УДК 004.021

МЕТОД ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ИНТЕРНЕТ СТРАНИЦ

Миргород В.С., Личканенко И.С., Мазур Д.М., Родригес Залепинос Р.А.
Донецкий национальный технический университет
кафедра компьютерных систем мониторинга
E-mail: innapainter@rambler.ru

Предложенный в статье метод использует 11 характеристик интернет страниц, в том числе Google Page Rank, рейтинг Yandex, Alexa Trafic Rank, рейтинг закладок Delicious и количество ссылок в Twitter за последний месяц. Между показателями производится поиск и анализ взаимозависимостей. При этом определяется влияние отдельных характеристик и их групп на общий рейтинг интернет страницы. Метод реализован на языке R. Приведены результаты анализа характеристик 46 интернет страниц предложенным методом. Обнаружено сильное влияние на рейтинг закладок Delicious группой двух показателей: количеством ссылок в Twitter и рейтингом посещаемости интернет страницы.

Введение

Сегодня одним из перспективных направлений интеллектуального анализа данных является разработка методов для поиска закономерностей, которые влияют на посещаемость интернет страниц. Используя полученные знания, можно добиться повышения посещаемости интернет страницы, и, следовательно, увеличения продаж товаров и услуг, предлагаемых на них.

В настоящее время для анализа интернет страниц широко применяются рейтинговые показатели и критерии, которые используются для повышения эффективности функционирования и оптимизации структуры интернет страницы. Компанией IBM разработано приложение SpeedTracer для анализа использования интернет ресурсов [1]. SpeedTracer отслеживает поведение пользователей для усовершенствования структуры интернет страницы и навигации. Программное обеспечение использует методы вывода для возобновления путей обхода пользователей. Алгоритмы интеллектуального анализа данных интернет страниц определяют закономерности движения пользователей по страницам. Результатом является набор шаблонов просмотра, который способствует лучшему пониманию поведения пользователей.

Проводятся исследования в сфере веб-аналитики, основной задачей которой является мониторинг посещаемости интернет страниц. На основании собранных данных изучается поведение посетителей для принятия решений по развитию и расширению функциональных возможностей интернет ресурса. Сервис веб-аналитики SpyBOX позволяет записывать и анализировать действия посетителей на интернет странице и оказывать влияние на их поведение [2].

В.В. Хайловой разработана система анализа поведения посетителей интернет страницы с использованием методов интеллектуального анализа данных и наглядной

интерпретацией полученных результатов [3]. Система снабжена интеллектуальными функциями: кластеризацией посетителей относительно выделенного целевого атрибута при помощи правил ДСМ (генерируются в программной среде QuDA) и оценкой качества интернет страниц методом нечеткого вывода.

Предложенный в работе метод анализирует интернет страницы по одиннадцати показателям. Из них были выбраны два показателя, у которых среднее геометрическое значение максимально: YLD (количество ссылающихся интернет страниц – Yahoo Links Domain) и DR (рейтинг социальных закладок Delicious). Проанализированы зависимости между ними и остальными показателями. Самая высокая корреляция была обнаружена с количеством ссылок в Twitter и рейтингом посещаемости интернет страницы. Это хорошо прослеживается на графиках зависимостей между показателями. Обнаружено их значительное влияние на популярность интернет страницы.

Интернет страницы для исследования и их характеристики

Для анализа было рассмотрено 46 новостных интернет страниц разных стран. Для получения их характеристик они были просканированы дополнением браузера Mozilla Firefox SEO Quake, сервисами Alexa (http://www.alexa.com) и http://www.cy-pr.com. Получены данные по различным показателям: GPR (Рейтинг Google), YR (Рейтинг Yandex), ATR (Рейтинг посещаемости –Alexa Trafic Rank), DR (Рейтинг закладок Delicious), TPLM (Количество ссылок из Twitter за последний месяц), GI (Количество страниц в индексе Google), YTIC (Тематический Индекс Цитирования Yandex), ASLI (Количество ссылающихся интернет страниц – Alexa Sites Linking in), YLD (Количество ссылок на интернет страницу – Yahoo Links Domain), PS (Размер главной страницы, кб), DA (Год создания домена). Примеры интернет страниц и полученных показателей приведены в таблице 1.

Таблица 1. Примеры интернет страниц и некоторых полученных показателей

Интернет страница	GPR	GI	YTIC	YR	ATR	YLD	DR	TPLM
http://lenta.ru	7	744000	19000	6	418	7727666	670	24445
http://www.segodnya.ua/	4	295000	6200	6	11261	303027	35	1316
http://www.unian.net/	7	269000	4200	6	12253	720950	42	2601
http://www.zn.ua/	6	133000	4900	6	36646	80947	31	62
http://kp.ua/	6	654000	4700	6	17994	1133334	9	1822
http://ura-inform.com/	5	403000	2200	5	42283	189989	5	1282
http://www.gazeta.ru/	7	483000	18000	6	1105	1041205	421	17961
http://www.rbc.ru/	8	134000	20000	6	448	2310582	359	1924
http://www.dni.ru/	6	59400	6900	6	5271	292374	72	1011
http://www.washingtonpost.com/	9	2560000	2500	5	358	6251265	7689	130518
http://www.nytimes.com/	9	19400000	5000	6	83	27885953	33386	296594
http://www.economist.com/	8	332000	1400	5	1557	3961499	11233	55952
http://www.cnn.com/	10	82100	4600	6	47	16309792	29572	250056
http://www.aolnews.com/	8	1310000	220	4	783	10689243	69	9699

Метод интеллектуального анализа интернет страниц

Разработан метод, позволяющий оценить зависимости между вышеуказанными показателями.

Из файла считываются данные в таблицу D_{ij}. Удаляется первый столбец с названием исследуемых объектов. Определяется номер строки нулевого объекта *z*. Приводятся показатели к шкале относительно нулевого объекта $D_{ij} = D_{ij} / D_{zi}$. Трансформируются показатели для последующего определения нелинейных зависимостей (возводятся в степени 2, -1, -2 с вычитанием логарифма, при этом *data_p2 = data^2*, *data_pm1 = data*^-1, *data_pm2 = data*^-2, *data_log = log*(*data*)). Определяются индексы D_{ix} и D_{iy} среди показателей для последующего построения функции $T_i = D_{ix} * D_{iy}$. Определяется геометрическая вероятность $W_i = T_i / \max(T)$. Вычисляется логарифм функции *W*. Определяется корреляция между функцией *W* (log(*W*)) и трансформированными показателями. Строится таблица корреляций для функций *W* и *W_log*, где строки — это название трансформации, а столбцы — показатели. Вычисляется среднее геометрическое G среди максимальных значений для каждого показателя в каждом столбце таблицы корреляций. Определяются индексы D_{ix} и D_{iy}, для которых значение *G* максимально, повторив операции, начиная с определения индексов D_{ix} и D_{iy} для построения функции Ti и закачивая высчитыванием среднего геометрического *G*, при различных значениях D_{ix} и D_{iy}.

Предложенный метод был реализован на языке *R*, который удобен для интеллектуального анализа данных.

Результаты

Для примера были подобраны индексы D_{ix} и D_{iy}, для которых среднее геометрическое значение *G* максимально: YLD (количество ссылающихся интернет страниц – Yahoo Links Domain) и DR (рейтинг социальных закладок Delicious). Затем установлены корреляции между геометрической вероятностью *W*, log(*W*) и показателями их трансформациями. Они представлены в таблицах 2 и 3.

Из полученных данных были выбраны зависимости с высокой корреляцией, по которым построены графики, представленные ниже.

Таблица 2. Корреляции между W и показателями и их трансформациями

	GPR	GI	YTIC	YR	ATR
data	0.41	0.86	0.10	0.31	0.07
data_pm1	0.29	0.04	0.12	0.23	0.81
data_p2	0.47	0.88	0.02	0.33	0.04
data_pm2	0.23	0.04	0.07	0.19	0.70
data_log	0.35	0.32	0.26	0.28	0.46
	ASLI	PS	DA	TPLM	
data	0.89	0.07	0.14	0.91	
data_pm1	0.21	0.036	0.14	0.05	
data_p2	0.97	0.10	0.14	0.98	
data_pm2	0.13	0.04	0.14	0.05	
data_log	0.52	0.01	0.14	0.42	

Таблица 3. Корреляции между log(W) и показателями и их трансформациями

	GPR	GI	YTIC	YR	ATR
data	0.70	0.37	0.16	0.46	0.36
data_pm1	0.58	0.01	0.59	0.58	0.58
data_p2	0.74	0.33	0.12	0.39	0.25
data_pm2	0.50	0.03	0.51	0.60	0.41
data_log	0.65	0.39	0.47	0.53	0.87
	ASLI	PS	DA	TPLM	
data	0.67	0.13	0.45	0.65	
data_pm1	0.83	0.02	0.45	0.09	
data_p2	0.50	0.07	0.45	0.51	
data_pm2	0.70	0.02	0.45	0.09	
data_log	0.94	0.11	0.45	0.73	

Таким образом, были обнаружены зависимости между: количеством ссылок на интернет страницу (YLD), рейтингом интернет страницы в социальных закладках Delicious (DR), количеством ссылок в публикациях в твиттер (TPLM) и рейтингом посещаемости интернет страницы (ATR).

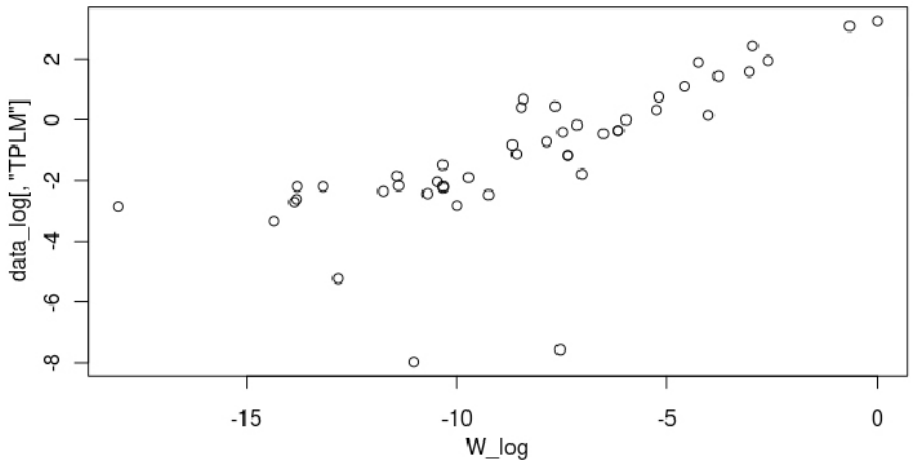

Рисунок 1. График корреляции между log(W) и показателем TPLM

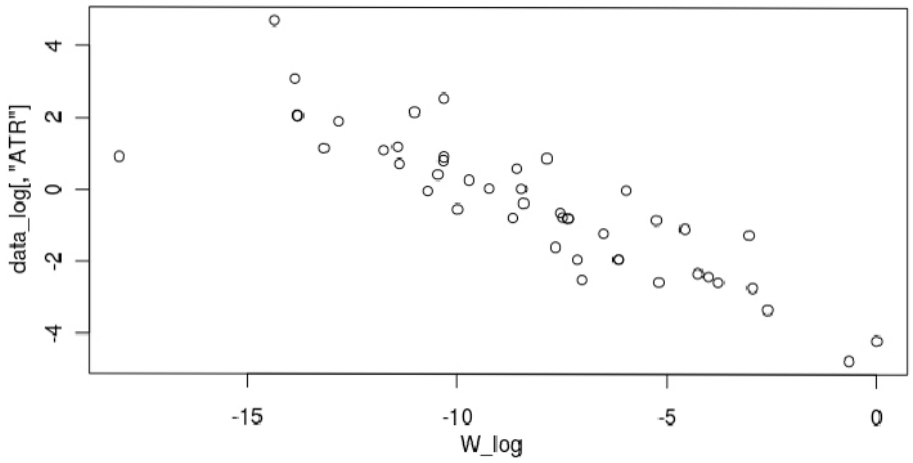

Рисунок 2. График корреляции между log(W) и показателем ATR

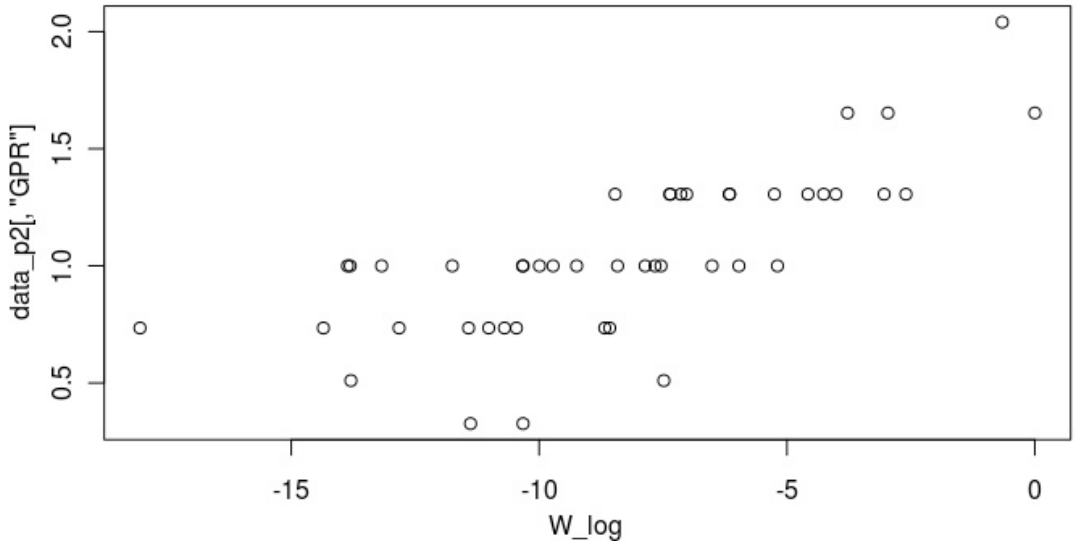

Рисунок 3. График корреляции между log(W) и показателем GPR

Выводы

Высокая корреляция количества ссылок на интернет страницу (YLD) и рейтинга в социальных закладках Delicious (DR) обнаружена с количеством ссылок в публикациях в твиттер (TPLM) и рейтингом посещаемости (ATR). Установлено, что наибольшее влияние на рейтинг интернет страницы оказывают количество ссылок из Twitter и рейтинг посещаемости интернет страницы. Найденные зависимости являются логичными, т.к. высокое количество ссылок из Twitter повышает посещаемость интернет страницы, а посещаемость непосредственно влияет на ее популярность. Предложенный метод можно использовать для поиска характеристик интернет страницы, которые необходимо улучшить для повышения ее посещаемости.

В дальнейшей работе планируется разработать и включить в общий анализ численные характеристики дизайна интернет страниц для поиска закономерностей влияния дизайна на посещаемость и другие важные показатели. Примерами показателей могут служить графические параметры (например, основные цвета интернет страницы и их сочетание, а также яркость и контрастность) и параметры навигации по интернет странице.

Перечень источников

[1] Web Mining [Электронный ресурс] – Режим доступа: http://www.galeas.de/webmining.html (30.03.2012)

[2] Д. Мелихов, И. Сарматов. Анализ сайта: справочник веб-аналитика. – К., 2011.

[3] В.В. Хайлова. Анализ эффективности работы Web-сайта с применением методов ИАД //Десятая национальная конференция по искусственному интеллекту с международным участием КИИ-2006 (25-28 сентября 2006 г., Обнинск): Труды конференции. В 3-т., М: Физматлит, 2006.