итерации, этот процесс повторяется для каждого из трех оставшихся треугольников. Продолжая описанную процедуру до бесконечности, образуется множество, называемое салфеткой Серпинского.

Очевидно, фрактальная размерность салфетки Серпинского составляет:

$$d_c = -\frac{\log\left(N(\delta)/N(\delta')\right)}{\log(\delta/\delta')} = \frac{\log 3}{\log 2} \approx 1{,}58. \qquad (3.12.12)$$

Этот фрактал интересен тем, что занимаемая им площадь равна нулю. Для обоснования этого факта подсчитаем суммарную площадь частей, исключенную при построении. На первом шаге выбрасывается четвертая часть площади исходного треугольника, на втором шаге у каждого из трех треугольников удаляется четвертая часть площади и т.д. Таким образом, полная удаленная площадь вычисляется как сумма ряда (площадь исходного треугольника принята равной единице):

$$S = \frac{1}{4} + \frac{3}{4}\cdot\frac{1}{4} + \frac{3}{4}\cdot\frac{3}{4}\cdot\frac{1}{4} + \ldots =$$
$$= \frac{1}{4}\cdot\left[1 + \left(\frac{3}{4}\right) + \left(\frac{3}{4}\right)^2 + \left(\frac{3}{4}\right)^3 + \ldots\right] = \qquad (3.12.13)$$
$$= \frac{1}{4}\cdot\frac{1}{1-3/4} = 1.$$



Построение салфетки Серпинского

Таким образом, исключенная площадь совпадает с площадью исходного треугольника.

Алгоритм построения множества Мандельброта (рис. 29) основан на итеративном вычислении по формуле:

$$Z_{i+1} = Z_i^2 + C, \quad i = 0,\ 1,\ 2,\ \ldots, \qquad (3.12.14)$$

где Z_{i+1}, Z_i и C - комплексные переменные.

Итерации выполняются для каждой стартовой точки C прямоугольной или квадратной области – подмножестве комплексной плоскости. Итерационный процесс длится до тех пор, пока Z_i не выйдет за пределы окружности заданного

радиуса, центр которой лежит в точке (0,0), или после достаточно большого количества итераций. В зависимости от количества итераций, в течение которых Z_i остается внутри окружности, устанавливаются цвета точки C (рис. 3.12.1). Если Z_i остается внутри окружности в течение достаточно большого количества итераций, то эта точка растра окрашивается в черный цвет. Множеству Мандельброта принадлежат именно те точки, которые в течение бесконечного числа итераций не уходят в бесконечность. Так как количество итераций соответствует номеру цвета, то точки, которые находятся ближе к множеству Мандельброта, имеют более яркую окраску.

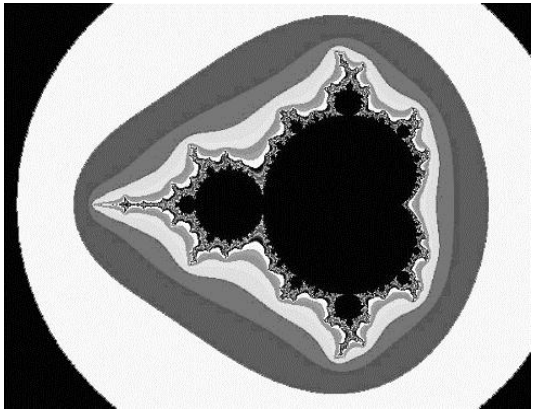


Рис. 3.12.1 – Множество Мандельброта

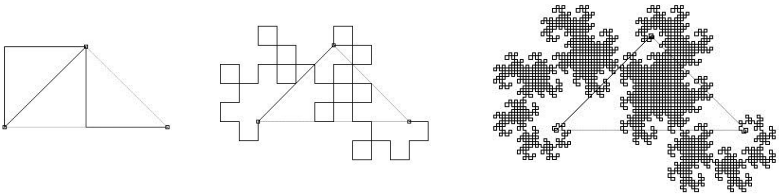
В 80-х годах прошлого столетия появился метод "Систем Итерационных Функций" (Iterated Functions System - IFS), представляющий собой простое средство получения фрактальных структур. IFS представляет собой систему функций, которые отображают одно многомерное множество на другое. Наиболее простая реализация IFS представляет собой афинные преобразования плоскости:

$$X' = AX + BY + C, \qquad (3.12.14)$$

$$Y' = DX + EY + F,$$

где X, Y – предыдущие значения координат, X', Y' – новые значения, A, B, C, D, E и F –коэффициенты.

В качестве примера использования IFS для построения фрактальных структур, можно привести "дракона" Хартера-Хейтуея (3.12.2), сформированного с помощью Java-апплета, приведенного в Интернет по адресу http://www.fractals.nsu.ru/fractals.chat.ru/ ifs2.htm (рис. 30). Использование IFS для сжатия обычных изображений, например, фотографии основаны на выявлении локального самоподобия (в отличие от фракталов, где наблюдается глобальное самоподобие).



а б в

Рис. 3.12.2 – "Дракон" Хартера-Хейтуея: а) – второй шаг итерации; б) – шестой шаг; в) – двенадцатый шаг

В 80-х годах М. Барнсли (M. Barnsley) и А. Слоан (A. Sloane) предложили идею сжатия и хранения графической информации, основанную на соображениях теории фракталов и динамических систем. На основе этой идеи был создан алгоритм фрактального сжатия информации, который позволяет сжимать некоторые образцы графической информации в 500-1000 раз. При этом каждое изображение кодируется несколькими простыми афинными преобразованиями. По алгоритму Барнсли в изображении происходит выделение пар областей, меньшая из которых подобная большей, и сохранение нескольких коэффициентов, которые кодируют преобразование, переводящее большую область в меньшую. При этом требуется, чтобы множество таких областей покрывало все изображение.

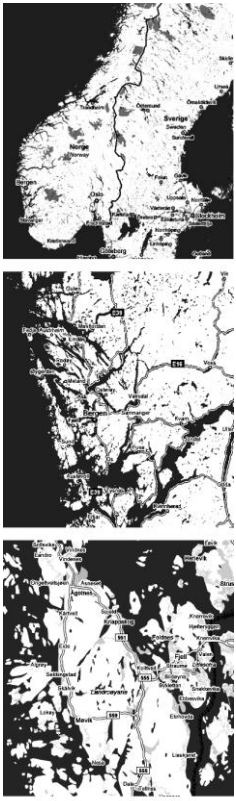
Один из лучших примеров проявления фракталов в природе - структура береговых линий. Действительно, на километровом отрезке побережье выглядит настолько же порезанным, как и на стокилометровом. Опыт показывает,

что длина береговой линии L зависит от масштаба l, которым проводятся измерения, и увеличивается с уменьшением последнего по степенному закону $L = \Lambda l^{1-\alpha}$, $\Lambda = const.$ Так, например, фрактальная размерность береговой линии Великобритании (рис. 31) составляет $\alpha \approx 1.52$.

3.12.3. Информационные потоки и фракталы

Применение теории фракталов при анализе информационных потоков позволяет с общей позиции взглянуть на закономерности, которые составляют основы информатики. Известно, что многие информационно-поисковые системы, включающие элементы кластерного анализа, позволяют автоматически обнаруживать новые классы и распределяют документы по этим классам. Соответственно, показано, что тематические информационные массивы представляют собой самоподобные развивающиеся структуры, однако их самоподобие справедливо лишь на статистическом уровне (например, распределение тематических кластеров документов по размерам).

Чем же определяется природа фрактальных свойств информационных потоков, порождаемых такими кластерными структурами? С одной стороны, параметрами ранговых распределений, а с другой стороны, механизмом развития информационных кластеров. Появление новых публикаций увеличивает размеры уже существующих кластеров и является причиной образования новых.



Побережье Норвегии в разных масштабах (по информации maps.google.com)

Фрактальные свойства характерны и для кластеров информационных веб-сайтов, на которых публикуются документы, соответствующие определенным тематикам.

Объемы сообщений в тематических информационных потоках образуют временные ряды, исследования которых все чаще используется теория фракталов.

Изучение характеристик временных рядов, порождаемых информационными потоками, сообщения которых отражают процессы, происходящие в реальном мире, дает возможность прогнозировать их динамику, выявлять скрытые корреляции, циклы и т.п.

В качестве иллюстраций приведены результаты реальных численных экспериментов. Как база для исследования фрактальных свойств рядов, отражающих интенсивность публикаций тематических информационных потоков, использовалась система контент-мониторинга новостей с веб-сайтов сети Интернет InfoStream. Тематика исследуемого информационного потока определялась запросом к этой системе. Для иллюстрации был проведен анализ сообщений онлайновых СМИ – массива из 14069 документов, опубликованных с 1 января 2006 г. по 31 декабря 2007 г., по тематике компьютерной вирусологии, удовлетворяющих запросу:

«компьютерный вирус» OR «вирусная атака» OR (антивирус AND (программа OR утилита OR Windows OR Linux)),

определенный суточной дискретностью (рис. 3.12.3).

Остановимся подробнее на некоторых методах анализа подобного типа временных рядов, порождаемых, в частности, информационными потоками.

3.12.4. Метод DFA

Один из универсальных подходов к выявлению самоподобия основывается на методе DFA (Detrended Fluctuation Analysis) – универсальном методе обработки рядов измерений.

Метод DFA представляет собой вариант дисперсионного анализа, который позволяет исследовать эффекты долговременных корреляций в нестационарных рядах. При

этом анализируется среднеквадратичная ошибка линейной аппроксимации в зависимости от размера отрезка аппроксимации.

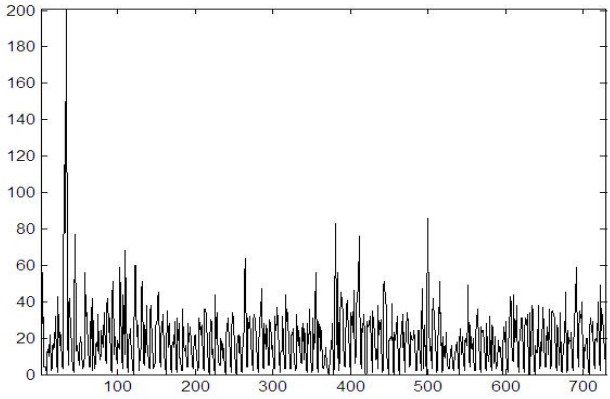


Рис. 3.12.3 – Количество тематических публикаций (ось ординат) в разрезе дат – порядковый номер дня (ось абсцисс)

Peng C.-K., Havlin S., Stanley H.E., Goldberger A.L. Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series // Chaos. - **5**. - 1995. - pp. 82.

Один из универсальных подходов к выявлению самоподобия основывается на методе DFA (Detrended Fluctuation Analysis) – универсальном методе обработки рядов измерений.

Метод DFA представляет собой вариант дисперсионного анализа, который позволяет исследовать эффекты долговременных корреляций в нестационарных рядах. При этом анализируется среднеквадратичная ошибка линейной аппроксимации в зависимости от размера отрезка аппроксимации. В рамках этого метода сначала осуществляется приведение данных к нулевому среднему (вычитание среднего значения $\langle F \rangle$ из временного ряда F_n, $n = 1, \ldots, N$):

$$y(k) = \sum_{i=1}^{k} \left[F_i - \langle F \rangle_N \right]. \qquad (3.12..16)$$

Потом ряд значений $y(k)$, $k=1,...,N$ разбивается на неперекрывающиеся отрезки длины n, в пределах каждого из которых методом наименьших квадратов определяется уравнение прямой, аппроксимирующей последовательность $y(k)$. Найденная аппроксимация $y_n(k)$ ($y_n(k)=a_nk+b_n$) рассматривается как локальный тренд.

Далее вычисляется среднеквадратичная ошибка линейной аппроксимации $D(n)$ при широком диапазоне значений n:

$$D\ n\ =\sqrt{\frac{1}{N}\sum_{k=1}^{N} y(k)-y_n(k)^{\ 2}}\ . \qquad (3.12..17)$$

В случае, когда зависимость $D(n)$ имеет степенной характер $D\ n\ \sim n^{\alpha}$, т.е. наличия линейного участка при двойном логарифмическом масштабе $\ln D\sim\alpha\ln n$, можно говорить о существовании скейлинга.

Как видно по рис. 3.12.4, значения $D(n)$ для выбранного информационного потока степенным образом зависят от n, т.е. в двойном логарифмическом масштабе эта зависимость близка к линейной.

5.12.5. Корреляциионный анализ

$$m=\frac{1}{N}\sum_{t=1}^{N}Y_t$$

Если обозначить через Y_t член ряда количества публикаций (количества электронных сообщений, поступивших, например, в день t, $t=1,...,N$), то автокорреляционная функция для этого ряда Y определяется как:

$$F(k)=\frac{1}{(N-k)}\sum_{t=1}^{N-k}(Y_{t+k}-m)(Y_t-m)= \\ =\frac{1}{N-k}\sum_{t=1}^{N-k}X_{t+k}X_t, \qquad (3.12.18)$$

где m – среднее значение ряда Y.

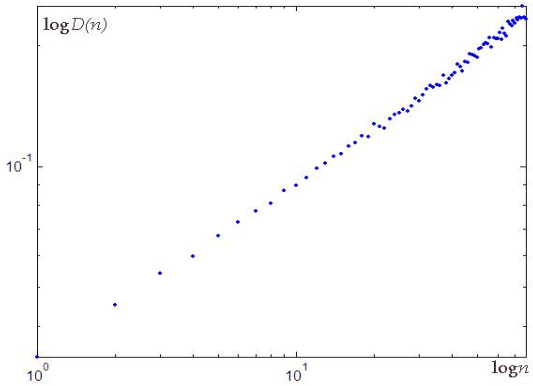


Рис. 3.12.4 – Зависимость D(n) ряда наблюдений (ось ординат) от длины отрезка аппроксимации n (ось абсцисс) в логарифмической шкале (α =0.7)

Предполагается, что ряд Y может содержать скрытую периодическую составляющую.

Известно, что функция автокорреляции обладает тем свойством, что если скрытая периодическая составляющая существует, то ее значение асимптотически приближается к квадрату среднего значения исходного ряда.

Бриллюэн Л. Наука и теория информации. – М.: Гос. изд. физ.-мат. лит., 1960.

Известна теорема, что если рассматриваемый ряд периодический, т.е. может быть представлен как:

$$X_t = u_0 \cos(\omega t + \varphi), \qquad (3.12.19)$$

то его автокорреляционная функция будет равна:

$$F(k) = \frac{1}{2} u_0^2 \cos(\omega k), \qquad (3.12.20)$$

т.е. автокорреляционная функция периодического ряда также является периодической, имеет ту же частоту, но без фазового угла φ.

Рассмотрим числовой ряд X, являющийся суммой некоторой содержательной составляющей N и синусоидальной сигнала S:

$$X_t = N_t + S_t. \qquad (3.12.21)$$

Найдем функцию автокорреляционную функцию для этого ряда (значения приведены к среднему $m = 0$ и разделены на среднеквадратичные отклонения):

$$F(k) = \frac{1}{N-k}\sum_{t=1}^{N-k} X_{k+t}X_t =$$
$$= \frac{1}{N-k}\sum_{t=1}^{N-k}(N_{k+t} + S_{k+t})(N_t + S_t) =$$
$$= \frac{1}{N-k}\sum_{t=1}^{N-k} N_{k+t}N_t + \frac{1}{N-k}\sum_{t=1}^{N-k} S_{k+t}S_t + \frac{1}{N-k}\sum_{t=1}^{N-k} N_{k+t}S_t + \frac{1}{N-k}\sum_{t=1}^{N-k} S_{k+t}N_t.$$

(3.12.22)

Очевидно, первое слагаемое – это функция непериодическая, асимптотически стремящаяся к нулю. Так как взаимная корреляция между N и S отсутствует, то то третье и четвертое слагаемое также стремятся к нулю. Таким образом, ненулевой вклад составляет второе слагаемое – автокорреляция сигнала S. Т.е. функция автокорреляции ряда X остается периодической.

В качестве иллюстрации рассмотрим модель информационного потока, в рамках которой рассматривается временной ряд, соответствующий количеству новых сообщений в сети. Предполагается, что ежедневное количество сообщений в сети растет по экспоненциальному закону (с очень небольшим значением экспоненциальной степени), и на это количество накладываются колебания, связанные с недельной цикличностью в работе информационных источников. Также принимается во внимание некоторый элемент случайности, выраженный соответствующими отклонениями.

Для получения соответствующего временного ряда были рассмотрены значения функции:

$$y = ae^{0.001x} + \sin(\pi x/7 + a), \qquad (3.12.23)$$

которая реализует простейшую модель информационного потока – экспонента отвечает за рост количества публикаций во времени (общая тенденция), синус – за недельную

периодичность, параметр a – за случайные отклонения. Количество публикаций y не может быть отрицательным числом. На рис. 3.12.5 представлен график модели .

Исходный ряд был обработан: приведен к нулевому среднему и нормирован (каждый член разделен на среднее). После этого были рассчитаны коэффициенты корреляции, которые для рядов измерений X длиной N рассчитываются по формуле:

$$\sigma^2 = \frac{1}{N}\sum_{t=1}^{N}(Y_t - m)^2$$

$$R(k) = \frac{F(k)}{\sigma^2}, \qquad (3.12.24)$$

где $F(k)$ – функция автокорреляции; σ^2 – дисперсия.

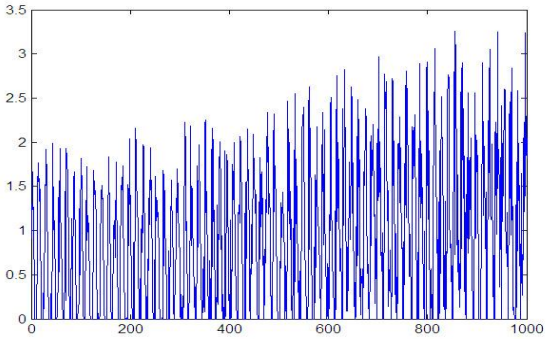


Рис. 3.12.5 – Модель потока с экспоненциальным ростом (ось абсцисс – переменная x – день, ось ординат – переменная y – количество публикаций)

На рис. 3.12.6 приведен график значений коэффициентов корреляций (ось абсцисс – переменная k, ось ординат – коэффициент корреляции $R(k)$.

Графическое представление коэффициента корреляции для ряда наблюдений, соответствующего динамике реального информационного потока веб-публикаций свидетельствует о неизменности корреляционных свойств по дням недели (рис. 3.12.7). Вместе с тем коэффициенты корреляции ряда наблюдений, усредненного по неделям, аппроксимируются гиперболической функцией, которая характеризует долгосрочную зависимость членов исходного ряда (рис. 3.12.8).

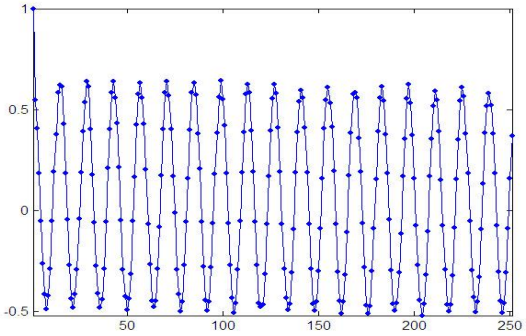


Рис. 3.12.6 – Значения коэффициентов корреляции модели

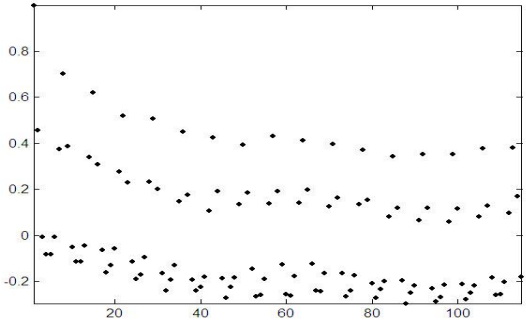


Рис. 3.12.7 – Коэффициенты корреляции ряда наблюдений $R(k)$ (ось ординат) в зависимости от k (ось абсцисс)

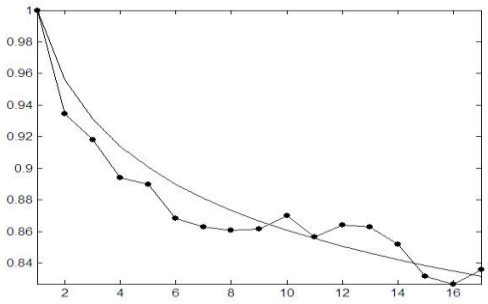


Рис. 3.12.8 – Коэффициенты корреляции ряда наблюдений $R(k)$ (ось ординат), усредненного по неделям в зависимости от k (ось абсцисс)

3.12.6. Фактор Фано

Для изучения поведения процессов принято использовать еще один показатель – индекс разброса дисперсии (IDC), так называемый фактор Фано. Эта величина определяется как отношение дисперсии количества событий $\sigma^2(k)$ на заданном окне наблюдений k к соответствующему математическому ожиданию $m(k)$:

$$\Phi(k) = \sigma^2(k) / m(k). \qquad (3.12.25)$$

Для самоподобных процессов выполняется соотношение:

$$\Phi(k) \approx 1 + Ck^{2H-1}, \qquad (3.12.26)$$

где C и H – константы.

Уго Фано (1912-2001)

Fano U. Ionization field of radiations. II. The fluctuations of the number of ions. Phys. Rev., 1947. – N 72.

3.12.7. R/S-анализ. Показатель Херста

Показатель Херста (H.E. Hurst) – H связан с коэффициентом нормированного размаха R/S, где R – вычисляемый определенным образом «размах» соответствующего временного ряда, а S – стандартное отклонение. Херст экспериментально обнаружил, что для многих временных рядов справедливо: $R/S = (N/2)^H$.

Доказано, что показатель Херста связан с традиционной «клеточной» фрактальной размерностью D простым соотношением:

$$D = 2 - H. \qquad (3.12.27)$$

Условие, при котором показатель Херста связан с фрактальной «клеточной» размерностью в соответствии с приведенной формулой, определено Е. Федером следующим образом: «… рассматривают клетки, размеры которых малы по сравнению как с длительностью процесса, так и с диапазоном изменения функции; поэтому соотношение справедливо, когда структура кривой, описывающая

фрактальную функцию, исследуется с высоким разрешением, т.е. в локальном пределе». Еще одним важным условием является самоаффинность функции.

Показатель Херста характеризует персистентность – склонность процесса к трендам (в отличие от обычного броуновского движения). Значение $H > ½$ означает, что направленная в определенную сторону динамика процесса в прошлом, вероятнее всего, повлечет продолжение движения в том же направлении. Если $H < ½$, то прогнозируется, что процесс изменит направленность. $H = ½$ означает неопределенность – броуновское движение.

Гарольд Эдвин Херст (*1880-1978*)

Hurst H. Long Term Storage Capacity of Reservoirs // Transactions of the American Society of Civil Engineers, 1951. – № 116.

В частности, для изучения фрактальных характеристик тематических информационных потоков для временных рядов $F(n),\ n=1,...,N,$ составленных из количества сообщений, опубликованных за промежуток времени от $n-1$ до n, изучалось значение показателя Херста, определяемое из соотношения:

$$R/S=\left(N/2\right)^{H},\quad N>>1. \qquad (3.12.28)$$

Здесь S – стандартное отклонение:

$$S=\sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(F(n)-\left\langle F\right\rangle_{N}\right)^{2}}, \qquad (3.13.29)$$

$$\left\langle F\right\rangle_{N}=\frac{1}{N}\sum_{n=1}^{N}F(n), \qquad (3.13.30)$$

а R - так называемый размах:

$$R(N)=\max_{1\le n\le N}X(n,N)-\min_{1\le n\le N}X(n,N), \qquad (3.12.31)$$

где

$$X(n,N)=\sum_{i=1}^{n}\left(F(i)-\left\langle F\right\rangle_{N}\right). \qquad (3.12.32)$$

Исследования фрактальных свойств рядов измерений, получаемых в результате мониторинга тематических информационных массивов из Интернет, свидетельствуют о

том, что показатель H принимает значения в диапазоне 0.65 ÷ 0.75, т.е. намного превышает ½. Поэтому можно утверждать, в этом случае обнаруживается персистентность (существование долговременных корреляций, которые могут быть связаны с проявлением детерминированного хаоса). Оказывается, что ряд $F(n)$ имеет фрактальную размерность D, равную $D = 2 - H \approx 1.35 \div 1.25$.

Исследования тематических информационных потоков подтверждают предположение о самоподобии и итеративности процессов в веб-пространстве. Републикации, цитирование, прямые ссылки и т. п. порождают самоподобие, проявляющееся в устойчивых статистических распределениях и известных эмпирических законах.

3.13. Законы Парето-Ципфа

Вильфредо Парето

(1848 – 1923)

Анализируя общественные процессы, В. Парето рассмотрел социальную среду как пирамиду, на вершине которой находятся некоторые люди, представляющие элиту. В результате исследований он математически сформулировал зависимость между величиной дохода и количеством лиц, которые его получают. Парето в 1906 году установил, что около 80 % земли в Италии принадлежит лишь 20 % ее жителей. Он пришел к заключению, что параметры полученного им распределения приблизительно одинаковы и не различаются принципиально в разных странах и в разное время. Точно такая же закономерность по Парето наблюдается и в распределении доходов между людьми.

Распределение доходов по Парето описывается уравнением $N \approx A/X^p$, где X – величина дохода, N – количество людей с доходом, равным или превышающим X, A и p - параметры распределения. В математической статистике это распределение получило имя Парето, при этом предполагаются естественные ограничения на параметры: $X \geq 1,\ p > 1$.

Распределению Парето присуще свойство устойчивости, т.е. сумма двух случайных переменных, которые имеют распределение Парето, также будет соответствовать этому распределению. Замеченное правило, называемое "законом Парето" или "принципом 80/20", применимо в очень многих областях. Например, при информационном поиске достаточно определить 20% важнейших ключевых слов, чтобы найти 80% необходимых документов, а затем расширить поиск или воспользоваться опцией "найти похожие" для полного решения задачи. Еще один пример: 80% посещений веб-сайта приходится лишь на 20% его веб-страниц.

При построении систем массового обслуживания, в том числе и информационно-поисковых систем, необходимо учитывать тот факт, что наиболее сложным функциональным возможностям системы, на реализацию которых уходит 80 и больше процентов трудозатрат, будут пользоваться не более чем 20 процентов пользователей данной системы.

В строгой формулировке этот эффект носит название принципа Парето. Предположим, что последовательность x_1, x_2, ...x_n, ... соответствует размерам доходов отдельных людей. После ранжирования этой последовательности по убыванию получается новая последовательность x_1, x_2, ..., x_r, ... (элементы x_r расположены в порядке убывания).

Предположим, что N – общее число людей, у которых доход составляет не менее $x_{(r)}$, т.е. $N = r$. Тогда правило Парето можно переписать в таком виде:

$$r = \frac{A}{x_{(r)}{}^{p}}. \qquad (3.13.1)$$

Откуда:

$$x_{(r)} = \left(\frac{A}{r}\right)^{\frac{1}{p}}. \qquad (3.13.2)$$

Рассматривается сумма первых n ($n=1,\ 2,\ ...,\ N$) значений величины x_r, т.е. общая величина дохода наиболее богатых людей - $m\ n$:

$$m\ n\ =\sum_{r=1}^{n}x_r\ =\sum_{r=1}^{n}\left(\frac{A}{r}\right)^{1/p}=\sum_{r=1}^{n}\frac{C}{r^{\gamma}},\qquad (3.13.3)$$

где $\gamma=1-1/\ p;\ \ C=A^{1/p}.$

Переходя от дискретных величин к непрерывным (предполагая, что $n>>1$), имеем:

$$m\ n\ \approx\int_{1}^{n}\frac{C}{r^{\gamma}}dr\approx\frac{C}{1-\gamma}n^{1-\gamma}.\qquad (3.13.4)$$

В безразмерных переменных $\mu=m\ n\ /m\ N$ - и $\nu=n/N$ последнее равенство имеет вид (см. рис. 23):

$$\mu=\nu^{1-\gamma}.\qquad (3.13.5)$$

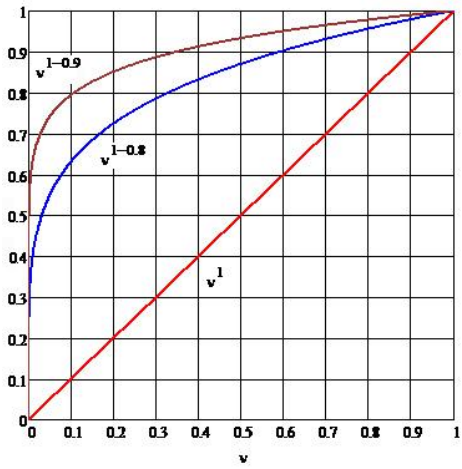


Рис. 3.13.1 – Распределение Парето для различных значений параметров: зависимость $\mu=\nu^{1-\gamma}$ для трех случаев: $\gamma=0$, $\gamma=0.8,\ \ \gamma=0.9$

Величина μ – в нашем примере – относительное количество дохода, получаемого первыми по рангу n людьми, доля которых (относительно всех людей) равна ν.

Джордж Ципф (1902 -1950)

Manning C.D., Schütze H. Foundations of Statistical Natural Language Processing - Cambridge, Massachusetts: The MIT Press, 1999.

Дж. Ципф изучил использование статистических свойств языка в текстовых документах и выявил несколько эмпирических законов, которые представил как эмпирическое доказательство своего «принципа наименьшего количества усилий». Он экспериментально показал, что распределение слов естественного языка подчиняется закону, который часто цитируется как первый закон Ципфа, относящийся к распределению частоты слова в тексте. Этот закон можно сформулировать таким образом. Если для какого-нибудь довольно большого текста составить список всех слов, которые встретились в нем, а потом ранжировать эти слова в порядке убывания частоты их появления в тексте, то для любого слова произведение его ранга и частоты появления будет величиной постоянной: $f \times r = c$, где f - частота встречаемости слова в тексте; r - ранг слова в списке; c - эмпирическая постоянная величина (коэффициент Ципфа). Для славянских языков, в частности, коэффициент Ципфа составляет приблизительно 0,06-0,07.

Приведенная зависимость отражает тот факт, что существует небольшой словарь, который составляет большую часть лексем текста. Это главным образом служебные слова. Например, приведенный в монографии К.Д. Маннинга и Г. Шютце анализ романа «Том Сойер», позволил выделить 11.000 английских слов. При этом было обнаружено двенадцать слов (the, and, и др.), каждое из которых охватывает более 1 % лексем в романе.

Ципф объяснял гиперболическое распределение «принципом наименьшего количества усилий» предполагая что при создании текста меньше усилий уходит на повторение некоторых слов, чем на использование новых, т.е. на обращение к «оперативной памяти, а не к долговременной».

Ципф сформулировал еще одну закономерность, которая состоит в том, что частота и количество слов, которые входят в текст с данной частотой, также связанны подобным соотношением, а именно:

$$N(f)=\frac{B}{f^{\beta}},$$

(3.13.6)

где $N(f)$ – количество различных слов, каждое из которых используется в тексте f раз, B – некоторая константа нормирования.

Лексема – слово как абстрактная единица морфологического анализа. В одну лексему объединяются различные словоформы одного слова. Например, ***словарь****,* ***словарем****,* ***словарю*** *– это формы одной и той же лексемы, по соглашению пишущейся как* ***словарь****.*

Существует простая количественная модель определения зависимости частоты от ранга. Предположим, что генерируется случайный текст обезьяной на пишущей машинке. С вероятностью p генерируется пробел, а с вероятностью $(1-p)$ – другие символы, каждый из которых имеет равную вероятность. Показано, что полученный таким образом текст будет давать результаты, близкие по форме распределению Ципфа.

Более сложную модель генерации случайного текста, удовлетворяющего второму закону Ципфа, предложил Г.А. Саймон в 1955 г. В соответствии с этой моделью, если текст достиг размера в n слов, тогда то, каким будет $(n+1)$-е слово текста определяется двумя допущениями:

1. Пусть $N(f,n)$ - количество разных слов, каждое из которых использовалось f раз среди первых n слов текста. Тогда вероятность того, что $(n+1)$-ым окажется слово, которое до того использовалось f

раз пропорционально $f \cdot N(f,n)$ - общему количеству появления всех слов, каждое из которых до этого использовалось f раз.

2. С вероятностью δ $(n+1)$-ым словом будет новое слово.

Из допущения 1 следует:

$$N(f,n+1)-N(f,n)=K(n)\,(f-1)\cdot N(f-1,n)-f\cdot N(f,n)\ , \quad (3.13.6)$$

где $K(n)$ - коэффициент пропорциональности.

Аналогично допущение 2 приводит к уравнению:

$$N(1,n+1)-N(1,n)=\delta-K(n)\cdot N(1,n). \quad (3.13.7)$$

Из условия того, что вероятность генерации $(n+1)$-го слова равна 1, имеем:

$$\sum_f K(n)\cdot f\cdot N(f,n)+\delta=K(n)\sum_f f\cdot N(f,n)+\delta=1\,. \quad (3.13.8)$$

Учитывая то, что $\sum_f f\cdot N(f,n)+\delta=n,$ имеем:

$$K(n)=\frac{1-\delta}{n}. \quad (3.13.9)$$

Кроме того, вводится еще одно допущение, состоящее в том, что для всех f выполняется:

$$\frac{N(f,n+1)}{N(f,n)}=\frac{n+1}{n}. \quad (3.13.10)$$

Из последнего допущения следует, что

$$\frac{N(f,n)}{N(f-1,n)}=\frac{N(f,n+1)}{N(f-1,n+1)}=\varphi(f). \quad (3.13.11)$$

При этом $\varphi(f)$ не зависит от n и, с учетом предыдущих уравнений:

$$\varphi(f)=\frac{(1-\delta)(f-1)}{1+(1-\delta)f)}. \quad (3.13.12)$$

Переходя к функции $N(f)=N(f,n)/n$ для $\varphi(f)$, имеем:

$$\frac{N(f)}{N(f-1)}=\varphi(f). \quad (3.13.13)$$

Использовав последнее уравнение $f-1$ раз получаем:

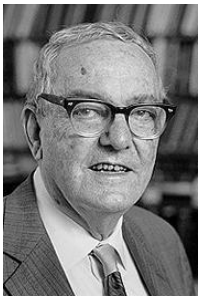
$$N(f)=\varphi(f)N(f-1)= \\ =\varphi(f)\varphi(f-1)\cdots\varphi(2)N(1). \qquad (3.13.14)$$

Введя обозначение $\rho=1/(1-\delta)$ последнее уравнение можно переписать:

$$N(f)=\frac{(f-1)(f-2)\cdots 2\cdot 1}{(f+\rho)(f+\rho-1)\cdots(\rho+2)}N(1)=\frac{\Gamma(f)\Gamma(f+2)}{\Gamma(f+\rho+1)}N(1). \qquad (3.13.15)$$

Учитывая то, что $\frac{\Gamma(f)}{\Gamma(f+\rho)}\sim f^{-\rho}$ при $f\to\infty$ и обозначив $\beta=1+\rho$, имеем окончательно:

$$N(f)\sim f^{-\beta}. \qquad (3.13.16)$$



Г.А. Саймон (Herbert A. Simon, 1916-2001)

Распределение Ципфа часто искажается на практике ввиду недостаточных объемов текстовых корпусов, что приводит к проблеме оценки параметров статистических моделей. С другой стороны, соотношение между рангом и частотой была взята Солтоном в 1975 г. [116] как отправная точка для выбора терминов для индексирования. Далее им рассматривалась идея сортировки слов в соответствии с их частотой в корпусе. Как второй шаг высокочастотные слова могут быть устранены, потому что они не являются хорошими различительными признаками для документов коллекции. На третьем шаге термы с низкой частотой, определяемой некоторым порогом (например слова, которые встречаются только единожды или дважды) удаляются, потому что они встречаются так нечасто, что редко используются в запросах пользователей. Используя этот подход, можно значительно уменьшить размер индекса поисковой системы. Более принципиальный подход к

подбору индексных термов – учет их весовых значений. В весовых моделях среднечастотные термы оказываются самыми весомыми, так как они являются наиболее существенными при отборе того или иного документа (наиболее частотные слова встречаются одновременно в большом количестве документов, а низкочастотные могут не входить в документы, интересующие пользователя).

Еще один эмпирический закон, сформулированный Ципфом состоит в том, что количество значений слова коррелирует с квадратным корнем его частоты. Подразумевалось, что нечасто используемые слова менее неоднозначны, а это подтверждает то, что высокочастотные слова не подходят для внесения в индексы информационно-поисковых систем.

Ципф также определил, что длина слова обратно пропорциональна его частоте, что может быть легко проверено путем простого анализа списка служебных слов. Последний закон действительно служит примером принципа экономии усилий: более короткие слова требуют меньше усилий при воспроизведении, и таким образом, используются более часто. Этот «закон» можно подтвердить, рассматривая приведенную выше модель генерации слов обезьяной. Легко видеть, что вероятность генерации слова уменьшается с длиной, вероятность слова из n непробельных символов равна:

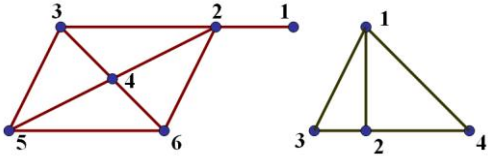
$$(1-p)^n \cdot p, \qquad (3.13.17)$$

где p - вероятность генерации пробела.

Хотя закон Ципфа дает интересные общие характеристики слов в корпусах, в общем случае замечены некоторые ограничения его применимости при получении статистических характеристик коллекций документов, состоящих из множества независимых документов разных авторов. Законам Ципфа удовлетворяют не только слова из одного текста, но многие объекты современного информационного пространства.

Вопросы для контроля

1. Привести примеры графов, которые а) можно и б) нельзя обойти, не проходя по кокой-либо связи дважды.
2. Сформулировать понятие социальной сети.
3. Перечислить основные характеристики сложных сетей.
4. Какие характеристики сетей Эрдеша-Реньи вы знаете?
5. Что такое средний минимальный путь (SP)?
6. Как зависит от размера сети средний минимальный путь (SP) для а)квадратной сетки, б)сети Эрдеша-Реньи, в)малого мира Ваттса-Строгаца?
7. Что такое коэффициент кластеризации данного узла? Какое свойство сети он характеризует?
8. Найти коэффициент кластеризации для узлов изображенных графов



9. В каком случае действенность (Efficiency) более удачно характеризует свойства сети, чем средний кратчайший путь (SP)? Приведите конкретный пример.
10. Как вычисляется средний кратчайший путь (SP) для сети с весом связей?
11. Как связано число ближайших m-ных соседей z_m с z_1 и z_2?
12. Сформулировать критерий Моллоя-Рида.
13. Записать функцию распределения степеней узлов для сети Эрдеша-Реньи в случае большого числа узлов.
14. Записать функцию распределения степеней узлов для масштабно-инвариантной (SF) сети в случае большого числа узлов. В чем особенность нормировки этой функции?
15. Приведите пример детерминированной масштабно-инвариантной сети.
16. Нарисуйте несколько шагов построения (u,v) - цветка для (1,2), (1,3), (1,4), (2,2) , (2,3), (3,3).

17. Чем отличаются алгоритмы построения малого мира (small world) в моделях Ваттса-Строгаца и Ваттса-Строгац-Ньюмена?

18. Алгоритм построения перколяционной сети.

19. Что такое порог протекания?

20. Чему равен порог протекания для одномерной решетки? Для бесконечно мерной?

21. Что общего и различного в свойствах бесконечного (гигантского) кластера в перколяционной и сложной сети?

22. Найти Паде-аппроксиманты для следующих функций:

$ctg(x), \quad x \approx 0$

$1/\cos(x), \quad x \approx \pi/2$

$1/\sin(x), \quad x \approx 0, \quad x \approx \pi$

23. Вывести распределение Коши – (3.4.5).

24. Показать, что формула Планка для теплового излучения (3.5.12) удовлетворяет скейлинговому соотношению (закону Вина) (3.5.14).

25. Что характеризует параметр порядка?

26. Сформулировать основные положения феноменологической теории фазовых переходов второго рода.

27. Дайте определение бесконечного кластера.

28. Изобразите (качественно) зависимость плотности бесконечного кластера от концентрации связей (узлов).

29. Какие есть основания утверждать, что перколяционный переход есть фазовый переход второго рода?

30. Основы кластерного анализи. Пояснить алгоритм иерархического группирования-объединения.

31. Основы кластерного анализа. Метод *k-means*. Какая функция качества кластеризации максимизируется этим алгоритмом?

32. Пояснить основные концепцию клеточных автоматов. Привести алгоритм модели диффузии информации.

33. Привести определение и алгоритм расчета показателя Херста. Как фрактальная размерность самоподобного временного ряда соотносится с показателем Херста?

34. Дать определение задачи линейной классификации. Написать и пояснить формулу Роккио для расчета вектора-профайла категории C_i.

35. Привести формулу и алгоритм вычисления автокорреляционной функции.

36. Привести основные метрики расстояния между документами, используемые в кластерном анализе.

37. Булевая модель поиска. Дать определение ДНФ та привести к єтой форме запрос – логическое выражение: $q = a \wedge (b \vee \neg c)$. Сколько операций дизъюнкции будет использовано в полученной ДНФ?

38. Написать формулу вычисления меры близости документа и запроса в соответствии с векторно-пространственной моделью поиска.

39. Привести байесовский критерий принадлежности документа к определенной категории.

40. Вывести формулу вычисления поискового статуса в соответствии с вероятностной моделью поиска.

41. Метод опорных векторов (SVM). Разделяющая полоса задается системой неравенств:

$$3x - 4y > 4;$$
$$3x - 4y < 6.$$

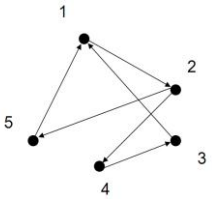
Чему равна ширина этой полосы?

42. Показатели качества поиска/классификации. Написать формулы для расчета полноты, точности и F-меры информационного поиска. Рассчитать F-меру, если значение полноты составляет 0,8, а точности – 0,6.

43. Метод латентного семантического индексирования (LSA/LSI). Что такое сингулярное разложение матрицы и его применение в методе LSA/LSI

44. Написать систему уравнений для

рекурсивного подсчета коєффициентов авторства и посредничества в алгоритме HITS, а также провести расчет (4 итерации) для сети:



45. Привести имитационную имитационную модель та итеративную формулу для расчета параметра PageRank.

46. Класссическое определение $TF \cdot IDF$. Терм w_i встречается в документе $d^{(j)}$ с нормализованной частотой, равной 0,5. Число документов в информационном массиве, содержащих терм w_i равно 20, а общее количество документов в массиве – 1000. Чему равно значение $TF \cdot IDF$?