В.С. Левченков, Л.Г. Левченкова

УПОРЯДОЧЕНИЕ СТРАНИЦ WWW ПО ИНФОРМАЦИИ О ВЗАИМНЫХ ЦИТИРОВАНИЯХ[1]

Введение

При построении поисковых систем для информационных массивов с внутренними связями между элементами – например, для поиска в множестве страниц World Wide Web (WWW) – важное значение приобретает проблема упорядочения фрагментов информации не только по степени релевантности запросу пользователя, но и по тем качественным характеристикам, которые описывают взаимную связь этих фрагментов. В широко известной поисковой системе Google соответствующее упорядочение строится на основе специального количественного критерия (Google PageRank [1]), вычисляемого на основе характеристик специальной динамической системы – Вероятностной Марковской Цепи (ВМЦ), – сопоставляемой матрице взаимных цитирований страниц WWW. Однако, в определение переходных вероятностей ВМЦ авторы этого подхода [1] включили хоть и малый, но произвольный фактор, позволяющий существенно упростить численную процедуру нахождения скалярного критерия, и, одновременно, искажающий реальные связи между страницами WWW. В [2] был разработан иной подход к этой проблеме, который использовал методы символической динамики для построения соответствующего упорядочения. В настоящей работе мы проводим детальное исследование видов возникающих при этом упорядочений и той роли какую играет интерперетация матрицы связей в выборе одного из них для построения успешно работающей поисковой системы.

1. Упорядочение страниц WWW и структура матрицы взаимных цитирований

Для упорядочения страниц WWW рассмотрим матрицу цитирований $L = (l_{ij})_{i,j=1}^{n}$, строки и столбцы которой занумерованы элементами множества $\mathcal{P}$ всех страниц WWW. Матрица L состоит из нулей и единиц, причем матричный элемент l_{ij} принимает значение 1, когда на странице

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

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). Эта матрица неотрицательна и перестановкой рядов может быть приведена к *нормальной форме* (рис. 1).

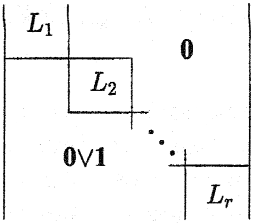


Рис. 1. Нормальная форма матрицы L

Ее блочная структура может быть получена следующим образом (см., например, [3]-[5]). Используем граф $G(L)$, строящийся по матрице L: число его вершин равно n, из вершины i ведет ориентированная дуга в вершину j т.и.т.т., когда $l_{ij} = 1$. На его основе введем следующее *бинарное отношение достижимости* R_a на множестве $\mathcal{P}$. Для произвольных страниц i и j полагаем, что страница j достижима из страницы i (обозначение jR_ai), если в графе $G(L)$ существует ориентированный путь, соединящий вершины i и j.

С помощью этого отношения из множества $\mathcal{P}$ можно выделить подмножество $\mathcal{P}_0$ так называемых *возвратных* элементов. Элемент $i \in \mathcal{P}_0$ называется *возвратным*, если он удовлетворяет условию iR_ai, т.е. в графе $G(L)$ можно найти такой путь, который, стартуя со страницы i, вернется опять на эту же страницу. Элементы множества $\mathcal{P} \setminus \mathcal{P}_0 = \{j_1, ..., j_s\}$ назовем *невозвратными*.

На множестве $\mathcal{P}_0$ рассмотрим новое отношение E_a, являющееся симметричной частью отношения R_a

$$\forall i, j \in \mathcal{P}_0 \quad iE_aj \Leftrightarrow iR_aj \& jR_ai. \qquad (1)$$

Это отношение рефлексивно, симметрично и транзитивно, т.е. является эквивалентностью. Значит, множество $\mathcal{P}_0$ можно единственным образом представить в виде разбиения $\mathcal{P}_0 = \bigcup\limits_{i=1}^{k} \mathcal{P}_i$ на непересекающиеся множества $\mathcal{P}_i$ эквивалентных между собой элементов. Используем это разбиение и совокупность невозвратных элементов $\{j_1, ..., j_s\}$ для построения разбиения всего множества $\mathcal{P}$,

$$\mathcal{P} = \bigcup_{l=1}^{r} Q_l, \tag{2}$$

где $Q_l \in Q = \{\mathcal{P}_1, ..., \mathcal{P}_k, \{j_1\}, ..., \{j_s\}\}$, а $r = k + s$.

Элементы множеств Q_l могут быть весьма прихотливым образом связаны друг с другом отношением достижимости. Для выяснения этой связи используем следующую конструкцию. Назовем *"хвостом" множества* Q_l такую совокупность множеств

$$t(Q_l) = \{Q_{j_1}, Q_{j_2}, ..., Q_{j_q}\}, \tag{3}$$

что выполнено

1) $j_1 = l$;

2) $\forall i \in Q_{j_s},\ \forall j \in Q_{j_{s+1}} \quad iR_aj \quad (s = 1, ..., q-1)$.

Число q будем называть *длиной* "хвоста" $t(Q_l)$.

Если $T(Q_l)$ – совокупность всех возможных "хвостов" множества Q_l, а $h(l)$ – максимальное значение q по всем элементам из $T(Q_l)$, то положим $h = \max\limits_{1 \le l \le r} h(l)$. Перенумеруем элементы множества Q так, что совокупность $\tilde{Q}_1 = \{Q_1, ..., Q_{r_1}\}$ – все те элементы из $\{Q_l\}_{l=1}^{r}$, максимальная длина "хвоста" которых равна h; для $\tilde{Q}_2 = \{Q_{r_1+1}, ..., Q_{r_2}\}$ эта длина равна $h-1$, и т.д.; наконец, для $\tilde{Q}_h = \{Q_{r_{h-1}+1}, ..., Q_r\}$ – максимальная длина "хвоста" равна 1. В результате элементы множества $Q = \{Q_1, Q_2, ..., Q_r\}$ оказываются занумерованными так, что если $k < s$, то ни одна страница $i \in Q_k$ не содержит цитирований страниц, принадлежащих множеству Q_s.

Занумеруем теперь – последовательно от 1 до $|\mathcal{P}|$ – страницы, принадлежащие множествам совокупности Q. Если теперь эту нумерацию использовать для строк и столбцов матрицы цитирований L, то мы получим ее канонический вид (рис. 1).

Из этого рассмотрения вытекает следующий вывод: существует два типа отношений, упорядочивающих страницы WWW. Первое отношение связано с упорядочением множества Q и порождается слабым порядком (обозначим его w), заданным на элементах $\{Q_1, Q_2, ..., Q_r\}$

$$\begin{aligned} &\forall s \quad \forall Q_i, Q_j \in \tilde{Q}_s, \quad Q_iwQ_j, \\ &\forall s, k \quad \forall Q_i \in \tilde{Q}_s \quad \forall Q_j \in \tilde{Q}_k, \quad s < k \Rightarrow Q_iwQ_j. \end{aligned} \tag{4}$$

В силу выбора нумерации множества Q отношение w – слабый порядок (т.е. рефлексивно, связно и транзитивно).

Второе отношение определяет упорядочение элементов внутри каждого класса эквивалентности $Q_k \in Q$. Оно связано с системой взаимных цитирований страниц $i, j \in Q_k$, информация о которых содержится в сужении L_k матрицы L на множество Q_k.

Согласно [2], соответствующее упорядочение $p^{(k)}$ $(\forall i, j \in Q_k \quad ip^{(k)}j \Leftrightarrow \sigma^{(k)}(i) \geq \sigma^{(k)}(j))$ вычисляется на основе скалярного критерия $\sigma^{(k)}$, удовлетворяющего системе соотношений

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

где λ_0 – максимальное собственное значение матрицы $T = (t_{ij})_{i,j\in Q_k}$, называемой *матрицей переходов* и имеющей вид

$$t_{ij} = \begin{cases} l_{ij}, & \text{если} \quad i \neq j \\ \sum\limits_{s\in Q_k} l_{si}, & \text{если} \quad j = i. \end{cases} \tag{6}$$

Если $|Q_k| = 1$, то по определению положим $i \in Q_k \Rightarrow \sigma^{(k)}(i) = 1$.

Оба типа упорядочения определяют естественный порядок на $\mathcal{P}$, если $\forall s \quad |\tilde{Q}_s| = 1$, поскольку в этом случае отношение w является линейным порядком (т.е. оказывается еще и антисимметричным) и строго ранжирует классы страниц. При этом отношение R на $\mathcal{P}$ вида

$$\begin{aligned} &\forall i, j \in Q_k, \quad iRj \Leftrightarrow ip^{(k)}j \\ &\forall i \in Q_s, \quad \forall j \in Q_k, \quad s < k \Rightarrow iRj \end{aligned} \tag{7}$$

является слабым порядком (этот результат в других терминах был сформулирован в [2]).

Однако, при наличии $\tilde{Q}_k$, для которых $|\tilde{Q}_k| > 1$, объединение двух типов отношений может быть проведено несколькими способами, порождающими различные слабые порядки на $\mathcal{P}$.

Заметим, что отношение w на каждом множестве $\tilde{Q}_s$ является отношением эквивалентности, причем все элементы из $\tilde{Q}_s$ образуют один класс, который можно строго упорядочить $|\tilde{Q}_s|!$ способами. Пусть $R^{(s)}$ $((1 \leq s \leq h)$ одно из таких ранжирований. Тогда определим отношение

R на $\mathcal{P}$ формулой

$$\forall i,j \in Q_k, \quad iRj \Leftrightarrow ip^{(k)}j$$

$$\forall Q_l, Q_k \in \tilde{Q}_s \quad \forall i \in Q_l \quad \forall j \in Q_k, \quad Q_l R^{(s)} Q_k \Rightarrow iRj \tag{8}$$

$$\forall Q_l \in \tilde{Q}_s \quad \forall Q_k \in \tilde{Q}_m \quad \forall i \in Q_l \quad \forall j \in Q_k, \quad s < m \Rightarrow iRj.$$

Мы получаем целое семейство отношений R слабого порядка вида (8), допускающее произвол в выборе ранжирования элементов в каждой совокупности $\tilde{Q}_s$, содержащей более одного элемента. Этот произвол частично можно преодолеть, если использовать метод, позволяющий в соответствии с (5) и (6) упорядочить элементы внутри каждой совокупности $\tilde{Q}_s$.

Чтобы применить соответствующий подход, заметим, что множества, входящие в совокупность Q, по-разному связаны между собой. Именно, будем говорить, что два различных множества Q_s и Q_l из Q связаны друг с другом *отношением цитирования* (обозначение $Q_s \leftrightarrow Q_l$), если найдется такая пара страниц $i \in Q_s$ и $j \in Q_l$, что либо i цитирует страницу j, либо j цитирует страницу i.

Пусть теперь фиксирован один из классов $\tilde{Q}_s$ $(s = 1, ..., h)$ эквивалентных по w множеств. Для удобства будем также обозначать его элементы латинскими буквами $x, y, z, ...$ Множество $Q_l \in Q \setminus \tilde{Q}_s$ порождает на $\tilde{Q}_s$ бинарное отношение w_l, характеристическая функция которого $w_l(x,y)$ $(xw_ly \Leftrightarrow w_l(x,y) = 1)$ имеет вид

$$\forall x, y \in \tilde{Q}_s$$
$$w_l(x,y) = \begin{cases} 0, \text{ если } Q_l \leftrightarrow y \text{ и не имеет места } Q_l \leftrightarrow x \\ 1 \text{ во всех остальных случаях.} \end{cases} \tag{9}$$

Совокупность бинарных отношений (9) при всех $Q_l \in Q \setminus \tilde{Q}_s$ образует профиль отношений $\{w_l\}_{Q_l \in Q \setminus \tilde{Q}_s}$, характерный для информации, используемой в задачах голосования. Для упорядочения элементов множества $\tilde{Q}_s$ согласно этому профилю используем правило самосогласованного выбора для мультиотношений (см. [3], глава 5). Согласно этому методу, отношения $w_l(x,y)$ подвергаются процедуре z-нормализации, на основе которой возникает новый профиль, состоящий из индивидуальных мультиотношений $m_l(x,y)$

$$\forall Q_l \in Q \setminus \tilde{Q}_s \quad m_l(x,y) = \frac{2w_l(x,y)}{w_l(x,y) + w_l(y,x)}.$$

Эти отношения сворачиваются в групповое мультиотношение

$$m(x,y) = \sum_l m_l(x,y), \tag{10}$$

на основе которого строится соответствующая ему матрица переходов $\hat{T} = (t_{xy})_{x,y\in\tilde{Q}_s}$

$$t_{xy} = \begin{cases} m(y,x), & \text{если} \quad x \neq y \\ \sum\limits_{z\in\tilde{Q}_s\setminus\{x\}} m(x,z), & \text{если} \quad y = x. \end{cases} \tag{11}$$

Если эта матрица неразложима, то упорядочение элементов множества $\tilde{Q}_s$ проводится на основе скалярного критерия $\gamma^{(s)}(x)$, вычисляемого по системе уравнений, аналогичных (5)

$$\gamma^{(s)}(x) = u(x)v(x); \quad \sum_{x\in\tilde{Q}_s} \gamma^{(s)}(x) = 1; \quad \hat{T}u = \mu_0 u, \quad v\hat{T} = \mu_0 v, \tag{12}$$

где μ_0 – максимальное собственное значение матрицы $\hat{T}$.

Предложение 1. Матрица $(m(x,y))_{x,y\in\tilde{Q}_s}$, порождаемая согласно (10) отношениями профиля (9), при $s < h$ является неразложимой.

Доказательство. Предположим, что эта матрица разложима, и $\tilde{Q}_0$ такое подмножество $\tilde{Q}_s$ $(s < h)$, для которого справедливо $\forall x \in \tilde{Q}_0$, $\forall y \in \tilde{Q}_s \setminus \tilde{Q}_0$, $\quad m(x,y) = 0$ или, эквивалентно, $\forall x \in \tilde{Q}_0$, $\forall y \in \tilde{Q}_s \setminus \tilde{Q}_0$, $\forall l \quad w_l(x,y) = 0$. Заметим теперь, что согласно построению системы множеств $\tilde{Q}_s$, при $s < h$ у элемента $x \in \tilde{Q}_s$ длина его "хвоста" больше 1, т.е. существует такой элемент $Q_l \in Q \setminus \tilde{Q}_s$, который связан с x, значит $\forall y \in \tilde{Q}_s \setminus \tilde{Q}_0 \quad w_l(x,y) = 1$. Полученное противоречие завершает доказательство. *Q.E.D.*

Предложение 2. Если матрица $(m(x,y))_{x,y\in\tilde{Q}_h}$ разложима, то множество $\tilde{Q}_h$ разбивается на два непересекающихся подмножества X и Y, $\tilde{Q}_h = X \cup Y$, таких что $\forall x \in X$, $\forall y \in Y \quad m(x,y) = 0$, причем ни один элемент из X не связан ни с каким элементом из $Q \setminus \tilde{Q}_h$, а $\forall y \in Y$, $\forall Q_l \in Q \setminus \tilde{Q}_h \quad Q_l \leftrightarrow y$.

Доказательство. Пусть X – множество тех элементов из $\tilde{Q}_h$, которые не связаны ни с одним элементом из $Q \setminus \tilde{Q}_h$. Если матрица $(m(x,y))_{x,y\in\tilde{Q}_h}$ разложима, то это множество не пусто (если X пусто, то согласно (10) $\forall x,y \in \tilde{Q}_h \quad m(x,y) > 0$). Пусть, далее, Y подмножество $\tilde{Q}_h$, состоящее из элементов, связанных с каждым элементом из $Q \setminus \tilde{Q}_h$. Тогда, очевидно,

$\forall x \in X, \forall y \in Y$ справедливо $(m(x,y)) = 0$;
$\forall x, y \in X \quad m(x,y) = r_{h-1}$;
$\forall x, y \in Y \quad m(x,y) = r_{h-1}$.

Теперь остается только показать, что $\tilde{Q}_h = X \cup Y$. Допустим, что это неверно. Тогда $\forall z \in \tilde{Q}_h \setminus (X \cup Y)$ существуют такие $Q_l, Q_k \in Q \setminus \tilde{Q}_h$, что Q_l связано с z, а Q_k нет. Значит, $\forall x \in X \;\; m(x,z) > 0$ и $m(z,x) > 0$ (поскольку $w_k(x,z) = w_k(z,x) = 1$), а также $\forall y \in Y \;\; m(y,z) > 0$ и $m(z,y) > 0$ (т.к. $w_l(y,z) = w_l(z,y) = 1$). Из этих соотношений легко видеть, что матрица $(m(x,y))_{x,y\in\tilde{Q}_h}$ оказывается неразложимой. Полученное противоречие завершает доказательство. *Q.E.D.*

Из предложений 1 и 2 вытекает, что множества $\tilde{Q}_s$ $(s < h)$, а также подмножества X и Y $(X \cup Y = \tilde{Q}_h)$ могут быть упорядочены согласно (12). Оказывается, что вид этого упорядочения можно оценить, не решая систему уравнений (12). Чтобы сформулировать точное утверждение, введем функцию $\rho_s: \tilde{Q}_s \to Z_+$ (Z_+ – множество целых неотрицательных чисел), такую что

$$\forall x \in \tilde{Q}_s \quad \rho_s(x) = |\{Q_l \in Q \setminus \tilde{Q}_s : \quad Q_l \leftrightarrow x\}|, \tag{13}$$

т.е. $\rho_s(x)$ равно числу тех элементов из множества $Q \setminus \tilde{Q}_s$, которые связаны с $x \in \tilde{Q}_s$ отношением цитирования. Эта функция однозначно определяется системой множеств

$$Y_l = \{x \in \tilde{Q}_s : \quad x \leftrightarrow Q_l\}, \tag{14}$$

поскольку

$$\rho_s(x) = \sum_{Q_l \in Q\setminus\tilde{Q}_s} \chi^{Y_l}(x), \tag{15}$$

где $\chi^{Y_l}(x)$ – характеристическая функция множества Y_l

$$\chi^{Y_l}(x) = \begin{cases} 1, \; x \in Y_l \\ 0, x \notin Y_l. \end{cases}$$

Легко заметить, что

$$\forall x, y \in \tilde{Q}_s \quad m_l(x,y) = 1 + \chi^{Y_l}(x) - \chi^{Y_l}(y), \tag{16}$$

а

$$m(x,y) = n_s + \rho_s(x) - \rho_s(y), \tag{17}$$

где $n_s = |Q \setminus \tilde{Q}_s|$.

Теорема 1. Скалярный критерий $\gamma^{(s)}(x)$ (12), упорядочивающий элементы множества $\tilde{Q}_s$ $(s < h)$ удовлетворяет следующему условию

$$\forall x, y \in \tilde{Q}_s \quad \rho_s(x) \geq \rho_s(y) \Rightarrow \gamma^{(s)}(x) \geq \gamma^{(s)}(y). \tag{18}$$

Доказательство. Согласно (17), элементы $m(x, y)$ подчинены условию

$$\forall x, y \in \tilde{Q}_s \quad m(x, y) + m(y, x) = 2n_s, \tag{19}$$

и в силу предложения 1 порождают неразложимую матрицу.

Матрица переходов $\hat{T}$ находится по ним согласно соотношениям (11) и имеет постоянную строчную сумму

$$\sum_{y \in \tilde{Q}_s} t_{xy} = \sum_{z \in \tilde{Q}_s \setminus \{x\}} m(x, z) + \sum_{z \in \tilde{Q}_s \setminus \{x\}} m(z, x) = 2n_s(m_s - 1),$$

где $m_s = |\tilde{Q}_s|$.

Согласно теореме Фробениуса-Перрона [5], в этом случае максимальное собственное значение матрицы $\hat{T}$ равно $\mu_0 = 2n_s(m_s - 1)$, а соответствующий ему правый собственный вектор пропорционален единичному вектору. Таким образом, критерий $\gamma^{(s)}(x)$ в этом частном случае удовлетворяет системе уравнений

$$\sum_{z \in \tilde{Q}_s} \gamma^{(s)}(z) t_{zx} = 2n_s(m_s - 1)\gamma^{(s)}(x),$$

из которой с учетом (19) можно выделить при $x \neq y$ следующие два уравнения

$$\begin{aligned} &\sum_{z \neq x, y} t_{zx}\gamma^{(s)}(z) + t_{yx}\gamma^{(s)}(y) = \sum_{z \neq x} t_{xz}\gamma^{(s)}(x) \\ &\sum_{z \neq x, y} t_{zy}\gamma^{(s)}(z) + t_{xy}\gamma^{(s)}(x) = \sum_{z \neq y} t_{yz}\gamma^{(s)}(y). \end{aligned} \tag{20}$$

Заметим теперь, что если для $x \neq y$ выполнены условия

(a) $t_{yx} \geq t_{xy}$,

(b) $\forall z \in \tilde{Q}_s \setminus \{x, y\} \quad t_{zx} \geq t_{zy}$,

то из (20) вытекает соотношение $\gamma^{(s)}(x) \geq \gamma^{(s)}(y)$.

Действительно, вычитая в (20) из первого уравнения второе, получим с учетом условия (b) неравенство

$$\gamma^{(s)}(x) \left(\sum_{z \neq x} t_{xz} + t_{xy} \right) \geq \gamma^{(s)}(y) \left(\sum_{z \neq y} t_{yz} + t_{yx} \right)$$

Поскольку $t_{xy} \le t_{yx}$, а

$$\sum_{z\neq x} t_{xz} = \sum_{z\neq x}(2n_s - t_{zx}) =$$
$$= -t_{yx} + \sum_{z\neq x,y}(2n_s - t_{zx}) \le -t_{xy} + \sum_{z\neq x,y}(2n_s - t_{zy}) = \sum_{z\neq y} t_{yz},$$

то

$$\gamma^{(s)}(x) \ge \gamma^{(s)}(y).$$

То обстоятельство, что предположение $\rho_s(x) \ge \rho_s(y)$ при учете (17) приводит к выполнению условий (a) и (b), завершает доказательство теоремы. *Q.E.D.*

Следствие 1. Для элементов из $\tilde{Q}_h = X \cup Y$ (где X и Y те же множества, что и в предложении 2) справедливо

$$\forall x \in X \quad \rho_h(x) = 0,$$

$$\forall y \in Y \quad \rho_h(y) = r_{h-1}.$$

Поскольку сужения матрицы $m(x,y)$ на X и на Y являются неразложимыми матрицами с элементами, равными r_{h-1}, то для $\tilde{Q}_h$ выполнено

$$\gamma_X^{(h)}(x) = const; \quad \gamma_Y^{(h)}(x) = const,$$

где, например, $\gamma_X^{(h)}$ – вектор, удовлетворяющий (12), с матрицей T построенной по сужению матрицы $m(x,y)$ на множество X.

Таким образом, множество $\tilde{Q}_h$ можно упорядочить следующим образом:

1) каждый элемент из Y считается лучше любого элемента из X;
2) все элементы из X (Y) эквивалентны друг другу.

Следствие 2. Если на каждом $\tilde{Q}_s$ сужение матрицы $(m(x,y))_{x,y\in\tilde{Q}}$ неразложимо и скалярный критерий $\gamma^{(s)}(x)$ из (12) является строгим, т.е. $\forall x,y \in \tilde{Q}_s\ x \neq y \Rightarrow \gamma^{(s)}(x) \neq \gamma^{(s)}(y)$, то существует слабый порядок R на множестве $\mathcal{P}$ всех WWW страниц, строящийся аналогично (8), в котором отношение $R^{(s)}$ определяется скалярным критерием $\gamma^{(s)}$.

2. Роль матрицы цитирований при построении упорядочения страниц WWW

Матрица цитирований $L = (l_{ij})_{i,j \in \mathcal{P}}$ формально описывает прямую связь между WWW страницами, обнаруживаемую присутствием на странице i указателя на страницу j. Эта связь носит направленный (асимметричный) характер (от i к j) и при динамической интерпретации L приводит к дуге на графе $G(L)$, представляющем собой графическую интерпретацию матрицы L. Однако, наличие цитирования показывает также, что страницы i и j связаны между собой другим отношением, которое качественно описывает степень тематической близости между страницами. Это отношение уже симметрично, поскольку если тема страницы i близка теме страницы j, то верно и обратное. Использование этого отношения весьма существенно, так как при анализе информации, помещенной на странице j для пользователя важно не только на какие другие страницы ссылается j, но и какие страницы $i \in \mathcal{P}$ ссылаются на j. Поэтому в ходе поиска ему следует просматривать страницы не только по направлению ссылок, но и в обратном направлении, пользуясь информацией о страницах, ссылающихся на текущую. Формально это можно выразить путем введения в рассмотрение не только матрицы L, но и симметричной матрицы $C = (c_{ij})_{i,j \in \mathcal{P}}$, элементы которой также имеют значения только 0 или 1. При этом $c_{ij} = 1$, если страницы i и j связаны цитированием, т.е. либо i цитирует j, либо j цитирует i, т.е.

$$c_{ij} = l_{ij} \vee l_{ji}, \tag{21}$$

где $\vee$ – операция дизъюнктивного сложения, определенная на множестве чисел $\{0, 1\}$:

$$1 = 1 \vee 1 = 1 \vee 0 = 0 \vee 1; \quad 0 \vee 0 = 0.$$

Таким образом, сопоставление страниц друг с другом следует вести не только по матрице L, но и по матрице C. Графически, это означает. что парное соотношение страниц i и j должно описываться не только дугой из i в j (если страница i цитирует страницу j), но и двумя дополнительными дугами, связывающими i и j в цикл (рис. 2)

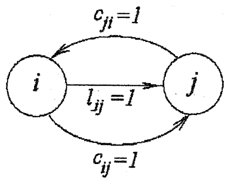


Рис. 2

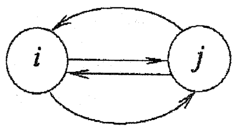


Рис. 3

При этом, если страница j также цитирует страницу i, то поскольку это обстоятельство ничего нового не добавляет к факту уже существующей (из-за цитирования j на странице i) тематической связи страниц i и j, фрагмент графа для этой пары вершин должен иметь вид, представленный на рис. 3.

В результате, исходной информацией для анализа упорядочения страниц WWW становится матрица $B = (b_{ij})_{i,j\in\mathcal{P}}$

$$b_{ij} = c_{ij} + l_{ij} \tag{22}$$

и отвечающий ей мультиграф, возможные фрагменты которого представлены на рис. 2,3.

Используя новую матрицу связей B вместо L, применим процедуру, изложенную в первом разделе для построения соответствующего упорядочения.

На основе отношений достижимости и эквивалентности, соответствующих новому графу $G(B)$, мы придем к разбиению множества страниц $\mathcal{P}$ на систему непересекающихся подмножеств $Q' = \{Q'_l\}_{l=1}^{s}$

$$\mathcal{P} = \bigcup_{l=1}^{s} Q'_l. \tag{23}$$

Это разбиение в силу свойств матрицы B имеет весьма специальную структуру.

Предложение 3. Для любого множества $Q'_l \in Q'$ длина его "хвоста" равна 1.

Доказательство. Заметим, что согласно определению матрицы B справедливо $b_{ij} > 0 \Rightarrow b_{ji} > 0$. Действительно, легко проверить, что

$$b_{ij} = 2l_{ij} + l_{ji}(1 - l_{ij}). \tag{24}$$

Тогда условие $b_{ij} > 0$ эквивалентно условию $l_{ij} \vee l_{ji} > 0$, откуда в силу симметрии следует $b_{ji} > 0$. Значит, отношение достижимости, построенное по матрице B будет симметричным, т.е. "хвост" любого Q'_l содержит только само это множество. *Q.E.D.*

Следствие. Любые два различных множества Q'_i и Q'_j не связаны друг с другом отношением цитирования.

Таким образом, разбиение (23) состоит из таких множеств Q'_l, страницы которых связаны между собой внутренними цитированиями, но не имеют указаний на страницы других множеств и не цитируются ни на одной странице, лежащей вне множества Q'_l.

Причины отсутствия взаимных цитирований между классами $\{Q'_l\}_{l=1}^s$ могут быть различными. Например,

1) на любых страницах $i \in Q'_l$ и $j \in Q'_k$ ($l \neq k$) находится совершенно различная информация;

2) на страницах из Q'_l и Q'_k ($l \neq k$) находится близкая по темам информация, но создатели этих страниц не имеют представления о существовании друг друга;

3) создатели страниц $i \in Q'_l$ и $j \in Q'_k$ ($l \neq k$) знают о существовании близкой по теме страницы, но не считают нужным (по разным причинам) делать ссылку.

Если реализуется первый случай, то нужная пользователю информация будет лежать только в одном из классов $\{Q'_l\}_{l=1}^s$.

Во втором случае с течением времени возможно слияние классов Q'_l и Q'_k, когда авторы страниц обнаружат близкую им страницу и сделают соответствующую ссылку.

В третьем случае пользователь будет стоять перед необходимостью искать нужную информацию как в Q'_l, так и в Q'_k.

Из этого анализа следует естественный вывод, что в процессе поиска следует предоставить пользователю возможность поэтапного поиска в каждом из тех классов $\{Q'_l\}_{l=1}^s$, где есть страницы, релевантные запросу по ключевым словам. Для этого достаточно провести отдельное упорядочение элементов каждого класса поотдельности. Для построения соотвествующего упорядочения применима система уравнений, аналогичная (5). Единственное отличие связано с иным видом матрицы $\tilde{T} = (t_{ij})_{i,j \in Q'_k}$, $|Q'_k| \geq 2$, которая получается теперь преобразованием элементов матрицы B, отвечающим множеству Q'_k,

$$\forall i, j \in Q'_k \quad t_{ij} = \begin{cases} b_{ij}, \text{ если } i \neq j \\ \sum\limits_{s \in Q'_k} b_{si}, \text{ если } j = i, \end{cases}$$

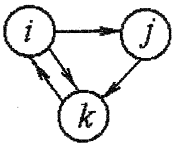
что с учетом (24) дает следующую связь матрицы $\tilde{T}$ с матрицей L

$$t_{ij} = \begin{cases} 2l_{ij} + l_{ji}(1 - l_{ij}), \text{ если } i \neq j \\ 2 \sum\limits_{s \in Q'_k} l_{si} + \sum\limits_{s \in Q'_k} l_{is} - (L^2)_{ii}, \text{ если } j = i, \end{cases} \tag{25}$$

где $(L^2)_{ii}$ – диагональный элемент матрицы L^2.

Для иллюстрации соответствующего упорядочения рассмотрим следующие примеры [2].

Пример 1. Пусть матрица связей и ее граф имеют вид (рис. 4)



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

Рис. 4

В работе [2] показано, что поскольку матрица L неразложима, то упорядочение, находящееся по (5), характеризуется матрицей

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

Максимальное собственное значение этой матрицы $\lambda_0 \cong 2.84$.

Элементы $\{i, j, k\}$ упорядочены скалярным критерием, компоненты которого пропорциональны вектору $(0.1, 0.04, 0.2)$, т.е. имеет место ранжирование kij ($\xi \cong (0.4, 0.2, 0.4)$, $\eta \cong (0.3, 0.2, 0.5)$).

Если вместо L использовать матрицу B (24)

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

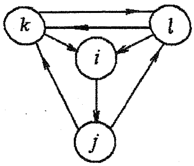
то соответствующая ей по (25) матрица

$$\tilde{T} = \begin{pmatrix} 3 & 2 & 2 \\ 1 & 3 & 2 \\ 2 & 1 & 4 \end{pmatrix}$$

имеет максимальное собственное значение $\lambda_0 \cong 6.7$, ее левый и правый собственные вектора $\xi \cong (0.4, 0.3, 0.4)$, $\eta \cong (0.3, 0.3, 0.4)$. Тогда вектор

$(\xi_i\eta_i) \cong (0.7, 0.5, 1)$ порождает то же самое ранжирование множества $\{i, j, k\}$.

Пример 2. Рассмотрим множество из четырех страниц $\{i, j, k, l, \}$ с матрицей связей L и ее графом $G(L)$, представленными на рис. 5



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

Рис. 5

Матрица

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

полученная преобразованием L согласно (6), имеет $\lambda_0 \cong 3.5$, $\xi \cong (0.16, 0.24, 0.3, 0.3)$ и $\eta \cong (0.34, 0.14, 0.26, 0.26)$. Скалярный критерий $(\xi_i\eta_i) \cong (0.05, 0.03, 0.08, 0.08)$ приводит к упорядочению $(kl)ij$, где элементы k и l эквивалентны друг другу и лучше элементов i и j.

Матрица B (24) для этого случая имеет вид

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

Ей отвечает согласно (25) матрица

$$\tilde{T} = \begin{pmatrix} 5 & 2 & 1 & 1 \\ 1 & 4 & 2 & 1 \\ 2 & 1 & 5 & 2 \\ 2 & 1 & 2 & 5 \end{pmatrix},$$

с максимальным собственным значением $\lambda_0 \cong 9.5$, $\xi \cong (0.21, 0.24, 0.27, 0.27)$ и $\eta \cong (0.28, 0.19, 0.26, 0.26)$. Их покомпонентное произведение $(\xi_i\eta_i) \cong (0.06, 0.05, 0.07, 0.07)$ приводит к тому же самому упорядочению $(kl)ij$.

Таким образом, мы видим, что использование вместо L другой матрицы B приводит для неразложимых матриц к схожим упорядочениям элементов. Однако, в случае разложимых матриц классы эквивалентности для матриц L и B будут, вообще говоря, различными, и для матрицы B их количество будет существенно меньшим, поскольку различными будут только классы, вообще не связанные друг с другом цитированиями. Эти классы и станут теми основными структурными единицами многообразия всех страниц WWW, в которых поиск релевантной информации следует вести независимо.

Литература

1. Page L., S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web (http://pr.efactory.de/e-references.shtml, PDF, 1998).
2. Левченков В.С., Левченкова Л.Г. Методы символической динамики и проблема моделирования поиска информации в Интернете. Прикладная математика и информатика. М: МГУ, 2004, N16, с. 74-89.
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. I. Введение в эргодическую теорию. - М.: ВМиК МГУ, 1997.
5. Гантмахер Ф.Р. Теория матриц. М.: Наука, 1967.