УДК 004.67

В.А. Силич, А.О. Савельев, А.В. Марчуков, А.А. Алексеев

Методы оценки актуальности научных публикаций, на основе анализа интернет-обращений к научным порталам

Рассматриваются недостатки методов оценки научных публикаций, основанных на индексе цитируемости. Предлагается уточняющий метод оценки научных публикаций на основе мониторинга обращений к научным веб-порталам и серверам, основанный на наивном байесовском классификаторе. Выделены независимые признаки, характеризующие поведение веб-пользователя.

Ключевые слова: индекс цитируемости, веб-аналитика, веб-журнал, задача классификации, наивный байесовский классификатор.

Введение

Одним из основных показателей деятельности ученого является рейтинг цитируемости его статей в научной периодике, т.е. ссылки на его публикации в печатном виде. Но данный метод требует больших затрат, не оперативен и обладает рядом недостатков.

К числу недостатков методов оценки, основанных на научном цитировании, относятся [1]:

- зависимость от конъюнктуры – популярные работы цитируются лучше, легче, чем пионерские;
- индекс цитируемости зависит не только от научного уровня, но и от PR-активности ученого (конференции, контакты);
- индекс цитируемости имеет разную цену для разных областей науки;
- критика ошибочных работ приносит немало цитирований;
- проблема соавторов;
- самоцитирования – ими можно набрать немалый индекс – сотни. Отсеять их технически тяжело;
- своевременность – расчет индексов цитирования напрямую зависит от состояния и полноты данных в базах цитируемости. Организация подобного рода баз данных и актуальность содержащейся в них информации целиком зависят от организаций, сопровождающих их.

Для преодоления существующих проблем постоянно разрабатываются новые и совершенствуются существующие методики расчета индексов цитируемости и основанных на них показателей.

Существует также иной подход к оценке научных публикаций, основанный на анализе данных о посещаемости научных веб-сайтов и порталов. Следует отметить, что за рубежом, в частности в США, интернет-публикация получает поддержку государства, фонды, финансирующие исследования, требуют обязательного размещения электронных копий результатов работ в открытом доступе в глобальной сети [2].

Основываясь на всевозрастающей роли Интернета в обеспечении доступности научных статей и разработок, а также на существующих инструментальных средствах, более чем логично выглядит организация систем и комплексов по анализу данных о посещаемости веб-ресурсов.

Преимущество веб-анализа популярности научных статей имеет ряд неоспоримых преимуществ [3]:

- записи о посещаемости веб-серверов ведутся сразу же после публикации статьи на веб-ресурсе;
- число записей о посещаемости веб-ресурсов значительно превышает объем данных о цитируемости;
- учет активности большего числа людей: авторов, практикующих ученых, заинтересованной публики;
- охват практически всех направлений в науке;
- выявление связей между научными направлениями;
- построение карт науки;
- оценка перспективности новых научных и технических направлений, методов и технологий в любых сферах науки и т.д.

Крайне важны математические методы и алгоритмы обработки массивов записей о посещениях. Это обусловлено многими факторами: повышением точности результатов обработки, определением социального состава посетителей, интереса, географии и профессиональных интересов и демографических характеристик.

Для наиболее эффективного анализа наряду с математическими методами, как правило, это методы обработки и анализа DataMining, необходима специальная организация веб-ресурса. К примеру, самым простым способом является разбиение одной статьи на несколько веб-страниц, таким образом, данные о последовательном обращении к страницам, являющимся частью одной статьи, иллюстрируют однозначный интерес пользователя к опубликованному материалу.

В данной статье мы предлагаем алгоритм организации систем анализа актуальности научных публикаций, основанный на методе сегментации посетителей научных веб-ресурсов по степени интереса к опубликованным материалам.

1. Исходные данные

В качестве исходных данных для анализа были выбраны веб-журналы научных серверов [4].

Веб-журнал содержит следующую информацию:

- дата и время выполнения транзакции;
- время, затраченное на выполнение транзакции;
- количество байт, переданное клиентом серверу;
- количество байт, переданное сервером клиенту;
- IP-адрес клиента;
- URI-запрос;
- адрес предыдущей посещенной клиентом веб-страницы.

2. Формулировка задачи

Имеется n классов посетителей научных веб-ресурсов. Каждый посетитель ресурса может быть представлен как исследуемый объект I_j, который характеризуется набором переменных:

$$I_j = \{X_1, X_2, X_3, \ldots X_m\}. \quad (1)$$

Каждая из переменных X_k может принимать значения из множества

$$C = \{C_1, C_2 \ldots\}. \quad (2)$$

Таким образом, задача может быть представлена в виде задачи классификации [5].

В данной работе рассмотрим метод решения данной задачи, основанный на наивном байесовском классификаторе.

3. Независимые переменные

1. Количество обращений к веб-страницам одной тематики за исследуемый промежуток времени, в дальнейшем SN.

Переменная SN может быть представлена числом повторных обращений к определенной тематике. При этом возникает вопрос – что можно считать повторным обращением?

К примеру, пользователь перешел на страницу Page1 в 14:00, 14:05, 14:20, 15:10, 15:17 в течение одного дня наблюдений. Что следует считать повторным обращением к странице?

На это будет влиять ряд условий.

а) Допустим, страница Page1 содержит ссылки на страницы Page2, Page3, Page4 и маршрут перемещения пользователя выглядит согласно табл. 1.

В таком случае, исходя из малого времени, проведенного пользователем на странице Page1, переходов со страницы Page1 на другие страницы и регулярных возвратах на Page1 может быть сделан вывод о том, что пользователя на странице Page1 интересовала только навигация. Очевидно, что считать каждое обращение к Page1 основанием для приращения значения SN не следует.

б) Маршрут перемещения пользователя выглядит согласно табл. 2.

В данном случае, пользователь снова и снова возвращался к странице Page1, что может трактоваться как проявление интереса к содержимому страницы.

Отметим также, что, как правило, маршруты пользователей выглядят еще более запутанно.

В контексте вышеназванного для наиболее точного определения значения переменной SN предлагаем следующий поход.

Введем следующие переменные:

T – время между переходом пользователя на страницу и уходом с нее; $\max T$ – максимальное время, необходимое для изучения содержимого веб-страницы; $\min T$ – минимальное время, необходимое для определения содержимого веб-страницы.

Таблица 1

Модель №1а перемещения посетителя

Время	Веб-страница
14:00	**Page1**
14:02	Page2
14:05	**Page1**
14:07	Page3
14:20	**Page1**
14:23	Page4
15:10	**Page1**
15:14	Page3
15:17	**Page1**

Таблица 2

Модель №1б перемещения посетителя

Время	Веб-страница
14:00	**Page1**
14:02	Page2
14:05	**Page1**
14:20	**Page1**
14:23	Page4
15:10	**Page1**
15:17	**Page1**
15:30	Page3
15:34	Page5
15:38	**Page1**
15:42	Page2

Таким образом, если выполняется условие

$$\min T < T < \max T, \quad (3)$$

то можно говорить о повторном обращении к странице Page1.

2. Длина «переходов» в рамках тематики, в дальнейшем TL.

Предполагается, что последовательное обращение и переходы «статья – статья» в рамках одного направления исследований (допустим, «Беспроводные каналы связи») является дополнительным аргументом в пользу заинтересованности пользователя тематикой в целом. При этом учитываются только уникальные обращения к страницам тематики.

В качестве примера, допустим, что веб-страницы Page1, Page2, Page3, Page4 и Page5 можно отнести к тематике «Беспроводные средства связи», а страницы Page6, Page7, Page8 – к тематике «Архитектура ПК», итак, следующий маршрут перемещения пользователя – **Page1** > **Page3** > **Page5** > Page6 > Page7 > **Page1** > **Page4** > > **Page3** > Page7 > Page8> **Page2** > Page6 можно трактовать следующим образом.

Для тематики «Беспроводные сети» за исследуемый промежуток времени значение $TL = 5$.

3. Среднее время просмотра страницы, в дальнейшем AT.

Допустим, что пользователь просмотрел i страниц за общее время T. В таком случае среднее время просмотра страницы

$$AT = \frac{T}{i}. \quad (4)$$

4. Коэффициент активности пользователя, в дальнейшем CA.

Применение данного коэффициента основано на предположении, что пользователь, если тематика статьи его действительно заинтересовала, для получения полной информации готов выполнить несложное действие.

Иными словами, на веб-странице размещаются Web beacons (специальные элементы, фиксирующие данные), которые задействуются в случае определенных пользовательских действий и заносят соответствующую запись в веб-журнал.

Для расчета CA введем следующие переменные: $cAct$ – число Web beacons, задействованных пользователем; $pAct$ – общее число Web beacons на страницах исследуемой тематики, просмотренных пользователем.

К примеру, допустим, в статье имеется иллюстрация. Предлагается не размещать полноразмерную картинку на странице, а выводить ее в случае нажатия пользователем уменьшенного ее изображения на странице. Также можно разбить одну статью на несколько веб-страниц. Тогда переход пользователя от одной части статьи к другой также будет основанием для приращения переменной $cAct$.

Следует также отметить, что использование подобного подхода применимо в случае анализа интереса пользователя к публикациям. В случае анализа иного рода веб-ресурса, к примеру интернет-магазина, использование Web beacon не желательно, т.к. в этом случае пользователь, как правило, предпочтет другой интернет-магазин с более «легким» интерфейсом, при прочих равных условиях.

Таким образом, наблюдение Y, характеризующее посетителя, будет выглядеть так

$$Y = \{SN, TL, AT, CA, C_h\}, \quad (5)$$

где C_h – переменная класса посетителя.

Таблица 3

Пример категоризации переменной SN

Значение SN	Категория SN
от 0 до 3	Эпизодическое
от 4 до 7	Частое
от 8 и более	Регулярное

Несмотря на то, что переменные, характеризующие посетителя, являются числовыми, нетрудно преобразовать их в категориальный вид, (табл. 3.) на примере переменной *SN*.

Аналогичным образом можно придать переменным *CL*, *AT*, *CA* категориальный вид, на основе экспертной оценки.

4. Обоснование независимости переменных

Для обоснования возможности применения наивного байесовского классификатора в качестве возможного метода решения задачи классификации пользователей по степени интереса к веб-материалам рассмотрим зависимость вышеобозначенных переменных друг от друга.

Для этого проанализируем переменные более подробно и определим, от каких именно показателей и условий будут зависеть их значения.

Переменная *SN* – данная переменная указывает лишь количество страниц, за рассматриваемый промежуток времени, которым пользователь уделил внимание. Исходя из условий приращения переменной (3) очевидно, что она зависит от величин $\min T$ и $\max T$, определяемых экспертным образом, и поведения пользователя. Учитываются повторные обращения пользователя к одной и той же странице.

Переменная *TL* – указывает количество уникальных статей, просмотренных пользователем за исследуемый промежуток времени в рамках определенной тематики. Зависит только от поведения пользователя и ссылочной структуры сайта. Соотнесение маршрута переходов пользователя с ссылочной структурой сайта, при учете времени переходов от страницы к странице, способно скорректировать значение переменной для более точного анализа.

Переменная *AT* – среднее значение времени, уделенного пользователем на просмотр одной страницы, следовательно, не зависит напрямую от переменных *SN* и *TL*.

Переменная *CA* – отношение задействованных пользователем Web beacon к их общему числу на просмотренных пользователем страницах.

Из вышеизложенного следует, что у каждой из этих переменных так или иначе наблюдается зависимость от значения просмотренных пользователем страниц, но напрямую переменные *SN*, *TL*, *AT*, *CA* не зависят от значений друг друга.

5. Байесовский классификатор

Таким образом, были соблюдены условия, позволяющие использовать наивный байесовский классификатор в качестве средства решения задачи сегментации посетителей веб-ресурсов и определить переменные, характеризующие веб-посетителя.

В результате мы имеем наблюдение Y, характеризующее веб-посетителя (5), и условная вероятность принадлежности посетителя к классу Ch будет рассчитываться по формуле [6]:

$$P(C_h \mid Y) = \frac{P(SN \mid C_h) \cdot P(CL \mid C_h) \cdot P(AT \mid C_h) \cdot P(CA \mid C_h) \cdot P(C_h)}{P(Y)}. \qquad (6)$$

Элементы системы анализа посещений были реализованы на базе веб-сервера «Orion». Сервер представлял промышленность Томской области в интернет-пространстве, период наблюдения – 1998–2005 гг. За годы наблюдения удалось выявить ряд особенностей продукции томских предприятий на внутреннем и внешнем рынках, сезонность спроса, географию спроса, ряд закономерностей поведения веб-заказчиков от прочтения информации до заключения контрактов. Выявление хай-тек продукции, пользующейся спросом на мировом рынке, дало совершенно неожиданные результаты: по многолетним наблюдениям первое место занимали исследования и продукция НИИПП в области арсенид-галлиевых электронных приборов. Установлена причина повышенного интереса зарубежных исследователей и компаний к данной продукции.

Заключение

Предложенный метод может служить уточняющим инструментом при оценке качества научных публикаций методами, основанными на индексе цитируемости. Применение нового метода в сочетании с уже известными позволит повысить обоснованность оценки научных публикаций, ослабить зависимость методов оценки от PR-активности ученого, получать первичные результаты оценки сразу после публикации статьи, вне зависимости от полноты баз цитируемости, устранить проблему «самоцитирования».

Литература

1. Недостатки индекса цитируемости [Электронный ресурс]. – Режим доступа: http://www.scientific.ru/whoiswho/roundtab/disadv.html, свободный (дата обращения 17.09.2010).

2. Интернет-активность как обязанность ученого [Электронный ресурс]. – Режим доступа: http://www.keldysh.ru/gorbunov/duty.htm, свободный (дата обращения 20.09.2010).

3. Clickstream Data Yields High-Resolution Maps of Science [Электронный ресурс]. – Режим доступа: http://www.plosone.org/article/info:doi/10.1371/journal.pone.0004803, свободный (дата обращения 20.09.2010).

4. Mining Significant Usage Patterns from Clickstream Data [Электронный ресурс]. – Режим доступа: http://www.citeulike.org/user/udamahan/article/2489574, свободный (дата обращения 22.09.2010).

5. Анализ данных и процессов: учеб. пособие / А.А. Барсегян, М. С. Куприянов, И.И. Холод и др. – 3-е изд., перераб. и доп. – СПб.: БХВ-Петербург, 2009. – 102 с.

6. Бизнес-аналитика: от данных к знаниям/ Н.Б. Паклин, В.И. Орешков. – 2-е изд., перераб. и доп. – СПб.: Питер, 2010. – 425 с.

Силич Виктор Алексеевич
Зав. каф. оптимизации систем управления (ОСУ) Института кибернетики ТПУ
Тел.: (382-2) 42-07-60
Эл. почта: vas@tpu.ru

Савельев Алексей Олегович
Аспирант каф. ОСУ Института кибернетики ТПУ
Тел.: 8-909-540-63-78
Эл. почта: sava@cc.tpu.edu.ru

Марчуков Артур Викторович
Зав. лаб. сетей ЭВМ и телекоммуникаций каф. ОСУ Института кибернетики ТПУ
Тел.: (382-2) 42-05-40
Эл. почта: orion@cc.tpu.edu.ru

Алексеев Александр Александрович
Студент 5-го курса каф. ОСУ Института кибернетики ТПУ
Тел.: 8-923-604-05-86
Эл. почта: frt@tpu.ru

Silich V.A., Saveliev A.O., Marchykov A.V., Alexeev A.A.

Estimation methods of scientific publications based on analysis of the internet references to scientific portals

The drawbacks of the scientific publications estimation methods based on the citation index are considered. A specifying method for estimation of scientific publications, which relies upon monitoring of references to scientific web-portals and servers and uses the naive Bayesian classifier, is suggested. The independent features characterizing a web-user behavior are extracted.
Keywords: citation index, web-mining, web log, classification problem, naive Bayesian classifier.
