УДК 378.1:681.5

М.Ю. Воеводина

Харьковская национальная академия городского хозяйства, Харьков

ОБ ОДНОМ ИЗ АСПЕКТОВ ПРИМЕНЕНИЯ Э-МЕТРИК В ДИСТАНЦИОННОМ ОБРАЗОВАНИИ

В статье рассмотрены электронные методики измерений (э-метрики) активности пользователей образовательного сайта. Статистическими методами решается задача определения рационального размера файла, содержащего учебные материалы.

дистанционное образование, электронные методики измерений, сайт, учебный материал

Введение

Формулировка проблемы. Организация сайта – одна из основных составляющих дистанционного образования. Решение этой задачи связано с множеством проблем как технического, так и организационного характера. Это и создание программных средств, предназначенных для организации тестирования через InterNet, и разработка тестовых заданий, и организация «обратной связи» со слушателями при обсуждении результатов. Статья посвящена вопросу определения рационального размера файла, содержащего учебные материалы.

Анализ последних исследований. Большое количество публикаций в настоящее время посвящено проблеме организации дистанционного образования. Отличие этой публикации состоит в применении методов э-метрик для повышения эффективности работы образовательного сайта.

Цели статьи и формулировка задачи исследования. Целью данной статьи является определение рационального размера файла, содержащего учебные материалы. В настоящей работе описано решение задачи статистическими методами. Рассмотрены также электронные методики измерений (э-метрики), которые относятся одновременно как к объекту исследования (электронные документы и операции), так и к инструментарию (сбор статистики методами информационно-коммуникационных технологий) и позволяют интенсифицировать работу образовательного сайта.

Изложение основного материала исследований

Современная система дистанционного образования должна включать такие компоненты:

– развитую систему учебно-методической литературы в электронном виде, организованную в единую базу данных со средствами поиска на уровне гипертекстовых связей, тезаурусов, реферативных ссылок и т.п.;

– систему поддержки взаимодействия слушателей с преподавателями через средства электронных коммуникаций;

– систему организация консультирования в реальном времени;

– систему тестирования для организации текущего контроля остаточных знаний;

– средства организации коллективных обсуждений и общения слушателей друг с другом (создание медиа-центров).

Наличие указанных компонент характерно для достаточно развитой системы дистанционного образования [3]. В настоящее время поддержку такой системы может позволить себе далеко не каждое учебное заведение. Как правило, в реальной жизни большинство действующих систем дистанционного образования включают лишь некоторые из перечисленных компонент. Но без файлов, содержащих учебные материалы, еще никто не обошелся.

Материалы на сайте должны быть не только содержательны, структурированы, красиво и на-

глядно оформлены. Не менее важно сделать так, чтобы заинтересованный пользователь мог успешно скачать их без особых усилий! Система управления взаимоотношениями с пользователем (УВП) предполагает измерение и регулирование корреляций эффективности работы образовательного сайта.

Метрики, как стандарты измерений, имеют свой электронный формат, который мы по аналогии называют э-метриками [1]. В ныне существующей глобальной сетевой среде можно выделить два важных аспекта (если угодно, две грани), которыми они определяются и одновременно отличаются от других типов измерений. Электронные методики измерений (э-метрики) относятся одновременно как к объекту исследования (электронные документы и операции), так и к инструментарию (сбор статистики методами информационно-коммуникационных технологий).

Наиболее простой формой э-метрик являются замеры количества посещений веб-сайта. Такого рода замеры могут содержать следующую информацию: сколько раз посещается какая-то конкретная страница, или количество просмотров веб-страниц в какой-то конкретный день или за неделю. Более общее и более полное по охвату определение э-метрик включает также исследование активности и характера использования сетевых ресурсов. Сюда входят данные по количеству уникальных идентифицированных посетителей веб-сайта, полное число поисков, выполненных в течение месяца, по какому-то конкретному запросу (термину или фразе) в базе данных, количество (и результат) скачиваний учебных файлов, регистрация переписки с тьютором и т.п.

Остановимся на простой и часто встречающейся задаче: каков рациональный размер файла, размещаемого на INTERNET-сайте? Зачастую качество INTERNET не дает возможности пользователю скачать достаточно большой файл целиком. Плохое качество INTERNET, низкая квалификация пользователей не всегда позволяют это сделать. Успех этого действия сильно зависит от размеров скачиваемого файла: чем больше его объем, тем труднее его добыть. С другой стороны, делая файлы ничтожно малого объема, мы рискуем утратить интерес пользователей из-за их малой информативности.

Для анализа был выбран учебный файл, размещенный на образовательном сайте. Размер файла составляет 3,5 мегабайта. В качестве метода исследования был выбран статистический анализ данных, которые фиксировались с октября 2001 г. по сентябрь 2006 г. За указанный период зафиксировано 6585 обращений к файлу. Статистика нашего сервера позволяет определять IP-адрес пользователя и объем переданной информации. Кроме того, можно определить, использовалась ли одна из специальных программ (их называют программами «докачки»), позволяющая скачивать файл небольшими частями.

Следует отметить, что сам по себе статистический анализ не дает ответа вопрос, какой должен быть оптимальный размер INTERNET-страницы, он позволяет лишь оценить вероятность скачивания файла в зависимости от его размера [2]. На основании этой информации (а также своего опыта и интуиции) оптимальный размер файла должен определять человек.

На основании статистической информации можно классифицировать (может быть, приблизительно) пользователей по качеству их подключений к INTERNET. Мы разбили всех пользователей на три категории следующим образом.

I категория. Пользователи, которые не использовали программы докачки и не смогли скачать файл до конца. Качество INTERNET у таких пользователей мы считали плохим.

II категория. Пользователи, которые смогли скачать тестовый файл с использованием специальных программ докачки. Качество INTERNET у них нам не известно, однако само по себе это качество в данном случае не имеет значения, так как квалификация пользователя позволяет нормально работать даже с низким качеством связи.

III категория. Пользователи, которые смогли скачать тестовый файл за один раз. Мы считаем, что INTERNET у таких пользователей хороший.

В табл. 1 представлены поквартальные статистические данные (количество и процентное соотношение пользователей для каждой категории)

Таблица 1

Поквартальные статистические данные

Год/ Квартал	**Количество пользователей**			**Процентное соотношение**		
	I	**II**	**III**	**I**	**II**	**III**
2001/4	22	9	23	40.7	16.7	42.6
2002/1	13	2	15	43.3	6.7	50.0
2002/2	27	6	22	49.1	10.9	40.0
2002/3	6	1	14	28.6	4.8	66.7
2002/4	20	5	21	43.5	10.9	45.7
2003/1	19	3	19	46.3	7.3	46.3
2003/2	31	5	22	53.4	8.6	37.9
2003/3	12	1	16	41.4	3.4	55.2
2003/4	20	5	23	41.7	10.4	47.9
2004/1	24	4	20	50.0	8.3	41.7
2004/2	22	5	28	40.0	9.1	50.9
2004/3	10	4	20	29.4	11.8	58.8
2004/4	24	5	23	46.2	9.6	44.2
2005/1	21	4	27	40.4	7.7	51.9
2005/2	23	6	28	40.4	10.5	49.1
2005/3	7	1	16	29.2	4.2	66.7
2005/4	26	5	26	45.6	8.8	45.6
2006/1	24	8	28	40.0	13.3	46.7
2006/2	17	6	29	32.7	11.5	55.8
2006/3	9	7	20	25.0	19.4	55.6
За весь период наблюдений	**377**	**92**	**440**	**41.4**	**10.1**	**48.5**

Проанализируем данные из табл. 1. На рис. 1 представлено процентное соотношение пользователей, имеющих хороший INTERNET в каждом квартале исследуемого периода.

Уравнение линии тренда y=0,39x+46,0 показывает, что число таких пользователей увеличивается приблизительно на 2,8% в год.

Из всего вышесказанного можно сделать следующие выводы:

1. Наши статистические данные опровергают тезис о стремительном улучшении качества INTERNET. Если существующие тенденции сохранятся, то процент пользователей с плохим INTERNET станет равным 0 только в 2023 году, это дает серьезные основания надеяться на то, наши исследования будут актуальными еще несколько лет.

2. Статистический анализ в данном случае не позволяет однозначно формально оценить процентные соотношения пользователей различных категорий. Из многих вариантов мы выбрали достаточно консервативный – выбрали средние значения за весь период наблюдений. Таким образом, мы считаем, что доли пользователей категорий I, II, III составляют соответственно 0,41, 0,10 и 0,49.

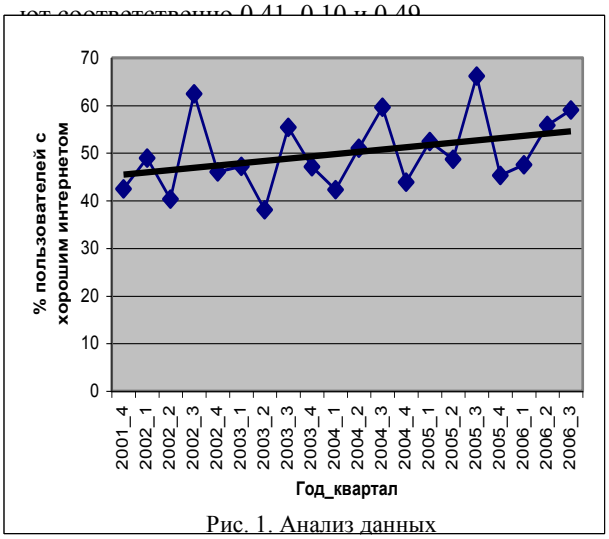

Рис. 1. Анализ данных

На данном этапе исследований мы определим вероятность скачивания тестируемого файла пользователями с плохим качеством INTERNET. Мы выбрали данные для исследования следующим образом. Длина тестируемого файла – 3,5 МБ. При каждом прерывании процесса скачивания файла фиксировалось значение случайной величины (с.в.) X, численно равное длине той части файла, которую удалось скачать (в байтах). Объем полученной выборки – 910 элементов. Для исследования закона распределения с.в. X весь интервал, в котором лежат значения с.в., мы разбили на 20 равных частей. Границы интервалов разбиения представлены в табл. 2.

Далее определяется количество попаданий с.в. в каждый из интервалов разбиения и соответствующие частоты. Вид распределения эмпирических частот позволяет выдвинуть гипотезу об экспоненциальном законе распределения с.в. X со средним выборки $\tilde{m} = 403042$. Для проверки выдвинутой гипотезы табулируем теоретическую плотность распределения в точках, являющихся серединами интервалов разбиения, и вычисляем значение критерия χ^2.

Значение критерия χ^2=1,0678 не входит в критическую область, следовательно, гипотеза об экспоненциальном законе распределения с.в. X принимается (с вероятностью 0,95). Построение теоретического закона распределения в виде плотности f(x) с.в. X позволят вычислить вероятность скачивания файла пользователем INTERNET плохого качества в зависимости от его длины. В нашем случае «плохой» INTERNET действительно очень плох, например, файл около 180 килобайт пользователь скачает лишь с вероятностью 0,2.

Таблица 2

Границы интервалов разбиения

№	Границы интервалов		Кол-во попаданий в инт.	Частота попаданий в инт.	Плотность распределения (теор.)	Значение критерия χ^2
	мин.	макс.				
1	0	183 500	506	0,5560	0,3626	0,1032
2	183 500	367 000	182	0,2000	0,2300	0,0039
3	367 000	550 500	48	0,0527	0,1459	0,0594
4	550 500	734 000	27	0,0297	0,0925	0,0427
5	734 000	917 500	36	0,0396	0,0587	0,0062
6	917 500	1 101 000	13	0,0143	0,0372	0,0141
7	1 101 000	1 284 500	16	0,0176	0,0236	0,0015
8	1 284 500	1 468 000	14	0,0154	0,0150	0,0000
9	1 468 000	1 651 500	6	0,0066	0,0095	0,0009
10	1 651 500	1 835 000	7	0,0077	0,0060	0,0005
11	1 835 000	2 018 500	12	0,0132	0,0038	0,0230
12	2 018 500	2 202 000	8	0,0088	0,0024	0,0167
13	2 202 000	2 385 500	6	0,0066	0,0015	0,0166
14	2 385 500	2 569 000	6	0,0066	0,0010	0,0324
15	2 569 000	2 752 500	4	0,0044	0,0006	0,0231
16	2 752 500	2 936 000	1	0,0011	0,0004	0,0013
17	2 936 000	3 119 500	5	0,0055	0,0002	0,1106
18	3 119 500	3 303 000	6	0,0066	0,0002	0,2625
19	3 303 000	3 486 500	4	0,0044	0,0001	0,1844
20	3 486 500	3 670 000	3	0,0033	0,0001	0,1647
Σ			**910**	**1,0000**	**0,9913**	**1,0678**

Рассмотрим вероятность скачивания статических и динамических страниц. Теперь мы можем ответить на вопрос: с какой вероятностью ***любой*** пользователь INTERNET скачает нужный файл (длиной не менее x). Воспользуемся формулой полной вероятности. Определим все необходимые

случайные события и их вероятности. Событие A – пользователь скачал нужный файл (длиной не менее x). Вероятность этого события нам нужно определить. Событие H_1 – пользователь имеет возможность скачать нужный файл (длиной не менее x) с первой попытки. Будем считать, что $P(H_1) = p$ – вероятность события H_1.

Событие H_2 – пользователь не имеет возможность скачать нужный файл (длиной не менее x) с первой попытки и скачивает его с некоторой вероятностью, которая зависит от x. В нашем случае события H_1 и H_2 образуют полную группу, поэтому вероятность события H_1 равна

$$P(H_2) = 1 - P(H_1) = 1 - p .$$

Условная вероятность события A при наступлении события H_1 равна 1, то есть $P(A/H_1) = 1$.

Рассмотрим условную вероятность события A при наступлении события H_2. Это вероятность скачивания файла пользователем с плохим INTERNET, она зависит от длины файла x. Ранее мы уже определили закон распределения (экспоненциальный) длины скачиваемого файла. Однако нас в данном случае интересует не конкретное значение длины файла x, а значение длины файла не менее x. Поэтому от функции плотности $f(x)$ перейдем к интегральной функции распределения $F(x)$. Таким образом,

$$P(A/H_2) = \int_x^{\infty} f(x)dx = 1 - F(x) .$$

Теперь можно применить формулу полной вероятности

$$P(A) = P(H_1)P(A/H_1) + P(H_2)P(A/H_2)$$

или

$$P(A) = p \cdot 1 + (1-p)(1 - F(x)) .$$

Для практического применения полученной формулы нам осталось только определить величину p. Для этого воспользуемся классификацией пользователей INTERNET.

В сети INTERNET принято различать динамические и статические страницы. В рамках нашего исследования нас интересует только тот факт, что статические страницы можно скачивать по частям (с помощью программ докачки), а динамические – нет.

1 случай. Статические страницы. Их наверняка скачают пользователи категорий II (используют программы докачки) и III (хороший INTERNET). В этом случае $p = 0{,}49 + 0{,}10 = 0{,}59$, а $1 - p = 0{,}41$.

2 случай. Динамические страницы. Их наверняка скачают только пользователи категорий III. Смогут ли скачать файл пользователи категории II, нам неизвестно, сделаем предположение, что INTERNET у них плохой и файл они скачивают точно так же, как пользователи категории I. В этом случае $p = 0{,}49$, а $1 - p = 0{,}51$.

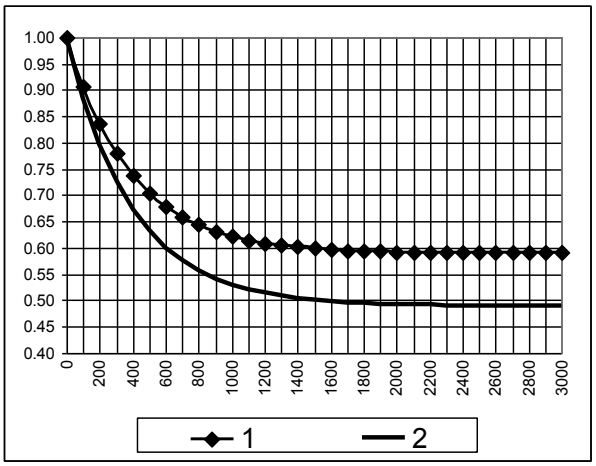

Рис. 2. Зависимость вероятности скачивания от размера файла

На рис. 2 представлена зависимость вероятности скачивания от размера файла (в килобайтах). Кривая 1 выражает зависимость для процесса скачивания статической страницы, а кривая 2 – для динамической. Вероятность скачивания файла размером не менее 50 килобайт в обоих случаях приблизительно равна 0,95. Дальше кривые расходятся. Для файла не менее 100 КБ вероятности скачивания равны 0,92 (для статических страниц) и 0,87 (для динамических). Для файла размером не менее 1000 КБ – 0,62 и 0, 53. Для файлов размером больше 3000 КБ обе кривые проходят параллельно оси X, вероятности скачивания таких файлов приблизительно равны 0,59 и 0,49.

Выводы и перспективы дальнейших исследований в данном направлении

Глубокое изучение и аккуратное применение э-метрик поможет их использованию для сбора и анализа данных. Методика э-метрик помогает определять успешность веб-сайта, базируясь на изучении практических действий пользователя, его умения осуществлять навигацию по материалам сайта. Методы э-метрик могут также лежать в основе измерений качества работы сайта.

Список литературы

1. Уайт Э., Камаль Э.Д. Статистические методы работы с электронными документами в библиотечной сфере, или Э-метрики. – М.: Омега-Л, 2006. – 395 с.

2. Феллер В. Введение в теорию вероятностей ее приложения. Т. 2. – М.: Мир, 1967. –752 с.

3. Бочаров Б.П., Рябченко Н.И, Донец Л.Ю., Воеводина М.Ю. Опыт использования системы тестирования знаний в InterNet // Сб. научн. тр. 7-й межд. конф. Украинской ассоциации дистанционного образования. – Харьков-Ялта: УАДО, 2003. – С. 308-311.

Поступила в редколлегию 7.12.2007

Рецензент: д-р техн. наук, проф. Г.Н. Жолткевич, Харьковский национальный университет им. В.Н. Каразина, Харьков.