В.С. Левченков, Л.Г. Левченкова

МЕТОДЫ СИМВОЛИЧЕСКОЙ ДИНАМИКИ И ПРОБЛЕМА МОДЕЛИРОВАНИЯ ПОИСКА ИНФОРМАЦИИ В ИНТЕРНЕТЕ [1]

Введение

В связи с ростом числа сайтов в Интернете чрезвычайно актуальной проблемой математического моделирования является решение задачи эффективного обнаружения адекватной запросу информации. Существенным шагом в ее решении стал подход, реализованный в поисковой системе Google [1]. В ней при оценке значимости страниц Интернета используется их ранжирование, возникающее на основе рассмотрения специальной динамической системы (вероятностной марковской цепи), сопоставляемой процессу поиска на основе матрицы взаимных цитирований. Однако, использованная в Google интерпретация "random surfer" [2] аппелирует к механистической модели пользователя, "прыгающего" с одной страницы на другую без определенной цели. Главная проблема для него – наличие новых ссылок на той странице, на которой он оказался в данный момент. Поскольку a priori нельзя гарантировать, что при таком поиске на текущей странице Интернета либо вообще есть хоть одна ссылка, либо есть ссылка на страницы, которые еще не встречались "по пути", то Google дополняет процесс поиска возможностью случайного перехода на любую страницу. Математически это выражается добавлением в матрицу цитирований некоторой "фоновой" компоненты (демпфирующего фактора, по терминологии [2]), которая искажает реальную структуру связей страниц.

В настоящей работе показывается, что задача оценки страниц Интернета адекватна проблеме самосогласованного выбора в теории голосования [3], [4] и может быть решена без привлечения демпфирующего фактора.

1. Ранжирование страниц в Интернете на основе Google Pagerank

Поисковая система Google получила в последнее время весьма широкое распространение благодаря новому методу оценки соответствия

[1]Работа выполнена при поддержке Гранта Президиума РАН по проекту ИКС.

между запросом пользователя и теми страницами в Интернете, которые были найдены поисковой системой по его ключевым словам. Именно, Google упорядочивает предъявляемые пользователю ссылки на страницы в соответствии со специальным числовым показателем, так называемым Google's Pagerank. Его обозначают PR и вычисляют следующим образом.

Пусть $\mathcal{P}$ – множество страниц в Интернете ($|\mathcal{P}| = n$), размещенных на совокупности сайтов S ($|S| = m$). Существенным обстоятельством, используемым в поисковой системе Google, является наличие на страницах Интернета ссылок на его другие страницы, содержащие дополнительную к рассматриваемой странице информацию. Эти ссылки могут быть легко организованы в матрицу ссылок $L = (l_{ij})_{i,j=1}^{n}$, строки и столбцы которой занумерованы соответствующими страницами из $\mathcal{P}$. Матрица L состоит из нулей и единиц, причем матричный элемент l_{ij} принимает значение 1, когда на странице i есть ссылка на страницу j, и равен 0 в противном случае (подчеркнем, что диагональные элементы матрицы L равны 0, поскольку страница не ссылается сама на себя). Согласно подходу Google, ранг $PR(i)$ страницы i из $\mathcal{P}$ вычисляется на основе формулы

$$PR(i) = \frac{(1-d)}{n} + d\sum_{j=1}^{n} \frac{PR(j)l_{ji}}{O(j)}, \tag{1}$$

где $O(i) = \sum\limits_{j=1}^{n} l_{ij}$ – число ссылок со страницы i на все остальные страницы Интернета, d – положительное число из отрезка $[0,1]$, смысл применения которого мы обсудим ниже.

Если ввести вектор $\gamma = (\gamma_i)_{i=1}^{n}$ с компонентами $\gamma_i = PR(i)$, а также две матрицы: $\tau = (\tau_{ij})_{i,j=1}^{n}$, состоящую из элементов $\tau_{ij} = 1/n$, и $\tilde{L} = (\tilde{l}_{ij})_{i,j=1}^{n}$, $\tilde{l}_{ij} = l_{ij}/O(i)$, то соотношение (1), определяющее $PR(i)$, можно переписать в матричном виде

$$\gamma = (1-d)\gamma\tau + d\gamma\tilde{L}. \tag{2}$$

Заметим теперь, что матрица

$$P = (1-d)\tau + d\tilde{L} \tag{3}$$

является стохастической. Действительно, ее элементы $p_{ij} = (1-d)\tau_{ij} + d\tilde{l}_{ij}$ неотрицательны и удовлетворяют соотношению

$$\sum_{j=1}^{n} p_{ij} = (1-d) + d\frac{1}{O(i)}\sum_{j=1}^{n} l_{ij} = 1.$$

Если считать, что $0 < d < 1$, то элементы матрицы P будут строго положительными, $p_{ij} > 0$, значит, P – неразложимая матрица. Хорошо известно [5], что тогда у P есть единственный нормированный на 1 положительный левый собственный вектор, отвечающий собственному значению 1. Таким образом, вектор γ в (2) при $0 < d < 1$ определяется единственным (с точностью до нормировки) образом. Из теории конечных марковских цепей известно, что этот вектор задает асимптотическое значение частоты появления любого элемента из $\mathcal{P}$ почти во всякой реализации случайного марковского процесса с матрицей переходов (3). Эта математическая структура – вероятностная марковская цепь (ВМЦ) – послужила в [2] основой для интерпретации значений $PR(i)$, как средних частот выбора пользователем страниц Интернета, если он с вероятностью p_{ij} переходит со страницы i на страницу j (так называемая модель "random surfer" – RS).

Анализ выражения для p_{ij} показывает, что оно состоит из двух частей: постоянной части, определяемой матрицей τ и части, задаваемой матрицей $\tilde{L}$, элементы которой связаны с количеством дуг, соединяющих страницы i и j, дополнительно подвергнутых "нормировке" с помощью фактора $O(i)$. Матрицы τ и $\tilde{L}$ "смешаны" друг с другом в виде линейной комбинации с коэффициентами $(1 - d)$ и d. Таким образом, фактор d – это вероятность, определяющая долю постоянной матрицы τ в вероятности перехода между страницами Интернета. Поскольку матрица $\tilde{L}$, вообще говоря, разложима, именно использование фактора d приводит к неразложимости "смешанной" матрицы P. В результате, вероятность перехода p_{ij} содержит две компоненты: одна связана с "нормированным" числом ссылок из i в j, а вторая не связана со ссылками, а представляет собой постоянный "шум" (демпфирующий фактор, в терминологии [2]). Авторы модели RS интерпретируют этот "шум" как элемент усталости пользователя, когда ему надоедает прыгать по ссылкам и он "щелкает мышкой" по произвольной странице в Интернете.

Роль демпфирующего фактора d в модели RS чрезвычайно важна. Если в соотношении (3) положить $d = 1$, то вероятности переходов в такой марковской цепи будут определяться только матрицей ссылок Интернета. В случае, если эта матрица разложима (структуру матрицы такого вида мы подробно обсудим в следующем разделе работы), то хорошо известно, что среди страниц Интернета можно выделить множество $G_0 \subset \mathcal{P}$, из страниц которого не будет ссылок на страницы, не принадлежащие G_0. Таким образом, пользователь, выбрав для рассмотрения некоторую

страницу $i \in G_0$ и пользуясь ссылками, никогда не попадет ни в какую страницу из множества $\mathcal{P} \setminus G_0$. Наличие неравного единице фактора d позволяет избежать этой проблемы, поскольку появляется ненулевая вероятность p_{ij}, $(p_{ij} \geq \frac{1-d}{n})$, перехода в любую страницу Интернета j. С вычислительной точки зрения, случай $0 < d < 1$ приводит к задаче нахождения собственного вектора неразложимой матрицы, отвечающего *ее максимальному собственному значению, а эта задача решается на* основе быстро сходящегося алгоритма, что важно ввиду большого количества элементов в множестве $\mathcal{P}$.

Однако, наличие произвольного фактора ставит вопрос о том, как следует выбирать его величину, чтобы обеспечить решение поставленной задачи, а именно, предоставить пользователю наиболее эффективный способ поиска требуемой ему информации. Важность выбора конкретного значения d очевидна хотя бы из того, что если положить $d = 0$, то модель RS сведется к равновероятному выбору любой страницы из $\mathcal{P}$, что эквивалентно "слепому" поиску нужной информации. Авторы Google полагают $d = 0.85$ [1], опираясь на опыт вычислений для реальной матрицы связей в Интернете. Однако, без точного понимания соответствия между видом матрицы L и оценкой важности страницы Интернета с точки зрения поиска требуемой информации, роль d понять нельзя. Подчеркнем однако, что если матрица L неразложима, то необходимость в факторе d отпадает, поскольку в этом случае пользователь и так посетит все страницы Интернета, пользуясь матрицей переходных вероятностей (3), вычисляемой только по существующим ссылкам.

Суммируя сказанное выше, модель RS поиска по Google можно описать так.

Страницы Интернета в силу специальной структуры связей между ними оказываются неэквивалентными. Эта неэквивалентность численно оценивается величиной $PR(i)$, $i \in \mathcal{P}$, показывающей частоту посещения страницы i в модели вероятностной марковской цепи с переходными вероятностями вида (3). Значения переходных вероятностей содержат две компоненты: одна из них обусловлена реальными связями между страницами, а другая – представляет собой "фоновый" параметр, обеспечивающий равновероятный переход во все страницы Интернета. Авторы Google считают, что если запрос пользователя "высветил" некоторое подмножество страниц $Q \subset \mathcal{P}$, то релевантную информацию в первую очередь следует искать в тех страницах $i \in Q$, для которых $PR(i)$ принимает наибольшее значение.

В следующем разделе мы обсудим эти выводы, рассмотрев структуру матрицы связей L и ее роль в оценке значимости страницы.

2. Иерархия классов эквивалентности в множестве страниц Интернета

При описании связей в Интернете возникает два вида графов. Один из них – граф сайтов Интернета $M(S)$ – содержит в качестве вершин элементы множества сайтов S ($|S| = m$). Два сайта $x, y \in S$ связаны друг с другом совокупностью n_{xy} направленных дуг, $n_{xy} = \sum\limits_{\substack{i \in x \\ j \in y}} l_{ij}$, показывающих количество цитирований страниц, принадлежащих сайту y, на всех страницах, принадлежащих сайту x. При наличии ссылок в страницах сайта x на страницу этого же сайта в графе возникают петли. Таким образом, граф сайтов математически представляет собой структуру, называемую псевдографом (ввиду наличия множества петель и дуг, соединяющих его вершины). Роль этого графа в процессе поиска информации в Интернете мы обсудим в следующем разделе.

Другой граф – граф страниц Интернета $M(\mathcal{P})$ – более прост. Он содержит в качестве вершин множество страниц $\mathcal{P}$, при этом из страницы i на страницу j проводится направленная дуга, если страница i цитирует страницу j. В этом графе уже нет петель, поскольку страница не ссылается сама на себя. Изучим в этом разделе структуру графа $M(\mathcal{P})$. Этот граф однозначно связан с неотрицательной матрицей $L = (l_{ij})_{i,j \in \mathcal{P}}$. Хорошо известно [5], что соответствующей перестановкой строк и столбцов она может быть приведена к каноническому виду, так называемой нормальной форме (см. рис. 1).

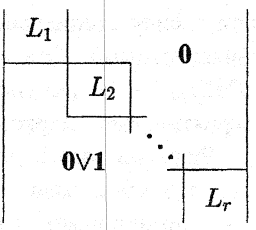


Рис. 1. Нормальная форма матрицы L

На блочной диагонали нормальной формы матрицы стоят квадратные матрицы L_k ($k = \overline{1, r}$), причем матрицы размерностью больше 1 являются неразложимыми. Выше блочной диагонали в матрице стоят

только нулевые элементы, а под диагональю могут быть как нули, так и единицы.

Пусть множество $Q_k \subset \mathcal{P}$ нумерует строки матрицы L_k. Тогда из рис. 1 видно, что при $k < l$ из группы Q_k страниц Интернета нет ни одной ссылки в группу страниц Q_l, а из Q_l в Q_k ссылки могут быть. Если при $l > k$ из какой-то страницы группы Q_l есть ссылка на какую-то страницу *группы Q_k, то будем говорить, что группа страниц Q_l связана с группой Q_k*.

Что дает описанная структура для возможности оценки информации, помещенной на некоторую страницу из $\mathcal{P}$?

Рассмотрим сначала страницы из группы Q_k ($|Q_k| > 1$), связи между которыми описываются матрицей L_k. Поскольку число страниц в этой группе больше единицы, матрица L_k будет неразложимой, а отвечающий ей граф $M(L_k)$ – связный. Это означает, что любые две страницы i, j из Q_k можно соединить последовательностью ссылок, ведущей со страницы i на страницу j. В результате можно сказать, что все страницы из Q_k тематически связаны друг с другом, т.е. каждая страница содержит, по крайней мере, часть информации, родственной информации хотя бы еще на одной странице из Q_k. Нельзя сказать, что все страницы из Q_k содержат родственную информацию, поскольку среди Q_k могут быть "синтетические" (многоплановые) страницы, содержащие разнородную информацию (например, страницы новостей информационного агенства), ссылки из которых тематически многообразны и ведут во многие элементы из Q_k. О чем говорит наличие ссылки со страницы i на некоторую другую страницу j? Ссылка – это предложение пользователю уйти со страницы i на страницу j для просмотра информации, дополняющей информацию, имеющуюся на странице i. Чем больше ссылок на странице i, тем больше соблазна у пользователя заглянуть на другие страницы из Q_k. Какой фактор уравновешивает это стремление и заставляет пользователя более подробно рассмотреть страницу i, а не перескочить сразу на другие страницы? Очевидно, это количество ссылок на страницу i со стороны других страниц из Q_k. Строя свою оценку страницы i, пользователь учитывает как количество новых возможностей для просмотра (показывающих неполноту информации на этой странице), так и количество ссылок на i из других страниц (показывающих важность информации на i с точки зрения других страниц). Баланс этих двух факторов, а не только учет спектра возможностей для перехода на другие страницы, и должен учитываться при построении показателя (назовем его

степенью значимости страницы), характеризующего страницу из Q_k на фоне других страниц из этого множества.

Задачи такого рода впервые рассматривались в теории голосования (см., например, [3], [6], [7]), где было показано, что показатель типа "степень значимости" следует находить следующим образом.

Матрице связей $L_k = (l_{ij})_{i,j \in Q_k}$ сопоставляется матрица переходов $T = (t_{ij})_{i,j \in Q_k}$, имеющая вид

$$t_{ij} = \begin{cases} l_{ij}, & \text{если} \quad i \neq j \\ \sum\limits_{s \in Q_k} l_{si}, & \text{если} \quad j = i \end{cases} \tag{4}$$

и характеризуемая псевдографом $M(T)$, отличающимся от графа $M(L_k)$ наличием в нем петель, число которых определяется значениями диагональных элементов t_{ii} матрицы T. Матрица переходов позволяет построить динамическую систему – топологическую марковскую цепь (ТМЦ), на основе которой можно выяснить, с какой частотой $\sigma_k(i)$ элементы $i \in Q_k$ встречаются почти во всех путях графа $M(T)$ (оценка ведется по некоторой мере, вычисляемой по параметрам ТМЦ). Эти частоты $\sigma_k(i)$ и являются численным выражением "степень значимости" страницы $i \in Q_k$. Значения $\sigma_k(i)$ находятся из следующей системы уравнений

$$\sigma_k(i) = \xi(i)\eta(i); \quad \sum_i \sigma_k(i) = 1; \quad T\xi = \lambda_0 \xi; \quad \eta T = \lambda_0 \eta, \tag{5}$$

где λ_0 – максимальное собственное значение матрицы T.

Если матрица T обладает постоянной строчной суммой, т.е.

$$\forall i \quad \sum_j t_{ij} = \lambda_0, \tag{6}$$

то система уравнений упрощается и принимает вид

$$\sigma_k T = \lambda_0 \sigma_k, \quad \sum_i \sigma_k(i) = 1. \tag{7}$$

Этот случай возникает, например, при выполнении специального условия на элементы матрицы цитирований

$$\forall i, j \in Q_k \quad l_{ij} + l_{ji} = 1, \tag{8}$$

которое означает, что любые две страницы $i, j \in Q_k$ обязательно связаны ровно одной ссылкой.

Очевидно, произвольно взятая матрица связей не обязана удовлетворять условиям (6) или (8). Однако, ее всегда можно преобразовать так, чтобы например, условия (6) выполнялись. Эта операция в теории мультиотношений [4] называется операцией z-нормализации.

Именно, пусть нам задана некоторая неразложимая матрица $T = (t_{ij})_{i,j=1}^{q}$, *элементы которой – целые неотрицательные числа.* Рассмотрим новую матрицу $\tilde{T} = (\tilde{t}_{ij})_{i,j=1}^{q}$, связанную с T соотношениями

$$\tilde{t}_{ij} = zt_{ij} / \sum_{k=1}^{q} t_{ik}, \tag{9}$$

где целое число z является общим кратным чисел $s_i = \sum\limits_{k=1}^{q} t_{ik}$ $(i = \overline{1,q})$.

Очевидно, числа $\tilde{t}_{ij}$ – целые и удовлетворяют условию

$$\forall i \quad \sum_{j=1}^{q} \tilde{t}_{ik} = z.$$

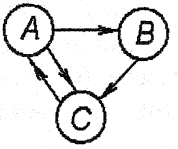
Таким образом, если в (5) вместо T использовать z-нормализацию, $\tilde{T}$, то процедура нахождения критерия $\sigma_k(i)$ сведется к системе уравнений вида (7).

Этот прием и использовал Google при нахождении $PR(i)$. Действительно, система уравнений (1) при $d = 1$ дает

$$PR(i) = \sum_{j \in \mathcal{P}} \frac{l_{ji} PR(j)}{O(j)}. \tag{10}$$

Пусть z – общее кратное чисел $O(i)$ $(i = \overline{1,n})$, $\gamma = PR(i))$, $\hat{L} = (\hat{l}_{ij})_{i,j \in \mathcal{P}}$, $\hat{l}_{ij} = z\frac{l_{ij}}{O(i)}$, тогда (10) переписывается в виде $\gamma \hat{L} = z\gamma$, совпадающем с (7). В результате, при проведении расчетов $PR(i)$ в системе Google фактически используется не исходная матрица связей L, а измененная – $\hat{L}$. Численная величина $PR(i)$ показывает не частоту появления страницы i при связях, описываемых L, а другую величину: частоту выбора страниц Интернета при динамическом процессе "перепрыгивания" со страницы на страницу согласно ВМЦ с матрицей переходов (3). Эта величина соответствует некоторой частоте появления страниц, но вычисленной не по исходной матрице связей, а по другой матрице, $\hat{L}$.

Пример 1. Наглядная иллюстрация нахождения $PR(i)$ состоит в рассмотрении трех страниц A, B и C, с матрицей связей L и соответствующим графом, представленными на рис. 2.



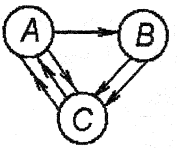
$$L = \begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

Рис. 2

Легко подсчитать (согласно (1) с $d = 1$), что $PR(A) = PR(C) = 0.4$; $PR(B) = 0.2$. Вычислим теперь частоту посещения вершин A, B и C согласно (5), положив там матрицу $T = L$. Легко подсчитать, что $\xi = a(\lambda_0^2, 1, \lambda_0)$, $\eta = b(\lambda_0, 1, \lambda_0^2)$, $\sigma = ab(\lambda_0^3, 1, \lambda_0^3)$; здесь a и b – нормировочные константы.

Мы видим, что $\sigma(A) = \sigma(C)$; $\sigma(B) = \frac{1}{\lambda_0^3}\sigma(A)$. Максимальное собственное значение матрицы L находится из уравнения $\lambda_0^3 = \lambda_0 + 1$ и удовлетворяет условию $1 < \lambda_0 < 2$. Значит, $\sigma(B) = \frac{1}{\lambda_0+1}\sigma(A) < \frac{1}{2}\sigma(A)$. Таким образом, $\sigma(A) > PR(A)$, а $\sigma(B) < PR(B)$.

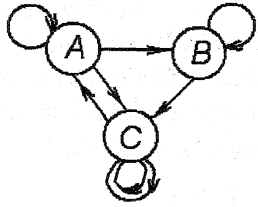
В этом нет ничего удивительного, потому что $PR(i)$ это частоты посещения вершин для графа, приведенного на рис. 3. Здесь же представлена z-нормализация, $\hat{L}$, для $z = 2$.



$$\hat{L} = \begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 2 \\ 2 & 0 & 0 \end{pmatrix}$$

Рис. 3

Оба использованных метода приписывают страницам A и C равную частоту, большую, чем частота страницы B. Значит ли это, что страницы A и C по степени важности для пользователя эквивалентны, а страница B хуже? Посмотрим на рис. 2, обе страницы, и A, и B, имеют ссылки на страницу C, признавая тем самым, что на ней помещена важная с их точки зрения информация, в то же время, важность A подтверждается только ссылкой C, а важность B – ссылкой A. Можно ли считать, что в таких условиях страницы A и C одинаково важны? Мы уже говорили выше, что в теории выбора степень значимости страницы нужно вычислять иным образом. Как следует из [7], метод расчета (5) нужно применять не к матрице L, а к матрице T, содержащей по сравнению с матрицей L еще и ненулевые диагональные элементы. Граф соответствующей матрицы T и ее вид представлены на рис. 4.



$$T = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 2 \end{pmatrix}$$

Рис. 4

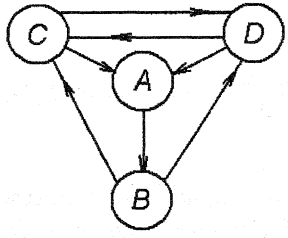
Расчет по (5) с матрицей T этого вида показывает, что теперь

$$2 < \lambda_0 < 3, \quad \sigma = ab\left(1, \frac{1}{\lambda_0}, \frac{\lambda_0 - 1}{\lambda_0 - 2}\right),$$

т.е. $\sigma(C) > \sigma(A) > \sigma(B)$, что количественно подтверждает качественное рассмотрение графа на рис. 2, приведенное ранее.

Одинаковый ранг по Google могут иметь страницы, связанные с остальными страницами весьма непохожим образом.

Пример 2. Пусть есть четыре страницы A, B, C и D с матрицей связей и соответствующим графом, изображенными на рис. 5.



$$L = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix}$$

Рис. 5

Матрица связей L, согласно (1) с $d = 1$, преобразуется в стохастическую матрицу

$$\tilde{L} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{pmatrix},$$

которая оказывается дважды стохастической, т.е. сумма ее элементов по столбцам также равна единице. Значит, левый

собственный вектор матрицы $\tilde{L}$ пропорционален единичному, то есть $PR(A) = PR(B) = PR(C) = PR(D)$. Это весьма странный факт, поскольку элементы A и B весьма несимметричным образом связаны с остальными элементами. То, что такая несимметрия может быть учтена, подтверждается рассмотрением матрицы L на основе соотношений (5). Матрица T, строящаяся по L согласно (5), имеет вид

$$T = \begin{pmatrix} 2 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 2 & 1 \\ 1 & 0 & 1 & 2 \end{pmatrix}.$$

Ее правый и левый собственные вектора равны

$$\xi = a\left(1, (\lambda_0 - 2), (\lambda_0 - 3)^{-1}, (\lambda_0 - 3)^{-1}\right),$$

$$\eta = b\left((\lambda_0 - 1), 1, (\lambda_0 - 3)^{-1}, (\lambda_0 - 3)^{-1}\right);$$

λ_0 удовлетворяет уравнению $\frac{2}{\lambda_0 - 3} = (\lambda_0 - 1)(\lambda_0 - 2)$ и лежит в интервале $3 < \lambda_0 < 4$. Тогда

$$\sigma = ab\left((\lambda_0 - 1), (\lambda_0 - 2), (\lambda_0 - 3)^{-2}, (\lambda_0 - 3)^{-2}\right).$$

Поскольку

$$(\lambda_0 - 3)^{-2} - (\lambda_0 - 1) = \frac{(\lambda_0 - 1)(\lambda_0 - 2)}{2(\lambda_0 - 3)} - (\lambda_0 - 1) = (\lambda_0 - 1)\frac{4 - \lambda_0}{2(\lambda_0 - 3)} > 0,$$

то $\sigma(C) = \sigma(D) > \sigma(A) > \sigma(B)$.

Таким образом, правило (5) одинаково оценивает симметрично расположенные страницы C и D, однако различает A, B и C.

В заключение этого раздела обсудим роль демпфирующего фактора d в случае, когда матрица связей неразложима. Система уравнений для $\gamma = (PR(i))_{i=1}^{n}$ имеет вид (2). Пусть демпфирующий фактор задан рациональным числом, $d = \frac{r}{k}$, а целое число z делится на любое число $O(j)$, $j = \overline{1, n}$. Умножая систему (2) на целое число nzk, приходим к матричному уравнению

$$nzk\gamma = (zk - r)\gamma\tau + nz\gamma\tilde{L} = \gamma((zk - r)\tau + nz\tilde{L}) = \gamma F, \qquad (11)$$

где матрица

$$F = (zk - r)\tau + nrz\tilde{L} \qquad (12)$$

содержит целочисленные элементы, а система (11) по виду совпадает с системой (7).

В силу того, что обычно $z \ll n$ (поскольку число ссылок, выходящих из страницы много меньше общего числа страниц Интернета), второе слагаемое в (12) существенно больше первого и значит, при неразложимой $\tilde{L}$ значения компонент вектора γ будут в основном определяться элементами матрицы $nrz\tilde{L}$. Таким образом, малый по величине демпфирующий фактор может оказать существенное влияние только тогда, когда матрица $\tilde{L}$ разложима.

3. Роль расположения классов эквивалентности для оценки релевантности информации

Соотношения (5) позволяют найти степень значимости каждой страницы в выделенном классе Q_k. Однако, классов эквивалентности в нормальной форме матрицы L может быть несколько. Как они соотносятся друг с другом с точки зрения процесса поиска релевантной информации? Важную роль в ответе на этот вопрос играет структура связей между классами. Предположим, что класс Q_l связан с классом Q_k. Это означает, что существует такая страница $i \in Q_l$, которая содержит ссылку на страницу $j \in Q_k$, но ни одна страница из Q_k не ссылается на страницы из Q_l. С точки зрения состава информации связь такого рода подразумевает, что на странице j содержится информация, дополняющая и уточняющая информацию (или ее часть), содержащуюся на странице i. Более того, на странице i нет информации, дополняющей или уточняющей информацию на j в каком-то аспекте, поскольку ни одна из страниц в Q_k не ссылается не только на j, но на любую другую страницу из Q_l. Отметим, что принадлежность страницы какому-то классу означает, что страницы этого класса связаны друг с другом последовательностями ссылок, т.е. информация на них (или некоторые части ее) информационно однородны и содержат общие темы. Таким образом, при поиске нужной информации на двух страницах i и j, следует в первую очередь обратиться к странице j, а уж потом к странице i. Рассматривая далее связь класса Q_k с другими классами, можно найти другой класс Q_s, с которым связан класс Q_k, и обнаружить страницу $r \in Q_s$, на которую ссылается некоторая страница $j' \in Q_k$. Поскольку j и j' лежат в одном классе Q_k, то существует последовательность ссылок, ведущих из j в j', и далее в r. Таким образом, информация в r дополняет и уточняет информацию на j', а значит, в определенном смысле дополняет и информацию на i. Конечно, многоплановость информации, зачастую помещаемой на страницах, делает

эту связь весьма условной, так как в результате последовательности цитирований может измениться тематическое содержание информации, если по пути встретится многотематическая страница (типа страницы новостей). Поэтому в общем случае, можно лишь сказать, что страница $r \in Q_s$ дополняет и уточняет информацию для какой-то части страниц из Q_k. Однако, если мы точно не знаем, какая часть страниц из Q_k представляет интерес, то разумнее вначале просмотреть страницы из Q_s, чтобы оперировать с более детальной (а значит, и более узкой по теме) информацией, нежели сразу обратиться к анализу информации из Q_k.

Естественный вывод, который вытекает из проведенного выше анализа, таков: нормальная структура матрицы качественно выделяет ту последовательность (сверху вниз) просмотра информации, которая (с точностью до пропущенных цитирований) показывает внутреннюю иерархию информационных связей страниц. Классы, которые информационно не связаны с другими классами, следует просматривать в первую очередь. Конечно, некоторые страницы могут быть составлены либо некомпетентными, либо "специально ориентированными" авторами, претендующими на уникальность и самодостаточность представленной информации и поэтому не ссылающихся на другие страницы (или ссылающихся на узкий круг "элитарных" страниц). Однако, такие страницы (и классы) легко отсеять, оценив общее количество ссылок на них со стороны других страниц Интернета. В идеальном случае, когда все страницы верно указывают ссылки на более детальную или дополняющую информацию, упорядочение страниц должно строиться на основе двух критериев: один определяется номером класса Q_k, которому принадлежит страница i Интернета в структуре нормального разложения матрицы связей; а другой характеризуется величиной степени значимости $\sigma_k(i)$ этой страницы в классе Q_k и вычисляется на основе соотношений (5).

Вид такого бинарного отношения при упорядочении сайтов Интернета мы рассмотрим в следующем разделе.

4. Упорядочение сайтов Интернета по информации о взаимных цитированиях

Матрица цитирований множества S ($|S| = m$) сайтов Интернета $N = (n_{x,y})_{x,y \in S}$ отличается от матрицы связей страниц тем, что в ее строках могут стоять не единицы и нули, а произвольные неотрицательные числа. Как мы указывали ранее, эта матрица порождает псевдограф $M(S)$, и с точки зрения попарного сравнения сайтов друг с другом порождает так

называемое мультиотношение [4]

$$m: \quad S^2 \to Z_+, \qquad \forall x, y \in S \quad m(x,y) = n_{xy}. \tag{13}$$

На основе мультиотношения можно осуществить такую же процедуру разбиения на классы элементов из S, как это было проведено для страниц Интернета. Именно, приведем матрицу N к нормальной форме (рис. 6).

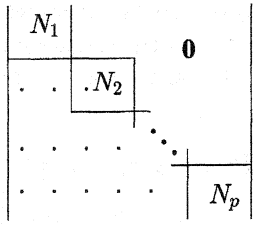


Рис. 6. Нормальная форма матрицы N

Пусть $N_1, \ldots, N_p$ – квадратные диагональные блоки нормальной формы матрицы N, а $S_1, \ldots, S_p$ – элементы множества S $(S = \bigcup_{i=1}^{p} S_i)$, нумерующие строки совокупности матриц $\{N_i\}_{i=1}^{p}$, соответственно. Согласно построению нормальной формы матрицы, выполнено

a) если $|S_i| \geq 2$, то матрица N_i неразложима;

b) если существуют $x \in S_i$, $y \in S_j$ такие, что $n_{xy} > 0$, то $i > j$.

Для нахождения степеней значимости элементов из S по информации (13) используем правило самосогласованного выбора по мультиотношениям [3], [4], [6], согласно которому в каждом множестве S_k, $(|S_k| \geq 2)$ его элементы упорядочены согласно численному критерию $\sigma_k(x)$

$$\sigma_k(x) = v_k(x)u_k(x); \quad \sum_k \sigma_k(x) = 1; \quad Mu_k = \lambda_0 u_k; \quad v_k M = \lambda_0 v_k, \tag{14}$$

а матрица $M = (m_{xy})_{x,y \in S_k}$ находится по (13) согласно соотношениям

$$\forall x, y \in S_k \qquad m_{xy} = \begin{cases} m(x,y), \text{ если } x \neq y \\ \sum\limits_{z \in S_k} m(z,x), \text{ если } y = x. \end{cases} \tag{15}$$

Множества элементов, принадлежащих различным классам S_i, S_j $(i \neq j)$ упорядочены относительно друг друга согласно правилу: элемент

$x \in S_i$ считается не хуже элемента $y \in S_j$, если не существует последовательности элементов $z_0, ..., z_{k+1} \in S$ такой, что

$$z_0 = x; \quad z_{k+1} = y; \quad \prod_{i=0}^{k} m(z_i, z_{i+1}) > 0. \tag{16}$$

Если любой элемент из S_i не хуже любого элемента из S_j, то будем также говорить, что группа сайтов S_i *не связана* с S_j. Содержательно, это означает, что не существует последовательности ссылок, соединяющих произвольно выбранные из S_i и S_j элементы друг с другом.

Таким образом, все сайты Интернета оказываются упорядочены бинарным отношением R, вычисляемым по матрице цитирований на основе (14) и (16):

I. $\forall k \quad \forall x, y \in S_k \quad xRy \Leftrightarrow \sigma_k(x) \geq \sigma_k(y)$;

II. $\forall i, j (i \neq j) \quad \forall x \in S_i \quad \forall y \in S_j \quad xRy$, если группы сайтов S_i и S_j не связаны друг с другом.

Построенное таким образом отношение R представляет собой слабый порядок, т.е. является рефлексивным, связным и транзитивным отношением.

Действительно, в силу I для любого сайта справедливо xRx (в случае, если $x \in S_i$ с $|S_i| = 1$, будем считать, что xRx имеет место по определению). Далее, для любых двух различных элементов x, y, лежащих в одном классе, например, S_k, согласно условию I имеет место xRy или yRx. Если же x и y лежат в различных классах, $x \in S_i$ и $y \in S_j$, $i \neq j$, то либо S_i не связано с S_j (т.е. справедливо xRy), либо S_j не связано с S_i (справедливо yRx), как это следует из нормальной формы матрицы N. Таким образом, отношение R связно. Его транзитивность доказывается аналогично.

В заключение отметим, что аналогичным образом строится отношение, упорядочивающее и страницы Интернета.

Литература

1. Brin S., and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine (http://pr.efactory.de/e-references.shtml, PDF, 1998).

2. Page L., S. Brin, R. Motwani, and T. Winograd. The Page Rank Citation Ranking: Bringing Order to the Web

   (http://pr.efactory.de/e-references.shtml, PDF, 1998).

3. Левченков В.С. Два принципа рациональности в теории выбора: Борда против Кондорсе. - М.: Издательский отдел ф-та ВМиК МГУ, 2002.

4. Левченков В.С. Элементы эргодической теории с приложениями к проблемам выбора. II. Приложение эргодической теории к задачам выбора. - М.: ВМиК МГУ, 1997.

5. Гантмахер Ф.Р. Теория матриц. М.: Наука, 1967.

6. Левченков В.С. Аксиоматический подход к самосогласованному выбору. ДАН, 1993, т.330, N2, 173-176.

7. Левченков В.С. Игровое обоснование правила самосогласованного выбора. Вестник Московского университета 15. Вычислительная математика и кибернетика, 2000, N1, с. 30-34.