УДК 004.738.5

Огийко А. А.

Применение метода Борда в ранжировании веб-сайтов

Рекомендовано к публикации доцентом Печниковым А. А.

1. Введение. На фоне стремительного развития и внедрения информационных технологий одним из важнейших показателей деятельности крупных организаций становится качество их представления в сети Интернет. Грамотно оформленное содержимое веб-сайтов и ссылки с авторитетных ресурсов обеспечивают удобный доступ к важной информации, как для пользователей, так и для потенциальных спонсоров, а также влияют на положительную оценку организации-владельца общественностью. Это касается различных учреждений, в числе которых вузы, исследовательские лаборатории, медицинские институты. Рост внимания к качеству специализированных веб-ресурсов привел к появлению вебометрических рейтингов, в рамках которых ресурсы ранжируются на основе некоторых показателей, таких как, объем информации, цитируемость и т.д.

На сегодняшний день не существует единого подхода к созданию вебометрических рейтингов. При выборе набора показателей и объединяющей их формулы ранжирования, исследовательские группы руководствуются различными принципами. В среднем, используется от 6 до 8 показателей, а основным методом подсчета рейтингов является суммирование показателей с определенными коэффициентами (линейная функция). При этом, иногда, перед вычислением общего рейтинга проводится нормировка показателей различными способами (так как данные, полученные с помощью поисковых систем, могут быть ошибочны). Таким образом, остается актуальной задача поиска других существующих математических методов, которые могут применяться в вебометрическом ранжировании. Одним из таких методов является метод Борда, который нашел широкое применение в создании социологических и экономических рейтингов.

Огийко Анна Алексеевна – аспирант, Институт прикладных математических исследований КарНЦ РАН; e-mail: ogiiko@krc.karelia.ru, тел.: +7(931)702-58-75

Работа выполнена при финансовой поддержке РГНФ, грант № 12-03-12001

2. Использование метода Борда при ранжировании по нескольким показателям. Задача ранжирования веб-сайтов может быть представлена в форме задачи о групповом выборе. Согласно одному из существующих определений, приведенному Угольницким [1], стандартная постановка задачи группового выбора предполагает получение функции, которая должна строить групповую ранжировку для некоторого множества альтернатив на основе индивидуальных упорядочений. Применительно к веб-ресурсам и показателям их оценки, ранжируемыми альтернативами в такой задаче могут быть сами веб-сайты, а индивидуальные ранжировки можно построить на основе показателей, в роли которых выступают различные данные, получаемые с помощью поисковых систем (количество страниц на сайте, количество полнотекстовых файлов и т.д.).

Среди существующих функций группового выбора, для решения такой задачи можно выбрать правило Борда, одним из достоинств которого является возможность получения групповой ранжировки при любых условиях.

Для того чтобы использовать правило, необходимо вначале вычислить рейтинги участников ранжирования по каждому показателю, используя их порядковые номера в упорядоченных списках. Полученный результат будет являться набором индивидуальных ранжировок. Далее для каждого участника необходимо определить число Борда, по следующей формуле:

$$B(a) = B_1(a) + ... + B_n(a),$$

где a — сайт-участник, $B_i(a)$ — число сайтов, расположенных ниже него в i-той индивидуальной ранжировке. Групповая ранжировка строится по следующему правилу: сайт a имеет больший ранг, чем сайт b тогда и только тогда, когда $B(a) > B(b)$.

Очевидно, что в такой модели не будут учитываться порядки различия значений показателей для набора веб-сайтов. Например, два участника могут занимать близкие позиции в общем рейтинге при сильно различающихся значениях одного из показателей. Исследования, проведенные на множестве веб-сайтов научных учреждений России, показали, что для достаточно большого количества единиц анализа (веб-сайтов) характерно экспоненциальное распределение значений показателей. Таким образом, можно сделать вывод о необходимости предварительной обработки данных, подлежащих ранжированию методом Борда.

3. Нормирование показателей. Одним наиболее эффективных и простых в исполнении методов нормирования показателей, подлежащих ранжированию, является стратификация — разбиение по корзинам (группам). Идея разбиения участников на группы по каждому показателю уже была использована группой профессора Антопольского при составлении рейтинга РИВНОУ [2] (однако, далее использовалось ранжирование простым суммированием индивидуальных рейтингов).

Для применения разбиения необходимо определить на множестве значений каждого показателя N интервалов одинаковой длины и для каждого веб-сайта определить порядковый номер интервала, в который попадает соответствующее ему значение показателя. В зависимости от числа корзин, порядок разности между значениями одного показателя для разных участников будет учитываться в большей или меньшей степени. Если взять достаточно большое число корзин, такое что в каждую корзину попадает не более одного участника, мы получим индивидуальные ранжировки, аналогичные тем, что могут быть рассчитаны без предобработки.

4. Эксперимент. В ходе эксперимента были использованы данные о 398 сайтах научных учреждений России, входящих в состав РАН, полученные с помощью поисковых систем Yandex и Google, а также поискового робота BeeCrawler [3], в июне 2013 года. Были рассмотрены следующие показатели:

- S_Y — количество страниц, индексируемых поисковой системой Яндекс на заданном сайте;
- S_G — количество страниц, обнаруживаемое поисковой системой Google на заданном сайте;
- S_B — количество страниц, обнаруживаемое на первых восьми уровнях сайта инструментом BeeCrawler [3];
- V_G — количество гипертекстовых ссылок с других веб-ресурсов на заданный сайт, обнаруживаемое Google;
- $V_B(links) * V_B(sites)$, где
  - $V_B(links)$ — количество гипертекстовых ссылок, на заданный сайт с других сайтов целевого множества, обнаруживаемых BeeCrawler;

  - $V_B(sites)$ — количество сайтов целевого множества, с которых сделаны гипертекстовые ссылки, на заданный сайт, обнаруживаемое BeeCrawler;

- R_Y — суммарное количество файлов с расширениями PDF, DOC, PS и др., обнаруживаемое Яндекс на заданном сайте;

- R_G — суммарное количество файлов с расширениями PDF, DOC, PS и др., обнаруживаемое Google на заданном сайте;

- Sc_{GS} — количество ссылок на сайт, обнаруживаемых сервисом Google Scholar.

Значения первых трех показателей соответствуют одному свойству веб-сайта — его объему, который выражается в количестве индексируемых поисковыми системами веб-страниц. Тем не менее, из-за специфики алгоритмов различных поисковых сервисов, эти показатели могут сильно различаться, что было подтверждено ранее в работе [4]. Таким образом, возникает необходимость проведения процедуры сглаживания. Под данной процедурой подразумевается вычисление правдоподобных значений показателей вместо измеренных и представляющихся ошибочными значений с использованием данных о сайте, обнаруживаемых поисковым роботом BeeCrawler (алгоритм работы которого известен в отличие от алгоритмов других используемых поисковых систем). Более подробно методы сглаживания ошибок измерений для S_Y и S_G описаны в работе [4].

Следующий шаг — это непосредственно подсчет рангов сайтов участников различными способами, которые были описаны ранее:

- с использованием метода Борда без разделения на корзины,

- с предварительным разделением на 50/100/200 корзин и применением метода Борда.

Для того, чтобы сравнить между собой полученные ранжировки использовалось расстояние Кемени—Снелла, которое во многих исследованиях, связанных с теорией выбора и экспертного оценивания, выполняет роль количественного показателя различия между двумя индивидуальнымя экспертными ранжированиями [5]. Меньшее значение расстояния Кемени—Снелла соответствует большей схожести между двумя ранжировками. Данная мера эквивалентна удвоенному значению нормированного расстояния Хэмминга, которое также

использовуют для измерения различий между двумя ранжировками [6]. В рассмативаемом случае оно вычисляется по формуле:

$$d_H = \frac{1}{n(n-1)} \sum_{i,j=1}^{n} |r_{ij} - s_{ij}|,$$

где n — число сайтов-участников, r_{ij} и s_{ij} — элементы матриц бинарных отношений, соответствующих ранжированиям R и S.

Таблица 1. Значения расстояний Кемени—Снелла при различном числе корзин

	50 корзин	**100 корзин**	**200 корзин**	**Без группировки**
50 корзин	0	0,36	0,49	0,73
100 корзин	0,36	0	0,20	0,65
200 корзин	0,49	0,20	0	0,66
без группировки	0,73	0,65	0,66	0

Основываясь на полученных значениях парных расстояний между ранжировками, можно сделать следущие выводы:

- наиболее схожими оказались результаты ранжирования с разбиением на 100 и 200 корзин, а значит, с ростом числа корзин растет сходство между ранжировками с близким по количеству групп нормированием;
- между ранжировкой без нормирования и ранжировкой с разделением на группы (в том числе, на достаточно большое количество групп: 200 корзин при 398 участниках) существуют значительные различия (расстояние Кемени—Снелла $> 0{,}5$);
- из рассмотренных в ходе эксперимента вариантов ранжирования с практической точки зрения интересен метод, включающий нормировку с разделением на 200 корзин, так как суммарное расстояние от каждого рассмотренного ранжирования до данного сравнительно невелико, и вместе с тем сохраняется достаточное количество позиций в рейтинге (271 позиций для 397 учреждений).

Полученные выводы можно расширить, исследовав большее количество модификаций, как для метода Борда, так и для других методов многокритериального ранжирования. Кроме того, можно сравнить аналогичным образом ранжирования, полученные по значениям каждого из показателей, с групповыми ранжировками, чтобы определить наиболее согласованное с каждым отдельным критерием ранжирование.

Литература

1. Угольницкий Г. А. Модели конфликтов. М.: Вузовская книга, 2012. 320 с.

2. Антопольский А. Б., Поляк Ю. Е., Усанов В. Е. О российском индексе веб-сайтов научно-образовательных учреждений // Информационные ресурсы России. 2012. № 4. С. 2–7.

3. 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.

4. Печников А. А. Об измерениях вебометрических индикаторов // Международный журнал экспериментального образования. 2013. № 10. С. 400–404.

5. Kemeny J. G., Snell J. L. Mathematical Models in the Social Sciences. Boston.: Ginn, 1962. 145 p.

6. Миркин Б. Г., Орлов М. А. Методы многокритериальной стратификации и их экспериментальное сравнение. Препринт WP7/2013/06. М.: Изд. дом Высшей школы экономики, 2013. 32 с.