

**Тезисы доклада**

Начало формы

1. **НАЗВАНИЕ ДОКЛАДА:**

Об исследованиях веб-графа сайта c использованием имитационной модели

On the research of site web-graph using the simulation model

1. **АВТОРЫ:**

Д. И. Чернобровкин

D. I. Chernobrovkin

1. **ОРГАНИЗАЦИЯ (полное наименование, без аббревиатур):**

Факультет прикладной математики и процессов управления, Санкт-Петербургский государственный университет

Faculty of Applied Mathematics and Control Process, Saint Petersburg State University

1. ГОРОД:

Санкт-Петербург

Saint Petersburg

1. ТЕЛЕФОН: +79112426802
2. ФАКС:
3. E-mail:denis\_univer@mail.ru
4. **ТЕКСТ ТЕЗИСОВ ДОКЛАДА:**

С развитием Интернета веб-мастерам приходится все больше внимания уделять не только качественному наполнению и дизайну, но и структурной организации самого сайта. Если проводить параллели с домом, где дом – это веб-сайт, комнаты – веб-страницы, а коридоры – это ссылочная структура сайта,то можно было бы сказать, что сейчас важно не только, чтобы комнаты в доме были красиво и комфортно обустроены, но чтобы и коридоры между ними были расположены удобно и интуитивно понятно. И зачастую довольно тяжело предсказать, как ведет себя пользователь внутри сложной структуры веб-сайта. Структуре сегментов Веба посвящено большое количество работ [1, 2], но относительно немного работ [3], которые бы изучали структуру веб-графа сайтов. В данном докладе мы рассмотрим имитационную модель поведения пользователя при навигации по веб-графу сайта.

Прежде чем переходить к глубинному рассмотрению вопроса структуры веб-графа, стоит разобраться, чем же отличается веб-граф сайта от веб-графа сегмента. Основные различия заключатся в следующем:

1. Точка входа. В отличие от веб-графа сегмента Веба, в веб-графе сайта есть точка входа – это его стартовая (главная) страница. Чаще всего пользователь начинает работу с сайтом именно с неё.
2. Уровневая структура. Так как есть точка входа, то соответственно можно выделить и уровневую систему: чем больше переходов от точки входа до выбранной страницы, тем ниже её уровень. У стартовой страницы уровень 0, у всех страниц, на которые сделана ссылка с главной страницы уровень 1 и т.д.
3. Веб-граф сайта всегда связный. Веб-граф всегда будет связным и почти всегда сильно связным. Связным веб-граф будет по следующей причине: если страница не достижима с главной страницы, то и причислить её к данному сайту довольно тяжело, т.к. пользователь на неё никогда не попадет. Сильно связным веб-граф будет в силу того факта, что чаще всего сайты на каждой своей странице имеют ссылку на начальную страницу, а если такого нет, то можно ввести искусственное ребро – кнопка «назад» в браузере.

Для изучения структуры веб-графа и модели поведения пользователя были созданы модули, которые расширили и дополнили BeeBot – уже существующий комплекс программ-инструментов для изучения Веба [4]. Модуль BeeDrone предназначен для сбора информации о структуре веб-графа сайта. С помощью специально разработанного модуля AntHill реализуется имитационная модель поведения пользователя. Рассмотрим её в подробнее.

Рассматривая поведение пользователя на сайте, можно говорить о случайности данного процесса (при условии, что мы отбрасываем визуальную и семантическую составляющую расположения ссылок). Процесс навигации пользователя можно описать в виде марковского процесса.

Пусть $W=\{w\_{1},w\_{2},w\_{3}…w\_{n}\}$ – множество страниц исследуемого сайта, $w\_{1}$ – стартовая страница. Вектор начальных состояний $R=\{r\_{1},r\_{2},r\_{3}…r\_{n}\}$, определяет вероятность выбора пользователем страницы $w\_{i}$ из множества $W$ при начале просмотра сайта.

Набор ссылок между страницами сайта обозначим через $L=\{L(i,j)\}^{n,n}$, где *L(i,j)* – количество ссылок на странице *i*, указывающих на страницу *j*. Если ссылки нет, то *L(i,j)*=0, если ссылка одна, то *L(i,j)*=1, две – *L(i,j)* =2 и т.д.

У каждой ссылки *L(i,j)* есть вероятность $P(i,j,a)$, что пользователь по ней перейдет. Сумма $\sum\_{1}^{m}P\_{n\_{m}}+d+b=1$ для каждого *n*. Здесь *d* – своеобразный дампинг фактор для сайта (вероятность, что пользователь покинет сайт), *b* – фактор рестарта сессии. Дампинг фактор на каждом шаге будет изменяться (соответствующим образом будут пересчитываться вероятности), фактор рестарта сессии будет постоянным.

В нашем случае будем считать, что $P\_{n\_{m}}$зависит исключительно от количества ссылок на странице $w\_{n}$. Вероятность перехода с *i*-й страницы на *j*-ю:

$$P\left(i,j\right)=\frac{L\left(i,j\right)(1-d-b)}{\sum\_{k=1}^{n}L(i,k)}$$

Имея матрицу переходов, можно проводить эксперименты с переходами пользователя по ссылкам. На первом шаге выбирается произвольная страница с вероятностью, взятой из вектора начальных состояний. На каждом последующем шаге будет выбираться случайная страница из списка страниц, на которые есть ссылка с текущей.

Приведем пример апробации данной модели на сайте Института прикладных математических исследований Карельского Научного Центра РАН (*http://mathem.krc.karelia.ru*). Результаты, полученные для веб-графа, построенного на 1500 страницах сайта с 0-го по 5-й уровень получились довольно интересными:

1. Главная страница получает около 35% процентов всех посещений.
2. Среди полутора тысяч наибольшее число посещений получают страницы, которые находятся в «меню» сайта, т.е. в тех панелях, которые видит пользователь при переходе на любую страницу. Почти все пункты меню получают примерно по 2% посещений каждый.
3. Средняя длина сессии пользователя – 4 перехода после главной страницы.

Довольно ожидаемые результаты на деле говорят о том, что, в рамках сформированной веб-мастером структуры сайта, на главную страницу и страницы, соответствующие пунктам меню получают обращается наибольшее внимание пользователя. Один из интересных выводов: если вы хотите выделить какую-то определенную страницу и направить внимание пользователя к ней (например рассказать о конференции которая будет проходить при поддержке вашего университета), то стоит вынести ссылку на эту страницу отдельным пунктом в меню.

Работа выполнена при поддержке Программы стратегического развития Петрозаводского государственного университета на 2012 - 2016 годы.

Литература.

1. Broder A., Kumar R., Maghoul1 F., Raghavan P., Rajagopalan S., Stata R., Tomkins A., Wiener J. Graph structure in the web // Journal of Computer Networks. 2000. №33(1-6). Р. 309-320.
2. Печников А.А. Модель университетского Веба // Вестник Нижегородского университета им. Н.И. Лобачевского. №6. 2010. C. 208-214.
3. Горбунов А. Л. Марковские модели посещаемости веб-сайтов / А. Л. Горбунов // Интернет-математика 2007: сб. работ участников конкурса науч. проектов по информ. поиску. – Екатеринбург : Изд-во Урал.ун-та, 2007. – С. 65–73.
4. Печников А. А., Чернобровкин Д. И. Адаптивный краулер для поиска и сбора внешних гиперссылок // Управление большими системами. Выпуск 36. – М.: ИПУ РАН, 2012. С.301-315.