УДК 004

*З.В. ДУДАРЬ, М.В. ЗБИТНЕВА*

# МАРКОВСКИЕ МОДЕЛИ ДЛЯ ОЦЕНИВАНИЯ РЕЙТИНГА ВЕБ-САЙТА

Предлагается структура веб-агента, основанного на модели и цели. В качестве модели выбрана марковская модель, отражающая частоту и длительность посещаемости веб-сайта. Модели строятся для обобщенной и наиболее типичной структуры веб-сайта. В качестве цели выбрано повышение рейтинга веб-сайта. Веб-агент решает задачи анализа качества страниц веб-сайта, а именно выделяет наиболее приемлемые страницы для размещения рекламы, а также страницы, обладающие малой степенью полезности.

## 1. Введение

Качество контента определяется его актуальностью, отсутствием большого количества рекламы, стилем. Если в текстах присутствуют устаревшие сведения или информация о ваших товарах или услугах не соответствует действительности - это также говорит о том, что качество контента оставляет желать лучшего. О его качестве свидетельствует и то, насколько учтены особенности написания веб-текста.

Рейтинг сайта - комплексное понятие, привязанное к определённому отрезку времени. Под рейтингом сайта в поисковой системе понимается позиция, занимаемая сайтом по результатам запросов поисковой системы по конкретному ключевому слову или ключевой фразе. Под рейтингом сайта в Интернет-ресурсе, главная или одна из основных задач которого – ранжирование сайтов по определённым критериям, понимается рейтинг сайта по числу его посетителей в день.

Критерии, по которым производится ранжирование сайтов (расстановка, упорядочение, оценка), могут быть следующие:

– по числу посетителей сайта на конкретный час дня;

– по числу просмотренных веб-страниц сайта на конкретный час дня;

– по числу ссылок на сайт с других сайтов (“индекс цитирования”, “индекс популярности”);

– расстановка сайтов по алфавиту;

– расстановка сайтов по дате добавления в каталог;

– по оценке модераторов (экспертов) того каталога, в котором данный сайт зарегистрирован.

*Целью* данной работы является исследование первого критерия - числа посетителей сайта на конкретный час дня. Поэтому к задачам исследования относятся:

– построение полумарковской модели обобщенной структуры сайта;

– построение полумарковской модели наиболее типичной структуры сайта.

Так как закон перехода на страницы не является одним и тем же, следовательно, все модели считаются полумарковскими.

**2. Суть исследования**

В качестве автоматического средства выберем веб-агента. Наиболее употребимы несколько типов интеллектуальных агентов [4]: простые рефлексные агенты; рефлексные агенты, основанные на модели; агенты, основанные на цели; агенты, основанные на полезности.

Простейшим видом агента является простой рефлексный агент. Подобные агенты выбирают действия на основе текущего акта восприятия, игнорируя всю остальную историю актов восприятия. Данный тип работает с полностью наблюдаемой средой.

Наиболее эффективный способ организации работы в условиях частичной наблюдаемости - это применение агентов, основанных на модели. Такой агент поддерживает внутреннее состояние, отражающее акты восприятия. Для обновления информации необходимы сведения о том, как мир изменяется независимо от агента и как влияют на мир собственные действия агента. В случае необходимости при выборе решения учета цели добавляется учет цели в виде следующего типа агента, который определяет, достигнуто решение или нет. Это добавляет агенту гибкости, так как знания, на которые он опирается, представлены явно и могут быть модифицированы. Добавляя полезность, которая обозначает соответствующую степень m-ное значение «удовлетворенности», получают агентов, основанных на модели и на полезности. Полная спецификация функций полезности дает возможность выбирать решение при наличии конфликтующих целей или нескольких целей, каждая из которых не может быть достигнута со всей определенностью. В этом случае осуществляется способ взвешенной оценки вероятности успеха.

В качестве структуры агента для данной работы выбран агент, основанный на модели и цели (рис.1). В качестве модели выступают марковские модели. В качестве цели – реструктуризация структуры веб-сайта.

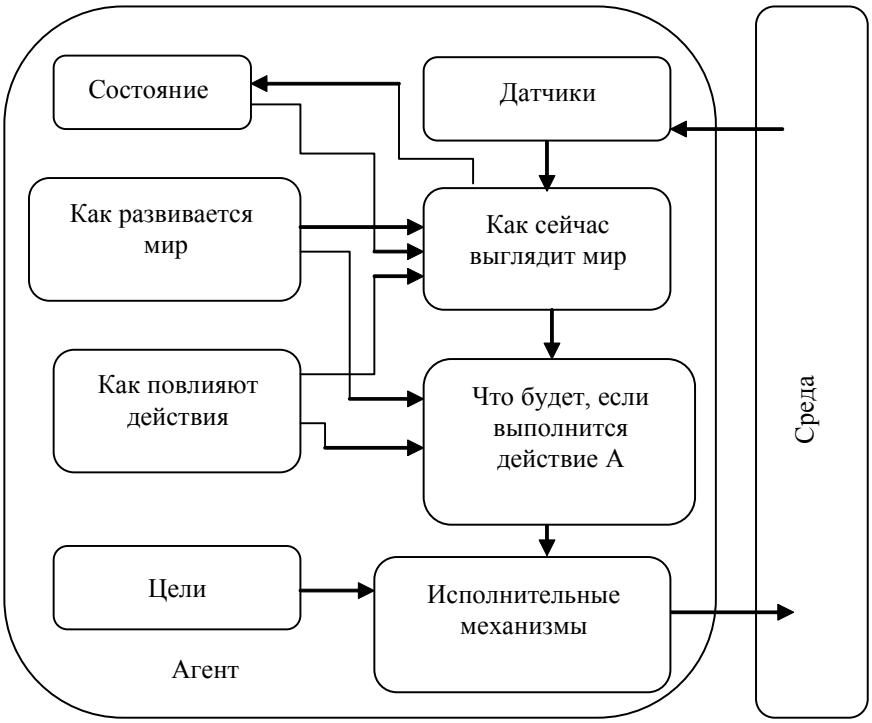


Рис. 1. Структуры агентов, основанные на модели и цели

Имеется класс математических моделей, аппроксимирующих широкий спектр случайных процессов. Такой класс составляют марковские процессы. Основные виды классифицируются в соответствии с задаваемыми значениями на временных и числовых множествах. Марковские цепи – это вид марковских процессов, которые описывают дискретный процесс с дискретным временем [5].

Рассмотрим посещение пользователей сайта как марковский процесс. Опишем его в виде обобщенной полумарковской модели.

Для структуры сайта, в котором из каждой страницы можно попасть на каждую, строится обобщенная полумарковская модель, состояния которой являются сообщающимися, а граф переходов – компонентой сильной связности. Граф состояний представлен на рис. 2. $S_0$ – первая или главная страница сайта, которая имеет 0-й уровень. $S_{11}$ … $S_{1n}$ – страницы, доступные для перехода с главной, описывают 1 уровень. $S_{111}$ … $S_{11n}$ – страницы, доступные для перехода с $S_{11}$, описывают уровень 2 т.д. Количество уровней вложенности и количество страниц разного уровня равно n. Таким образом, множество состояний образуют $\{S_0, S_{11}, S_{111} \dots S_{11n}, \dots S_{1n}, \dots S_{nn}\}$. Вектор начальных вероятностей (1, 0, …, 0). Рейтинг сайта, как и модели для него, как правило, оценивается за период времени – за месяц.

Рассмотрим первый показатель, применяемый для оценки рейтинга – число просмотренных веб-страниц. В качестве его первой составляющей являются переходы между страницами. Пусть количество страниц на одном уровне равно m, количество уровней - n.

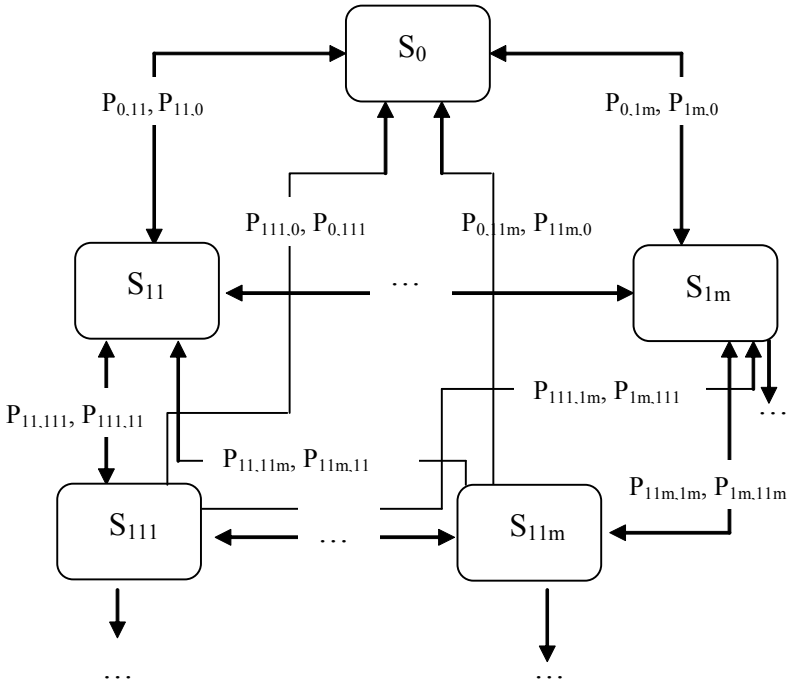


Рис. 2. Граф обобщенной полумарковской цепи веб-сайта

Матрица вероятностей переходов для данной модели имеет следующий вид:

$$P = [p_{ij}] = \begin{array}{c|cccccccc} & S_0 & S_1 & S_{11} & \cdots & S_{1m} & \cdots & S_n \\ S_0 & p_{0,0} & p_{0,1} & p_{0,11} & \cdots & p_{0,1m} & \cdots & p_{0,n} \\ S_1 & p_{1,0} & p_{1,1} & p_{1,11} & \cdots & p_{1,1m} & \cdots & p_{1,n} \\ S_{11} & p_{11,0} & p_{11,1} & p_{11,11} & \cdots & p_{11,1m} & \cdots & p_{11,n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ S_{1m} & p_{1m,0} & p_{1m,1} & p_{1m,11} & \cdots & p_{1m,1m} & \cdots & p_{1m,n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ S_n & p_{n,0} & p_{n,1} & p_{n,11} & \cdots & p_{n,1m} & \cdots & p_{n,n} \end{array},$$

где $P_{i,j}$ – вероятность перехода из состояния i в состояние j или вероятность перехода со страницы i на страницу j:

$$P_{ij} = \frac{k_j}{N},$$

здесь $k_j$ – количество посещений j страницы; N – количество посещений всех страниц сайта за определенный промежуток времени.

Второй составляющей ячейки матрицы является длительность посещения страницы пользователем за определенный промежуток времени. Вероятность для нее рассчитывается по следующей формуле:

$$Q = \frac{d_j}{N_d},$$

где $d_j$ – длительность посещения j страницы; $N_d$ – суммарная длительность посещения всех страниц.

Сумма вероятностей дуг, исходящих из одной вершины, равняется единице, т.е. для каждой строки матрицы сумма вероятностей равна 1:

$$\sum_{S_{ij}} P_{ij} = 1.$$

Построим полумарковскую модель для сайта с наиболее типичной многоуровневой структурой (рис. 3). С каждой страницы есть возможность вернуться на главную. Движение между страницами заключается в движении внутри одного уровня и на уровень выше/ниже.

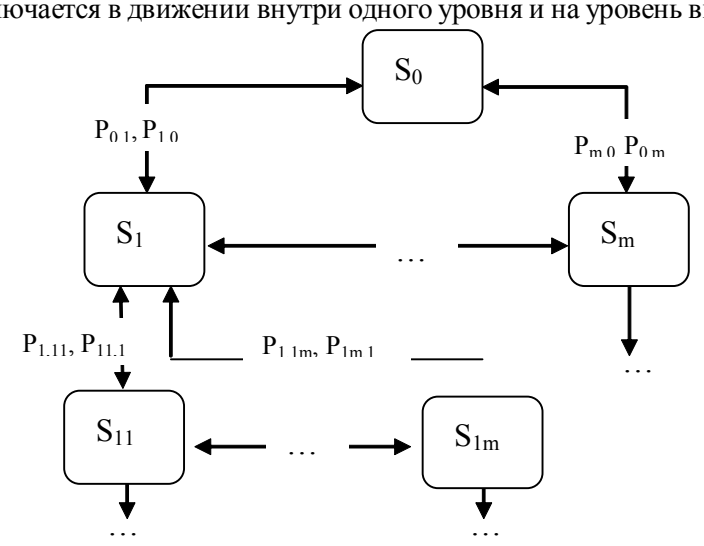


Рис. 3. Граф типичной полумарковской модели

Матрица вероятностей переходов для данной модели имеет следующий вид:

$$P = [p_{ij}] = \begin{array}{c|ccccccccc} & S_0 & S_1 & S_2 & S_{11} & \dots & S_{1n} & \dots & S_n \\ S_0 & p_{0,0} & p_{0,1} & p_{0,2} & p_{0,11} & \dots & p_{0,1n} & \dots & 0 \\ S_1 & p_{1,0} & p_{1,1} & p_{1,2} & p_{1,11} & \dots & p_{1,1n} & \dots & 0 \\ S_2 & p_{2,0} & p_{2,1} & p_{2,2} & p_{2,11} & \dots & p_{2,1n} & \dots & 0 \\ S_{11} & p_{11,0} & p_{11,1} & 0 & p_{11,11} & \dots & p_{11,1n} & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ S_{1n} & p_{1n,0} & p_{1n,1} & 0 & p_{1n,11} & \dots & p_{1n,1n} & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ S_n & p_{n,0} & 0 & 0 & 0 & \dots & 0 & \dots & p_{n,n} \end{array}$$

**Выводы**

К результатам исследования относятся построенные полумарковские модели для оценки рейтинга веб-сайта при использовании в качестве составляющих показателя числа просмотренных веб-страниц. Предложена программно-компонентная структура в виде веб-агента для решения поставленной цели. Научная новизна и практическая значимость выражается в разработанных моделях, а также во введенных составляющих критерия оценки рейтинга веб-сайта по числу посетителей.

Можно просчитывать горизонтальный и вертикальный показатель вероятности. Горизонтальный – по страницам в рамках одного уровня. Вертикальный – по заданным индексам страниц разных уровней, например, для просмотра всех страниц по определенной теме.

Для оценки рейтинга сайта предлагается вычислить экспериментальным путем значения следующих показателей: $P_{\delta min}$ – значение вероятности, полученное экспериментальным путем, ниже значения которого страница считается малопосещаемой, обладающей малой степенью полезности, качества. Действия, которые необходимо предпринять – это реформировать контент или объединить с другой страницей, или удалить, или поднять/опустить страницу на уровень выше/ниже; $P_{\delta max}$ – показатель вероятности, выше значения которого страница считается достаточно посещаемой для размещения на ней рекламы, для поддержания рейтинга сайта.

На данный момент проведены исследования применения марковских процессов посещения сайтов и поведения посетителей в работах Дишпэйнда и Кариписа, Андерсона и Домингоса, Янсена и Сараевича. Дишпэйнд и Карипис исследуют проблему предсказания поведения посетителя веб-сайта с помощью марковских моделей [1]. Андерсон, Домингос и Велд [2] описывают посетительское поведение посредством «относительных» марковских моделей. Янсен, Спинк, Сараевич конструируют модель поведения системы пользователь – сайт. Такой подход не привязан к специфике сайта [3].

В качестве перспектив развития данного исследования можно отметить выявление законов формирования вероятностей для построенных моделей, уточнение типа моделей, а также моделирование сочетаний слов в тексте при учете качества веб-контента.

**Литература**: **1.** *Deshpande M., Karypis G.* Selective Markov models for predicting Web page accesses. ACM Transaction on Internet Technology (TOIT). May 2004, Volume 4, Issue 2, Pages 163-184. ACM Press, NY, USA. **2.** *Jansen B., Sprink A., Saraevic T.* Real Life, Real users and Real needs: a Study and Analysis of User Queries on the Web. Information Processing and Management. 36: 2000 Elsevier. P. 207-227. **3.** *Anderson C., Domingos P., Weld D.* Relational Markov models and their application to adaptive web navigation // Proc. of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining table of contents, 2002. P. 143-152. ACM Press, NY, USA. **4.** *Рассел Стюарт, Норвиг Питер.* Искусственный интеллект: современный подход, 2-е изд.: Пер.с англ. М.: Издательский дом «Вильямс», 2006. 1408 с. **5.** *Андронов А.М., Копытов Е.А., Гринглаз Л.Я.* Теория вероятностей и математическая статистика: Учебник для вузов. СПб.: Питер, 2004. 461 с.

*Поступила в редколлегию 17.06.2009*

**Дударь Зоя Владимировна,** канд. техн. наук, профессор кафедры ПОЭВМ ХНУРЭ. Научные интересы: математическое и программно-техническое обеспечение взаимодействия крупномасштабных систем баз данных в динамическом окружении. Адрес: Украина, 61166, Харьков, пр. Ленина, 14, e-mail: software@kture.kharkov.ua.

**Збитнева Майя Вячеславовна,** канд. техн. наук, доцент кафедры ПОЭВМ ХНУРЭ. Научные интересы: интеллектуальные агенты. Адрес: Украина, 61166, Харьков, пр. Ленина, 14, e-mail: mayazbt@yandex.ru.