

Цепи Маркова и их применения

Творческая по программированию

Выполнил: **Дубатовкин Даниил**

Группа: **10иЗ**

1 Введение

Сколько раз надо перетасовать колоду, чтобы карты находились в абсолютно случайном порядке? Сколько урана нужно, чтобы сделать ядерную бомбу? Как предсказать следующее слово в предложении? Откуда поисковики знают, какую страницу выдать по запросу?

Мы знаем ответы на эти вопросы, потому что 120 лет назад между двумя русскими математиками разгорелся ожесточенный спор, который породил один из важнейших алгоритмов на сегодняшний день – Цепи Маркова.

2 Конфликт и зарождение идеи

2.1 Спор Некрасова и Маркова

В 1905 году, произошла всем нам известная революция. Восставшие требовали свержения самодержавия, становления парламентской республикой, улучшения условий труда, а также установление равноправия среди граждан. Соответственно с одной стороны выступали монархисты, а с другой социалисты. Это разделение пронизывало общество и дошло до всех сфер, в том числе и науки.

Так, на стороне монархистов выступал математик Павел Некрасов, которого некоторые считали гением теории вероятности. Он был религиозным человеком, и к 1905 году уже прожил и проработал 52 года, что позволило занимать высокое положение в обществе. Он считал, что математика может объяснить свободу воли и даже волю Божью. Его оппонентом из противоположного лагеря был Андрей Марков. Он был атеистом и не терпел недобросовестных ученых, к числу которых относил Некрасова, потому что считал, что математика не имеет ничего общего с религией и вопросом свободы воли, поэтому он публично критиковал работы Некрасова и называл их злоупотреблением математикой. Они спорили по поводу принципа, который использовался в теории вероятности уже две сотни лет.

Этот принцип - закон больших чисел, доказанный Бернулли в 1713 году. Он объясняет, почему случайные события имеют предсказуемый результат при большом количестве испытаний. То есть по факту было введено понятие математического ожидания некоей величины. Закон больших чисел оставался ключевой концепцией теории вероятностей вплоть до Маркова и Некрасова.

Но Бернулли доказал этот закон только для независимых друг-от-друга событий, таких как бросок монеты или попытки разных людей угадать стоимость предмета. А что, если люди будут высказывать свои предположения вслух и слышать ответы друг друга? Допустим, первый человек решил, что предмет дорогой и выкрикивает "сто тыщ рублей". Тогда остальные, услышав его ответ, вероятно, будут отталкиваться от этой первой оценки. Их предположения уже не будут независимыми. Среднее значение не будет стремиться к реальной цене. Оно застрянет где-то в районе самого первого предложения. И вот на протяжении 200 лет теория вероятности опиралась на утверждение, что для соблюдения закона больших чисел необходима независимость событий. Именно это и стало предметом споров Некрасова и Маркова.

Statistical Regularity and Free Will

Table 1
Numbers of Belgian Marriages Classified by Age of Bride and Year of Marriage, for Bridesmaids of Age 30 and Under.

		Year				
		1841	1842	1843	1844	1845
Age of Bride	≤ 30	12,788	12,422	12,368	13,024	13,157
	(30, 45]	2,630	2,626	2,406	2,375	2,438
	(45, 60]	93	121	125	129	102
	> 60	7	6	8	5	5
Total Marriages (Brides and Grooms of all ages)		29,876	29,023	28,220	29,326	29,210

Рис. 1: Статистика браков в Бельгии

Некрасов как и Бернулли считал, что закон требует независимости событий друг от друга. Но он вывел из этого еще и обратную связь. Если соблюдается закон больших чисел, значит рассматриваются независимые события.

Для этого Некрасов взял статистику браков в Бельгии за пять лет (рис. 1). Мы видим, что среднее значение всегда находится в районе 29 тысяч. Казалось бы, значения остаются на одном уровне, а значит, подчиняются закону больших чисел. Когда Некрасов изучил другие социальные показатели, уровень преступности и рождаемость, он обнаружил аналогичную закономерность. Эти данные в тотальном большинстве случаев отражают решение людей вступить в брак, совершить преступление или завести ребенка. Некрасов закономерно сделал вывод, раз статистика подчиняется закону больших чисел, значит и лежащие в ее основе решения должны быть независимыми. Другими словами, это доказывает, что свобода воли существует, ведь люди независимо ни от чего приняли свое решение.

Для Некрасова она была не философской концепцией, а измеряемой величиной, научным фактором. Но Марков считал выводы Некрасова неразумными, а попытку связать математическую независимость со свободой воли абсурдом. Поэтому он поставил перед собой задачу доказать, что зависимые события тоже подчиняются закону больших чисел, и теория вероятностей работает даже на них.

2.2 Эксперимент с текстом

Ему требовалась модель, где каждое последующее событие зависит от предыдущего. И тогда он понял, что именно так устроен текст. Выбор следующей буквы, согласной или гласной, напрямую зависит от текущей. Чтобы проверить эту гипотезу, Марков обратился к Евгению Онегину. Он взял первые 20 тысяч букв, убрал все пробелы и знаки препинания и получил непрерывную цепочку символов. Ручной подсчет показал, что 43 % букв в цепочке гласные, 57 % согласные. Затем Марков разбил эту цепочку на пересекающиеся пары. Возможных комбинаций было 4. Гласная-гласная, согласная-согласная, гласная-согласная и согласная-гласная.

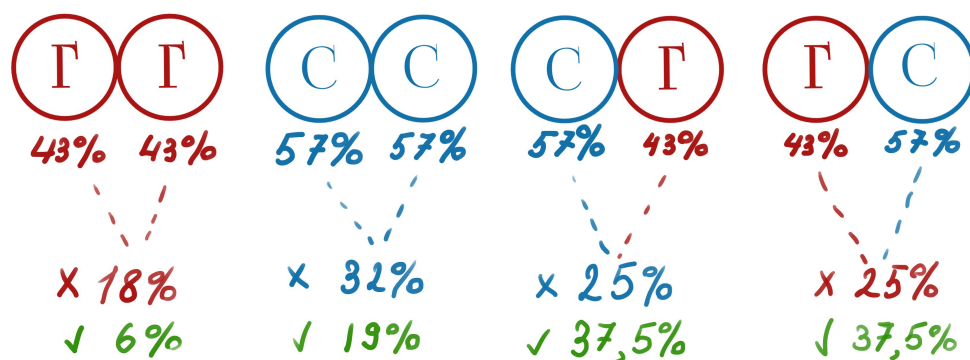


Рис. 2: Вероятности сочетаний

Если бы буквы не зависели друг от друга (рис 2), вероятность сочетания гласная-гласная равнялась бы вероятности гласной в квадрате, то есть около 18 % . Однако, когда Марков провел подсчет, оказалось, что такие сочетания встречаются лишь в 6 % случаев, значительно реже, чем при условии независимости. Когда он проверил другие комбинации, то обнаружил, что все реальные значения существенно отличаются от прогноза, основанного на предположении о независимости. Так Марков доказал зависимость букв. Теперь, чтобы взять верх над Некрасовым, ему оставалось показать, что эти буквы все же подчиняются закону больших чисел.

Для этого он создал своего рода машину предсказаний. Сначала Марков изобразил гласную и согласную в виде двух кружков. Это были состояния системы. Допустим, мы начинаем с гласной. Следующий может быть как гласная, так и согласная. Эти возможные переходы он обозначил стрелками. Какова вероятность переходов? Марков знал, что случайно выбранная буква с вероятностью 43 % будет гласной, а пара из двух гласных встречается примерно в 6 % случаев. Чтобы узнать вероятность перехода от одной гласной к другой гласной, он разделил 0,06 на 0,43 и получил 0,13. Мы точно знаем, что дальше будет какая-то буква, поэтому сумма вероятности всех переходов из одного состояния равна единице. Следовательно, вероятность перехода к гласной это $1 - 0,13$, то есть 0,87. Затем он выполнил те же расчеты и для согласных.

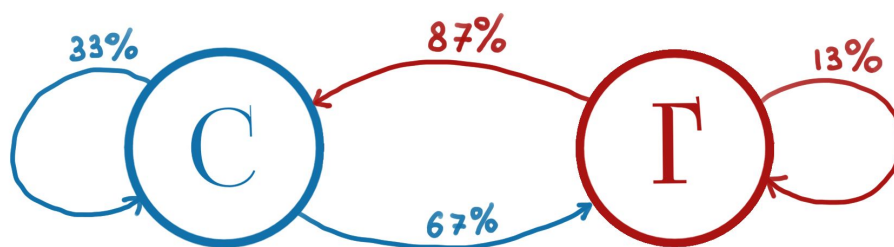


Рис. 3: Цепь Маркова для букв

Так как все же это работает? Начнем с гласной. Выберем случайное число от 0 до 1. Если число меньше 0,13, следующей буквой будет гласная, если больше — согласная. Допустим выпало 0,78, значит переходим к согласной. Затем выбираем новое число, если оно меньше 0,67, переходим к гласной. Предположим у нас 0,21, значит следующая буква — гласная. Продолжаем выполнять эти операции и следим за соотношением гласных и согласных. Поначалу оно будет очень нестабильным, но со временем начнет все больше смещаться к соотношению 43 на 57, ровно к тому, которое Марков вычислил вручную.

Ему удалось создать систему зависимых величин, цепочку событий, в которой соблюдается закон больших чисел. Он доказал, что сходимость значений в социальной статистике не гарантирует независимости рассматриваемых событий. То есть, статистика не доказывает свободу воли. Марков осознавал, что разгромил аргументацию Некрасова, поэтому свою работу он завершил язвительным замечанием: "Для теории вероятности свобода воли не обязательна. Необязательна и независимость событий." Этот способ работать с вероятностями зависимых событий впоследствии назовут цепью Маркова.

Это был настоящий прорыв, потому что в реальном мире практически каждое событие зависит от другого. Погода завтра зависит от условий сегодня. Распространение болезни зависит от того, кто заражен сейчас. Даже поведение частицы зависит от поведения других частиц вокруг нее. Многие из этих процессов можно смоделировать с помощью цепей Маркова.

Однако громкой сенсации не получилось. Да и сам Марков, похоже, совсем не задумывался о возможном практическом применении своего метода. Он говорил, что его интересует лишь решение аналитических задач, а к прикладным вопросам он был довольно таки безразличен. Он и не подозревал, что его вклад в теорию вероятности вскоре сыграет ключевую роль в одном из важнейших событий. Утром 16 июля 1945 Соединенные Штаты испытали "The Gadget первую в мире ядерную бомбу. Детонация шестикилограммового плутониевого заряда была эквивалентна примерно 25 килотоннам тротила. Ядерное испытание стало кульминацией манхэттенского проекта. В течение трех лет над ним работали величайшие умы эпохи, среди которых были Роберт Оппенгеймер, Джон фон Нейман и малоизвестный тогда математик Станислав Улам.

3 Цепи Маркова и метод Монте-Карло

3.1 Проблема Манхэттенского проекта

Даже после окончания войны Улам продолжал исследовать поведение нейтронов в ядерной бомбе. Принцип ее работы можно описать примерно так. Допустим, у нас есть ядро урана-235. Когда с ним сталкивается нейтрон, оно распадается, высвобождая энергию, и что важнее, несколько своих нейтронов. Если эти нейтроны затем расщепят в среднем хотя бы по два атома каждый, начнется неконтролируемая цепная реакция, то есть, ядерный взрыв. Добыть уран-235 было трудно, и пришлось рассчитывать минимальное количество необходимое для взрыва бомбы. Это и пытался сделать Улам, изучая нейтроны. Но в январе 1946 года исследование пришлось прервать. Врачи диагностировали у него тяжелое воспаление головного мозга и едва спасли ученого. Выздоровление шло медленно. Большую часть времени Улам проводил в постели. Чтобы скоротать время, он раскладывал простенький пасьянс. За этим занятием, пока пасьянс то сходился, то нет, он задумался, как вычислить вероятность того, что при случайном порядке карт в колоде пасьянс сойдется. Задача оказалась на удивление

сложной. В его колоде было 52 карты, а их порядок после каждой перетасовки получался уникальным. Количество возможных раскладов $52!$, то есть примерно $8 \cdot 10^{67}$. Решить эту задачу аналитически было невозможно. Уламу пришла в голову мысль. Сыграть сотни раз, посчитать количество побед и таким образом получить статистическую оценку искомой вероятности.

Тем временем в Лос-Аламосе ученые бились над задачами посложнее пасьянса. Например, изучали поведение нейтронов в активной зоне ядерной реакции, там, где триллионы частиц непрерывно взаимодействуют с атомами. И напрямую просчитать все возможные исходы таких взаимодействий просто невозможно. Когда Улам вернулся к работе, его вдруг осенило. А что, если моделировать эти процессы, генерируя множество случайных исходов, как он делал с пасьянсом? Он поделился этой идеей с фон Нейманом, который мгновенно оценил ее потенциал, но сразу выявил ключевую проблему. В пасьянсе каждая игра независима. Расклад карт в одной партии не влияет на следующую. Но с нейтронами все иначе. Их поведение зависит как от текущего положения, так и от предыдущих взаимодействий. Поэтому выбирать состояние случайным образом, как в пасьянсе, нельзя. Нужно было создать такую цепочку событий, где каждый шаг определял бы последующий. Фон Нейман понял, что для этого нужна цепь Маркова. И ее построили.

3.2 Синтез идей Улама и фон Неймана

В упрощенном виде она работает примерно так. Начальное состояние — это нейтрон, движущийся через активную зону. И тут могут произойти три вещи. Он может отскочить от атома и продолжить движение, тогда стрелка возвращается обратно в то же состояние. Он может покинуть систему без каких-либо взаимодействий, например, если его поглотит атом, не расщепляющегося изотопа урана, и его цепь Маркова завершится. Или он может столкнуться с атомом урана-235, что спровоцирует реакцию деления и высвобождение новых нейтронов, которые затем начнут собственные цепи. Однако вероятности переходов в цепи не фиксированы. Они зависят от положения нейтрона, его скорости и энергии, а также от строения и массы урановой ядерной мишени. Так, например, быстрый нейтрон отскочит в 30 % случаев, пропадет или будет поглощен в 50 %, а деление запустят в 20 %, но для медленного нейтрона вероятности будут другими. Ученые запустили эту цепь на первом в мире электронном компьютере ENIAC. Он случайным образом задавал начальные условия нейтрона и последовательно проходил по цепи, отслеживая среднее количество нейтронов, которые образуются за один заход, так называемый коэффициент умножения. Таким образом, если в среднем один нейтрон производит два других нейтрона, коэффициент равен двум. Если каждые два нейтрона производят три нейтрона, то $3/2$ и так далее. Затем, после полного прохождения цепи для заданного числа шагов, мы вычисляем средний коэффициент и записываем его в гистограмму. Этот процесс повторяли сотни раз, а результаты подсчитывали, получая статистическое распределение исходов. Если в большинстве случаев коэффициент меньше единицы, реакция затухает. Если равен единице, возникает самоподдерживающаяся стабильная цепная реакция. А если коэффициент больше единицы, реакция нарастает экспоненциально и происходит взрыв.

Таким образом, у фон Неймана и Улама появился статистический метод для определения количества производимых нейтронов, который не требовал точных вычислений. То есть они нашли приближенное решение системы дифференциальных уравнений, которые невозможно решить аналитически. Оставалось только придумать название для нового метода. Дядя Улама очень любил казино, и случайная выборка с высокими ставками напомнила Уламу об одном казино в Монако — Монте-Карло. Название прижилось.

Так появился метод Монте-Карло. Он оказался настолько успешным, что вскоре о нем узнали и другие ученые. Уже в 1948 специалисты Аргонской лаборатории в Чикаго с его помощью исследовали разные конструкции реакторов. Методу находили все новые применения. Позже Улам отмечал: "С тех пор я не перестаю удивляться тому, как простенькая схемка может изменить ход истории человечества".

4 Алгоритм PageRank или почему цепи Маркова основа интернета

4.1 Проблема раннего интернета

В 1993 году интернет стал общедоступным и набирал популярность. К середине 90-х ежедневно появлялись тысячи новых страниц, и их количество только увеличивалось. Это породило совершенно новую проблему. Как в этом бесконечно расширяющемся море информации найти хоть что-нибудь? В 1994-м два аспиранта Стэнфорда, Джерри Янг и Дэвид Файло, создали поисковую систему Yahoo, чтобы решить эту проблему. Однако им требовалось финансирование. В следующем году они договорились о

встрече с японским миллиардером Масайоси Соном, известным как японский Билл Гейтс. Они рассчитывали привлечь в свой стартап 5 миллионов долларов. Но у Сона были другие планы. Он предложил инвестировать целых 100 миллионов в 20 раз больше запрашиваемой суммы. Джерри Янг отказался, нам столько не нужно. Но Сон парировал. "Джерри, 100 миллионов нужны всем". И прежде, чем его собеседники успели ответить, Сон перехватил инициативу. "Кто ваши главные конкуренты? Excite и Lycos — ответили они. Сон приказал помощнику записать эти названия и заявил, "Если вы не примите мои инвестиции в Yahoo, я отдам деньги им, и для вас все закончится". Сон понял кое-что важное. Ни одна из ведущих поисковых систем того времени не обладала технологическим превосходством. Все они действовали по одной и той же схеме. Ранжировали страницы по частоте упоминания на них поискового запроса. Поэтому битву за звание лучшей поисковой системы выиграет тот, кто привлечет больше пользователей и больше вложит в маркетинг. Но маркетинг требовал финансов. А раз у Сона были деньги, значит, именно он решал, кто станет победителем. Основатели Yahoo поняли, что им остается только принять его инвестиции.

И всего через 4 года Yahoo стал самым посещаемым сайтом на планете, обрабатывая до 19 миллионов поисковых запросов. Состояние каждого из основателей оценивалось в 120 миллионов долларов. Но у Yahoo была фатальная уязвимость. Поиск по ключевым словам было легко обмануть. Чтобы поднять страницу в результатах, достаточно было напечатать ключевые слова сотни раз белым цветом на белом фоне.

В те годы отсутствовало понятие качества результатов. Обращали внимание только на релевантность, то есть соответствует ли тот или иной документ теме запроса, который интересует пользователя. Нужен был метод ранжирования страниц одновременно по релевантности и качеству. Но как измерить качество веб-страницы? Чтобы это понять, давайте посмотрим, как работают библиотеки. В библиотечных книгах есть бумажная карточка с датами возврата. Берешь книгу, видишь кучу штампов и думаешь, ага, наверное, хорошая книга. А если карточка пустая, что ж, возможно, не самая лучшая. Отметки работали как рекомендации. Чем их больше, тем лучше книга. Эту же идею можно применить и к интернету. Два аспиранта Стэнфорда, Сергей Брин и Ларри Пейдж, работали как раз в этом направлении.

4.2 Снова цепи Маркова

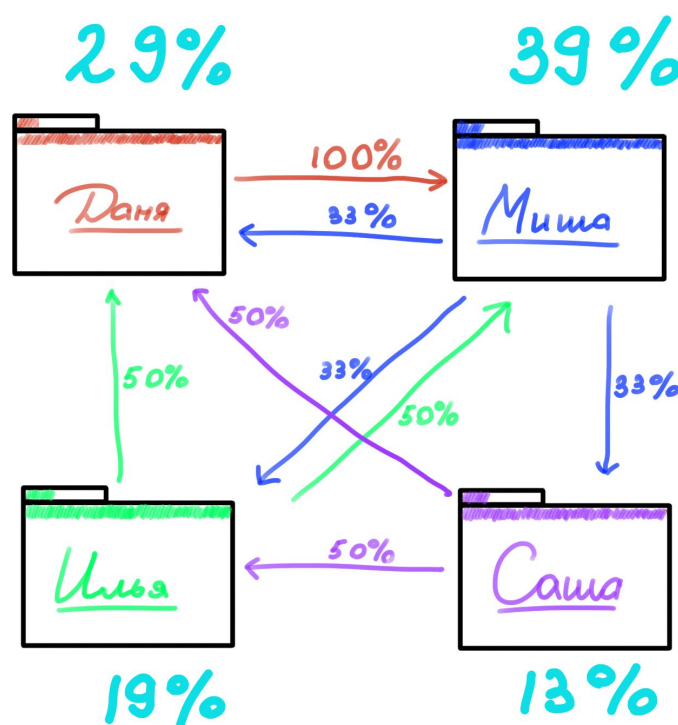


Рис. 4: Цепи Маркова в PageRank

Они поняли, что каждая ссылка на страницу может считаться рекомендацией, но чем больше ссылок на одном ресурсе, тем ниже ценность каждой из них. Они смоделировали интернет как цепь Маркова. Чтобы понять, как это работает, представим, что в интернете всего четыре сайта. Ими владеют Даня, Миша, Илья и Саша. Это наше состояние. Обычно веб-страница содержит ссылки на другие страницы, позволяя переключаться между ними. Это наши переходы. В этой модели Даня ссылается только на Мишу, поэтому вероятность перехода от Дани к Мише составляет 100%. Миша ссылается на Даню, Илью и Сашу, поэтому вероятность перехода на любую из этих страниц — 33%. Аналогичным образом мы можем заполнить вероятности остальных переходов. Теперь запустим эту цепь и посмотрим, что произойдет. Представим, что мы бродим по этой сети. Начинаем со случайной страницы, допустим, Дани, и переходим по ссылкам, отслеживая, как часто мы оказываемся на каждом сайте, в процентах. Со временем соотношение стабилизируется, и мы видим относительную важность каждого из них. Чаще всего мы попадали на страницу Миши, поэтому он на первом месте, за ним Даня, затем Саша и на последнем Илья. Может показаться, что систему легко обмануть. Создаете 100 страниц со ссылкой на ваш сайт, получаете 100 голосов, и вот вы на первом месте. Но это не так. Первое время у вас будет высокий рейтинг. Однако другие сайты на эти 100 страниц не ссылаются, и через множество шагов эти ранние переходы перестанут играть значимую роль. Ссылок может быть и много, но они некачественны и не влияют на результат.

Однако остается еще одна проблема. Не все страницы связаны между собой. В таких сетях можно застрять в цикле и никогда не попасть в остальные части. Чтобы это исправить, ввели правило. В 85% случаев следуем по ссылкам, как обычно, но примерно в 15% будем переходить на случайную страницу. Так можно гарантировать, что мы исследуем все части интернета и нигде не застрянем. С помощью цепей Маркова Пейдж и Брин создали более совершенную поисковую систему и назвали ее PageRank. Потому что она описывает, как взаимодействуют веб-страницы, а страница по-английски Page. К тому же название отсылает и к фамилии одного из основателей, Ларри Пейджа. PageRank стал выдавать более релевантные результаты, часто находя нужный сайт с первой попытки. Хотя некоторым эта идея казалась откровенно плохой. Если пользователь сделает несколько запросов, то и рекламу можно будет показать несколько раз. А если он сразу получит ответ, то тут же уйдет с сайта. Хороший поиск только навредит. Но Пейдж и Брин стояли на своем. Они были уверены, пользователи не проигнорируют действительно качественный продукт. И вот в 98-м году они запустили свою поисковую систему, чтобы бросить вызов Yahoo. Изначально они называли ее BackRub, потому что она анализировала бэклинки, но потом поняли, что это название не звучит. Планы были грандиозные, проиндексировать все страницы интернета, и название нужно было не менее масштабное. Тогда они взяли самое большое число, какое могли придумать — 10^{100} , или Googole. Но при регистрации домена случайно сделали опечатку. Так родился Google. За следующие четыре года Google отнял пальму первенства у Yahoo и стал самой популярной поисковой системой. Сегодня Alphabet, материнская компания Google, оценивается примерно в 2 триллиона долларов. Любое малейшее изменение в алгоритмах Google грозит серьезными последствиями. Они сейчас на коне, и причина их успеха в точности поиска. И в основе этого алгоритма на триллион долларов лежит цепь Маркова, которая, учитывая только текущее состояние, предсказывает последующие события.

5 Современные применения и границы модели

5.1 Языковые модели и предсказание текста

В 1940-х Клод Шеннон, отец теории информации, размышлял над другим вопросом. Он вернулся к истокам цепей Маркова и тоже работал с текстом, но ему было интересно рассчитать вероятности для конкретных букв. И он подумал, а что, если учитывать не одну предыдущую букву, а сразу две? С этим подходом он получил вот такой текст:

"IN NO IST LAT WHEY CRATICT FROURE GROCID PONDENOME OF DEMONSTURES OF THE REPTAGIN IS REGOACTIONA OF CRE."

Пока это мало похоже на что-то осмысленное, но можно разобрать какие-то слова, whey, of и артикль the. Однако Шеннон был уверен, что результат можно улучшить. И тогда за основу вместо букв он взял целые слова. Получился такой текст

"THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED"

примерно следующего содержания. "Наперед и атака на английского писателя, потому что важен персонаж, а значит другой подход к буквам, что время тех, кто сформулировал проблему, неожиданно".

Конечно, этот текст полный бред, но Шеннон заметил, что цепочки из четырех слов часто выглядят осмысленными. Например, фраза "атака на английского писателя" звучит нормально. Ученый пришел к заключению, что если учитывать больше предыдущих слов, то можно точнее предсказывать, каким будет следующее. Очень похоже на то, как Gmail подсказывает, что написать дальше. И это не случайность. Алгоритмы, отвечающие за такие предсказания, основаны на цепях Маркова. Не обязательно использовать буквы. Используют так называемые токены. И это могут быть буквы, слова, знаки препинания или что-нибудь еще. Таким образом, получается набор гораздо шире алфавита. И все очень просто. У нас есть последовательность, ну, предположим, из 30 токенов. И нужно узнать, какова вероятность, что следующим будет тот или иной токен.

Современные большие языковые модели обрабатывают токены по-разному. В отличие от простых цепей Маркова, они могут определять, на какие элементы следует обращать внимание. Например, если рядом с фразой "устройство клетки" будут слова "кровь" или "митохондрия" то модель поймет, что речь о биологии, а не о клетке для животных. Так модель корректирует свои предсказания. Но чем языковые модели популярнее, тем чаще созданные ими тексты попадают в интернет и становятся обучающими материалами для новых моделей. Если за этим не следить, система быстро деградирует. И мы придем к крайне примитивному статическому состоянию, будет повторяться одно и то же. И языковые модели подвержены такому сценарию.

5.2 Границы применимости

Систему с обратной связью вообще сложно моделировать с помощью цепей Маркова. Возьмем, к примеру, глобальное потепление. Когда в воздухе увеличивается количество углекислого газа, средняя температура Земли повышается. Чем больше температура атмосферы, тем больше водяного пара, одного из парниковых газов. А чем больше водяного пара, тем сильнее растет температура, из-за чего становится еще больше пара. Так возникает положительная обратная связь, из-за чего становится трудно предсказать последующие изменения. Таким образом, существуют системы, где цепи Маркова неприменимы, но во многих случаях они остаются эффективным вероятностным инструментом.

Самое удивительное, что с одной стороны можно заглянуть в прошлое таких систем, проследить каждую букву в тексте, каждое взаимодействие нейтрона или изменение погоды за месяц. Однако, как обнаружил сам Марков и другие, во многих случаях почти всей этой историей можно пренебречь. Достаточно учитывать только текущее состояние. И можно игнорировать все остальное. Получается, что у цепей Маркова нет памяти, и в этом секрет их эффективности. Это свойство позволяет радикально упрощать даже чрезвычайно сложные системы, сохраняя возможность эффективно делать прогнозы.

Как заметил один из следователей, решение задачи часто сводится к созданию подходящей цепи Маркова. Для меня все еще удивительно, что к такому математическому открытию привел весьма абсурдный спор, который, казалось бы, не имеет к нему никакого отношения. Однако, похоже, все указывает на то, что главным мотивом Маркова было желание переспорить Некрасова. Вот тебе и русский менталитет.

6 Заключение

Но на один вопрос мы так и не ответили. Когда Улам играл в пасьянс, как он определял, что колода действительно хорошо перетасована? Сколько раз нужно перемешать карты, чтобы добиться полностью случайного порядка?

7. Перетасовку карт можно представить как цепь Маркова, где порядок карт — это состояние, а каждая перетасовка — переход. Для колоды из 52 карт достаточно 7 раз, чтобы каждый возможный порядок карт стал практически равновероятным, то есть, по сути, случайным (важная ремарка что мешать надо так называемой riffle shuffle).

Но я например так не умею. Я использую классический метод (overhand shuffle). В таком случае придется перемешать колоду около двух тысяч раз.

Так что в следующий раз, когда кто-нибудь возьмётся тасовать перед игрой колоду, проследите, чтобы он сделал это правильно, и семь раз.

Иначе не считается.

Список литературы

Цепь Маркова Википедия. URL: <https://clck.ru/3QiNMu>

Башарин, Г. П., Ланнэ, А. А., Юркевич, И. А. Андрей Андреевич Марков (1856–1922): Жизнь, творчество, личность. – СПб.: Изд-во С.-Петерб. ун-та, 1997.

Grinstead, C. M., Snell, J. L. Introduction to Probability. – 2nd ed. – American Mathematical Society, 1997. – (Гл. 11: Markov Chains).

Häggström, O. Finite Markov Chains and Algorithmic Applications. – Cambridge University Press, 2002. – (Введение и исторические заметки).

Ulam, S. M. Adventures of a Mathematician. – New York: Charles Scribner's Sons, 1976.

Metropolis, N., Ulam, S. The Monte Carlo Method // Journal of the American Statistical Association. – 1949.

Eckhardt, R. Stan Ulam, John von Neumann, and the Monte Carlo Method // Los Alamos Science. – 1987. – Special Issue (15).

Page, L., Brin, S., Motwani, R., Winograd, T. The PageRank Citation Ranking: Bringing Order to the Web // Stanford InfoLab Technical Report. – 1999.

Langville, A. N., Meyer, C. D. Google's PageRank and Beyond: The Science of Search Engine Rankings. – Princeton University Press, 2006.

Jurafsky, D., Martin, J. H. Speech and Language Processing. – 3rd ed. – Prentice Hall, 2023. – (Главы про языковые модели n-грамм).

Norris, J. R. Markov Chains. – Cambridge University Press, 1997. – (Введение).