О выборе потенциальных ключевых слов

Шумейко А.А., Шумейко М.А.

Поисковая оптимизация сайта, является важным элементом “Semantic WEB”, позволяющая повысить эффективность ведения бизнеса с использованием возможностей Internet. Одним из узловых элементов поисковой оптимизации является подбор ключевых слов, которые соответствуют продукту или услуг, размещенных на продвигаемом сайте. Традиционно в выборку ключевых слов включаются, как высокочастотные запросы, так и группы ключевых слов, которые относятся к низкочастотным запросам. Например, слово «игра» является высокочастотным запросом, а словосочетание «on-line игра-стратегия империя» относится к низкочастотным. Ясно, что, несмотря на то, что высокочастотные запросы приводят к большему трафику (количеству заходов клиентов на сайт), они в целом, неэффективны, так как пришедшие по таким запросам клиенты, не формируют устойчивую целевую аудиторию сайта. Низкочастотные запросы не могут похвастаться высоким трафиком, однако имеются все предпосылки, что посетители, которые пришли по низкочастотным запросам, приобретут товар или услугу, предлагаемые на продвигаемом сайте.

Не существует единого универсального метода оптимизации подбора ключевых слов, например, для одностраничного приложения (каковым является, например, on-line игра), нужно использовать такие ключевые слова, которые максимально отражают тему веб-сайта, а для мультистраничного, например, для интернет-магазина, необходимо выбирать такие ключевые слова, которые соответствуют как содержанию каждой отдельной страницы сайта, так и общей теме сайта. Так, страницы с профильным контентом (содержанием) лучше всего оптимизировать узкоспециализированными ключевыми словами, а общие ключевые слова подходят для оптимизации страниц с обобщенным содержанием. При этом главная страница сайта должна быть первой в результатах поиска по отношению к другим страницами сайта, что достигается за счет использования общих и популярных ключевых слов. Остальные страницы сайта оптимизируют по средне- и низкочастотным поисковым запросам.

Данная работа посвящена прогнозированию рейтинга нового ключевого слова основываясь на статистике существующего множества ключевых слов, описывающих возможности информационного ресурса.

На данный момент для прогнозирования рейтинга ключевых слов используется следующая схема.

Для получения статистических данных об актуальных ключевых словах и потенциальных ключевых словах используется сервис SensorTower.

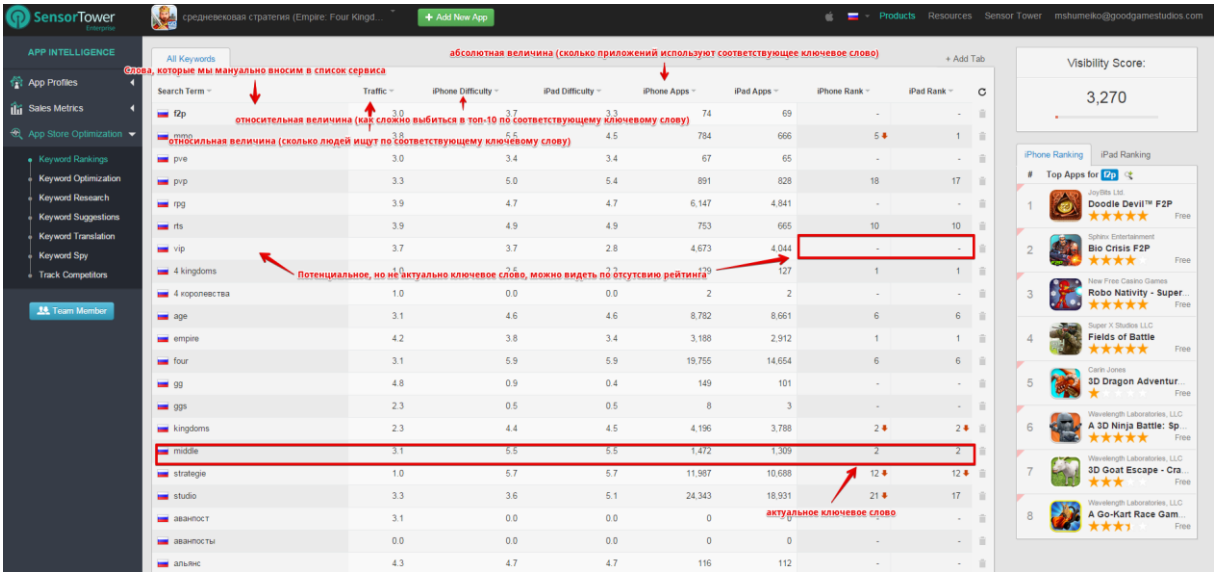


Рис.1. Получение статистики ключевых слов от сервиса SensorTower.

Этот сервис так же дает возможность получить отчет по рейтингу за последние 5 недель.

По полученной информации о рейтингах, получаем среднее значение рейтинга за последние 5 недель. Следующим шагом является получение линейной регрессионной модели только по актуальным ключевым словам, так как у них имеется рейтинг (зависимая переменная).

Term	Traffic	Difficulty	Competing Ap	Ranking Me
4 kingdoms	2.5	4.1	135	1
age	5.2	5.7	9442	30
alliance	4.8	6.4	1097	13
alliances	3.8	6	1096	3
archer	5.6	5.4	709	13
army	5.5	5.4	4390	39
army strateg	1.7	5.1	464	14
attack	4.7	6.1	7260	33
batalla	3.3	6.6	1921	42
battle	4.6	6.8	12016	72
battle for en	1	4.6	456	2
battles	4.2	7.6	11993	29
bottlenose	3.3	6.1	12007	69
camelot	4.8	5.1	266	9
castillo	3.6	7.3	981	14
castle	5.8	7.8	4211	14
castle app	1.8	6	4189	24
castle battle	1.9	5.8	506	10
castles	4.6	5.7	4211	35
conquer	4.3	5.1	344	9
empire	5.6	4.3	3330	1
empire apps	1.5	5.7	3323	10
empire game	3.8	4.7	2876	4
empires	3.7	5.1	3315	15
f2p	4.2	5	77	5
fight	4.8	6.2	9053	54
fortress	4.8	5.2	550	12
fortresses	2.7	5.2	548	2
free strategy	1	4.1	137	1
game	5.6	7.3	473713	590
godgame	2.8	1.9	19	1
guild	4.4	5.3	422	13
king	5.2	4.6	9311	20
kingdom	4.9	5.6	4566	3
kingdom bat	1	4.4	617	7
kingdom gar	2.8	5.5	3596	10
kingdoms	4	4.7	4540	3
kings	5.4	6.7	9286	42

AUSGABE: ZUSAMMENFASSUNG

Regressions-Statistik

Multipler Korrelationskoeffizient	0.567075256
Bestimmtheitsmaß	0.321574346
Adjustiertes Bestimmtheitsmaß	0.318343748
Standardfehler	106.7075648
Beobachtungen	423

ANOVA

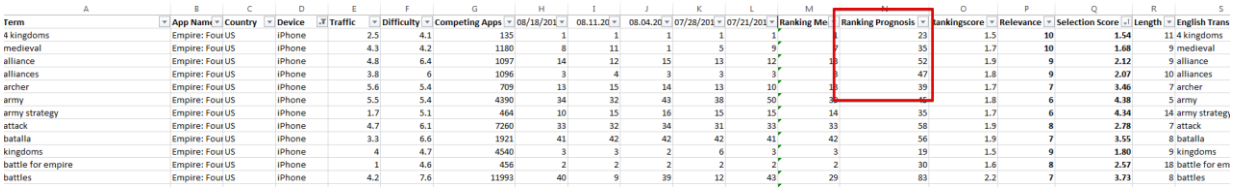
	Freiheitsgrade (df)	*Quadratsummen (SS)*	*Mittlere Quadratsumme (MS)*	*Prüfgröße (F)*	*F krit*
Regression	2	2266829.43	1133414.715	99.54018162	4.1255E-36
Residue	420	4782331.845	11386.50439		
Gesamt	422	7049161.275			

	Koeffizienten	*Standardfehler*	*t-Statistik*	*P-Wert*	*Untere 95%*	*Obere 95%*	*Untere 95.0%*	*Obere 95.0%*
Schnittpunkt	-25.76846438	15.78925433	-1.632025417	0.103423631	-56.8042693	5.26734053	-56.8042693	5.26734053
Difficulty	11.87376346	3.551988469	3.342849664	0.000903542	4.89187446	18.8556525	4.89187446	18.8556525
Competing Apps	0.001521646	0.000119784	12.70328271	1.63117E-31	0.0012862	0.0017571	0.0012862	0.0017571

Data before change | Data after change (03.09) | Translation | Regression | Relevance | Analysis before change | Analysis after change

Рис.3. Построение регрессионной модели.

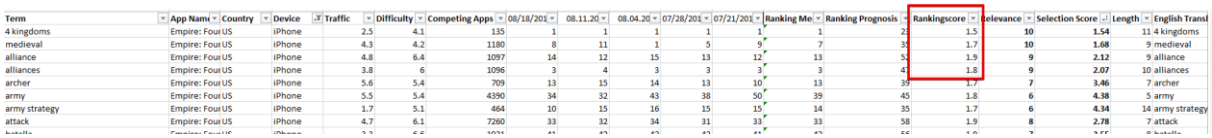
С помощью найденных коэффициентов строится прогноз рейтинга для каждого из потенциальных ключевых слов.



Term	App Nam	Country	Device	Traffic	Difficulty	Competing Apps	08/18/201	08.11.20	08.04.20	07/28/201	07/21/201	Ranking Me	Ranking Prognosis	Rankingscore	Relevance	Selection Score	Length	English Trans
4 kingdoms	Empire: Four US		iPhone	2.5	4.1	135	1	1	1	1	1	1	23	1.5	10	1.54	11	4 kingdoms
medieval	Empire: Four US		iPhone	4.3	4.2	1180	8	11	1	5	9	7	35	1.7	10	1.68	9	medieval
alliance	Empire: Four US		iPhone	4.8	6.4	1097	14	12	15	13	12	13	52	1.9	9	2.12	9	alliance
alliances	Empire: Four US		iPhone	3.8	6	1096	3	4	3	3	3	3	47	1.8	9	2.07	10	alliances
archer	Empire: Four US		iPhone	5.6	5.4	709	13	15	14	13	10	13	39	1.7	7	3.46	7	archer
army	Empire: Four US		iPhone	5.5	5.4	4390	34	32	43	38	50	39	45	1.8	6	4.38	5	army
army strategy	Empire: Four US		iPhone	1.7	5.1	464	10	15	16	15	15	14	35	1.7	6	4.34	14	army strategy
attack	Empire: Four US		iPhone	4.7	6.1	7260	33	32	34	31	33	33	58	1.9	8	2.78	7	attack
batalla	Empire: Four US		iPhone	3.3	6.6	1921	41	42	42	42	41	42	56	1.9	7	3.55	8	batalla
kingdoms	Empire: Four US		iPhone	4	4.7	4540	3	3	2	6	3	3	19	1.5	9	1.80	9	kingdoms
battle for empire	Empire: Four US		iPhone	1	4.6	456	2	2	2	2	2	2	30	1.6	8	2.57	18	battle for em
battles	Empire: Four US		iPhone	4.2	7.6	11993	40	9	39	12	43	29	83	2.2	7	3.73	8	battles

Рис.4. Получение прогноза по линейной регрессионной модели.

Ввиду того, что полученное значение прогноза может выйти за границы шкалы (от 0 до 10), то на следующем шаге проводится нормировка данных и оценка рейтинга по 10-ти бальной шкале.



Term	App Nam	Country	Device	Traffic	Difficulty	Competing Apps	08/18/201	08.11.20	08.04.20	07/28/201	07/21/201	Ranking Me	Ranking Prognosis	Rankingscore	Relevance	Selection Score	Length	English Transl
4 kingdoms	Empire: Four US		iPhone	2.5	4.1	135	1	1	1	1	1	1	23	1.5	10	1.54	11	4 kingdoms
medieval	Empire: Four US		iPhone	4.3	4.2	1180	8	11	1	5	9	7	35	1.7	10	1.68	9	medieval
alliance	Empire: Four US		iPhone	4.8	6.4	1097	14	12	15	13	12	13	52	1.9	9	2.12	9	alliance
alliances	Empire: Four US		iPhone	3.8	6	1096	3	4	3	3	3	3	47	1.8	9	2.07	10	alliances
archer	Empire: Four US		iPhone	5.6	5.4	709	13	15	14	13	10	13	39	1.7	7	3.46	7	archer
army	Empire: Four US		iPhone	5.5	5.4	4390	34	32	43	38	50	39	45	1.8	6	4.38	5	army
army strategy	Empire: Four US		iPhone	1.7	5.1	464	10	15	16	15	15	14	35	1.7	6	4.34	14	army strategy
attack	Empire: Four US		iPhone	4.7	6.1	7260	33	32	34	31	33	33	58	1.9	8	2.78	7	attack

Рис.5. Получение прогноза по нормированной шкале

Таким, образом, единственным инструментом, используемым для построения прогноза, является линейная регрессия. Данный подход не есть исключением. Построение регрессионных моделей на сегодняшний день, без всяких сомнений, является наиболее распространенным методом многомерного статистического анализа данных. Подавляющее большинство статей, анализирующих эмпирические данные, основаны на использовании регрессионных моделей. Вместе с тем многие особенности и ограничения регрессионных моделей обычно остаются вне сферы внимания исследователей, что, подчас, приводит к неточным, а иногда, и просто ошибочным результатам.

Применительно к рассматриваемой задаче, традиционная модель множественного линейного регрессионного анализа подразумевает поиск показателей, определяющих значение отдельной количественной переменной, обозначаемой y по множеству имеющихся данных

$$\mathbb{P} = \{P_i(x_i^0, x_i^1, x_i^2, y_i) \ \ (i = 0,1, \dots, n)\}\,.$$

Структура связи в данной модели предполагается линейной:

$$y = a^0x^0 + a^1x^1 + a^2x^2 + b$$

где b - остаточный член, фиксирующий ту часть информации y, которая не описывается переменными (x^0, x^1, x^2).

Регрессионный анализ показывает, во-первых, качество модели, то есть степень того, насколько данная совокупность $\mathbb{P} = \{P_i(x_i^0, x_i^1, x_i^2, y_i) \ \ (i = 0,1, \dots, n)\}$ объясняет y. Показатель качества называется коэффициентом детерминации R^2 и показывает, какой процент информации $Y = \{y_i\}$ можно объяснить поведением $X = \{x_i^0, x_i^1, x_i^2\}$. Во-вторых, регрессионный анализ вычисляет значения коэффициентов a^0, a^1, a^2, b, то есть определяет, с какой силой каждый из $X = \{x_i^0, x_i^1, x_i^2\}$ влияет на $Y = \{y_i\}$.

Методологическим недостатком такого подхода является то, что данная зависимость ищется единой для всей совокупности исходных данных. Иными словами, предполагаем, что для всех ключевых слов характер зависимости *Y* от *X* единый. В том случае, когда выборочная совокупность достаточно однородна, такого рода допущение имеет под собой определенные основания, однако, если анализируются, малое число «хороших» ключевых слов совместно с большим числом «плохих», то допущение об однородности данных выглядит не очень убедительным. Единая форма уравнения в этой ситуации сильно огрубляет реальную зависимость, качество модели неизбежно оказывается весьма низким, а смысл регрессионных коэффициентов, фиксирующих степень влияния множества X на Y, будет скорее отражать влияние «плохих» данных, чем «хороших».

Вполне очевидно, что гораздо разумнее строить отдельные модели для существенно различающихся между собой групп данных. Решению именной этой задачи и посвящена данная работа.

В нашем случае прогнозирование рейтинга ключевых слов можно формализовать следующим образом. Даны точки $P_i(x_i^0, x_i^1, x_i^2, y_i)$ $(i = 0,1, \dots, n)$, для $(\tilde{x}^0, \tilde{x}^1, \tilde{x}^2)$ нужно спрогнозировать значение $\tilde{y}$. По сути, точки $P_i(x_i^0, x_i^1, x_i^2, y_i)$ описывают тело, где (x_i^0, x_i^1, x_i^2) – его координаты, а y_i- плотность. Для более точного прогноза плотности тела в точке $(\tilde{x}^0, \tilde{x}^1, \tilde{x}^2)$ нужно использовать те точки (x_i^0, x_i^1, x_i^2) $(i = 0,1, \dots, n)$, которые лежат наиболее близко к прогнозируемой точке. Поэтому предлагается следующий алгоритм, основанный на том факте, для любого набора точек, существует регулярное (наиболее близкое к правильному) разбиение на симплексы. Такое симплициальное разбиение названо именем Делоне [1-2].

Множеством Делоне назовем множество $D \subset E^d$, для которого существуют положительные числа r и R такие, что для любого открытого d-шара B_r^0 и замкнутого шара B_R радиусов r и R соответственно, выполнены неравенства $|B_r^0 \cap D| \leq 1$ и $|B_R \cap D| \geq 1$, где $|Z|$ означает мощность множества Z. В общем смысле, никакое подмножество этого множества, состоящее из d+2 точек, не лежит на одной (d-1)-сфере.

Другими словами - что понимается под регулярной триангуляцией (под триангуляцией понимаем симплициальное разбиение), - через вершины невырожденного симплекса (в простейшем случае это невырожденный треугольник, то есть все вершины не лежат на одной прямой) можно провести одну и только одну сферу. Если триангуляция такова, что внутрь симплекса не попадает ни одна точка разбиения, то такое разбиение будем называть регулярным (множество Делоне). В этом случае, если точка $(\tilde{x}^0, \tilde{x}^1, \tilde{x}^2)$ лежит в j-м симплексе регулярной триангуляции, то вершины этого симплекса представляют собой точки $(x_i^0, x_i^1, x_i^2, y_i)$ наиболее «плотно» прилегающие к искомой точке $(\tilde{x}^0, \tilde{x}^1, \tilde{x}^2)$, и, соответственно, наилучшим образом могут ее описать, то есть прогноз y_i будет наилучшим, среди всех остальных.

Алгоритм. Итак, пусть дана точка $(\tilde{x}^0, \tilde{x}^1, \tilde{x}^2)$. Найдем четыре точки $(x_i^0, x_i^1, x_i^2)(i = \mathrm{k}, \mathrm{k}+1, \mathrm{k}+2, k+3)$ такие, что расстояние

$$\varepsilon_i = \sqrt{\sum_{m=0}^{2} (x_i^m - \tilde{x}^m)^2}$$

наименьшее и все эти точки не лежат на одной плоскости, то есть

$$\begin{vmatrix} x_{k+1}^0 - x_k^0 & x_{k+1}^1 - x_k^1 & x_{k+1}^2 - x_k^2 \\ x_{k+2}^0 - x_k^0 & x_{k+2}^1 - x_k^1 & x_{k+2}^2 - x_k^2 \\ x_{k+3}^0 - x_k^0 & x_{k+3}^1 - x_k^1 & x_{k+3}^2 - x_k^2 \end{vmatrix} \neq 0.$$

Проведем через эти точки сферу. Центр сферы (x^0, x^1, x^2) находится из системы линейных уравнений

$$(x_{k+1}^0 - x_k^0)(x^0 - (x_{k+1}^0 + x_k^0)/2) + (x_{k+1}^1 - x_k^1)(x^1 - (x_{k+1}^1 + x_k^1)/2) + (x_{k+1}^2 - x_k^2)(x^2 - (x_{k+1}^2 + x_k^2)/2) = 0,$$

$$(x_{k+2}^0 - x_k^0)(x^0 - (x_{k+2}^0 + x_k^0)/2) + (x_{k+2}^1 - x_k^1)(x^1 - (x_{k+2}^1 + x_k^1)/2) + (x_{k+2}^2 - x_k^2)(x^2 - (x_{k+2}^2 + x_k^2)/2) = 0,$$

$$(x_{k+3}^0 - x_k^0)(x^0 - (x_{k+3}^0 + x_k^0)/2) + (x_{k+3}^1 - x_k^1)(x^1 - (x_{k+3}^1 + x_k^1)/2) + (x_{k+3}^2 - x_k^2)(x^2 - (x_{k+3}^2 + x_k^2)/2) = 0.$$

А радиус

$$r^0 = \sqrt{\sum_{m=0}^{2} (x^m - x_k^m)^2}$$

Если найдется точка $(x_i^0, x_i^1, x_i^2)(\mathrm{i} \neq \mathrm{k}, \mathrm{k}+1, \mathrm{k}+2, k+3)$, которая лежит от центра этой окружности на расстоянии, меньше радиуса, то ее включаем в список прилегающих точек и строим новые четверки точек, пока не получим такую четверку, для которой сфера, построенная по этим точкам не будет содержать никакую другую точку. Понятное, что данный алгоритм не нацелен на построение триангуляции Делоне всего множества данных. Результатом является только один симплекс, содержащий данную точку $(\tilde{x}^0, \tilde{x}^1, \tilde{x}^2)$. Но нам, по большому счету, ничего иного и не нужно.

Любая точка симплекса может быть записана в барицентрических координатах

$P = \sum_{m=0}^{3} \lambda_m P_{k+m}$, где $\sum_{m=0}^{3} \lambda_m = 1$.

Барицентрические координаты λ_m $(m = 0,1,2,3)$ легко найти из системы

$$x^j = \sum_{m=0}^{3} \lambda_m x_{k+m}^j \; (j = 0,1,2).$$

И, наконец, прогноз рейтинга будет равен

$$\tilde{y} = \sum_{m=0}^{3} \lambda_m y_{k+m}.$$

Рассмотренная конструкция не дает возможности найти прогноз ключевых слов, не лежащих внутри облака исходных данных, другими словами, если точка, соответствующая ключевому слову, лежит внутри облака (тела), то прогнозирование ее значения можно описать термином «интерполяция», а вне – «экстраполяцией».

Для решения задачи экстраполяции данных, сведем ее к интерполяции, с этой целью для точки $\tilde{P}(\tilde{x}^0, \tilde{x}^1, \tilde{x}^2)$ найдем в множестве $\hat{\mathbb{P}} = \{\hat{P}_i(x_i^0, x_i^1, x_i^2)\ (i = 0,1, \dots, n)\}$ ближайшую к $\tilde{P}(\tilde{x}^0, \tilde{x}^1, \tilde{x}^2)$ точку $\hat{P}_{i0}(x_{i0}^0, x_{i0}^1, x_{i0}^2)$

$$\hat{P}_{i0} = arg\left\{min\{|\tilde{P} - \hat{P}_i||i = 0,1, \dots, n\}\right\}.$$

Рассмотрим множество

$$\tilde{\mathbb{P}} = \{\hat{\mathbb{P}} \backslash \hat{P}_{i0}\} \cup \tilde{P}$$

то есть, выбросим из множества исходных данных $\hat{\mathbb{P}} = \{\hat{P}_i(x_i^0, x_i^1, x_i^2)\ (i = 0,1, \dots, n)\}$ точку $\hat{P}_{i0}(x_{i0}^0, x_{i0}^1, x_{i0}^2)$ и добавим точку $\tilde{P}(\tilde{x}^0, \tilde{x}^1, \tilde{x}^2)$. Для полученного множества построим триангуляцию Делоне и выберем симплекс с вершиной в точке $\tilde{P}(\tilde{x}^0, \tilde{x}^1, \tilde{x}^2)$ внутри которого находится точка $\hat{P}_{i0}$. Такой симплекс существует и он единственный (что следует из теоремы о существовании триангуляции Делоне).

Тогда

$$\hat{P}_{i0} = \lambda_0\tilde{P} + \lambda_1\hat{P}_k + \lambda_2\hat{P}_{k+1} + \lambda_3\hat{P}_{k+2}, \text{ где } \sum_{m=0}^{3}\lambda_m = 1,$$

и, следовательно,

$$\tilde{P} = \frac{1}{\lambda_0}\left(\hat{P}_{i0} - \lambda_1\hat{P}_k - \lambda_2\hat{P}_{k+1} - \lambda_3\hat{P}_{k+2}\right).$$

Тогда прогноз рейтинга будет равен

$$\tilde{y} = \frac{1}{\lambda_0}(y_{i0} - \lambda_1 y_k - \lambda_2 y_{k+1} - \lambda_3 y_{k+2}).$$

Если $\lambda_0 = 0$, то точка $\hat{P}_{i0}$ лежит на гиперплоскости, образуемой точками $\hat{P}_k, \hat{P}_{k+1}, \hat{P}_{k+2}$, следовательно из множества $\hat{\mathbb{P}}$ нужно выбросить точку $\hat{P}_{i0}$ и начать все заново.

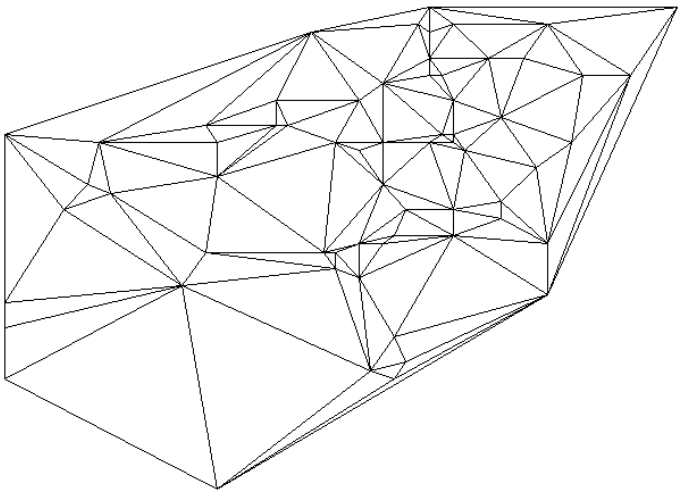


Рис.5. Триангуляция Делоне по множеству ключевых слов.

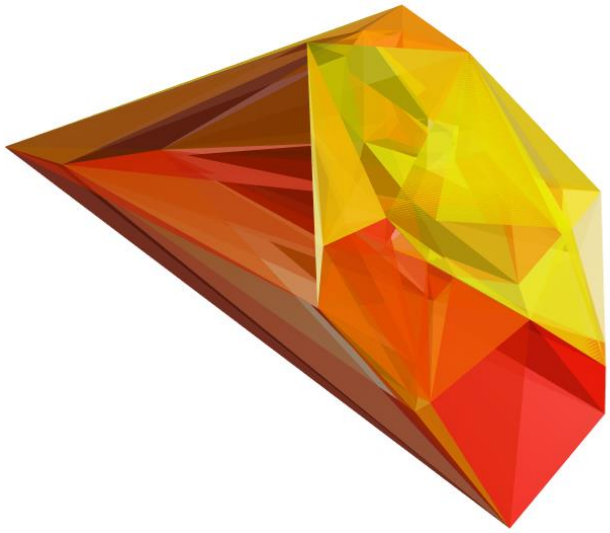


Рис.6. Разбиение нормированного в единичный куб множества ключевых слов на регулярные симплексы с учетом плотности каждого симплекса

В предложенном подходе есть «подводные камни». Ключевые слова, получаемые из сервисной службы, принадлежат разным группам потребителей, как по национальному, так и по территориальному признаку. Поэтому, одно и то же ключевое слово, набранное в разных странах, будет иметь разный рейтинг, что приводит к появлению в корпусе ключевых слов большого количества данных, которые можно классифицировать как шум. Выходом может быть, либо фрагментация статистических данных по тем или иным признакам, либо их фильтрация, позволяющая производить поиск только на корпусе ключевых слов, которые отражают интерес клиента к нашему ресурсу. С нашей точки зрения, второй подход более предпочтительный. Для этой цели приведем модификацию под наши задачи алгоритма кластеризации DBSCAN (Density Based Spatial Clustering of Application with Noise) [8]. Идея данного метода основана на гипотезе, что элементы одного кластера формируют область, плотность объектов внутри которого, по некоторому заданному порогу ε, превышает плотность за его пределами. По отношении к нашей задаче, в кластере должна присутствовать некоторая «непрерывность» данных, другими словами, если точки лежат недалеко друг от друга, то и значения функции в них не должны сильно отличаться.

Итак, пусть дано множество точек $\mathbb{P} = \{P_i(x_i^0, x_i^1, x_i^2, y_i)\ \ (i = 0,1, \dots, n)\}$ и $\mathbb{S}_r(P_i)$-сфера с центром в точке P_i радиуса r. Через $M_r(P_i)$ обозначим множество точек из $\mathbb{P}$, лежащих в сфере $\mathbb{S}_r(P_i)$, и пусть $|M_r(P_i)|$ их количество. Зафиксируем два числа – r и m и введем обозначения-

- Точку $P \in \mathbb{P}$ будем называть внутренней точкой кластера $K_{r,m}$, если $|M_r(P_i)| \geq m$.
- Точка $\tilde{P} \in \mathbb{P}$ непосредственно достижима по плотности от точки $P_i \in \mathbb{P}$, если $\tilde{P}$ лежит в сфере $\mathbb{S}_r(P_i)$, то есть $\tilde{P} \in M_r(P_i)$.
- Точка $\tilde{P} \in \mathbb{P}$ достижима по плотности от точки $\hat{P} \in \mathbb{P}$, если существует путь (упорядоченный набор точек) $\bar{P}_0, \bar{P}_1, \dots, \bar{P}_k$ $(\bar{P}_i \in \mathbb{P}, i = 0,1,2, \dots, k)$, где $\bar{P}_0 = \tilde{P}$ и $\bar{P}_k = \hat{P}$, все $\bar{P}_i$ при $i = 1,2, \dots, k$ являются внутренними точками кластера и каждая точка $\bar{P}_i$ непосредственно достижима по плотности от точки $\bar{P}_{i+1}$.

- Если точка $\tilde{P} \in \mathbb{P}$ достижима по плотности от внутренней точки $\hat{P} \in \mathbb{P}$, но выполняется $\left|M_r(\tilde{P})\right| < m$, то точка $\tilde{P}$ будет граничной точкой кластера.
- Если *P* является внутренней точкой, то она образует *кластер* вместе со всеми точками, которые от нее достижимы по плотности.

Достижимость по плотности – это транзитивное замыкание непосредственно достижимой по плотности точки. Если точка $\tilde{P}$ достижима по плотности из точки P_i, то это не говорит о том, что точка P_i достижима по плотности из точки $\tilde{P}$.

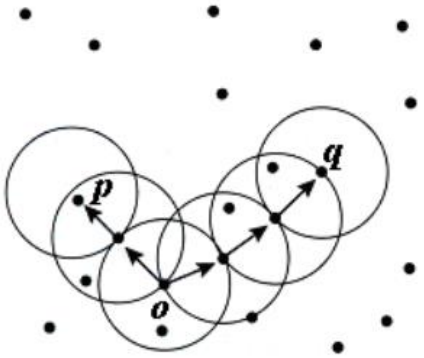


Рис.7. Иллюстрация достижимости по плотности с m=3.

Точка p достижима от o, но не наоборот.

Заметим, что для нашей задачи внутренней точкой кластера можно взять точку с максимальным рейтингом, а число m выбрать равным количеству вершин симплекса, то есть, равным 4. Выбор радиуса r обусловлен размером используемого корпуса ключевых слов, то есть, выберем значение радиуса таким, чтобы полученный кластер включил в себя все ключевые слова, для которых есть соответствие степени релевантности, и для прогноза потенциального ключевого слова будем использовать ***только*** слова полученного кластера.

Следующим этапом работы над потенциальными ключевыми словами является определение степени соответствия релевантности этого слова.

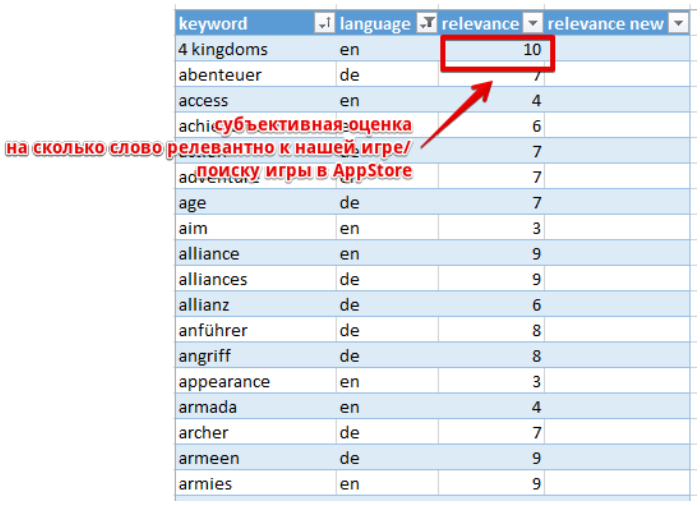
Релевантность ключевого слова– это содержательное соответствие поисковому запросу и ссылке, которая на нее ведет. Релевантность является важным критерием как в поисковой оптимизации, так и при размещении контекстной рекламы. Поисковым машинам важно, чтобы на продвигаемом ресурсе действительно присутствовал ключевой запрос, который пользователь ввел в строке поиска.

Значение релевантности можно записать в виде соотношения

$$r = P_r * (T + L),$$

где T – уровень соответствия внутренних критериев ресурса заданным требованиям поисковиков (релевантность текста), L – степень ссылочной ранжировки, т.е. степень совпадения текстов ссылок «на вход» поисковому запросу (ссылочная релевантность) и P_r – показатель внешнего мерила документа, который не зависит от запроса (авторитет ресурса).

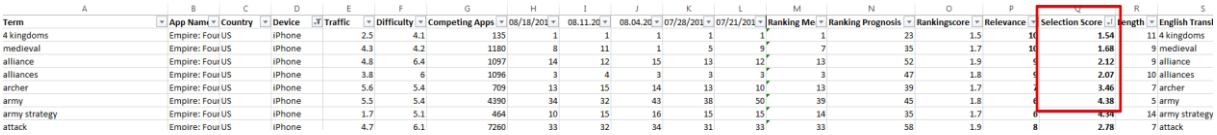
Так как используемые переменные достаточно субъективны, то будем считать, что релевантность определена по 10-бальной шкале, где 10 соответствует наивысшему уровню соответствия используемого ключевого слова нашему ресурсу.



keyword	language	relevance	relevance new
4 kingdoms	en	10	
abenteuer	de	7	
access	en	4	
achi		6	
		7	
adventure		7	
age	de	7	
aim	en	3	
alliance	en	9	
alliances	de	9	
allianz	de	6	
anführer	de	8	
angriff	de	8	
appearance	en	3	
armada	en	4	
archer	de	7	
armeen	de	9	
armies	en	9	

Рис. 8. Значения релевантности используемых ключевых слов

Опишем используемый критерий выбора ключевого слова с учетом его релевантности.



Term	App Nam	Country	Device	Traffic	Difficulty	Competing Apps	08/18/201	08.11.20	08.04.20	07/28/201	07/21/201	Ranking Me	Ranking Prognosis	Rankingscore	Relevance	Selection Score	Length	English Transl
4 kingdoms	Empire: Four	US	iPhone	2.5	4.1	135	1	1	1	1	1	1	23	1.5	10	1.54	11	4 kingdoms
medieval	Empire: Four	US	iPhone	4.3	4.2	1180	8	11	1	5	9	7	35	1.7	10	1.68	9	medieval
alliance	Empire: Four	US	iPhone	4.8	6.4	1097	14	12	15	13	12	13	52	1.9	9	2.12	9	alliance
alliances	Empire: Four	US	iPhone	3.8	6	1096	3	4	3	3	3	3	47	1.8	9	2.07	10	alliances
archer	Empire: Four	US	iPhone	5.6	5.4	709	13	15	14	13	10	13	39	1.7	7	3.46	7	archer
army	Empire: Four	US	iPhone	5.5	5.4	4390	34	32	43	38	50	39	45	1.8	6	4.38	5	army
army strategy	Empire: Four	US	iPhone	1.7	5.1	464	10	15	16	15	15	14	35	1.7	6	4.94	14	army strategy
attack	Empire: Four	US	iPhone	4.7	6.1	7260	33	32	34	31	33	33	58	1.9	8	2.78	7	attack

Рис.8. Вычисление критерия качества данного ключевого слова

Используемый критерий качества ключевого слова k с рейтингом R и релевантностью r вычисляется следующим образом

$$\varepsilon(k) = \sqrt{(10-r)^2 + (1-R)^2}$$

то есть, $\varepsilon(k)$ равна расстоянию от точки с координатами (R, r) до идеальной точки $(1{,}10)$- с наивысшим рейтингом и наилучшей релевантностью. Чем меньше значение $\varepsilon(k)$, тем лучше ключевое слово.

Заметим, что к достоинствам этой конструкции можно отнести только ее лаконичность.

Рассмотрим данную задачу подробнее. Так как, в данной постановке, между рейтингом и релевантностью существует обратная корреляционная зависимость, то критерием отбора ключевого слова должно служить увеличение по модулю коэффициента корреляции, то есть увеличение зависимости этих двух характеристик. То есть, если добавление конкретного ключевого слова увеличило зависимость, то есть коэффициент корреляции стал ближе к минус единице, то это слово хорошо описывает наш продукт, если коэффициент корреляции стал ближе к нулю, то это слово только ухудшает общую ситуацию, и его удаление только улучшит зависимость между рейтингом и релевантностью.

С другой стороны, если рейтинг слова низкий и релевантность близка к нулю, даже при высокой корреляционной зависимости такое слово никакой пользы не принесет,

например, при существующем множестве ключевых слов коэффициент корреляции равен -0.12230467, а при удалении слова с рейтингом 216, получаем -0.12122318, то есть корреляционная зависимость ухудшилась, но это говорит лишь о том, что низкая релевантность правильно соотносится с низким уровнем рейтинга и не более того.

Следовательно, выбор ключевого слова обусловлен выполнением, по крайней мере, двух условий, с одной стороны, повышением корреляционной связи между рейтингом и релевантностью, с другой, повышением среднего рейтинга набора ключевых слов. Таким образом, использование критерия качества (1) не позволяет решить ни одной из этих проблем.

Для решения поставленной задачи предложен следующий алгоритм.

Пусть каждому ключевому слову $k_i \in K$ соответствует пара (R_i, δ_i), где R_i – рейтинг этого слова и δ_i значение изменения коэффициента корреляции, равное

$$\delta_i = corr(K) - corr(K \backslash k_i),$$

где

$$corr(K) = \frac{cov(R, r)}{\sigma_R \sigma_r} = \frac{\sum_{i=1}^{n}(R_i - \bar{R})(r_i - \bar{r})}{\sqrt{\sum_{i=1}^{n}(R_i - \bar{R})^2 \sum_{i=1}^{n}(r_i - \bar{r})^2}}$$

и

$$\bar{R} = \frac{1}{n}\sum_{i=1}^{n} R_i\,, \qquad \bar{r} = \frac{1}{n}\sum_{i=1}^{n} r_i.$$

И пусть есть слово $\tilde{k}$ со значениями $(\tilde{R}, \tilde{\delta})$, претендующее стать ключевым словом.

Пусть, вначале, $p\left(\Delta^{+}(\tilde{\delta})\middle|R\right)$ – вероятность того, что при данном значении рейтинга R, встретится слово такое, что его значение изменения коэффициента корреляции будет больше $\tilde{\delta}$ и, соответственно, $p\left(\Delta^{-}(\tilde{\delta})\middle|R\right)$ – вероятность того, что при данном значении рейтинга R, значение изменения коэффициента корреляции будет меньше $\tilde{\delta}$.

Тогда задача определения рейтинга, при котором значение изменения коэффициента корреляции будет больше $\tilde{\delta}$, имеет вид

$$p\left(\Delta^{+}(\tilde{\delta})\middle|R\right) > p\left(\Delta^{-}(\tilde{\delta})\middle|R\right) \qquad (2)$$

В соответствии с теоремой Байеса,

$$p\left(\Delta^{+}(\tilde{\delta})\middle|R\right) = \frac{p\left(R\middle|\Delta^{+}(\tilde{\delta})\right)p\left(\Delta^{+}(\tilde{\delta})\right)}{p(R)}$$

где $p\left(R\middle|\Delta^{+}(\tilde{\delta})\right)$- вероятность встречи слова с рейтингом R при выполнении условия, что значение изменения коэффициента корреляции будет больше $\tilde{\delta}$, $p\left(\Delta^{+}(\tilde{\delta})\right)$- вероятность встречи ключевого слова, соответствующего условию $\Delta^{+}(\tilde{\delta})$ и $p(R)$- вероятность встречи слова с рейтингом R.

Тогда условие (2) примет вид

$$\frac{p(R|\Delta^+(\tilde{\delta}))p(\Delta^+(\tilde{\delta}))}{p(R)} > \frac{p(R|\Delta^-(\tilde{\delta}))p(\Delta^-(\tilde{\delta}))}{p(R)}$$

или, что то же самое,

$$p(R|\Delta^+(\tilde{\delta}))p(\Delta^+(\tilde{\delta})) > p(R|\Delta^-(\tilde{\delta}))p(\Delta^-(\tilde{\delta})).$$

В предположении, что слова по рейтингу распределены по нормальному закону,

$$p(R|\Delta^+(\tilde{\delta})) = \frac{1}{\sigma(R|\Delta^+(\tilde{\delta}))\sqrt{2\pi}} exp\left(-\frac{\left(R-\bar{R}^+(\tilde{\delta})\right)^2}{2\sigma^2(R|\Delta^+(\tilde{\delta}))}\right)$$

где

$$\sigma(R|\Delta^+(\tilde{\delta})) = \left(\sum\{1|i: \delta_i > \tilde{\delta}\}\right)^{-1} \sqrt{\sum\left\{\left(R_i-\bar{R}^+(\tilde{\delta})\right)^2 \middle| i: \delta_i > \tilde{\delta}\right\}}$$

и

$$\bar{R}^+(\tilde{\delta}) = \left(\sum\{1|i: \delta_i > \tilde{\delta}\}\right)^{-1} \sum\{R_i|i: \delta_i > \tilde{\delta}\}.$$

Аналогично,

$$p(R|\Delta^-(\tilde{\delta})) = \frac{1}{\sigma(R|\Delta^-(\tilde{\delta}))\sqrt{2\pi}} exp\left(-\frac{\left(R-\bar{R}^-(\tilde{\delta})\right)^2}{2\sigma^2(R|\Delta^-(\tilde{\delta}))}\right)$$

где

$$\sigma(R|\Delta^-(\tilde{\delta})) = \left(\sum\{1|i: \delta_i < \tilde{\delta}\}\right)^{-1} \sqrt{\sum\left\{\left(R_i-\bar{R}^-(\tilde{\delta})\right)^2 \middle| i: \delta_i < \tilde{\delta}\right\}}$$

и

$$\bar{R}^-(\tilde{\delta}) = \left(\sum\{1|i: \delta_i < \tilde{\delta}\}\right)^{-1} \sum\{R_i|i: \delta_i < \tilde{\delta}\}.$$

Соответственно,

$$p(\Delta^+(\tilde{\delta})) = \frac{1}{mes(K)}\sum\{1|i: \delta_i > \tilde{\delta}\}, \qquad p(\Delta^-(\tilde{\delta})) = \frac{1}{mes(K)}\sum\{1|i: \delta_i < \tilde{\delta}\},$$

где $mes(K)$- общее количество используемых ключевых слов множества K.

Используя полученные соотношения в (2), получаем неравенство

$$\frac{1}{\sigma(R|\Delta^+(\tilde{\delta}))\sqrt{2\pi}} exp\left(-\frac{\left(R-\bar{R}^+(\tilde{\delta})\right)^2}{2\sigma^2(R|\Delta^+(\tilde{\delta}))}\right)\frac{1}{mes(K)}\sum\{1|i: \delta_i > \tilde{\delta}\} >$$

$$\frac{1}{\sigma(R|\Delta^-(\tilde{\delta}))\sqrt{2\pi}} exp\left(-\frac{\left(R-\bar{R}^-(\tilde{\delta})\right)^2}{2\sigma^2(R|\Delta^-(\tilde{\delta}))}\right)\frac{1}{mes(K)}\sum\{1|i: \delta_i < \tilde{\delta}\},$$

решая его, получаем значение $R(\tilde{\delta})$, такое, что при $R > R(\tilde{\delta})$ принимаем решение, что корреляционная зависимость между рейтингом и релевантностью будет лучше, в противном случае, хуже. Обозначим через $\aleph(\tilde{\delta})$ множество ключевых слов k_i, таких, что $R_i < R(\tilde{\delta})$.

Пусть теперь $p(\Delta^{+}(\tilde{R})|\delta)$ – вероятность того, что при данном значении изменения коэффициента корреляции δ, встретится слово такое, что его рейтинг будет больше $\tilde{R}$ и, соответственно, $p(\Delta^{-}(\tilde{R})|\delta)$.

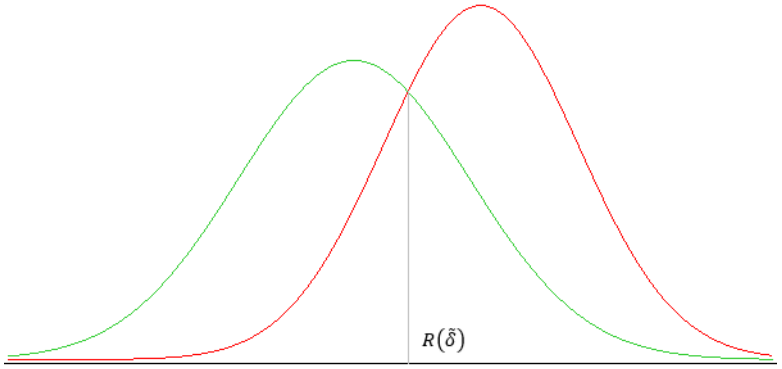


Рис.9. Иллюстрация нахождения точки $R(\tilde{\delta})$

Сделав аналогичные построения, получаем значение $\delta(\tilde{R})$, такое, что при $\delta > \delta(\tilde{R})$ принимаем решение, что рейтинг будет выше, а в противном случае, ниже. Через $\aleph(\tilde{R})$ обозначим множество ключевых слов k_i, таких, что $\delta_i < \delta(\tilde{R})$.

Результатом полученных построений является множество ключевых слов

$$\Re(\tilde{R}, \tilde{\delta}) = \aleph(\tilde{R}) \cap \aleph(\tilde{\delta}),$$

таких, что замена любого ключевого слова $k \in \Re(\tilde{R}, \tilde{\delta})$ ключевым словом $\tilde{k}$ только улучшит качество используемого корпуса ключевых слов. Выбор слова, удаляемого из используемого корпуса ключевых слов определяется необходимостью или увеличить средний рейтинг (попадание в верхнюю часть списка поиска по ключевым словам) или усилить корреляционную связь между релевантностью и рейтингом (соответствия ключевым словам содержанию ресурса).

Литература

1. Delone, B. N. Geometry of positive quadratic forms. *Uspekhi Matematicheskikh Nauk*, (3), 1937, с. 16-62.
2. Скворцов А.В. Триангуляция Делоне и ее применение.- Томск: Изд-во Том. Ун-та, 2002.-128 с.
3. Клячин, Владимир Александрович, Александр Александрович Широкий. "Триангуляция Делоне многомерных поверхностей." *Вестник Самарского государственного университета. Естественнонаучная серия* 4 (78), 2010 с. 51-55.

4. Хорошевский, Владимир Федорович. "Пространства знаний в сети Интернет и Semantic Web (Часть 1)." *Искусственный интеллект и принятие решений* 1 (2008): 80-97.
5. Абрамов, Егор Геннадьевич. "Подбор ключевых слов для научной статьи."*Научная периодика: проблемы и решения* 1.2 (2011): 35-40.
6. Шумейко А.А. Интеллектуальный анализ данных (введение в Data Mining)/ А.А.Шумейко, С.Л.Сотник.-Днепропетровск: Белая Е.А., 2012.-212 с.
7. NG, TS Eugene; ZHANG, Hui. Predicting Internet network distance with coordinates-based approaches. In: *INFOCOM 2002. Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*. IEEE, 2002. p. 170-179.
8. Martin Ester, Hans-Peter Kriegel, J&g Sander, Xiaowei Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, KDD'96. – P. 226-231