Александр Пиперски

Из мухи – слона

Окончание. Начало в № 2

Представьте себе, что мы играем однобуквенными словами. В «Грамматическом словаре» ровно 9 однобуквенных существительных: А, Е, И, О, У, Ы, Э, Ю, Я – названия гласных букв (слово Ё мы объединили со словом Е). Из любого слова к любому можно перейти за один ход, и это неинтересно. Если, напротив, играть в 10-буквенные слова, то от большинства слов – как, скажем, от слова БЕЗДЕЛЬНИК, – вообще нельзя никуда перейти. Получается, что игра интересна, когда выполнены два требования: пути между словами длинные и между двумя произвольно взятыми словами путь обычно существует (но не всегда – если успех гарантирован, играть тоже скучно).

Чтобы оценить среднюю длину пути, надо найти самые короткие пути между всеми парами точек, для которых путь есть, и вычислить среднее. Это трудоёмкий процесс, но есть алгоритмы, которые производят такое вычисление достаточно быстро. А сделав его, узнать вероятность успеха совсем просто: у нас есть 1712 4-буквенных слов, а значит, $1712 \cdot 1711$ упорядоченных пар (от любого из 1712 слов мы пытаемся пройти к любому из 1711 оставшихся); посчитаем, сколько из них соединены путём, и разделим на $1712 \cdot 1711$. Таких пар 1851342, а значит, вероятность успеха равна $1851342 / (1712 \cdot 1711) \approx 0{,}632$.

Можно сделать то же самое по-другому, опираясь на знания про размеры компонент связности. Возьмём произвольное начало цепочки; вероятность того, что оно попадёт в большую компоненту связности, составляет 1361/1712; возьмём теперь другое произвольное слово в качестве конца цепочки: в большой компоненте осталось 1360 слов, а всего в словаре – 1711, то есть вероятность того, что оно окажется связанным с началом, составляет 1360/1711. Тогда начало и конец цепочки попадут в большую компоненту с вероятностью $1361 \cdot 1360 / (1712 \cdot 1711)$. Во вторую по величине компоненту связности они попадут с вероятностью $11 \cdot 10 / (1712 \cdot 1711)$. Суммируя эти вероятности для всех компонент, получим то же число 0,632. Результа-

ты подсчётов для слов от 1 до 12 букв – в таблице:

Длина слова	Количество слов	Средняя длина пути	Вероятность успеха
1	9	1	1
2	70	3,6	1
3	490	4,4	0,866
4	1712	8,8	0,632
5	3642	8,6	0,076
6	5120	7,9	0,004
7	6584	8,9	0,002
8	6929	3,3	0,0001
9	6349	2,2	0,00008
10	5098	1,3	0,00003
11	3808	1,5	0,00004
12	2747	1,2	0,00002

Видно, что 4-буквенные слова действительно обеспечивают хорошую, но всё же не стопроцентную вероятность успеха и при этом создают достаточно длинные цепочки. А с 10-буквенными словами всё скучно: даже в тех редких случаях, когда путь есть, он обычно состоит из одного шага (ВОЗМЕЩЕНИЕ → → ВОЗМУЩЕНИЕ), редко – из двух или больше. Единственный путь из пяти шагов соединяет слова ПАССИРОВКА и ФАРШИРОВКА: ПАССИРОВКА → → МАССИРОВКА → МАСКИРОВКА → МАРКИРОВКА → МАРШИРОВКА → ФАРШИРОВКА.

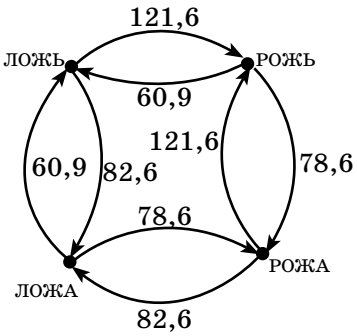
Эта цепочка напоминает ещё об одной проблеме, с которой надо разобраться: редкие слова. Попробуйте решить такую задачу: ДАЛЬ → ПАРИ. В этих словах совпадает буква А на втором месте, и есть надежда справиться за три шага. Действительно, такой путь есть: ДАЛЬ → ПАЛЬ → ПАЛИ → ПАРИ. Но едва ли он вас порадует: большинство людей не знают ни слова ПАЛЬ (выжженный участок в лесу или в степи), ни слова ПАЛИ (один из древних индийских языков). Но существует и путь из более нормальных слов: ДАЛЬ → → ДАНЬ → ДЕНЬ → ТЕНЬ → ТЕНТ →ТЕСТ → ТОСТ → → ПОСТ → ПОРТ → ПОРА → ПАРА →ПАРИ. Он длиннее, но в каком-то смысле лучше пути из трёх шагов. Попробуем формализовать эту идею.

Для этого нам понадобится ещё один словарь – частотный (я пользуюсь «Новым частотным слова-

рём русской лексики» О. Ляшевской и С. Шарова). В частотном словаре слова упорядочены по тому, насколько они употребительны в языке: 1-е место занимает союз И, на 2-м месте – предлог В, на 3-м – частица НЕ и т. д. А вот 10 самых частых 4-буквенных существительных с указанием их мест: 65) ДЕЛО, 71) ДЕНЬ, 74) РУКА, 104) ЛИЦО, 106) ДРУГ, 110) ГЛАЗ, 140) СИЛА, 191) ВОДА, 192) ОТЕЦ, 205) НОГА.

Слова типа ДЖИП и ТИТР стоят гораздо ниже: 6616) ДЖИП, 8086) ТИТР. Всего в этом словаре 52 138 слов, из них с «Грамматическим словарём» пересеклось 1140 4-буквенных существительных. А, например, слово ЛИВР (старая французская монета) даже не попало в частотный словарь. Условно присвоим ему и всем другим таким словам номер 100 000.

А теперь сделаем наш граф ориентированным и взвешенным. *Ориентированный* – это значит, что из вершины в вершину будут вести не отрезки, а стрелки, точнее пары стрелок: одна в одну сторону, другая в другую. А *взвешенный* значит, что каждой стрелке будет приписан вес: по одним стрелкам ходить будет дороже, а по другим дешевле. Веса припишем так: если стрелка ведёт в слово, которое занимает k-е место в частотном словаре, припишем ей вес $\sqrt{k}$ (мы могли бы выбрать и какую-нибудь другую функцию, но квадратный корень даёт результаты, которые кажутся подходящими для наших нужд). Например, все стрелки, ведущие в ДЕЛО, получают вес $\sqrt{65} \approx 8{,}1$, все стрелки, ведущие в ТИТР, – $\sqrt{8086} \approx$ $\approx 89{,}9$, а все стрелки, ведущие в ЛИВР, – $\sqrt{100000} \approx$ $\approx 316{,}2$. Длиной пути теперь будет не количество рёбер, а сумма чисел на стрелках, по которым мы двигались. Проходить редкие слова теперь невыгодно: ведущие в них стрелки весят слишком много.



Попробуем для примера найти оптимальный путь от слова ЛОЖЬ к слову РОЖА. Возьмём подграф, содержащий слова ЛОЖЬ, РОЖЬ, ЛОЖА и РОЖА, и разметим в нём веса.

Окажется, что путь ЛОЖЬ → РОЖЬ → РОЖА стоит 121,6 + 78,6 = 200,2, а путь ЛОЖЬ → ЛОЖА → → РОЖА стоит 82,6 + 78,6 = 161,2. Иначе говоря, выгоднее идти через более частое слово ЛОЖА, чем через более редкое РОЖЬ.

Но вернёмся к путям от ДАЛЬ к ПАРИ. Оказывается, более длинный по числу шагов путь выгоднее с точки зрения весов: ДАЛЬ $\xrightarrow{96,5}$ ДАНЬ $\xrightarrow{8,4}$ ДЕНЬ $\xrightarrow{36,4}$ $\xrightarrow{36,4}$ ТЕНЬ $\xrightarrow{140,9}$ ТЕНТ $\xrightarrow{65,8}$ ТЕСТ $\xrightarrow{81,6}$ ТОСТ $\xrightarrow{38,8}$ ПОСТ $\xrightarrow{58,4}$ $\xrightarrow{58,4}$ ПОРТ $\xrightarrow{16,6}$ ПОРА $\xrightarrow{27,9}$ ПАРА $\xrightarrow{126,2}$ ПАРИ (сумма 697,5); ДАЛЬ $\xrightarrow{316,2}$ ПАЛЬ $\xrightarrow{316,2}$ ПАЛИ $\xrightarrow{126,2}$ ПАРИ (сумма 758,6).

Так мы формализовали ощущение, почему длинная цепочка с известными словами лучше, чем короткая с неизвестными. Кстати, для задачи МУХА → → СЛОН цепочки без весов и с весами тоже разные. Без весов (10 шагов): МУХА → МУРА → МАРА → ПАРА → → ПАРК → ПАЕК → САЕК → СТЕК → СТЕН → СТОН → → СЛОН; с весами (13 шагов): МУХА → МУКА → → ЛУКА → ЛУПА → ЛАПА → ПАПА → ПАРА → ПАРК → → ПАЕК → САЕК → СТЕК → СТОК → СТОН → СЛОН.

Увы, и в цепочке с весами не удалось обойтись без малоизвестного слова САЕК (точнее, САЁК), которое значит «молодой олень». Но другие решения этой задачи включают в себя ещё и не такое. Самую короткую из известных цепочек придумал Владимир Гончаров; в ней 7 шагов: МУХА → МУЛА → КУЛА → → КИЛА → КИЛН → КИОН → СИОН → СЛОН. По нашим правилам она не подходит, потому что в «Грамматическом словаре» из 6 промежуточных слов есть только КИЛА (это разговорное название для грыжи), а если бы они и были, то с весами эта цепочка стоила бы очень дорого, ведь все промежуточные слова в ней редкие – да и разве интересна цепочка, в которой 6 из 6 промежуточных слов нам неизвестны?

Всё сказанное выше, с одной стороны, звучит неутешительно: игру «из мухи – слона» постигла судьба шахмат и шашек – компьютер играет в неё гораздо лучше человека. С другой стороны, не всё так плохо: компьютерный анализ позволил нам узнать много интересного и про саму игру, и про русский язык.

Художник Мария Усеинова