Решение задачи тематического ссылочного ранжирования Интернет-ресурсов

О.Г. Григорьев, М.А. Ширай

Аннотация: в данной работе изучаются современные проблемы информационно-поисковых систем (ИПС) сети Интернет, описывается постановка задачи тематического ранжирования web-сайтов. Рассматривается разработка модели тематического ранжирования на основе PageRank с использованием новых методов построения и хранения обратного ссылочного индекса. Рассматриваются различные методы решения СЛАУ, применительно к этой задаче. Производится выбор множества методов для реализации и сравнительная оценка полученных результатов.

Ключевые слова: Интернет, ИПС, PageRank, ссылочное ранжирование, обратный ссылочной индекс, СЛАУ, итерационные методы, экстраполяция.

Введение

Информационно-поисковые системы в настоящее время являются основным способом доступа к информации в сети Интернет [1]. Необходимость в службах, предоставляющих возможность поиска информации в Интернет, появилась сразу же после возникновения сети, и на данный момент по качеству поиска и количеству обработанных источников информационно-поисковые системы (ИПС) сети Интернет не имеют аналогов. Основным критерием качества работы поисковой системы является релевантность поиска — степень соответствия найденной информации запросу, то есть уместность результата.

Результатом работы поисковой системы — ответом на пользовательский запрос — является множество ссылок на интернет страницы различных web-сайтов. Чаще всего стандартный поисковый запрос выдает одну (реже несколько) страницу с сайта. Задача поиска информации в сети Интернет заключается в том, чтобы по запросу пользователя выбрать один десяток из многих миллионов страниц, найденных по запросу. Эта задача решается сортировкой страниц, по, поставленному им в соответствие, рангу. Ранг — вектор чисел, соответствующий множеству страниц. Элементы ранга определяют степень соответствия страницы-результата поисковому запросу. Существует множество различных критериев, называемых факторами ранжирования, по которым можно вычислить ранг множества страниц. Ранг, используемый при ранжировании, является суперпозицией таких рангов.

Факторы, оцениваемые методами ранжирования, принято разделять на внутренние и внешние. Внутренние факторы — это те, которые определяются содержанием страницы и характеристиками сайта, на котором она расположена, в данной работе они не рассматриваются. Внешние — все остальные [2]. На текущий момент основным внешним фактором ранжирования является количество и «качество» входящих ссылок со страниц других web-сайтов и количество исходящих ссылок. Оценка «качества» ссылок происходит на основе вероятности перехода пользователя по ним. Ранжирование, оценивающее этот фактор,

называется «ссылочное ранжирование» [3]. Ссылочное ранжирование имеет высокую трудоемкость, которая нарастает с увеличением числа страниц в сети и является сдерживающим фактором в ее индексации, но при этом его использование позволило поднять качество поиска на более высокий уровень.

В современном русскоязычном сегменте Интернета количество зарегистрированных доменных имен второго уровня зоны «.ru» в 2012 году превысило 3.5 миллиона [4]. Доля доменных имен регистрируемых исключительно с целью коммерческого использования возрастает по мере того как Интернет становится всемирной рекламной площадкой. Производители и поставщики товаров и услуг различного уровня активно переносят рекламные бюджеты в сеть Интернет. Главной ценностью в сети Интернет становится внимание пользователей. Оно распределяется поисковыми системами между сайтами, на которых пользователи находят необходимую им информацию. Позиции сайта в ранге поисковых систем в значительной мере определяются количеством входящих ссылок с других сайтов. Недобросовестные web-мастера пытаются получить прибыль из этого процесса, создавая бесполезные для пользователей сайты — сателлиты (поисковый спам). Создание таких сайтов не требует больших трудозатрат, что позволяет в краткие сроки извлечь большую прибыль от размещения рекламных материалов и ссылок. Основная задача web-мастеров, создающих сателлиты, состоит в том, чтобы добиться их присутствия в выдаче по различным запросам. Для этого они манипулируют факторами ранжирования, в том числе, покупая размещение ссылок на других ресурсах. В случае успеха на их сайты попадают пользователи, а поисковые системы учитывают их ссылки при ранжировании других сайтов. В итоге поисковый спам ухудшает релевантность выдачи ИПС и затрудняет поиск информации в сети Интернет. О важности проблемы говорит и то, что количество web-сайтов, зарегистрированных в бирже покупки и продажи ссылок «sape.ru», уже приближается к 500 000 [5]. Одним из способов борьбы с этой проблемой является применение тематического ссылочного ранжирования, то есть ранжирования, использующего суперпозицию рангов, построенных по отдельным тематикам.

Множество тематик определяется в соответствие с поставленной задачей ранжирования. Для организации поиска по запросам на естественном языке необходимо иметь каталог тематик, в равной степени соответствующий множеству поисковых запросов. Задача также может состоять в поиске всех наиболее релевантных документов для какой-то конкретной тематики, в данном случае ранжирование будет использовать только эту тематику.

Тематическое ранжирование повышает значение цепочек ссылок одной тематики и не учитывает цепочки ссылок различных тематик, за счет этого удается отсеять большое количество покупных и незначащих ссылок, что приводит к повышению релевантности поиска. Кроме того, ссылочное ранжирование может применяться для решения различных информационно-поисковых задач, например:

- создание каталога наиболее известных сайтов в каждой тематике;
- поиск сайтов повышенной социальной напряженности;
- поиск сайтов, обеспечивающий закрытый доступ к нелегальной информации;

В настоящий момент исследования, направленные на повышение эффективности работы и применение тематического ссылочного ранжирования, особенно актуальны.

Целью данной работы является разработка новых алгоритмов тематического ссылочного ранжирования и повышение эффективности их работы. При этом решаются следующие задачи:

- разработка новой модели ссылочного ранжирования;
- реализация различных методов решения систем линейных уравнений для построения ссылочного ранга;
- исследование эффективности их применения.

1. Тематическое ссылочное ранжирование Интернет-ресурсов

Ссылочное ранжирование рассматривает пользователей как случайно перемещающиеся по сети объекты. В сети интернет существует несколько способов совершить переход с одной

страницы на другую. Основной единицей разметки страницы, обеспечивающей такой переход, является гиперссылка, ссылающаяся на другую страницу. Гиперссылка обеспечивает непосредственный переход на указанную страницу средствами HTML разметки. Существуют и другие способы обеспечения перехода между страницами, все они производятся выполнением дополнительного программного кода сайта. Основной их задачей является выполнение тех или иных действий, направленных на сбор информации от пользователя, переход между страницами чаще всего носит утилитарный характер и не выражает мнения пользователя. Кроме того, вебмастер (создатель сайта) может намеренно использовать дополнительный программный код при создании ссылки для сокрытия ее от индексации любыми роботами поисковых систем. Поэтому известные алгоритмы ссылочного ранжирования учитывают в качестве средства перехода пользователей между страницами только гиперссылки (далее ссылки).

В известной модели ссылочного ранжирования PageRank [6] возможность перехода пользователя по той или иной ссылке на странице принимается равновероятной. Если же рассматривать реального пользователя сети, то такой подход можно улучшить. Пользователи в Интернет в каждый момент времени имеют цели, следуя которым, они ищут информацию и совершают перемещения между страницами. Следуя этой логике, можно сказать, что пользователь, перешедший на сайт по ссылке, с большей вероятностью перейдет с него по ссылке, отвечающей той же цели, то есть принадлежащей к той же тематике.

Ссылки в Интернет размечаются таким образом, чтобы отображать пользователю «анкор» — строку с описанием ссылки. «Анкоры» помогают пользователям делать выбор в посещении той или иной страницы. Исходя из этого, вероятность выбора пользователем ссылки можно оценить. Разбив информационное пространство на множество тематик, можно, используя «анкор», определить тематический вес ссылки в той или иной тематике [7]. Благодаря этому можно построить ссылочные ранги для каждой используемой тематики, включив в формулу ссылочного ранжирования коэффициент принадлежности к тематике. Таким образом, полученные ранги будут задавать степени соответствия страниц тематикам. При этом нет необходимости в лингвистическом анализе текста самой страницы, что позволяет значительно сократить использование вычислительных ресурсов. Множество таких рангов будем называть тематическим ссылочным рангом, а ранжирование, использующее его, — тематическим ссылочным ранжированием.

При организации тематического ранжирования встает задача об автоматическом определении тематики коротких фраз (от 1 до 10 слов), поисковых запросов и «анкоров» ссылок. За последнее время вышло большое количество статей, описывающих методы и алгоритмы решения этой задачи [8, 9]. Чаще всего для этого используется словарь слов и словосочетаний, наиболее точно характеризующий каждую из тематик. Составление такого словаря — непростая задача, обычно она решается с привлечением экспертов-лингвистов.

Для предоставления пользователю наиболее релевантной выдачи при тематическом ссылочном ранжировании определяется принадлежность запроса к тематикам. На ранжирование в значительной степени влияют ранги тех тематик, которым в большей степени принадлежит запрос. Тематическое ссылочное ранжирование не требует дополнительных посещений web-сайтов относительно ссылочного ранжирования. Использование тематического ссылочного ранжирования позволит давать пользователю результаты, как более соответствующие его цели. В зависимости от задачи, тематическое ранжирование можно применить как к отдельным страницам, так и к сайтам, оценивая их исходя из множества всех входящих и исходящих ссылок на страницах сайта. Количество страниц на несколько порядков больше количества сайтов, трудоемкость задачи ранжирования страниц соответственно выше.

При организации работы поисковой машины ссылочный фактор ранжирования является только одним из многих используемых факторов. При этом он является самым трудоёмким по количеству необходимых вычислительных затрат. Чаще всего поисковые машины выдают по одной странице с сайта, реже несколько.

Эти страницы можно получить ранжированием всех страниц сайта по внутренним факторам. Применяя ссылочное ранжирование к множеству сайтов, можно получить ранжирование сайтов между собой и, в совокупности с другими факторами ранжирования, результирующий ранг полученных страниц. Далее рассматривается тематическое ссылочное ранжирование множества сайтов, однако те же методы могут быть применены и при ранжировании отдельных страниц.

2. Постановка задачи тематического ссылочного ранжирования

Рассмотрим задачу построения тематического ссылочного ранга web-сайтов. Web-сайты и их ссылки можно представить в виде графа, в котором сайты — это вершины, а ссылки — дуги графа. Под моделью ссылочного ранжирования будем понимать функцию, вычисления ранга вершины. Для пресечения манипулирования выдачей все ИПС скрывают информацию о технологиях ранжирования, в публичном доступе есть только информация о моделях ссылочного ранжирования, дающих базовое представление о работе современных ИПС. Для разработки собственного тематического ранга требуется разработать подходящую модель ссылочного ранжирования. Описание моделей ссылочного ранжирования, представлено в [10].

Пусть $G(V,E)$ — орграф, где V — множество вершин, а E — множество дуг. ОСИ можно представить в виде разреженной матрицы S — матрицы смежности графа G, состоящей из элементов $S_{ij} = \begin{cases} 0, & (j,i) \notin E \\ 1, & (j,i) \in E \end{cases}, i \in V,\ j \in V$.

Стоит отметить, что в матрице могут быть как ссылки (i, j), так и (j, i) единовременно. Пусть Θ – некое множество тематик, функция $T(i,j,t)$ определяет тематический вес ссылки (j, i) в тематике $t \in \Theta$ одним из методов автоматического определения тематики коротких фраз в применении к «анкору» ссылки, $\deg\_o(i,t) = \sum_{j=0}^{N} T(i,j,t) \cdot S_{ij}$ — сумма тематических весов всех исходящих ссылок вершин i. Функция $P(i,j,t) = \dfrac{T(i,j,t)}{\deg\_o(i,t)}$ задает вероятность перехода пользователя по ссылке (i,j) с учетом ее тематики. $\forall t \in \Theta : S_t(i,j) = S_{ij} \cdot P(i,j,t)$ — матрица тематических весов для ссылок (i, j). Определим модель тематического ссылочного ранжирования с использованием этой матрицы и запишем систему линейных алгебраических уравнений (СЛАУ) в итерационном виде.

$$X_i^{k+1} = (1-d) + d \sum_{j=1,\, j \neq i}^{N} X_j^k \cdot S_t(j,i), \qquad (1)$$

где d — коэффициент затухания, k — номер шага итерации, X_j^k — значение ранга j-той вершины на итерации k.

Для построения ссылочного ранга в некоторой тематике с использованием указанной модели требуется решить систему уравнений, состоящую из уравнений вычисления ранга для каждой вершины сети. Для решения этой системы понадобится информация обо всех ведущих на web-сайт ссылках. Обратный ссылочный индекс (ОСИ) хранит и предоставляет такую информацию [3], разработка и реализация метода хранения такого индекса описана в [11].

3. Применение итерационных методов решения СЛАУ

Для матрицы системы уравнений (1), так же как и для модели PageRank, выполняются условия диагонального преобладания и обратимости [12]. Она является также хорошо обусловленной [13]. Это позволяет применять прямые и итерационные методы решения СЛАУ. Так как на текущий момент количество web-сайтов в русскоязычном сегменте интернета превышает 3.5 млн. [4], итерационные методы будут предпочтительны. Решение (1) возможно итерационными методами с применением современных средств организации параллельных вычислений [14] на классических вычислительных системах, так и итерационными методами, реализованными на графическом процессоре (GPU). Временем сходимости метода будем называть время работы итерационного алгоритма приво-

дящего к полной сходимости итерационной функции. Будем считать его критерием выбора метода, наравне с точностью получаемого решения в процентах ($\mathrm{d} = \frac{1}{n}\sum_{i}^{n}\frac{|X_i - Xe_i|}{Xe_i}$, где Xe_i - эталонное значение ранга). Рассмотрим методы решения матричного уравнения вида $Ax = b$ [15].

Метод Якоби. Для решения уравнения в матричной форме $A\vec{x} = b$, при заданном начальном приближении $\vec{x}_0$, итеративная процедура нахождения решения выглядит следующим образом:

$$\vec{x}^{k+1} = B\vec{x}^{k} + \vec{g},$$

где $B = E - D^{-1}A, \vec{g} = D^{-1}\vec{b}$, $E-$ единичная матрица, $D-$ диагональная подматрица матрицы A.

По методу Якоби итеративная формула ранга записывается так:

$$X_i^{k+1} = a_i \cdot \sum_{j=1, j\neq i}^{N} X_j^k \cdot S_t(i,j) + b_i, \qquad (2)$$

где $a_i = d, b_i = 1 - d$.

Метод Гаусса-Зейделя. Очевидный недостаток метода Якоби состоит в том, что при нахождении компонент X_i^{k+1} никак не используется информация о уже пересчитанных компонентах $X_1^{k+1},\ldots, X_{i-1}^{k+1}$. Исправить этот недостаток можно, переписав (2) в виде

$$X_i^{k+1} = a_i \cdot C_i^k + b_i,$$

$$C_i^k = \sum_{j=1}^{i-1} X_j^{k+1} \cdot S_t(i,j) + \sum_{j=i+1}^{N} X_j^k \cdot S_t(i,j). \qquad (3)$$

Такой вариант называется методом Гаусса-Зейделя, сходимость этого метода почти на 40% выше [16].

Метод секущих. Корень функции $f(x)$ может быть найден следующим образом. Выберем две начальные точки $C_1(x_1, y_1)$ и $C_2(x_2, y_2)$ и проведем через них прямую. Она пересечет ось абсцисс в точке $(x_3, 0)$. Теперь найдем значение функции для значения аргумента x_3. Временно будем считать x_3 корнем на отрезке $[x_1, x_2]$. Пусть точка C_3 имеет абсциссу x_3 и лежит на графике. Теперь вместо точек C_1 и C_2 возьмём точку C_3 и точку C_2. С этими двумя точками проделаем ту же операцию.ю и так далее, то есть будем получать две точки C_{n+1}, C_n и повторять операцию с ними. Эти действия нужно повторять до тех пор, пока мы не получим значение корня с нужным нам приближением. По сути метод секущих получается на базе метода касательных (метода Ньютона) заменой производной разностным приближением

$$f'(x^k) \approx \frac{f(x^k) - f'(x^{k-1})}{x^k - x^{k-1}}.$$

Метод секущих обладает рядом недостатков [17]. Он не позволяет находить корни, в которых функция не пересекает ось x, а лишь касается ее. Вместе с корнями определяются и точки разрыва функции, если при переходе через них функция меняет знак. Функция x_i^k асимптотически стремится к своему корню, производные в его окрестности принимают бесконечно малые значения, поэтому, при применении к этой функции, данный метод расходится.

Увеличение вычислительной эффективности на основе метода релаксации. Рассмотрим одну из вершин графа. Пусть x_i^k ранг в вершине i на k-той итерации. Web-граф $G(V,E)$ содержит большое число циклов (контуров), в результате чего ранг вершины на k-той итерации отчасти определяется рангом этой же вершины на предыдущей итерации. Если из web-графа удалить все содержащие вершину i контуры, то ранг вершины i полностью определяется рангами других вершин графа. Если бы web-граф не содержал контуров, его вершины можно было бы подвергнуть топологической сортировке и рассчитать их ранги за один проход. Суть метода заключается в том, чтобы рассчитать влияние вершины на саму себя для следующей итерации, на основе информации о циклах в графе и использовать его для ускорения сходимости. Рассмотрим модель PageRank:

$$X_i^{k+1} = (1-d) + d \cdot \sum_{j=1, j\neq i}^{N} \frac{X_j^k \cdot S(i,j)}{\deg\_o(j)}. \quad (4)$$

Выделим влияние вершины i на саму себя через простой контур $C_0 : i \to i_1 \to i$

$$X_i^{k+1} = (1-d) + d(1-d) + d^2 \cdot \frac{X_i^{k-1}}{\deg\_o(i) \cdot \deg\_o(i_1)} +$$

$$d \cdot \sum_{j=1, j\neq i, j\neq i_1}^{N} \frac{X_j^{k-1} \cdot S(i_1, j)}{\deg\_o(j)} + d \cdot \sum_{j=1, j\neq i, j\neq i_1}^{N} \frac{X_j^k \cdot S(i, j)}{\deg\_o(j)}$$

$$Kt_i^{C_0} = d^2 \cdot \frac{X_i^{k-1}}{\deg\_o(i) \cdot \deg\_o(i_1)}.$$

Исходя из условий сходимости $X_i^{k+1} \approx X_i^k$, поэтому для Ω_m — множества контуров длины m в которые входит вершина v, выполняется

$$\forall C \in \Omega_m : Kt_i^C = \left(\sum_{n=1}^{|C|} \frac{d^{n+1}}{\prod_{c=1}^{c<n} \deg\_o(c)} \right) \cdot X_i^k \quad (5)$$

$$\forall C \in \Omega_m : Kt_i^C = K_i^C \cdot X_i^k$$

$$K_i = \sum_{C \in \Omega_m} K_i^C.$$

Тогда (4) можно переписать в виде:

$$X_i^{k+1} = K_i \cdot X_i^k + B_i^k,$$

где B_i^k — влияние вершин не входящих в рассматриваемые контуры.

Исходя из условий сходимости, начиная с некоторой итерации $k : X_i^{k+1} \approx X_i^k$. Следовательно, для сходимости необходимо выполнение условия $X_i^{k+1} = K_i \cdot X_i^k + B_i^k$ то есть

$$X_i^k \approx \frac{B_i^k}{1-K_i} = \frac{X_i^{k+1} - K_i \cdot X_i^k}{1-K_i} \approx X_i^{k+1}. \quad (6)$$

Функция (6) имеет разрыв в окрестности точки $K_i = 1$. Для устранения разрыва рассмотрим ее предел.

$$\lim_{K_i \to 1} \frac{X_i^{k+1} - K_i \cdot X_i^k}{1-K_i} = X_i^k.$$

Это говорит о том, что ряд в этой окрестности сходится к X_i^k.

При этом $\begin{cases} K_i \in (0;1) : \dfrac{X_i^{k+1} - K_i \cdot X_i^k}{1-K_i} \\ K_i \approx 1 : X_i^k \end{cases}$

Применяя метод к (1), получаем аналогичный результат.

Увеличение вычислительной эффективности с помощью экстраполяции. Время сходимости (3) можно уменьшить. Применив методы экстраполяции, можно предугадывать значения функции на следующих шагах и использовать его в расчетах [18]. Для корректной работы методов экстраполяции функция должна быть дифференцируема на отрезке их применения. Итерационная функция асимптотически сходится к своему конечному значению. Однако на начальных итерациях ее значение далеко от конечного, а значения производных сильно варьируются. Поэтому метод можно применять, только начиная с некой *i*-той итерации, в которой будет известно *n* первых производных. Ввиду того, что вычисление экстраполяции будет проводиться для каждого из узлов на каждой итерации, необходимо разработать наименее трудоемкий алгоритм.

Наиболее простыми и распространенными являются методы полиномиальной интерполяции. Так как для каждого набора точек необходимо вычислить только одно значение экстраполяции, то методы построения интерполяционных многочленов в данной работе не рассматриваются. Рассмотрим метод конечных разностей [19]. Учитывая дискретный характер функции, *n*-ной производной f_i^n на *i*-той итерации будем называть конечные разности порядка *n*:

$$f_i^n = f_i^{n-1} - f_{i-1}^{n-1},$$

где f_i^0 — значение функции на -той итерации.

$$f_{i+1}^n = f_i^n + f_{i+1}^{n+1}. \quad (7)$$

Допустим, на шаге *i* известно *k* первых производных, примем $f_{i+1}^{n+1} \approx 0$, тогда, согласно (7), $f_{i+1}^k \approx f_i^k$. Рекурсивно подставляя (7) в себя, получим, что предполагаемый следующий элемент функции, при известных *k* производных, вычисляется так:

$$f_{i+1}^{0} \approx f_i^{0} + \sum_{n=1}^{k} f_i^{n} . \qquad (8)$$

Переходя к алгоритму, отметим, что для расчета необходимо хранить значения первых k производных на предыдущей итерации. Поэтому выбор параметра k производится исходя из имеющихся в наличии объемов оперативной памяти — чем он больше, тем выше точность получаемого решения (далее выбрано $k = 7$).

Функция ранга — постоянно возрастающая функция. Однако в случае, когда экстраполированное значение функции превышает конечное, значение функции может колебаться в окрестности результирующего значения — такая ситуация означает, что для выбранного параметра k функция достигла результирующего значения в некотором приближении. На малом количестве точек полиномиальная интерполяция не дает достаточной точности прогнозируемого решения, поэтому ее стоит применять только после того, как будет вычислено k первых производных.

Из (4) следует, что существует влияние вершины на саму себя через контур, в случае применения экстраполяции ее значение на i-той итерации можно представить в виде:

$$Xe_i = X_i + E(X_i, X_{i-1},...X_1),$$

где X_i — вычисленное значение, E — приращение от экстраполяции. Тогда $Xe_{i+1} = X_{i+1} + E(X_{i+1}, X_i + E(X_i, X_{i-1},...X_1),...X_1)$. Таким образом, начиная с третьего шага, экстраполяция производится не только от вычисленного значения, но и от приращения от экстраполяции на предшествующих шагах. В результате выходные значения алгоритма оказываются существенно завышенными. Наиболее очевидным решением введение в (8) коэффициента релаксации ω (значение можно вычислить экспериментально) для компенсации двойной экстраполяции:

$$f_{i+1}^{0} \approx f_i^{0} + \omega \cdot \sum_{n=1}^{k} f_i^{n} . \qquad (9)$$

Значения производных функции рангов на интервале итераций примерно [1,10] имеют тенденцию к возрастанию по модулю в зависимости от порядка производной $\left| f_i^{n+1} \right| > \left| f_i^{n} \right|$. В (9) каждая из производных входит с одинаковым коэффициентом, оказывая эквивалентное влияние на результат, что приводит к увеличению погрешности вычислений. Метод, задающий разную степень влияния для производных в зависимости от порядка, может иметь меньшую погрешность.

В связи с тем, что функция ранга, по результатам проведенного исследования, в интервале итераций [2, 10] и далее обладает достаточно высокой степенью гладкости (больше 7-ми), то (9) можно преобразовать в вычисление усеченного ряда Тейлора в окрестности точки i и применить его для экстраполяции в точке $i+1$. Для этого коэффициент ω следует внести под знак суммы, заменив вектором коэффициентов ряда Тейлора $\omega = \frac{h^n}{n!}$.

$$f_{i+1}^{0} \approx f_i^{0} + \sum_{n=1}^{k} \frac{f_i^{n} \cdot h^n}{n!}, \qquad (10)$$

где h — коэффициент шага функции для экстраполяции по равномерной сетке. Значение параметра h прямо пропорционально скорости сходимости и обратно пропорционально точности полученного решения. Наилучшие показатели работы алгоритма достигаются в диапазоне $0.2 < h < 1$ (раздел 4).

Кроме полиномиальных методов экстраполяции существует группа методов, основанная на применении сплайнов (spline). В основе сплайн-интерполяции лежит следующий принцип. Интервал интерполяции разбивается на небольшие отрезки, на каждом из которых функция задается полиномом третьей степени. Коэффициенты полинома подбираются таким образом, чтобы выполнялись определенные условия (какие именно, зависит от способа интерполяции). Общие для всех типов сплайнов третьего порядка требования — непрерывность функции и прохождение через указанные точки. Дополнительными требованиями могут быть линейность функции между узлами, непрерывность высших производных и другие. Наиболее распространены следующие типы сплайнов [20].

Линейный сплайн — это сплайн, составленный из полиномов первой степени, т.е. из отрезков прямых линий. Точность интерполяции

линейными сплайнами невысока, кроме того, они не обеспечивают непрерывности даже первых производных. Однако в некоторых случаях кусочно-линейная аппроксимация функции может оказаться предпочтительнее, чем аппроксимация более высокого порядка. Например, линейный сплайн сохраняет монотонность переданного в него набора точек.

Кубический сплайн. Кубическим сплайном называют такой сплайн, который получается, если потребовать непрерывности первой и второй производных. Кубический сплайн задается значениями функции в узлах и значениями производных на границе отрезка интерполяции (либо первых, либо вторых производных).

Интерполяция сплайнами третьего порядка — это быстрый и устойчивый способ интерполяции функций, сплайн-интерполяция является одной из альтернатив полиномиальной интерполяции. Системы линейных уравнений, которые требуется решать для построения сплайнов, очень хорошо обусловлены, что позволяет получать коэффициенты полиномов с высокой точностью. В результате даже при очень большом количестве точек вычислительная схема не теряет устойчивость [21]. Увеличение вычислительной эффективности с помощью экстраполяции сплайнами реализовано алгоритмами, взятыми из открытой онлайн библиотеки алгоритмов «Alglib» [21].

Адаптивный метод. Еще одним приемом, позволяющим увеличить быстродействие итерационных методов, является адаптивный метод [22], показывающий, что можно не пересчитывать ранги вершин, для которых итерационный процесс уже сошелся. Кроме того, существует некое количество вершин, множество исходящих дуг которых является пустым, такие вершины не оказывают влияния на значения рангов других вершин, следовательно, расчет их рангов можно отложить до конца итерационного процесса. А затем, рассчитать значения их рангов одним проходом.

4. Тестирование и сравнение

Тестирование проводилось на стандартном x86 сервере с двумя процессорами Intel Xeon 3.0, 4ГБ оперативной памяти, операционной системой Windows Server 2008. На сервере хранится обратный ссылочный индекс, сбор и хранение которого описаны в [11]. Приложение для тестирования описанных методов (далее LBRTest) написано на языке C#, базируется на программной платформе Microsoft.NET (для архитектуры x64). Все алгоритмы реализованы с использованием параллельных вычислений. В качестве тестовой задачи рассчитывается ранг для одной тематики. Алгоритм работы тестового приложения LBRTest.

- Загрузка обратного ссылочного индекса в память и расчет тематических весов ссылок.
- Расчет рангов алгоритмами, построенными с использованием описанных методов, замеры времени работы.
- Сравнение результатов (результаты приведены ниже в таблице).

Ранг узла считался сошедшимся тогда, когда $X_i - X_{i-1} \le \varepsilon$. Расчеты проводились для $\varepsilon = 0$. Результаты работы метода Якоби можно считать эталоном для сравнения точности вычислений других методов. Метод Гаусса-Зейделя сходится почти на 40% быстрее, погрешность вычислений мала (менее 0.01%) и зависит от параллельной реализации, количества ядер вычислительной машины и других характеристик вычислительной среды. Увеличение вычислительной эффективности на основе метода релаксации увеличивает сходимость метода Гаусса-Зейделя на 10%, при его применении дополнительных потерь точности вычислений практически не происходит. Увеличение вычислительной эффективности метода Гаусса-Зейделя с помощью простого полиномиального метода экстраполяции уменьшает время работы алгоритма в 5-6 раз, но при этом дает существенную погрешность.

Метод с применением экстраполяции усеченным рядом Тейлора (метод 5.2) показал лучшие, чем метод «4», результаты по точности вычислений, при близкой скорости сходимости (для $h = 0.4$). Самые лучшие результаты по точности вычислений из методов ускорения сходимости с применением экстраполяции показал метод 5 (отклонение около 0.0379%), сохранив при этом высокую скорость сходимости (ускорение метода Гаусса-Зейделя более чем в

Сравнение времени сходимости и точности некоторых методов, описанных в разделе 3

	Время работы алгоритма	Среднее отклонение значений от расчета методом Якоби (в процентах)
1. Метод Якоби	34785 мс	—
2. Метод Гаусса-Зейделя	20960 мс	<0.01%
3. Увеличение вычислительной эффективности на основе метода релаксации	18719 мс	<0.01%
4. Увеличение вычислительной эффективности с помощью полиномиального метода экстраполяции (9) ($\omega = 0.45$)	3663 мс	0.318%
5. Увеличение вычислительной эффективности с помощью экстраполяции рядом Тейлора (9) (h=0.2)	4354	0.0379%
5.1. Увеличение вычислительной эффективности с помощью экстраполяции рядом Тейлора (10) ($h = 0.3$)	4019	0.122%
5.2. Увеличение вычислительной эффективности с помощью экстраполяции рядом Тейлора (10) ($h = 0.4$)	3778	0.286%
5.3. Увеличение вычислительной эффективности с помощью экстраполяции рядом Тейлора (10) ($h = 0.5$)	3544	0.521%
6. Увеличение вычислительной эффективности с помощью экстраполяции линейными сплайнами ($\omega = 0.45$)	4237	0.387%
7. Увеличение вычислительной эффективности с помощью экстраполяции кубическими сплайнами ($\omega = 0.45$)	6521	0.165%

4-5 раз, для $h = 0.2$). Варианты метода 5 (5, 5.1-5.3) показывают, что с помощью изменения параметра h можно достичь положительных результатов по обоим критериям, что дает возможность применять метод совместно с другими методами увеличения скорости сходимости и точности. Методы увеличения вычислительной эффективность с помощью экстраполяции сплайнами не показали своих преимуществ, при решении данной задачи. Метод 6 имеет погрешность выше, а скорость сходимости ниже, чем методы 4 и 5-5.3. Метод 7 имеет неплохую точность, а при задании граничных условий [20], возможно, удастся полнее раскрыть его потенциал, но за счет вычислительной сложности (в сравнении с методами 4 и 5-5.3) скорость работы алгоритма оставляет желать лучшего.

При применении экстраполяции значения рангов вершин могут иметь положительную погрешность. В результате, больший, относительно базового метода, ранг получат те вершины, которые имеют большое количество входящих дуг. Погрешность может изменить некоторые позиции, однако общие тенденции ранжирования останутся без изменений. Уменьшения влияния большого количества дуг можно добиться варьированием значения коэффициента затухания, то есть можно скомпенсировать положительные изменения ранга и сделать его ближе к базовому рангу.

Заключение

Проведенные исследования (анализ возможных методов решения, выбор алгоритмов повышения производительности, их программная реализация и тестирование на сервере) показали, что наибольшую практическую ценность для применения имеет метод увеличения вычислительной эффективности с помощью экстраполяции рядом Тейлора (5), обладающий наилучшим показателем точности при сохранении высокой производительности. Погрешность метода составляет всего 0.0379 %, при почти 5 кратном сокращении времени решения задачи по сравнению с эталоном. Полученные результаты позволяют использовать этот метод при решении задач тематического ранжирования, построения различных систем отбора web-сайтов и web-страниц, а также при построении информационно-поисковых систем.

Тематический ссылочный ранг web-сайтов позволяет ранжировать выдачу поисковой системы с учетом тематики запроса. Ссылки, размещенные на страницах web-сайта, будут иметь высокий вес в том случае, если их тематика и тематика входящих на web-сайт ссылок соот-

ветствует тематике пользовательского запроса. Установка ссылок с коммерческой целью будет влиять на ранжирование только в том случае, когда тематика сайта источника соответствует тематике ссылки и тематике сайта приемника, то есть тогда, когда ссылка является полезной и для пользователя. Следовательно, при применении тематического ссылочного ранга релевантность поиска повысится. С помощью тематического ссылочного ранга можно решать и другие задачи, например, оценивать тематику web-сайта даже в том случае, когда информация, представленная на сайте, закрыта от индексирования. Это имеет место в задачах автоматической каталогизации информации, представленной в сети Интернет, а также для поиска закрытых от поисковых систем сайтов, распространяющих нелегальную информацию.

Литература

1. Aboba, Bernard. The Online User's Encyclopedia: Bulletin Boards and Beyond. Massachusetts : Addison-Wesley, November 1993. ISBN 0-201-62214-9.
2. Евдокимов Н.В. Основы контентной оптимизации. Эффективная Интернет-коммерция и продвижение сайтов в Интернет. б.м. : 000 "И.Д. Вильямс", 2007. ISBN 5-8459-1095-1.
3. Brin, S. and Page, L. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In: Seventh International World-Wide Web Conference (WWW 1998). April 14-18, 1998, 1998 г., Brisbane, Australia.
4. http://www.reg.ru. Крупнейший аккредитованный регистратор доменных имен в РФ.
5. http://www.sape.ru. Биржа покупки и продажи интернет-ссылок.
6. Page L., Brin S. «The PageRank Citation Ranking: Bringing Order to the Web». Stanford Digital Library Technologies Project, 1998 г., http://dbpubs.stanford.edu:8090/pub/1999-66.
7. Kellher D., Luz S. Automatic hypertext key-phrase detection. Proceeding of the nineteenth international joint conference on artifical intelegence. Edinburgh, Scotland, UK., 2005 г., p 1608-1610.
8. Andreas Heß, Philipp Dopichaj, Christian Maaß. Multi-Value Classification of Very Short Texts. KI 2008 Advances in Artificial Intelligence. 5243, 2008 г., Т. с 70-77.
9. Steven M. Beitzel, Eric C. Jensen, David D. Lewis, David D. Lewis Consulting, David D. Lewis Consulting. Automatic Classification of Web Queries Using Very Large Unlabeled Query Logs. ACM Transactions on Information Systems. 25, April 2007 г.
10. Ширай М.А., Григорьев О.Г. Исследование ранжирования интернет-ресурсов и методов построения обратного ссылочного индекса. Сборник трудов ИСА РАН под ред. В.Л. Арлазарова "Обработка информационных и графических ресурсов", 2010 г.
11. Ширай М.А., Григорьев О.Г. Разработка и реализация нового способа хранения обратного ссылочного индекса. Информационные технологии и вычислительные системы. 2011 г.
12. Haveliwala T.H., Kamvar S.D. The second eigenvalue of the Google matrix. Technical Report. Stanford University, 2003 г.
13. Hongbo Liu, Jiaxin Wang. A new way to enumerate cycles in graph. Telecommunications. AICT-ICIW '06. International Conference on Internet and Web Applications and Services/Advanced International Conference on (2006). 2006 г.
14. Ширай М.А. Обзор существующих методов и подходов к организации параллельных вычислений в базах данных на традиционных многопроцессорных системах. Системы компьютерной математики и их приложения: материалы X международной научной конференции. Смоленск: Изд-во СмолГУ, 2009 г., 10.
15. Баладин М.Ю., Шурина Э.П. Методы решения СЛАУ большой размерности. Новосибирск: НГТУ, 2000.
16. Berkhin, Pavel. A Survey on PageRank Computing. Internet Math. Volume 2, Number 1, 2005 г., стр. 73-120.
17. Зайцев В.В., Трещов В.М. Численные методы для физиков. Нелинейные уравнения и оптимизация: учебное пособие. Самара: Изд-во "Самарский университет", 2005. ISBN: 5-86465-240-7.
18. Haveliwala T.H., Kamvar S.D., Klein D., Manning C., Golub G.H. Computing PageRank using power extrapolation. Technical Report. Stanford University, July 2003 г.
19. Марчук Г. И., Агошков В. И. Введение в проекционно-сеточные методы. Москва: Наука, 1981.
20. Б.И., Квасов. Методы изогеометрической аппроксимации сплайнами. Москва: ФИЗМАТЛИТ, 2006.
21. http://alglib.sources.ru/interpolation/spline3.php. ALGLIB, онлайн библиотека математических алгоритмов.
22. Haveliwala T.H., Kamvar S.D., Golub G.H. Adaptive Methods for the Computation of PageRank. Technical Report. Stanford University, 2003 г.

Григорьев Олег Георгиевич. Заместитель директора ИСА РАН. Окончил Московский институт электронного машиностроения в 1980 году. Доктор технических наук. Автор свыше 35 печатных работ. E-mail: olegg@polikvart.ru

Ширай Михаил Александрович. Аспирант ИКТИ РАН. Окончил Московский государственный университет приборостроения и информатики в 2008 году. Автор 6 научных публикаций. Область научных интересов: поисковые системы сети Интернет, параллельное программирование, ссылочное ранжирование. E-mail: michael.sheerai@gmail.com