Семантический анализ текстов на примере интернет-запросов

Боровых К.О., студентка, Пермский национальный исследовательский политехнический университет, г. Пермь, Россия
Плотников А.В., к.э.н., Пермский национальный исследовательский политехнический университет, г. Пермь, Россия

Аннотация. В статье представлено описание усовершенствованного алгоритма поисковой выдачи для продвижения web-ресурсов на основе релевантности предоставляемого контента и ключевых слов. Описана методика индексирования LSI, где поисковые системы анализируют уникальность, насыщенность и содержание текста. В результате развития LSI большое внимание SEO-специалистов уделяется контент-анализу, где определяется плотность ключевых фраз в тексте, мета-тегах, анкорах, заголовках «Title» и H1-H3. Раскрыта тема анкоров, заключающее в себе семантическое ядро, которые являются основным элементом SEO-оптимизации сайта. В статье также представлен алгоритм TF-IDF, с помощью которого определяется значимость слова в документе относительно общей подборки в базе, и упрощается поиск путем выявления ключевых слов в документе, а также проводя индексацию стоп-слов и малозначимых фраз, имеющих низкое значение TF-IDF. Алгоритм TF-IDF продемонстрирован практическом примере.

Ключевые слова: интернет-маркетинг, поисковая оптимизация, контент-анализ, семантический анализ.

Semantic Text Analysis on the Example of Internet Queries

Borovykh K.O., student, Perm National Research Polytechnic University, Perm, Russia
Plotnikov A.V., Candidate of Economic Sciences, Perm National Research Polytechnic

University, Perm, Russia

Annotation. The paper considers the description of the improved algorithm of search delivery for promotion of web-resources based on relevance of the given content and keywords is presented. The technique of LSI indexing is described, where search engines analyze the uniqueness, richness and content of the text. Because of the development of LSI, a great deal of attention is paid to the SEO professionals content analysis, which determines the density of key phrases in the text, meta tags, anchors, title «Title» and H1-H3. The theme of anchors, which includes the semantic core, is the main part of search engine optimization of the site. The article also presents the TF-IDF algorithm, which helps to determine the significance of the word in the document relative to the general collection in the database and simplifies the search by identifying keywords in the document, and by indexing stop words and low-value phrases that have a low TF-IDF value. Algorithm TF-IDF is demonstrated a practical example.

Keywords: internet marketing, search engine optimization, content analysis, semantic analysis.

Введение

В настоящее время на просторах интернета сложились условия жесткой быстрорастущей конкуренции, что заставляет владельцев сайтов проводить качественную оптимизацию и буквально вступать «на тропу войны» за высокие показатели. Основным аспектом, определяющим эффективность площадки, является соответствие контента ключевым словам, по которым осуществляется продвижение web-ресурса.

Уровень соответствия искомого и найденного называют релевантностью. Поисковые системы в процессе индексации страниц сайтов выдают данные с помощью алгоритмов, которые ранжируют результаты поиска по принципу: вверху — наиболее релевантные ресурсы, ниже — страницы с меньшей степенью

смыслового соответствия пользовательскому запросу.

Сущность LSI

Благодаря непрерывным процессам совершенствования информационных систем на смену устаревшим алгоритмам поисковой выдачи пришли новые, в связи с этим для продвижения сайтов теперь всё чаще используется LSI-копирайтинг.

LSI с английского означает «латентное, скрытое семантическое индексирование» – это методика индексирования, в соответствии с которой поисковые системы анализируют не только уникальность и насыщенность ключевыми словами, но и само содержание текста. Таким образом, в результатах выдачи первые позиции занимают качественные статьи, соответствующие ключевому запросу. В LSI-копирайтинге особое внимание уделяется не наличию ключевых слов, а содержанию: чем интереснее и привлекательнее текст, тем выше качество контента в базах данных поисковых систем.

В одной из работ Chen C. M. [1], раскрывая сущность латентного семантического анализа (LSA [4]), отметил, что LSI [2] [3] – это подход к поиску информации, который считается наиболее эффективным инструментом корреляции с целью получения соответствующих документов. Латентное семантическое индексирование представляет собой систему и интегрирует такие элементы как: сбор документов и их первичная обработка; разделение сингулярных значений (SVD); многоязычная обработка; методика доступа на основе дерева запросов сходства. Несмотря на то, что большее внимание автора уделено изучению LSI, но внушительный объем материала посвящен алгоритмической (теоретической) основе модели. Практические же вопросы, касающиеся проблем внедрения, которые возникают при создании системы реализации, были недостаточно полно раскрыты.

BifetFiguerol A. C. [5] на примере работы Google (через API) проанализировал влияние функций страницы на ранжирование результатов

поисковой выдачи. Используя статистические методы, он исследовал несколько запросов по различным категориям. Автор назвал проблему изучения скрытых оценок проблемой двоичной классификации. Чтобы получить объективную оценку качества предиктора, BifetFiguerol A. C. разделил материал на набор тестов и обучающий набор. Как показали результаты, при использовании только наблюдаемых признаков, скоринговая функция не может быть хорошо аппроксимирована.

По мнению Saffer J., Gibbs A. и Ryley J. F. [6], LSI может применяться в патентном поиске, чтобы преодолеть недостатки булевого поиска и обеспечить более точный поиск. Установлено, что LSI соединяет векторную пространственную модель (VSM) извлечения документов с разложением одного значения (SVD), используя линейные методы алгебры, для выявления отношений словосочетаний и слов в текстовом материале. Результаты можно улучшить путем применения методик устранения неоднозначностей в языке, а также используя текстовую кластеризацию и настройки параметров SVD для соответствующего корпуса.

Контент–анализ

Достаточно давно началась борьба за качественное наполнение интернета. Поисковые системы отбирают площадки, которые имеют действительно полезную информацию, и ранжируют сайты с наиболее высоким уровнем качества. В связи с этим контенту сейчас уделяется довольно пристальное внимание. Анализ контента проводится SEO-специалистами с целью оценки наполнения и структуры сайта с учетом требований поисковых систем. Контент-анализ определяет плотность ключевых фраз в текстовом материале, мета-тегах, анкорах, заголовках «Title» и h1-h3. Затем осуществляется поисковая оптимизация страниц.

Особую роль в системе ранжирования страниц играет плотность или тошнота ключевых слов. Недостаточная тошнота ключевых слов приводит к выводу страниц на низких позициях в результатах поиска, а высокий процент

вхождений может вынудить системы наложить фильтры на сайт или даже исключить его из поисковой выдачи.

Анкоры

Одним из основных инструментов SEO-оптимизации сайта является анкор. Он представляет собой текст ссылки, который находится между открывающим и закрывающим тегами. Все сайты связаны гипертекстовыми ссылками, которые имеют формат:

текст ссылки.

Анкор заключает в себе семантическое ядро, в соответствии с которым сайту присваивается ранг в поисковых системах. К примеру, в процессе поиска сайтов в системе Яндекс часто встречаются площадки, в описании которых указано «Найден по ссылке». Если страница не соответствует пользовательскому запросу, но она релевантна ссылочному ранжированию, то система её всё равно будет отображать в поисковой выдаче.

Чтобы размещать анкоры на авторитетных ресурсах, необходимо грамотно составлять их. Для этого следует придерживаться рекомендаций:

- не допустимы грамматические ошибки;
- в анкоре необходимо избегать знаков препинания;
- не следует писать одинаковый текст в анкоре и рядом с ним;
- необходимо выделять слова прописными буквами;
- есть смысл иногда разбавлять анкоры другими словами;
- нельзя оставлять анкоры на сторонних площадках, не связанных с тематикой вашего сайта.

Виды анкоров

С учетом написания или типа вхождения выделяют несколько разновидностей анкоров. Неразбавленный анкор – это ключевая фраза, которая

имеет прямое вхождение и используется без других слов. Разбавленным анкором называют текст ссылки, который имеет и ключевые слова, и дополнительный текст. К примеру, для поискового запроса «кованые ворота», неразбавленный анкор – «кованые ворота», а «купить кованые ворота в Москве цена» и «кованые ворота в Москве недорого» – это разбавленные анкоры.

В соответствии с видами поисковых запросов и географии таргетинга анкоры разделяют на:

1. Высокочастотные (ВЧ) – имеют более тысячи запросов в месяц, состоят обычно из 1-2 слов, являются неразбавленными и характеризуются прямым вхождением ключевых слов, к примеру: «разработка приложения».

2. Среднечастотные (СЧ) имеют 100-1000 запросов в месяц, содержат от 2 до 4 слов и представляют собой разбавленные анкоры для ВЧ запросов: «разработка мобильного приложения».

3. Низкочастотные (НЧ) – до 100 запросов в месяц, это разбавленные ВЧ и СЧ анкоры, в которых более 4 слов, например: «стоимость разработки мобильного приложения».

Чтобы сэкономить бюджет на продвижение сайта, оптимизаторы стараются составить анкор из нескольких ключевых запросов.

Пассаж – это фрагменты текста, отделяющиеся от других текстов на странице html-тэгом или знаком препинания (например, «!», «?», «...» или точкой, после которой стоит пробел). Чтобы seo-тексты имели высокую эффективность, очень важно их составлять с учетом особенностей работы поисковиков с пассажами. Основное правило сводится к тому, что нельзя делить поисковый запрос на разные пассажи. В процессе индексации текстовой контент разбивается на пассажи, в которых система ищет фразы, соответствующие поисковому запросу. Если такие слова выявлены, и они недалеко друг от друга расположены, то текст классифицируется как релевантный.

Алгоритм TF-IDF

Алгоритм TF-IDF – это формула, по которой определяется значимость слова в документе относительно общей подборки в базе. TF – это частотность термина, которая характеризует плотность вхождений некоторого слова в отдельном документе и представляет собой отношение числа вхождений термина к общей сумме слов в документе.

$$tf(t,d) = \frac{n_i}{\sum_k n_k}$$

n_i – количество вхождений слова в документе,

$\sum_k n_k$ – суммарное число слов в этом документе.

IDF – inverse document frequency – обратная частота документа относительно запроса, то есть отношение всей подборки документов в поисковой базе к тем, что содержат в себе заданный термин. Оценка IDF снижает вес широко употребляемых слов и выявляет релевантность страницы ключевому запросу.

$$IDF(q_i) = log\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5}$$

где N – общее количество документов в поисковой базе,

$n(q_i)$ — количество документов, содержащих заданный термин «q_i».

В результате получаем оценку весомости конкретного термина в пределах одного документа.

$$tf - IDF(t,d,D) = tf(t,d) * IDF(t,D)$$

Данная формула позволяет упростить поиск путем выявления ключевых слов в документе, а также проводя индексацию стоп-слов и малозначимых фраз,

имеющих низкое значение TF-IDF [7].

Вычислим Вес слова на примере термина «закон». Исходные данные: страница сайта размещает 2 тысячи слов, среди которых термин «закон» встречается двадцать раз.

Таким образом, показатель TF равен 20/2000 = 0,01.

Если в интернете 8 млрд страниц, среди которых термин «закон» встречается на 4 млн страниц, тогда DF составит 4000000/8000000000 = 0,0005.

Далее нетрудно рассчитать Вес слова: TF/DF = 0,01/0,0005 = 20.

Этапы применения метода:

1. Поисковая система выдает топовые сайты, по которым проводится сопоставление основных показателей.

2. Затем необходимо рассчитать главные характеристики (Title, анкоры, короткие пассажи) документа и интернет-ресурсов из Top-10.

3. Формируются среднестатистические данные по всем словам, которые встречаются у конкурентов. Проводятся необходимые подсчеты, а показатели вхождения каждого термина в блоках и во всем документе переводятся в % значение.

4. На основании полученных данных создается техническое задание для копирайтера, в котором указываются требования по объему текста, пассажам, а также по количеству ссылок и точности вхождений в: текст, Title, description, заголовки h2 и h3.

5. Если требуется провести сравнение продвигаемого сайта с топовыми в результатах поисковой выдачи, то следует добавить его в это пространство.

6. Определять сопоставление следует по пассажам, ссылкам, тексту, а также по количеству знаков и слов в Title.

7. На основе полученной информации теперь можно проводить оценку оптимизации сайта под определенный запрос и создавать перечень рекомендаций по доработке страницы.

В рамках анализа взят ключевой запрос «разработка мобильных приложений стоимость». Первый этап исследования проводился в поисковой системе Яндекс, в качестве геопозиции выбран город Москва.

Анализируемые страницы, определенные в топ10 Яндекс:

http://woxapp.com/stoimost-razrabotki-prilozheniya/

http://livetyping.com/ru/blog/skolko-stoit-razrabotat-mobilnoe-prilojenie

http://habr.com/post/314916/

http://thebestapp.ru/calculator/

http://wnfx.ru/

http://g-application.com/

http://freelance.youdo.com/programming/mobile/

http://appfox.ru/razrabotka-prilozhenij/razrabotka-mobilnyh-prilozhenij.html

http://arcanite.ru/getprice/

URL для сравнения: app-android.ru

Таблица 1

Сравнение общих показателей в Яндекс и Google

	Title	**Ссылки**	**Пассажи**	**Текст**
Яндекс	8 слов 58 символов	102 штук 332 слов 1858 символов	105 (69) пассажей 425 (258) слов 3384 (1996) символов	16 (8) абзацев 758 (337) слов 6251 (2803) символов
Google	7(5) слов 57(46)символов	71 (68) штук 107 (39) слов 837 (267)символов	86 (69) пассажей 531 (258) слов 4117 (1996)символов	13 (8) абзацев 646 (337) слов 4950 (2803) символов

Таблица 2

Детализированный результат анализа по органической выдаче в Яндекс

Слово	**TF*IDF**	**IDF**	**Вхождений, %**	**Title**	**Анкоры**	**Пассажи**	**Текст**
			Медиана / среднее (сравниваемый URL)				
приложений	0.0237 (0.0273)	1.94	1.22 / 1.15 (1.41)	1 / 0.75 (1)	2 / 2.63 (0)	4 / 4.63 (4)	4 / 3.25 (4)
разработки	0.0149 (0.0025)	1.57	0.95 / 0.87 (0.16)	0 / 0.38 (0)	0 / 0.75 (0)	4 / 3.63 (0)	4 / 4.63 (1)

мобильных	0.0108 (0.0134)	1.72	0.63 / 0.73 (0.78)	1 / 0.88 (1)	1 / 1 (0)	4 / 2.5 (2)	2 / 1.63 (2)
стоимость	0.0083 (0.0028)	0.88	0.95 / 0.88 (0.31)	1 / 0.5 (0)	0 / 0.88 (0)	2 / 4.25 (1)	4 / 4 (1)
ios	0.0063 (0.006)	1.93	0.32 / 0.28 (0.31)	0 / 0.25 (0)	0 / 0.5 (1)	0 / 0.88 (1)	1 / 1.5 (0)
проекта	0.005 (0.002)	1.28	0.39 / 0.42 (0.16)	0 / 0 (0)	0 / 0.13 (0)	1 / 1.75 (1)	1 / 3.63 (0)
android	0.0046 (0.0177)	1.41	0.32 / 0.27 (1.25)	0 / 0.25 (0)	1 / 0.88 (0)	1 / 1 (7)	1 / 1.63 (1)
дизайн	0.0034 (0.0016)	1.01	0.33 / 0.35 (0.16)	0 / 0 (0)	1 / 0.88 (0)	2 / 2 (1)	0 / 1.75 (0)
программирование	0.0029 (0.0028)	1.82	0.16 / 0.17 (0.16)	0 / 0 (0)	0 / 0.25 (0)	0 / 1 (1)	0 / 1.25 (0)
веб	0.0028 (0)	1.98	0.14 / 0.15 (0)	0 / 0 (0)	0 / 0.5 (0)	0 / 0.63 (0)	0 / 1 (0)
компании	0.0026 (0.0028)	0.59	0.43 / 0.73 (0.47)	0 / 0 (0)	2 / 1.38 (1)	1 / 1.13 (0)	0 / 0.75 (2)
создание	0.0025 (0.0016)	1.05	0.24 / 0.24 (0.16)	0 / 0.13 (0)	0 / 0.63 (0)	1 / 0.75 (1)	1 / 1.38 (0)
прототипа	0.0023 (0)	2.47	0.09 / 0.09 (0)	0 / 0 (0)	0 / 0 (0)	0 / 0.13 (0)	0 / 0.88 (0)
проект	0.0023 (0)	1.15	0.2 / 0.17 (0)	0 / 0 (0)	0 / 0.13 (0)	2 / 1.25 (0)	1 / 1.38 (0)
интерфейса	0.0022 (0.0032)	2.04	0.11 / 0.1 (0.16)	0 / 0 (0)	0 / 0.13 (0)	1 / 0.88 (1)	0 / 0.5 (0)
платформ	0.0019 (0)	2.47	0.08 / 0.05 (0)	0 / 0 (0)	0 / 0 (0)	1 / 0.5 (0)	0 / 0.38 (0)
заказать	0.0018 (0)	0.93	0.2 / 0.15 (0)	0 / 0.38 (0)	0 / 0.63 (0)	0 / 0.5 (0)	0 / 0.25 (0)
разработчики	0.0016 (0)	2.09	0.08 / 0.06 (0)	0 / 0 (0)	0 / 0 (0)	0 / 0.13 (0)	1 / 0.88 (0)
технического	0.0016 (0.0054)	1.71	0.09 / 0.08 (0.31)	0 / 0 (0)	0 / 0.13 (0)	0 / 0.5 (1)	0 / 0.63 (1)
цена	0.0016 (0.0012)	0.74	0.22 / 0.38 (0.16)	0 / 0.38 (0)	0 / 0.38 (0)	0 / 0.38 (0)	1 / 0.75 (1)
этапах	0.0016 (0)	2.04	0.08 / 0.07 (0)	0 / 0 (0)	0 / 0 (0)	0 / 0.13 (0)	0 / 0.5 (0)
пользователей	0.0016 (0.0016)	1	0.16 / 0.1 (0.16)	0 / 0 (0)	0 / 0 (0)	1 / 0.5 (1)	1 / 1.13 (0)
тестирование	0.0015 (0.0061)	1.96	0.08 / 0.07 (0.31)	0 / 0 (0)	0 / 0 (0)	1 / 0.63 (2)	0 / 0.38 (0)
платформы	0.0015 (0)	1.95	0.08 / 0.06 (0)	0 / 0 (0)	0 / 0 (0)	0 / 0.38 (0)	0 / 0.5 (0)
требований	0.0015 (0)	1.57	0.1 / 0.1 (0)	0 / 0 (0)	0 / 0 (0)	0 / 0.25 (0)	1 / 1.13 (0)
заказчика	0.0014 (0.0028)	1.78	0.08 / 0.06 (0.16)	0 / 0 (0)	0 / 0.13 (0)	0 / 0.5 (1)	0 / 0.25 (0)
сколько	0.0014 (0)	0.97	0.14 / 0.12 (0)	0 / 0.13 (0)	0 / 0.13 (0)	1 / 0.88 (0)	0 / 0.75 (0)
задания	0.0014 (0.0053)	1.69	0.08 / 0.17 (0.31)	0 / 0 (0)	0 / 0.13 (0)	0 / 1.75 (1)	0 / 1 (1)
москве	0.0014 (0.0013)	0.84	0.16 / 0.35 (0.16)	1 / 0.5 (1)	0 / 0.25 (0)	0 / 0.25 (0)	0 / 0.5 (0)
windows	0.0013 (0)	1.42	0.09 / 0.08 (0)	0 / 0 (0)	0 / 0.5 (0)	0 / 0.5 (0)	0 / 0 (0)
поддержка	0.0013 (0.0018)	1.18	0.11 / 0.11 (0.16)	0 / 0 (0)	0 / 0.25 (0)	0 / 0.63 (1)	0 / 0.38 (0)
реализации	0.0012 (0)	1.45	0.08 / 0.06 (0)	0 / 0 (0)	0 / 0 (0)	0 / 0.25 (0)	0 / 0.63 (0)
прототип	0.0012 (0)	2.47	0.05 / 0.06 (0)	0 / 0 (0)	0 / 0 (0)	0 / 0.25 (0)	0 / 0.5 (0)

Релевантность = (exact * 0.3 + partial * 0.25 + term * 0.25 + word_count * 0.05 + w2v * 0.1 + gunning * 0.05) * 100%

exact = 0 - точные вхождения запроса (в т.ч. в других словоформах)

partial = 0.791803 - частичные вхождения запроса (через слово и части запроса)

term = 0.928572 - наличие в документе медианных лемм (топ-50 по tf_idf) из документов серпа

w2v = 0.96188 - похожесть среднего word2vec вектора

word_count = 0.997885 - похожесть распределения длин текста по типам фрагментов текста

gunning = 0.904574 - похожесть индекса туманности Ганнинга

Релевантность = (exact * 0.3 + partial * 0.25 + term * 0.25 + word_count * 0.05 + w2v * 0.1 + gunning * 0.05) * 100%

Релевантность по Google

http://woxapp.com/stoimost-razrabotki-prilozheniya/

http://stfalcon.com/ru/blog/post/how-much-to-develop-app

http://punicapp.com/calculator

http://magora-systems.ru/expertise/mobile-development/

http://livetyping.com/ru/blog/skolko-stoit-razrabotat-mobilnoe-prilojenie

http://appsgroup.ru/index.php/article/view/4/

http://apptractor.ru/info/articles/skolko-stoit-mobilnoe-prilozhenie-v-2017-godu.html

http://spaceshipapps.ru/

Таблица 3

Детализированный результат анализа по органической выдаче в Google

Слово	**TF*IDF**	**IDF**	**Вхождений, %**	**Title**	**Анкоры**	**Пассажи**	**Текст**
			Медиана / среднее (сравниваемый URL)				
приложений	0.0183 (0.0273)	1.94	0.94 / 0.94 (1.41)	0 / 0.38 (1)	2 / 2.63 (0)	4 / 4.38 (4)	4 / 3.63 (4)
разработки	0.0149 (0.0025)	1.57	0.95 / 0.81 (0.16)	0 / 0.13 (0)	0 / 0.88 (0)	6 / 5 (0)	4 / 5.38 (1)
приложения	0.0148 (0.0255)	1.48	1 / 1.41 (1.72)	0 / 0.38 (0)	1 / 1.38 (1)	6 / 7.63 (5)	6 / 7.25 (5)
мобильных	0.0108 (0.0134)	1.72	0.63 / 0.59 (0.78)	0 / 0.38 (1)	1 / 1.13 (0)	2 / 2.38 (2)	3 / 2.75 (2)
ios	0.0077 (0.006)	1.93	0.4 / 0.5 (0.31)	0 / 0 (0)	1 / 1.38 (1)	2 / 2.25 (1)	2 / 2.38 (0)
android	0.0067 (0.0177)	1.41	0.47/0.63 (1.25)	0 / 0 (0)	1 / 1.38 (0)	4 / 3.13 (7)	3 / 3.13 (1)
веб	0.0052 (0)	1.98	0.26 / 0.26 (0)	0 / 0 (0)	0 / 0.5 (0)	2 / 1.25 (0)	0 / 0.75 (0)
проекта	0.0051 (0.002)	1.28	0.4 / 0.38 (0.16)	0 / 0 (0)	0 / 0.13 (0)	3 / 2.13 (1)	2 / 3.5 (0)
стоимость	0.0046 (0.0028)	0.88	0.52 / 0.65 (0.31)	0 / 0 (0)	0 / 0.38 (0)	5 / 5.13 (1)	1 / 3.13 (1)
дизайн	0.0044 (0.0016)	1.01	0.43 / 0.36 (0.16)	0 / 0.13 (0)	1 / 0.88 (0)	2 / 2.13 (1)	2 / 2.25 (0)
iphone	0.0035 (0)	1.48	0.24 / 0.25 (0)	0 / 0 (0)	0 / 0.25 (0)	0 / 1.13 (0)	1 / 1.13 (0)
ui	0.0034 (0)	2.47	0.14 / 0.09 (0)	0 / 0 (0)	0 / 0 (0)	0 / 1 (0)	0 / 0.5 (0)
портфолио	0.0033 (0.012)	1.91	0.17 / 0.19 (0.63)	0 / 0 (0)	0 / 0.5 (4)	0 / 0.25 (0)	0 / 0.25 (0)
ipad	0.0026 (0)	1.68	0.16 / 0.3 (0)	0 / 0 (0)	0 / 0.38 (0)	0 / 1.63 (0)	0 / 0.5 (0)

создание	0.0025 (0.0016)	1.05	0.24 / 0.28 (0.16)	0 / 0 (0)	0 / 0.63 (0)	1 / 1.13 (1)	1 / 1 (0)
разработчиков	0.002 (0)	2.12	0.1 / 0.14 (0)	0 / 0 (0)	0 / 0.25 (0)	0 / 1 (0)	1 / 0.88 (0)
оценки	0.002 (0)	1.42	0.14 / 0.12 (0)	0 / 0 (0)	0 / 0.13 (0)	0 / 0.5 (0)	0 / 0.88 (0)
интерфейсов	0.0019 (0.0039)	2.47	0.08 / 0.05 (0.16)	0 / 0 (0)	1 / 0.5 (0)	0 / 0 (0)	0 / 0.38 (1)
бизнеса	0.0019 (0.0041)	1.3	0.14 / 0.13 (0.31)	0 / 0 (0)	0 / 0.38 (0)	0 / 0.5 (1)	0 / 0.5 (1)
проекты	0.0018 (0)	1.28	0.14 / 0.14 (0)	0 / 0 (0)	1 / 0.88 (0)	0 / 0.63 (0)	0 / 0.38 (0)
этапе	0.0018 (0.0024)	1.56	0.12 / 0.09 (0.16)	0 / 0 (0)	0 / 0 (0)	0 / 0.38 (0)	1 / 1.13 (1)
пользователей	0.0017 (0.0016)	1	0.17 / 0.16 (0.16)	0 / 0 (0)	0 / 0.13 (0)	1 / 1.13 (1)	1 / 1.38 (0)
целевой	0.0017 (0)	2.25	0.08 / 0.04 (0)	0 / 0 (0)	0 / 0 (0)	0 / 0.13 (0)	0 / 0.38 (0)
сколько	0.0017 (0)	0.97	0.17 / 0.13 (0)	0 / 0.38 (0)	0 / 0 (0)	1 / 1.38 (0)	0 / 0.63 (0)
дизайнер	0.0017 (0)	2.16	0.08 / 0.05 (0)	0 / 0 (0)	0 / 0.13 (0)	0 / 0.5 (0)	0 / 0.5 (0)
аудитории	0.0016 (0)	2.09	0.08 / 0.05 (0)	0 / 0 (0)	0 / 0 (0)	0 / 0.13 (0)	0 / 0.5 (0)
работать	0.0016 (0)	1.07	0.15 / 0.13 (0)	0 / 0 (0)	0 / 0 (0)	1 / 0.75 (0)	1 / 1.25 (0)
разработчики	0.0016 (0)	2.09	0.08 / 0.07 (0)	0 / 0 (0)	0 / 0 (0)	0 / 0.13 (0)	1 / 1 (0)
команда	0.0015 (0.0066)	1.41	0.11 / 0.19 (0.47)	0 / 0 (0)	1 / 0.88 (0)	0 / 0.25 (0)	1 / 0.75 (3)
бизнес	0.0015 (0.0051)	1.08	0.14 / 0.12 (0.47)	0 / 0 (0)	0 / 0.13 (1)	0 / 0.38 (1)	1 / 1 (1)
работы	0.0015 (0.0008)	0.49	0.31 / 0.27 (0.16)	0 / 0 (0)	0 / 0.38 (0)	1 / 1.88 (1)	1 / 2.88 (0)
продукта	0.0015 (0)	1.62	0.1 / 0.09 (0)	0 / 0 (0)	0 / 0.13 (0)	0 / 0.75 (0)	1 / 1.38 (0)
тестирование	0.0015 (0.0061)	1.96	0.08 / 0.09 (0.31)	0 / 0 (0)	0 / 0.13 (0)	1 / 0.63 (2)	0 / 0.5 (0)

Релевантность = (exact * 0.3 + partial * 0.25 + term * 0.25 + word_count * 0.05 + w2v * 0.1 + gunning * 0.05) * 100%

exact = 0 - точные вхождения запроса (в т.ч. в других словоформах)

partial = 0.911427 - частичные вхождения запроса (через слово и части запроса)

term = 0.95692 - наличие в документе медианных лемм (топ-50 по tf_idf) из документов серпа

w2v = 0.959804 - похожесть среднего word2vec вектора

word_count = 0.999696 - похожесть распределения длин текста по типам фрагментов текста

gunning = 0.892084 - похожесть индекса туманности Ганнинга

Таблица 4

Сравнение запросов по показателям

Запрос	**Релевантность**	**Похожесть word2vec**	**Похожесть туманности Ганнинга**
Yandex (Москва)	88.8%	96.2%	90.5%
Google (Москва)	94%	96%	89.2%

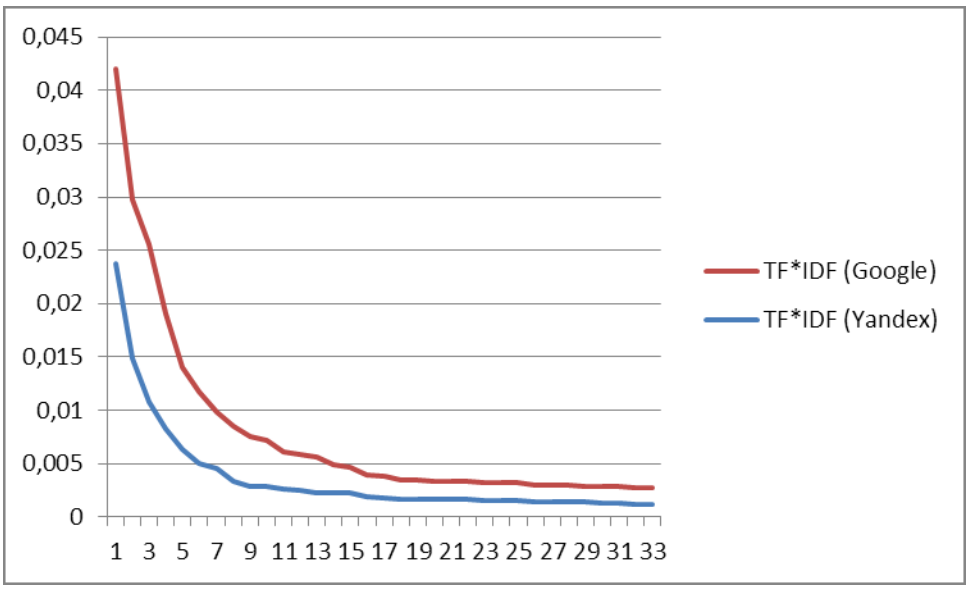

Рис. 1 – Сравнение TF*IDFGoogle и Яндекс

Библиографический список

1. Chen C.M. et al. Telcordia LSI engine: Implementation and scalability issues // Research Issues in Data Engineering, 2001. Proceedings. Eleventh International Workshop on. – IEEE, 2001. – С. 51-58.

2. Что такое LSI или латентно-семантический индекс для лучшего понимания контекста страницы https://seoprofy.ua/blog/wiki/what-is-lsi-keywords

3. Латентно-семантическое индексирование cropas.by/seo-slovar/lsi

4. Латентно-семантический анализ https://dic.academic.ru/dic.nsf/ruwiki/595989

5. BifetFiguerol A.C. et al. An analysis of factors used in search engine ranking. – 2005.

6. Ryley J. F., Saffer J., Gibbs A. Advanced document retrieval techniques for patent research //World Patent Information. – 2008. – Т. 30. – №. 3. – С. 238-243.

7. Проверка TF-IDFhttps://ru.megaindex.com/support/faq/tf-idf

8. Плотников А.В. Контент-анализ web-документов согласно поисковым запросам // Московский экономический журнал 5/2017 http://qje.su/otraslevaya-i-regionalnaya-ekonomika/moskovskij-ekonomicheskij-zhurnal-5-2017-3/

References

1. Chen C.M. et al. Telcordia LSI engine: Implementation and scalability issues // Research Issues in Data Engineering, 2001. Proceedings. Eleventh International Workshop on. – IEEE, 2001. – P. 51-58.

2. What is the LSI or latent-semantic index for better understanding of the context of the page https://seoprofy.ua/en/blog/wiki/what-is-lsi-keywords

3. Latent-semantic indexing of cropas.by/seo-slovar/lsi

4. Latent-semantic analysis https://dic.academic.ru/dic.nsf/ruwiki/595989

5. BifetFiguerol A.C. C. et al. An analysis of factors used in search engine ranking. – 2005.

6. Ryley J.F., Saffer J., Gibbs A. Advanced document retrieval techniques for patent research. World Patent Information. – 2008. – V. 30. – No. 3. – P. 238-243.

7. Check TF-IDF https: //ru.megaindex.com/support/faq/tf-idf

8. Plotnikov A.V. Content analysis of web-documents according to search requests // Moscow economic journal 5/2017 http://qje.su/otraslevaya-i-regionalnaya-ekonomika/moskovskij-ekonomicheskij-zhurnal-5-2017-3/