РАНЖИРОВАНИЕ ИНФОРМАЦИИ НА ОСНОВЕ ОЦЕНОК И ПОВЕДЕНИЯ ПОЛЬЗОВАТЕЛЕЙ

Брицов Роман Алексеевич,
аспирант, Российский университет дружбы народов, инженерный факультет, кафедра Инженерного бизнеса и управления предприятием, Ханты-Мансийский автономный округ – Югра, город Сургут, Россия, ***britsov@live.com***

Ключевые слова: *рейтинговая система, рейтингование, ранжирование информации, система управления базами данных.*

Статья посвящена проблемам обработки информации, в частности, процессам ранжирования цифровых данных, релевантность которых прямо или косвенно зависит от действий человека. Несмотря на то, что оценивание информации пользователями субъективно, важно то, что именно пользователи являются конечными потребителями оцениваемой информации, поэтому логично использовать именно их сознательные и бессознательные действия, как входящие данные для процедуры ранжирования.
Предметом исследования являются ключевые этапы ранжирования информации на основе оценок и поведения пользователей. Цель статьи – проанализировать и выявить существующие проблемы процесса ранжирования информации, параллельно предложив классификацию характеристик ранжирования, которая является научным новшеством. Перечислим некоторые выводы полученные в процессе исследования. Рейтинговая система является основой ранжировании информации. Идеальной рейтинговой системы не бывает. Тип оцениваемой информации и цели генератора (автора) контента обуславливают то, какая система оценивания информации будет наиболее оптимальной. В свою очередь от последней зависит выбор самой рейтинговой системы. Для оценивания информации на ознакомление с которой необходимо затратить некоторое количество времени (книга, фильм, спектакль, статья) лучше всего подходит пятибалльная или десятибалльная система оценивания. Их минусом является наличие невостребованных пользователями баллов: 2, 3, 4 – в пятибалльной системе и 2, 3, 4, 10 в десятибалльной системе. А их главный плюс - это возможность расчета интегрального показателя (в большинстве случаев основываясь на Байесовских рейтинг-системах), который понятен и прост для относительного сравнения. Что касается контента, на который пользователь затрачивает незначительное количество времени, то разумно использовать унарную или бинарную систему оценивания. Минусом данного подхода является эффект "богатые богатеют", устранение которого приводит к некоторым погрешностям. Однако плюсами являются простота реализации алгоритма. Ранжирование можно классифицировать по предметам оценки, по системам оценивания информации и поведения пользователей, по алгоритмам рейтинговых систем, по вариантам представления ранжированных данных целевому пользователю.

Для цитирования:
Брицов Р.А. Ранжирование информации на основе оценок и поведения пользователей // T-Comm: Телекоммуникации и транспорт. – 2016. – Том 10. – №1. – С. 62-66.

For citation:
Britsov R.A. Ranking data based on users' marks and behavior. T-Comm. 2016. Vol 10. No.1, pp. 62-66. (*in Russian*).

По оценке аналитического агентства IDC (International Data Corporation) каждые полтора года объем данных хранящихся в интернете удваивается (рис. 1). Ежедневно интернет генерирует один экзабайт (1024 млрд. Гб) информации [6]. Общество живет в условиях цифровой «информационной перегрузки», что в свою очередь усиливает потребность в ранжировании получаемой информации.

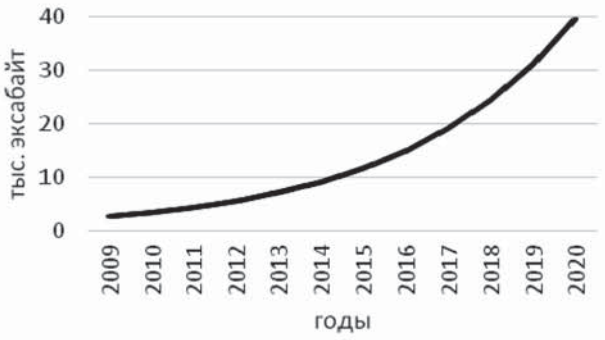


Рис. 1. Экспоненциальный рост объема информации в интернете

Рейтинговые системы тесно связаны с теорией вероятностей и математической статистикой. В первые научно задокументированная проблема рейтингования была поднята в 1654 г. в переписке между Блезом Паскалем и Пьером Ферма [5]. Существенный вклад в теорию рейтинговых систем внесли труды Томаса Байеса, опубликованные в 1763 г. [3], а также исследование Карла Гаусса в области нормального распределения [1]. Труды последних лежат в основе современных известных рейтинг-систем таких как модель Бредли-Терри [2], из которой выходит знаменитый рейтинг ЭЛО [4], лежащий в основе TrueSkill [7].

Несмотря на то, что вопросы рейтингования достаточно хорошо изучены, идеальной рейтинговой системы нет. Многие специалисты по-прежнему находят узкие места в существующих системах, вследствие чего появляются споры о том, какой алгоритм наиболее качественно и точно ранжирует данные той или иной направленности при определенных условиях.

Набор действий с момента оценивания пользователем до построения ранжированного ряда можно подразделить на несколько этапов:

1) Выбор предмета оценивания можно подразделить на две группы:

а. пользователь, оцениваемый сообществом;

б. статья, фильм, фотография, отзыв, товар, услуга и т.п.

Пользователь, как предмет оценивания выделяется в отдельную группу так как имеет возможность субъективного оценивания данных из группы б.

2) Выбор системы оценивания информации. Бальная классификация систем оценки информации пользователями приведена в табл. 1. Наиболее часто можно встретить систему оценки «плюс, минус и ноль», если пользователь воздержался от оценки или десятибалльную систему. Десятибалльная система соответственно подразумевает возможность оценки информации по шкале от 0 до 10.

Ноль в системе оценивая играет важную роль ввиду того, что этот результат оценки пользователями контента наиболее распространен. Большинство разработанных алгоритмов и формул для расчета интегральной величины оцениваемого объекта пользователями упускают возможность нулевой оценки. Количество пользователей, воздержавшихся от оценивания – это математический показатель, позволяющий дополнять точность результатов существующих алгоритмов, поэтому некоторые разработчики отдельно добавляют такой показатель, как количество уникальных запросов к контенту (просмотров, скачиваний и т.п.). В данном случае появляется сложность в сопоставлении двух показателей.

Таблица 1

Балльная классификация систем оценивания информации пользователями

б	Простая рейтинговая система (Плюс/минус)		Балльно-рейтинговая система (шкала)	
Баллы	**0, +1**	**-1, 0, +1**	**0…5**	**0…10**
Название	Унарная	Бинарная	Пятибалльная	Десятибалльная
Варианты визуальной реализации	Сердце, палец вверх, плюс, стрелочка вверх и т.п.	Плюс/минус, стрелки вверх/вниз, палец вверх/вниз	Шкала с элементами позволяющими оценить информацию в соответствии с мнением пользователя	
Примеры визуальной реализации	♥, ▲, +	+–, ▲▼	☺☺☺☺☺ или  (4 из 5)	☺☺☺☺☺☺☺☺☺☺ или ☺☺☺☺☺ (7 из 10)
Интернет-ресурсы и приложения*	github.com fb.com vk.com ok.ru coub.com	reddit.com youtube.com habrahabr.ru afisha.ru ozon.ru pikabu.ru 9gag.com fishki.net «Активный гражданин»	softodrom.ru tophotels.ru ozon.ru drom.ru auto.ru	imbd.com kinopoisk.ru market.yandex.ru amazon.com App Store Google Play TripAdvisor yelp.com

* Здесь и далее автор не несет ответственности за качество и безопасность ресурсов представленных в качестве примеров.

Оценивание объекта по системе от 0…5 будет наиболее близко для восприятия и сопоставления российскому человеку, ввиду того, что схожая система оценивая лежит в основе большинства российских образовательных учреждений. Однако статистически было определено, что в пятибалльной системе вторым после бездействия (оценка ноль) пользователь предпочитает ставить 5 баллов, а третьим 1. Так, на рис. 2 представлена структура оценок (от 1 до 5), которая была собрана роботом «YouTube» и опубликована в официальном блоге [8]. Именно по этой причине «YouTube» перешел на трехбалльную систему оценивания.

Тем не менее пятибалльные и десятибалльные системы наиболее подходят для оценки контента на который пользователю необходимо потратить некоторое количество времени (книга, фильм, игра). Но даже в таких случаях при использовании десятибалльной шкалы, когда количество

оценок стремится к бесконечности, будет образовываться коридор невостребованных баллов. Так для десятибалльного kinopoisk.ru основываясь на итоговых рейтингах к 12041 фильму, где свою оценку оставили более 500 человек невостребованным коридором является промежуток от 2 до 4 баллов, а также оценка в 10 баллов.

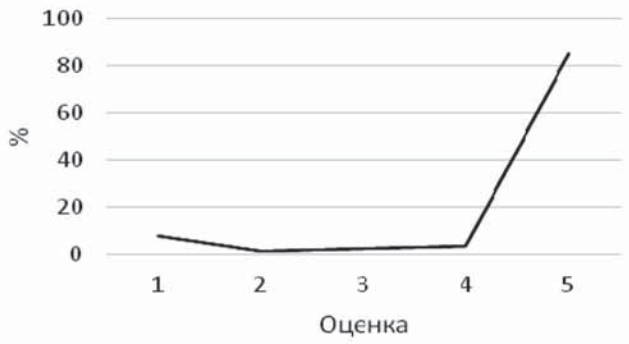


Рис. 2. Структура оценок видеоконтента пользователями «YouTube», до введения тройственной системы оценивания

В тоже время для контента на ознакомление с которым уходит незначительное количество времени рациональнее использовать унарную или бинарные системы оценки информации.

Стоить отметить, что существуют системы, чаще всего это форумы (например, cyberforum.ru, nowa.cc), для которых фактором оценки информации является количество просмотров, подписок, оставленных сообщений и/или опубликованных постов (например, livejournal.com). Некоторые файлообменники ориентируется на количество скачиваний контента (emule-project.net). При использовании такого алгоритма ранжирования особенно замечается эффект «богатые богатеют», когда наиболее раскрученный контент продолжает раскручиваться ещё больше (далее данный эффект будет рассмотрен более подробно).

В большинстве пиринговых (p2p) сетей отсутствует нарочное оценивание материала (rutraker.org, thepiratebay.vg). Основополагающим фактором выбора информации в таких системах является целенаправленный поиск необходимого контента пользователем. Таким образом реализуется органическое оценивание, где наиболее высокооцененный (социально-одобренный) контент обладает большим количеством «сидов» и «пиров» («личей»).

Также в некоторых системах ранжирование информации происходит по принципу естественного ранжирования (отсутствие оценивания): информация которая появилась последней по времени будет первой. Например, таковыми являются: контент размещаемый в группах «Вконтакте», большинство интернет-СМИ: «Коммерсантъ», «Ведомости», «Лента» или авторские ресурсы evanmiller.com, thisisindexed.com, crypto.com, moserware.com.

Таким образом система оценивания информации пользователем может быть:

а. Ручной. В таких системах пользователю необходимо осознано поставить субъективную оценку (см. табл. 1).

б. Автоматизированной (органической). Когда информация для ранжирования собирается роботом независимо от фактической оценки пользователем ресурса, классификация которой приведена в табл. 1. Робот может анализировать время просмотра определенной информации, фиксировать движение мышкой и т.п. Также определенное значение играет количество сообщений, просмотров, опубликованных постов и скачиваний. Целью сбора такой информации является как ранжирование, так и защита от виртуальных аккаунтов и других роботов: «спамеров», «накрутчиков» рейтинга.

в. Отсутствует система оценивания – естественное ранжирование по времени размещения данных, по цене, по количеству

г. Смешанной. Совмещение вышеперечисленных способов оценивания. Например, если рассматривать «Вконтакте» – систему с одной из наиболее сложных моделей оценивания и ранжирования, то при поиске пользователя предпочтение будет отдаваться аккаунтам с наибольшим числом подписчиков, друзей, а также совокупных положительных оценок опубликованного пользователем контента другими участниками сообщества.

3) Аккумуляция оценок пользователя. Чем больше оценок будет иметь контент, тем ниже дисперсия и, соответственно, большая вероятность объективного конечного результата, приносящего максимальную пользу.

4) Выбор рейтинговой системы. Рейтинговая система – это ключевой момент при расчете интегрального показателя. К самым распространенным относятся системы, в которых:

а. контент оценивается по количеству плюсов, просмотров, скачиваний и т.п.;

б. контент оценивается по количеству плюсов за вычетом минусов;

в. применяется балльно-рейтинговая система, а результат рассчитывается как среднее арифметическое, геометрическое или взвешенное;

г. применяются комплексные алгоритмы, основанные на расчете ожидания определенной оценки с учетом различной степени достоверности.

Плюсами всех трёх вариантов являются: простота реализации, минимальные нагрузки на систему и базу данных. Существенными минусами являются: равнозначность оценки опытного пользователя и новичка, а при варианте «а» и «б» в случае отсутствия «затухания» высоко оценённого контента, особо выраженным будет эффект «богатые богатеют». Данный эффект заключается в доминировании более старой и оцененной вышесреднего информации над новой. Когда пользователь сталкивается с большими массивами информации, физически сложно и в большинстве случаев невозможно исследовать данные в полном объеме, поэтому пользователь ограничивается той информацией, которая изначально была на верхних уровнях. Один из вариантов решения – автоматизированной экспоненциальное понижение рейтинга, зависящее от времени размещения информации или в качестве более простого решения – группировка контента по времени размещения.

Спорным вопросом при использовании варианта «б» является то, что должно иметь более высокий ранг: информация у которой много оценок, как положительных и отрицательных или только положительные оценки, но значительно меньшее количество. Например, «Информация А» имеет 70 плюсов и 30 минусов (70% положительных оценок), в то время как «Информация Б» имеет 9 плюсов и 1 минус (90% положительных оценок).

Также на отдельном примере рассмотрим проблему варианта «в». На одном информационном ресурсе «Информация А» имеет одну оценку в 5 баллов, а другая «Информация Б» множество оценок в 5 баллов, но также хотя бы одну оценку менее 5 баллов, следовательно, при прочих равных условиях рейтинг «Информации Б» будет ниже, чем у «Информации А».

Более сложными являются рейтинговые системы, направленные на решение минусов вышеупомянутых систем. Так решением проблем варианта «б» при учете только положительных и отрицательных оценок может выступать алгоритм Эдвина Уилсона [10]:

$$\frac{\hat{p}+\frac{z^2}{2n}\pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}+\frac{z^2}{4n^2}}}{1+\frac{z^2}{n}}, \qquad (1)$$

где $\hat{p}$ – доля положительных оценок, z – это $1-\frac{1}{2\alpha}$ или квантиль нормального распределения, α – вероятность ненаступления события, n – общее число оценок.

Для расчета нижней границы доли положительных оценок "±" нужно заменить на "–". При вероятности 95% определения доли положительных оценок, квантиль нормального распределения будет равен 0,975, соответственно z=1,96. Данный алгоритм позволяет сбалансировать положительные оценки и нивелировать разницу в количестве оценок.

Используя алгоритм Эдвина Уилсона рассмотрим вышеупомянутый спорный вопрос со значительной разницей в количестве голосов: 70 плюсов и 30 минусов, против 9 плюсов и 1 минуса: в первом случае нижняя граница вероятности положительных оценок составляет 60,42%, в то время как во втором 59,58%. Следовательно, первый вариант исходя из формулы 1 имеет более высокий рейтинг. Алгоритм Эдвина Уилсона используется redditblog.com и yelp.com [9].

Наиболее востребованными на сегодняшний день являются байесовские рейтинг-системы. Например, для бинарных классификаторов может использоваться метод парных сравнений Бредли-Терри [2] или, позволяющий решать более сложные задачи, запатентованный компанией «Microsoft» «TrueSkill» [7]. Оба рейтинга чаще всего используются в спортивных соревнованиях.

Следует отметить самую распространенную формулу, применяющуюся в пятибалльных и десятибалльных системах:

$$Rating=\left(\frac{V}{V+M}\right)R+\left(\frac{M}{V+M}\right)C, \qquad (2)$$

где V – совокупное количество оценок за оцениваемый контент; M – порог голосов, необходимый для участия в рейтинге (устанавливается по усмотрению разработчиков); R – среднее арифметическое оценок за оцениваемый контент; C – среднее значение рейтинга всего контента, имеющегося в системе.

Данную формулу применяют imdb.com, drom.ru, kinopoisk.ru.

5. Выбор варианта представления оцененной информации пользователю.

Рейтинговые системы не имели бы смысла, если бы результаты их работы не были доступны пользователю или не влияли на очередность показа того или иного контента.

Так оцененная информация может быть:

а. отсортированной по параметрам разработчика. Здесь без применения специальных знаний пользователь не будет видеть следы работы рейтинговой системы (kommersant.ru, vedomosti.ru, digg.com), однако разработчики её применяли. В большинстве таких систем пользователь может самостоятельно отсортировать информацию по новизне, объему (Мб), геолокационным показателям и т.п. Отдельно следует отметить сортировку по релевантности, где система отдает предпочтение контенту, который считает наиболее интересным (по ключевым словам, по общим интересам с взаимосвязанными пользователями).

б. отображаемой с рассчитанным интегральным показателем. При таком варианте представления к скрытым алгоритмам сортировки информации из пункта «а» добавляется открытая пользователю сортировка по рассчитаному рейтингу (kinopoisk.ru, market.yandex.ru, «App Store», youtube.com).

Если исключить из расчета коммерческую информацию, то изначально видео-контент представляемый youtube.com основывается на количестве просмотров и релевантности поискового запроса. Интересно то, что непосредственным оценкам пользователей уделяется второстепенное значение.

в. смешанной. При использовании такого варианта представления ранжирования информации на первый взгляд может показаться отсутствие работы рейтинговой системы. Однако при выборе контента становится очевидным, что существуют закономерность при ранжировании данных. Например, несмотря на то, что «Вконтакте» отказалась от открытого интегрального рейтингового показателя, поиск «Вконтакте» в первую очередь представит людей с наибольшим числом подписчиков, подарков и совокупным количеством одобренных постов. Так самые первые аккаунты попадают под эффект «богатые богатеют».

Таким образом процесс ранжирования информации на основе оценок и поведения пользователей состоит из пяти этапов. Некоторые этапы (характеристики ранжирования) возможно классифицировать по предметам оценки, по системам оценивания информации и поведения пользователей, по алгоритмам рейтинговых систем, по вариантам представления ранжированных данных целевому пользователю.

Литература

1. *Вентцель Е.С.* Теория вероятностей. – 10-е изд., стер. – М.: «Академия», 2005. – 576 с.
2. *Agresti, A* 2014. 'Categorical Data Analysis', John Wiley & Sons, pp. 436-439.
3. *Bayes, T, and Price, R,* 1763. 'An Essay towards solving a Problem in the Doctrine of Chance. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, M.A. and F. R.S., Philosophical Transactions of the Royal Society of London, vol. 53, pp. 370-418.
4. *Elo, A* 1978. 'The Rating of Chessplayers, Past and Present', Arco, 233 p.
5. *Ferma, P and Pascal, B.* 'Fermat and Pascal on Probability', University of York, the UK, viewed 12 November 2015, http://www.york.ac.uk/depts/maths/histstat/pascal.pdf
6. *Gantz, J and Reinsel, R* 2013. 'The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East', IDC Country Brief, p. 1.
7. *Graepel T., Minka T., Herbrich R* 2007. 'TrueSkill(TM): A Bayesian Skill Rating System', Advances in Neural Information Processing Systems, vol. 19 / Ed. by B. Scholkopf, J. Platt, T. Ho man. Cambridge, MA: MIT Press, pp. 569-576.
8. *Rajaraman, S* 2009. 'Five Stars Dominate Ratings', YouTube Official Blog, p. 1.
9. *Salihefendic, A.* 'How Reddit ranking algorithms work', viewed 19 November 2015, http://amix.dk/blog/post/19588.
10. *Wilson, E* 1927. 'Probable Inference, the Law of Succession, and Statistical Inference', Journal of the American Statistical Association, vol. 22, pp. 209-212.

RANKING DATA BASED ON USERS' MARKS AND BEHAVIOR

Roman Britsov, *aspirant, Department of Department of Engineering Business and Enterprise Management, Engineering faculty, People's Friendship University of Russia, Moscow, Russia,* **britsov@live.com**

Abstract

The article is concerned with the problems of information processing. In particular, the ranking digital data where the relevance depends on the human activity is discussed. Although the evaluation of information by users is subjective, this is the users who are the end consumers of ranked data. Therefore, it is logical to analyze the conscious and unconscious activity of users in the accumulation of incoming data for the ranking procedure. The subject of research is the key stages of ranking data based on estimates and user behavior. The purpose of the article is to analyze and identify the existing problems of the process of ranking data, simultaneously offering features ranging classification, which is a scientific innovation.

Let us list some of the findings obtained in the research process. The rating system is the basis of the information ranking. The ideal rating system does not exist. The type of the estimated information and goals of the content generator (author) determine what system of information estimation would be the most optimal. In turn, the selection of the most rating system depends on this system. It is better to use a five-point or ten-point grading system to estimate the information (a book, a film, a play, an article) that requires certain amount of time. The disadvantage of such systems is the presence of unclaimed points by users: 2, 3, 4 in a five-point system and 2, 3, 4, 10 in a ten-point system. But their main advantage is the possibility of calculating the integral indicator (in most cases based on Bayesian rating systems), which is clear and simple for relative comparisons. As regards the content on which the user spends a small amount of time, it is reasonable to use the unary or binary system of estimation. The downside of this approach is the effect of "the rich get richer", the elimination of which leads to some errors. However, the advantage is the simplicity of the algorithm implementation. The ranking can be classified according to the subjects of estimation, data estimation systems and user behavior, algorithms of ranking systems, options of presentation of ranked data to a target user.

Keywords: *rating system, data ranking, data management, database management system.*

References

1. *Venttsel E.* Probability theory. Moscow: 'Akademiya', 2005. 576 p. (*in Russian*)
2. *Agresti, A* 2014. 'Categorical Data Analysis', John Wiley & Sons, pp. 436-439.
3. *Bayes, T, and Price, R,* 1763. 'An Essay towards solving a Problem in the Doctrine of Chance. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, M. A. and F. R. S.', Philosophical Transactions of the Royal Society of London, vol. 53, pp. 370-418.
4. *Elo, A* 1978. 'The Rating of Chessplayers, Past and Present', Arco, 233 p.
5. *Ferma, P and Pascal, B.* 'Fermat and Pascal on Probability', University of York, the UK, viewed 12 November 2015, http://www.york.ac.uk/depts/maths/histstat/pascal.pdf.
6. *Gantz, J and Reinsel, R* 2013. 'The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East', IDC Country Brief, p. 1.
7. *Graepel T., Minka T., Herbrich R* 2007. 'TrueSkill(TM): A Bayesian Skill Rating System', Advances in Neural Information Processing Systems, vol. 19 / Ed. by B. Scholkopf, J. Platt, T. Homan. Cambridge, MA: MIT Press, pp. 569-576.
8. *Rajaraman, S* 2009. 'Five Stars Dominate Ratings', YouTube Official Blog, p. 1.
9. Salihefendic, A. 'How Reddit ranking algorithms work', viewed 19 November 2015, http://amix.dk/blog/post/19588.
10. *Wilson, E* 1927. 'Probable Inference, the Law of Succession, and Statistical Inference', Journal of the American Statistical Association, vol. 22, 209-212.