Раздел 1
МАТЕМАТИЧЕСКОЕ И КОМПЬЮТЕРНОЕ МОДЕЛИРОВАНИЕ ЭКОНОМИЧЕСКИХ ПРОЦЕССОВ

РАЗРАБОТКА АВТОМАТИЗИРОВАННОЙ СИСТЕМЫ РАНЖИРОВАНИЯ ПРИ ПРОВЕДЕНИИ ПРОЦЕССА РЕЦЕНЗИРОВАНИЯ НАУЧНЫХ СТАТЕЙ

Д. С. Алексеев

Саратовский национальный исследовательский государственный университет им. Н. Г. Чернышевского, Россия
E-mail: alekseevds@info.sgu.ru

В данной статье проведен краткий анализ существующих моделей рецензирования статей, а также представлен один из подходов автоматизации процесса рецензирования с использованием алгоритма PageRank. Приводится пример работы такого подхода и дается оценка его эффективности.

DEVELOPMENT OF AN AUTOMATED RANKING SYSTEM WHEN CARRYING OUT THE PROCESS OF REVIEW OF SCIENTIFIC ARTICLES

D. S. Alekseev

In the given article the brief analysis is conducted about existing models of reviewing articles, and represented approach which is used at automating the review process using the algorithm PageRank. Besides there given examples of application of this method and an assessment of its effectiveness.

В данной работе рассматривается процесс рецензирования и отбора научных статей для участия в конференции. Предполагается, что имеется n статей, которые нужно проранжировать для отбора лучших для участия в конференции. В распоряжении организационного комитета конференции имеется m рецензентов.

Заметим, что основной и наиболее распространенной на данный момент моделью является бально-рейтинговая модель оценки статей. Согласно этой модели, каждый рецензент (эксперт), входящих в пул рецензентов, получает некоторое количество статей для рецензирования и оценивает качество каждой статьи по 5-ти бальной шкале. Обычно каждая статья оценивается двумя или тремя рецензентами, и на основании средней их оценки формируется оценочный балл статьи. Статья, получившая большее количество баллов, считается лучше тех, чей средний балл ниже.

Главной проблемой при проведении отбора лучших статей и рецензировании по этой модели является человеческий фактор. Например, рецензент мо-

жет ошибочно поставить высокий балл статье с невысоким качеством, и наоборот. Кроме того, критерии каждого рецензента различны, и разные рецензенты могут поставить совершенно разные оценки одной и той же статье. В связи с этим предлагается разработать новую модель проведения отбора статей, основанную на принципах ранжирования.

В связи с этим в настоящей статье предлагается использовать другую модель для проведения процесса отбора статей, основанную на принципах ранжирования узлов графа, в котором узлами являются статьи, а дуги отражают отношение предпочтения (хуже, лучше) между статьями, высказанные рецензентами.

Основная идея предлагаемого подхода состоит в следующем. Рецензент получает несколько статей, которые он должен ранжировать от лучшей к худшей. Эта информация, полученная от всех рецензентов, позволяет сформировать граф, в котором узлами являются статьи, а дуга отражает отношение «лучше» между статьями.

После построения такого графа мы предлагаем применить хорошо известный алгоритм ранжирования вершин PageRank, который генерирует в результате своей работы список вершин, ранжированный от лучшей к худшей.

Заметим, что для корректной работы процедуры необходимо, чтобы между наборами статей, доставшихся разным рецензентам, были пересечения. Так, если некоторый набор статей достанется только одному рецензенту, то этот набор не будет оказывать влияние на другие «цепочки» предпочтений и участвовать в общем ранжировании.

PageRank – один из алгоритмов ссылочного ранжирования. Алгоритм применяется к коллекции документов, связанных гиперссылками, и назначает каждому из них некоторое численное значение, измеряющее его «важность» или «авторитетность» среди остальных документов. Алгоритм может применяться не только к веб-страницам, но и к любому набору объектов, связанных между собой взаимными ссылками, то есть к любому графу [1-3].

Рассмотрим граф из четырех веб-страниц: A, B, C и D. Ссылки со страницы на саму себя игнорируются. Множественные исходящие ссылки с одной страницы на другую рассматриваются как одна ссылка (только в этом примере). PageRank инициализируется одинаковым значением для всех страниц. Предполагается, что распределение вероятностей находится между 0 и 1. Следовательно, начальное значение для каждой страницы в этом примере равно 0,25. PageRank, передаваемый с одной страницы к тем, на которые есть исходящие ссылки на следующей итерации, делится поровну между всеми исходящими ссылками. Если бы единственные ссылки в системе были со страниц B, C и D на A, каждая ссылка передала бы 0,25 PageRank в A при следующей итерации, что в сумме составит 0,75.

Таким образом можно получить формулу для страницы A:

$$PR(A) = PR(B) + PR(C) + PR(D)$$

В общем случае значение PageRank любой страницы может быть вычислено по формуле:

$$\mathrm{PR}(u) = \sum_{v \in B_u} \frac{\mathrm{PR}(v)}{\mathrm{L}(v)},$$

где значение PageRank для страницы u зависит от значений PageRank для каждой страницы v, содержащейся в наборе B_u (набор, содержащий все страницы, ссылающиеся на страницу u), деленное на количество $\mathrm{L}(v)$ ссылок со страницы v.

Известно, что помимо перехода по ссылкам снова и снова, есть вероятность того, что этот процесс завершится. Вероятность того, что вычисления будут продолжаться, определяется демпфирующим фактором d. В различных исследованиях были проверены различные коэффициенты демпфирования, но обычно предполагается, что коэффициент демпфирования будет установлен около 0,85 [2]. Коэффициент демпфирования вычитается из 1 (и в некоторых вариантах алгоритма результат делится на количество документов N в коллекции), а затем этот член добавляется к произведению коэффициента демпфирования и суммы входящие оценки PageRank. Получится формула [1]:

$$PR(u) = \frac{1-d}{N} + d \sum_{v \in B_u} \frac{\mathrm{PR}(v)}{\mathrm{L}(v)}$$

Схема работы ранжирования и отбора статей имеет следующий вид. Статьи нумеруются от 0 до n-1, затем они распределяются по принципу, представленному в табл. 1.

Таблица 1

Схема распределения статей по рецензентам

	0	1	2	3	4	5	…	n-2	n-1
0	*	*	*	*	*				
1		*	*	*	*	*			
…									
m-2							*	*	
m-1							*	*	*

Заметим, что эта схема не обеспечивает одинаковое число оценок рецензентами для каждой статьи, однако это отдельная подзадача, которая требует изучения.

На основе этих оценок мы может составить граф связей между статьями и применить к нему PageRank. Далее необходимо проверить, как повлияет различное число связей в графе на результат работы алгоритма.

После вычисления эталонных значений PageRank в графе, функция reduce_graph_by_all, принимающая на вход число связей, которое нужно оставить, случайным образом прореживала исходный граф, то есть удаляла случайные дуги, так чтобы общее число дуг стало равно числу переданному в функцию. На полученном графе снова запускался PageRank и получались новые значения. Чтобы компенсировать случайность такого подхода, проводилось 10000 описанных выше получений новых графов. Так же на основе этих 10000 наборов ранков с помощью среднего арифметического по ранкам каждой вершины был получен список средних ранков.

Чтобы оценить эффективность алгоритма для данной задачи, были введе-

ны три характеристики:

1. Евклидово расстояние
2. Число неправильных позиций
3. Сумма модулей разности позиций для каждой вершины

Среди этих значений в каждом векторе характеристик вычислялись:

1. Максимальное
2. Минимальное
3. Математическое ожидание
4. Дисперсия

Тестирование проводилось на основе двух наборов данных. Первый – синтетический пример, в котором первая статья лучше всех; вторая лучше всех, кроме первой; третья лучше всех, кроме первой и второй и так далее; всего 16 статей и 120 связей. Второй пример был взят с соревнования по спорту, а именно с группового этапа для 16 команд, которые играли в 1 группе, что удобно, так как у нас есть оценка всех со всеми. В таком примере связь будет обозначать, что одна команда обыграла другую, что соответствует связи «лучше - хуже», связей так же 120. Обратимся к результатам.

В табл. 2 представлены результаты относительно евклидова расстояния.

Таблица 2

Результаты тестов относительно евклидова расстояния

Тест	Максимальное евклидово расстояние	Минимальное евклидово расстояние	Математическое ожидание евклидова расстояния	Дисперсия евклидова расстояния
1 эксперимент, 42 статьи	3,54	0,58	1,68	0,15
1 эксперимент, 54 статьи	2,44	0,46	1,21	0,07
1 эксперимент, 65 статей	1,69	0,35	0,92	0,04
1 эксперимент, 75 статей	1,44	0,21	0,71	0,02
1 эксперимент, 84 статьи	1,05	0,21	0,55	0,01
1 эксперимент, 92 статьи	0,88	0,17	0,44	0,01
2 эксперимент, 40 статей	37,31	6,22	14,09	7,51
2 эксперимент, 54 статьи	32,65	4,22	11,62	6,52
2 эксперимент, 65 статей	22,54	3,23	9,68	5,1
2 эксперимент, 75 статей	21,71	2,39	8,05	4,25
2 эксперимент, 84 статьи	19,18	2,51	6,63	3,29

В табл. 3 представлены результаты относительно числа неправильных позиций.

Таблица 3

Результаты тестов относительно числа неправильных позиций

Тест	Максимальное число неправильных позиций	Минимальное число неправильных позиций	Математическое ожидание числа неправильных позиций	Дисперсия числа неправильных позиций
1 эксперимент, 42 статьи	14	0	8,31	4,58
1 эксперимент, 54 статьи	13	0	7,44	4,58
1 эксперимент, 65 статей	13	0	6,58	4,64
1 эксперимент, 75 статей	13	0	5,58	4,41
1 эксперимент, 84 статьи	12	0	4,7	4,33
1 эксперимент, 92 статьи	11	0	3,78	3,94
2 эксперимент, 40 статей	16	7	14,26	1,73
2 эксперимент, 54 статьи	16	6	14,1	1,85
2 эксперимент, 65 статей	16	6	13,87	2,02
2 эксперимент, 75 статей	16	6	13,59	2,26
2 эксперимент, 84 статьи	16	6	13,28	2,5

В табл. 4 представлены результаты относительно суммы модулей разности позиций каждой вершины.

Таблица 4

Результаты тестов относительно суммы модулей разности позиций каждой вершины

Тест	Максимальная сумма модулей разности позиций каждой вершины	Минимальная сумма модулей разности позиций каждой вершины	Математическое ожидание суммы модулей разности позиций каждой вершины	Дисперсия суммы модулей разности позиций каждой вершины
1 эксперимент, 42 статьи	40	0	15,62	35,13
1 эксперимент, 54 статьи	36	0	12,55	24,61
1 эксперимент, 65 статей	28	0	10,09	18,19
1 эксперимент, 75 статей	24	0	8	13,16
1 эксперимент, 84 статьи	22	0	6,12	9,99
1 эксперимент, 92 статьи	18	0	4,64	7,53
2 эксперимент, 40 статей	126	6	78,85	220,64
2 эксперимент, 54 статьи	126	16	77,15	213,57
2 эксперимент, 65 статей	126	24	74,53	213,62
2 эксперимент, 75 статей	122	16	71,03	207,53
2 эксперимент, 84 статьи	118	18	67,55	202,07

Таким образом, результаты проведенных экспериментов показывают, что предлагаемый подход к отбору статей на основе их ранжирования обладает достаточной устойчивостью по отношению к распределению статей среди рецензентов.

СПИСОК ЛИТЕРАТУРЫ

1. *Brin S., Page L*. The anatomy of a large scale hypertextual // Web search engine. In: 7th International World-Wide Web Conference. Elsevier Press, Brisbane. 1998

2. *Lu P., Cong X*. The Research on Webpage Ranking Algorithm Based on Topic-Expert Documents. In: Unger H., Meesad P., Boonkrong S. (eds) // Recent Advances in Information and Communication Technology. Advances in Intelligent Systems and Computing, Springer, Cham 2015. Vol. 361.

3. *Ali M. Z. B., Nasser Y.* DistanceRank: An intelligent ranking algorithm for web pages // Information Processing & Management. 2008. Vol. 44(2). Pp. 877–892.