Анализ научного веб–пространства СО РАН методом PageRank*

Е. В. Константинова
Институт Математики им. С.Л. Соболева СО РАН, Новосибирск
e_konsta@math.nsc.ru

М. Ю. Савин
Новосибирский Государственный Университет, Новосибирск
buos91@yandex.ru

О. А. Клименко
Институт Вычислительных Технологий СО РАН, Новосибирск
klimenko@ict.nsc.ru

Аннотация. В работе представлен анализ научного веб–пространства Сибирского отделения Российской академии наук с использованием алгоритма ссылочного ранжирования PageRank, а также его модифицированной версии PageRankW. Полученные результаты позволяют выделить и проранжировать научные сайты СО РАН по степени их "важности". В работе также дается сравнительный анализ с другими системами ранжирования, в частности, методами вебометрики применительно к веб–графу научных сайтов СО РАН.

Ключевые слова. веб–графы, ссылочное ранжирование, вебометрика.

УДК 519-7

Введение

Алгоритм PageRank был разработан в 1998 г. Сергеем Брином и Ларри Пейджем [1] с целью ранжирования сайтов в поисковой системе Google. Он является одним из алгоритмов ссылочного ранжирования, применяемый для анализа информационных сетей, обладающих гиперссылками, в первую очередь, для анализа веб–пространств. Основная идея этого алгоритма состоит в том, чтобы проранжировать сайты по степени их *"важности"*: чем больше на сайт ссылаются с других сайтов и чем более "важными" эти сайты являются,

*Работа выполнена при финансовой поддержке Президиума СО РАН (Междисциплинарный интеграционный проект №21, 2012–2014) и РФФИ (проект 12-01-00448).

тем более важным считается исходный сайт. Каждому сайту алгоритм приписывает некоторое численное значение, называемое *весом* данного сайта, в соответствии с его "важностью". Таким образом, PageRank вычисляет вес сайта путём подсчёта важности ссылок на него. В целом, алгоритм может применяться к любому ориентированному графу с целью определения наиболее значимых вершин.

В данной работе предлагается модифицированная версия алгоритма PageRank для ориентированных взвешенных графов. Алгоритм PageRank и его модифицированная версия PageRankW используется для ранжирования научных сайтов Сибирского отделения Российской академии наук. Веб–граф научных сайтов СО РАН, построенный в Институте вычислительных технологий СО РАН [2], является ориентированным взвешенным графом, вершинами которого являются сайты научных организаций СО РАН, а дуги соответствуют наличию ссылок с одного сайта на другой. При этом дуга идет от вершины x к вершине y, если сайт, представленный в графе вершиной x, содержит хотя бы одну ссылку на сайт, представленный в графе вершиной y. Если два сайта содержат ссылки друг на друга, то между соответствующими им вершинами в графе будет две противоположно направленные дуги. Вес дуги определяется числом соответствующих ссылок. В [3] данный граф исследовался методами вебометрики с целью ранжирования научных сайтов СО РАН.

Статья устроена следующим образом. В следующих двух параграфах дается описание алгоритма PageRank, а также его модифицированной версии PageRankW. Затем для исследуемого графа приводится ранжирование научных сайтов СО РАН, полученное методами PageRank и PageRankW, и дается сравнительный анализ с ранжированием, полученным в [3] методами вебометрики.

1. Алгоритм PageRank

Дадим формальное описание алгоритма PageRank [1,4]. Пусть $\Gamma = (V, E)$ является ориентированным графом с множеством вершин V, $|V| = n$, и множеством дуг E. Определим для графа Γ квадратную $(n \times n)$–матрицу $H = (h_{ij})$, где

$$h_{ij} = \begin{cases} \frac{1}{q_i}, & \text{если существует дуга от вершины } i \text{ к вершине } j, \\ 0, & \text{в противном случае,} \end{cases}$$

где q_i соответствует числу дуг, выходящих из вершины i. Тогда вес π_j вершины j, $1 \leqslant j \leqslant n$, определяется следующим образом:

$$\pi_j = \sum_{i=1}^{n} \pi_i \, h_{ij},$$

или в матричном обозначении $\pi = \pi \, H$, где π является вектором весов вершин графа.

В данной модели имеются следующие две проблемы.

Проблема 1. В исследуемом графе могут присутствовать вершины, из которых не выходит ни одна дуга, такие вершины называются *висячими*. В этом случае соответствующая строка в матрице H будет содержать только нулевые элементы. Эту проблему решают следующим образом. Все нулевые строки в H заполняют одним и тем же значением $1/n$, где n равно числу вершин в графе. Полученную матрицу обозначим S.

Проблема 2. Как правило, в веб–графах имеется сильно связная компонента, внутри которой существует ориентированный путь между любой парой вершин. Следовательно, проходя по ссылкам соответствующих сайтов, можно обойти все вершины компоненты. Кроме этого, в веб–графах могут содержаться так называемые "ловушки", по своей структуре близкие к сильно связной компоненте, но с одной особенностью: из вершин "ловушки" не выходят дуги в вершины, которые не принадлежат "ловушке". Таким образом, попав в "ловушку" уже никогда не удается перейти в другие вершины графа. Эту проблему решают заменой матрицы S на *Google matrix* G следующего вида:

$$G = \alpha\, S + (1 - \alpha)\, E,$$

где $E = (\frac{1}{n})$ является квадратной $(n \times n)$–матрицей, а α является некоторым произвольным параметром, который называют *коэффициентом затухания*. Обычно, в алгоритме используют $\alpha = 0.85$. Детали об особенностях выбора α можно найти, например, в [4].

Тогда задача поиска вектора π весов вершин графа сводится к решению уравнения:

$$\pi = \pi\, G. \tag{1}$$

Обычно, вектор π нормализуется так, что $\sum_{i=1}^{n} \pi_i = 1$. В этом случае уравнение (1) принимает вид:

$$\pi = \alpha\, \pi\, S + (1 - \alpha)\, u, \tag{2}$$

где n–вектор u имеет вид $u = (\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n})$. Заметим, что полученная матрица G является стохастической, поскольку она является комбинацией двух статистических матриц S и E. Напомним, что *стохастической* называется матрица, чьи строки (или столбцы) дают в сумме единицу. Следовательно, уравнение (1) имеет хотя бы одно решение. Кроме этого, матрица G является *неприводимой*, поскольку она не содержит нулевых элементов, а значит, по Теореме Перрона–Фробениуса уравнение (1) будет иметь единственное решение.

Обычно для решения этого уравнению используют быстрый итерационный *степенной метод* [5], который состоит в следующем. Пусть $\pi^0 = u = (\frac{1}{n})$. Последовательно вычисляем $\pi^{k+1} = \pi^k\, G$ до тех пор, пока $\|\pi^{k+1} - \pi^k\| < \epsilon$, где $\|\cdot\|$ является расстоянием между двумя соседними векторами π, а ϵ задает нужную точность вычислений.

Данный метод можно интерпретировать, как случайное "хождение" по графу: случайный веб–серфер, начиная со случайного сайта, движется по случайным ссылкам [1]. Таким образом, значение PageRank для сайта определяется как относительная частота посещений данного сайта веб–серфером в предположении, что он движется бесконечно долго. Соответственно, Проблемы 1 и 2 можно интерпретировать следующим образом. Если веб–серфер попадает на "висячий" сайт, то он переходит на любой другой случайно выбранный сайт. В свою очередь, Google matrix задает вероятность α, с которой веб–серфер нажмет на случайную ссылку данного сайта, и вероятность $(1-\alpha)$, с которой он введет в строке браузера случайный URL другого сайта.

2. Модифицированный алгоритм PageRankW

Пусть $\Gamma = (V, E)$ является ориентированным взвешенным графом с множеством вершин V, $|V| = n$, и множеством дуг E, причем каждая дуга, идущая от вершины i к вершине j, имеет вес w_{ij}, равный числу ссылок, которые делаются с сайта i на сайт j. Зададим квадратную $(n \times n)$–матрицу $\widetilde{H} = (\widetilde{h}_{ij})$, где

$$\widetilde{h}_{ij} = \begin{cases} \frac{w_{ij}}{w_i}, & \text{если существует дуга от вершины } i \text{ к вершине } j, \\ 0, & \text{в противном случае,} \end{cases}$$

где $w_i = \sum_{j=1}^{n} w_{ij}$. Тогда вес $\widetilde{\pi}_j$ вершины j, $1 \leqslant j \leqslant n$, определяется следующим образом:

$$\widetilde{\pi}_j = \sum_{i=1}^{n} \widetilde{\pi}_i \, \widetilde{h}_{ij},$$

или в матричном обозначении $\widetilde{\pi} = \widetilde{\pi}\,\widetilde{H}$, где $\widetilde{\pi}$ является вектором весов вершин. Если в матрице $\widetilde{H}$ имеется нулевая i–строка (Проблема 1), то она заменяется строкой, состоящей из элементов $1/n$, где n равно числу вершин в графе. Полученную матрицу обозначим $\widetilde{S}$. Тогда результирующей матрицей, позволяющей избежать Проблемы 2, является матрица $\widetilde{G} = \alpha\,\widetilde{S} + (1-\alpha)\,E$, где $E = (\frac{1}{n})$ есть квадратная $(n \times n)$–матрица, а α есть коэффициент затухания. Задача поиска вектора $\widetilde{\pi}$ весов вершин графа в данном случае сводится к решению уравнения $\widetilde{\pi} = \widetilde{\pi}\,\widetilde{G}$. Если вектор $\widetilde{\pi}$ нормализуется так, что $\sum_{i=1}^{n} \widetilde{\pi}_i = 1$, то данное уравнение принимает вид $\widetilde{\pi} = \alpha\,\widetilde{\pi}\,\widetilde{S} + (1-\alpha)\,u$, где вектор u имеет вид $u = (\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n})$. Матрица $\widetilde{G}$ является стохастической и неприводимой с единственным решением, которое находится итерационным степенным методом, представленным в предыдущем параграфе.

3. Результаты

Алгоритм PageRank и его модифицированная версия PageRankW были использованы для ранжирования научных сайтов СО РАН на основе веб–графа, содержащего 88 вершин (сайтов) [2]. Полученные результаты с вычислительной точностью 10^{-8} представлены в Таблице 1, где deg^- и deg^+ обозначают число выходящих и входящих ссылок. Ранжирование организаций в Таблице дано в соответствии с данными алгоритма PageRankW.

В целом, оба алгоритма приводят к похожим результатам. Однако, следует отметить, что алгоритм PageRankW, учитывающий веса выходящих дуг, имеет бóльшую чувствительность к "важности" вершин. Особенно хорошо это видно на Рисунке 1. Первые два пика соответствуют Порталу СО РАН и Президиуму СО РАН, что вполне закономерно: именно эти сайты являются наиболее "важными" в научной сети СО РАН с административной точки зрения. Красная линия, соответствующая PageRankW, дает бóльший пик по сравнению с синей линией, соответствующей PageRank, поскольку PageRankW модифицирован специально для взвешенных ориентированных графов. Сайты на рисунке представлены в соответствии с [2]. Из научных сайтов, в первую очередь, выделяется сайт ИВТ СО РАН. В конце рейтинга оказываются сайты, на которые мало ссылаются и с которых имеется мало ссылок. Также в конец списка PageRankW попадают сайты, которые ссылаются на

“висячие” сайты, не являющиеся важными в данной сети, либо на него ссылаются с таких сайтов (например, сайт ЯНЦ СО РАН).

Полученные данные находятся в согласии с результатами, представленными в [3], где данный граф исследовался методами вебометрики с целью ранжирования научных сайтов СО РАН. Среди 20 лучших сайтов в рейтинге научных организаций, представленных в [3], 13 сайтов попадают в двадцать лучших в рейтинге алгоритма PageRankW, а именно: Портал СО РАН, Президиум СО РАН, ИВТ СО РАН, ГПНТБ СО РАН, ИМ СО РАН, ИЯФ СО РАН, ОУС СО РАН по НИТ, ИХКГ СО РАН, ИФПР СО РАН, ИНГГ СО РАН, ИВМиМГ СО РАН, ИЦиГ СО РАН, ИНХ СО РАН.

В 2010 г. независимые эксперты из институтов СО РАН разной тематической направленности изучили все сайты СО РАН, чтобы выделить группу институтов с лучшими сайтами [6]. По разным номинациям в группу лучших попали ИВТ СО РАН, ИМ СО РАН, ИК СО РАН, ИЦиГ СО РАН, ГПНТБ СО РАН, ИНХ СО РАН, ИВМиМГ СО РАН, сайт Объединенного ученого совета СО РАН по нанотехнологиям и информационным технологиям.

Существенное совпадение результатов трех независимых рейтингов (PageRankW, вебометрика, независимая экспертиза) говорит о правильности выбора критериев оценивания.

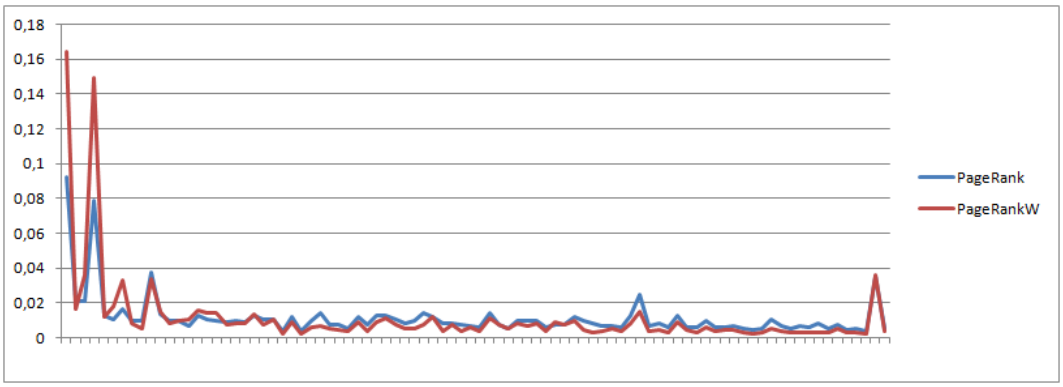

Рис. 1: Результаты алгоритмов PageRank и PageRankW на веб–графе СО РАН

ЛИТЕРАТУРА

[1]. Brin S., Page L. The anatomy of a large-scale hypertextual web search engine // Comput. Netw. ISDN Syst. 1998. Vol. 30 No. 1–7. P. 107–117.

[2]. Веб–граф организаций СО РАН. http://www.ict.nsc.ru/ranking/

[3]. Шокин Ю. И., Веснин А. Ю., Добрынин А. А., Клименко О. А., Рычкова Е. В., Петров И. С. Исследование научного веб–пространства Сибирского отделения Российской академии наук / представлено в печать.

[4]. Langville A.N., Meyer C. Google’s PageRank and Beyond: The Science of Search Engine Rankings / Princeton University Press.—NJ, USA, 2006.—234p.

[5]. R. von Mises, Pollaczek–Geiringer H. Praktische Verfahren der Gleichungsauflösung // ZAMM – Zeitschrift für Angewandte Mathematik und Mechanik. 1929. Vol. 9. P. 152–164.

[6]. http://www.sbras.ru/HBC/article.phtml?nid=562&id=14

Таблица 1: Данные веб–графа СО РАН и алгоритмов PageRankW, PageRank

Название организаций	deg^-	deg^+	PageRankW	PageRank
Портал СО РАН	10544	19937	0.16468	0.0921294
Президиум СО РАН	6303	15165	0.149067	0.0784779
ИВТ СО РАН	19297	3854	0.0361295	0.0208014
Отделение ГПНТБ СО РАН	13257	861	0.0358041	0.03529
ГПНТБ СО РАН	133	2078	0.0338478	0.0375937
ИМ СО РАН	324	2304	0.0326061	0.0164767
ИЯФ СО РАН	818	939	0.0181472	0.0102423
ОУС СО РАН по НИТ	4812	1249	0.0164676	0.0212415
ИХКГ СО РАН	99	493	0.015321	0.0123696
ИФПР СО РАН	2	290	0.0151103	0.013425
ИНЦ СО РАН	54	293	0.0151042	0.0244312
ИНГГ СО РАН	13	454	0.0141158	0.00990026
ИВМиМГ СО РАН	38	704	0.0138936	0.0104245
ИК СО РАН	399	657	0.0134709	0.012728
ИХБФМ СО РАН	93	775	0.0120586	0.0121437
ИЦиГ СО РАН	296	586	0.0117564	0.012574
ИСЭ СО РАН	123	483	0.0110488	0.0125004
ИДСТУ СО РАН	96	538	0.011009	0.0138236
ЛИН СО РАН	3	435	0.0107478	0.0103944
ИНХ СО РАН	5520	479	0.0105111	0.0066384
МТЦ СО РАН	28	197	0.00974846	0.0116713
ИТПМ СО РАН	303	543	0.00940108	0.0100319
ИФПМ СО РАН	77	336	0.00923991	0.0129145
ИМКЭС СО РАН	8	314	0.00923577	0.0124972
ИСЭМ СО РАН	303	385	0.00878714	0.0117728
ИФП СО РАН	23	647	0.00862628	0.0120326
ИГМ СО РАН	1010	235	0.00858894	0.0077112
ИФ СО РАН	694	393	0.00847579	0.00953941
ИИ СО РАН	2	529	0.00834924	0.00948687
ИАиЭ СО РАН	198	639	0.00820226	0.00855877

Название организаций	deg^-	deg^+	PageRankW	PageRank
НИОХ СО РАН	31	688	0.00810368	0.00977558
ИТ СО РАН	40	464	0.00796747	0.0100381
ИПРЭК СО РАН	230	212	0.00793664	0.012367
ИГСО РАН	85	358	0.00791316	0.00953433
ИЛ СО РАН	2	466	0.00771568	0.00856451
ИЛФ СО РАН	1	448	0.00768866	0.00777221
ИКЗ СО РАН	303	790	0.00747417	0.010141
ИЗК СО РАН	427	191	0.00729068	0.0107414
ИГД СО РАН	87	214	0.00723841	0.013968
ИАЭТ СО РАН	5	568	0.00719145	0.00813926
ИОА СО РАН	7	676	0.00713456	0.00767053
ТНЦ СО РАН	583	121	0.00645867	0.0139038
ИГХ СО РАН	28	253	0.00641652	0.00969615
СИФИБР СО РАН	2	193	0.00603319	0.00953433
ИВМ СО РАН	157	421	0.00584605	0.00968761
ИГиЛ СО РАН	47	297	0.00563962	0.00701016
НФ ИВЭП СО РАН	0	65	0.00548614	0.0106826
ИСиЭЖ СО РАН	0	242	0.00541454	0.00688851
ИЭОПП СО РАН	0	222	0.00533232	0.00713224
ИОЭБ СО РАН	0	185	0.00530576	0.00706676
ИБФ СО РАН	12	305	0.005284	0.00975704
ИХН СО РАН	5	227	0.00527855	0.00841373
ИСИ СО РАН	64	230	0.00498314	0.00942201
ЦСБС СО РАН	0	475	0.00483553	0.00528527
ОФ ИМ СО РАН	150	197	0.00448855	0.00998244
ТюмНЦ СО РАН	1077	167	0.00446569	0.00786293
ИХТТМ СО РАН	64	319	0.00439182	0.00770388
ГИН СО РАН	17	153	0.00432087	0.00646536
КТИ ВТ СО РАН	28	220	0.00431972	0.00564898
ИГАБМ СО РАН	36	150	0.00407029	0.00572216

Название организаций	deg^-	deg^+	PageRankW	PageRank
ИПХЭТ СО РАН	48	191	0.00404023	0.00570137
ИФЛ СО РАН	4	168	0.00402177	0.00656142
КНЦ СО РАН	190	131	0.00399293	0.00855118
КТФ ИГиЛ СО РАН	0	98	0.00390346	0.00687947
ИрИХ СО РАН	2	67	0.00389753	0.00627936
КТИ НП СО РАН	75	210	0.0038293	0.00602847
КемНЦ СО РАН	43	174	0.0037486	0.00572216
ИКФИА СО РАН	0	144	0.00374325	0.00528527
ИВЭП СО РАН	3	104	0.00353026	0.00737213
БИП СО РАН	0	132	0.00346999	0.00645068
ИБПК СО РАН	0	98	0.00343893	0.00528527
ИХХТ СО РАН	18	161	0.00331446	0.007337
ИМЗ СО РАН	0	131	0.00309612	0.0078776
ИУ СО РАН	0	96	0.00306697	0.00815335
ОНЦ СО РАН	7	159	0.00305808	0.00626157
ИГДС СО РАН	1	95	0.0029756	0.00572216
БНЦ СО РАН	4	224	0.00290553	0.00485527
ИПА СО РАН	0	82	0.00289001	0.00528527
ИПНГ СО РАН	0	72	0.0028547	0.00491937
ИМБТ СО РАН	0	69	0.0028267	0.0055838
ИППУ СО РАН	1	58	0.00272994	0.00528527
ЯНЦ СО РАН	446	74	0.00257741	0.00441838
ТФ ИТПМ СО РАН	1	65	0.00256565	0.00699055
ИФТПС СО РАН	0	79	0.00256037	0.00528527
ГС СО РАН	6	16	0.00240966	0.00464224
ИСЗФ СО РАН	0	1	0.00232515	0.0036219
ИПОС СО РАН	0	1	0.00232515	0.0036219
СКТБ Наука КНЦ СО РАН	3	1	0.00232515	0.0036219