ПРОГРАММА АВТОМАТИЗИРОВАННОЙ ГЕНЕРАЦИИ КЛЮЧЕВЫХ СЛОВ НА ОСНОВЕ КОЛЛЕКЦИИ ВЕБ-ДОКУМЕНТОВ

***Александр Сергеевич Журавель**, разработчик,*
E-mail: aszhuravel@yandex.ru,
ООО «Викимарт Технологии»

Автор статьи подробно рассматривает два известных подхода к генерации ключевых слов на основе коллекции веб-документов. На основе этого анализа формулируется новый подход.

Ключевые слова: контекстная реклама, генерация ключевых слов, TF-IDF, анализ веб-документа.


А.С. Журавель

Введение

Сеть Интернет стала одним из самых привлекательных каналов маркетинга. Поисковая реклама является эффективным методом привлечения клиентов для многих видов деятельности. Крупнейшие поисковые системы Яндекс и Google имеют сервисы, которые позволяют размещать платные рекламные объявления рядом с естественной выдачей. Такие объявления называются контекстной рекламой.

Рынок контекстной рекламы активно растет. По данным eLama.ru [1] в 2013 году бюджет на платную поисковую рекламу превысил аналогичные показатели по печатной, наружной рекламой и рекламой на радио. Это наглядно показано на рисунке 1.

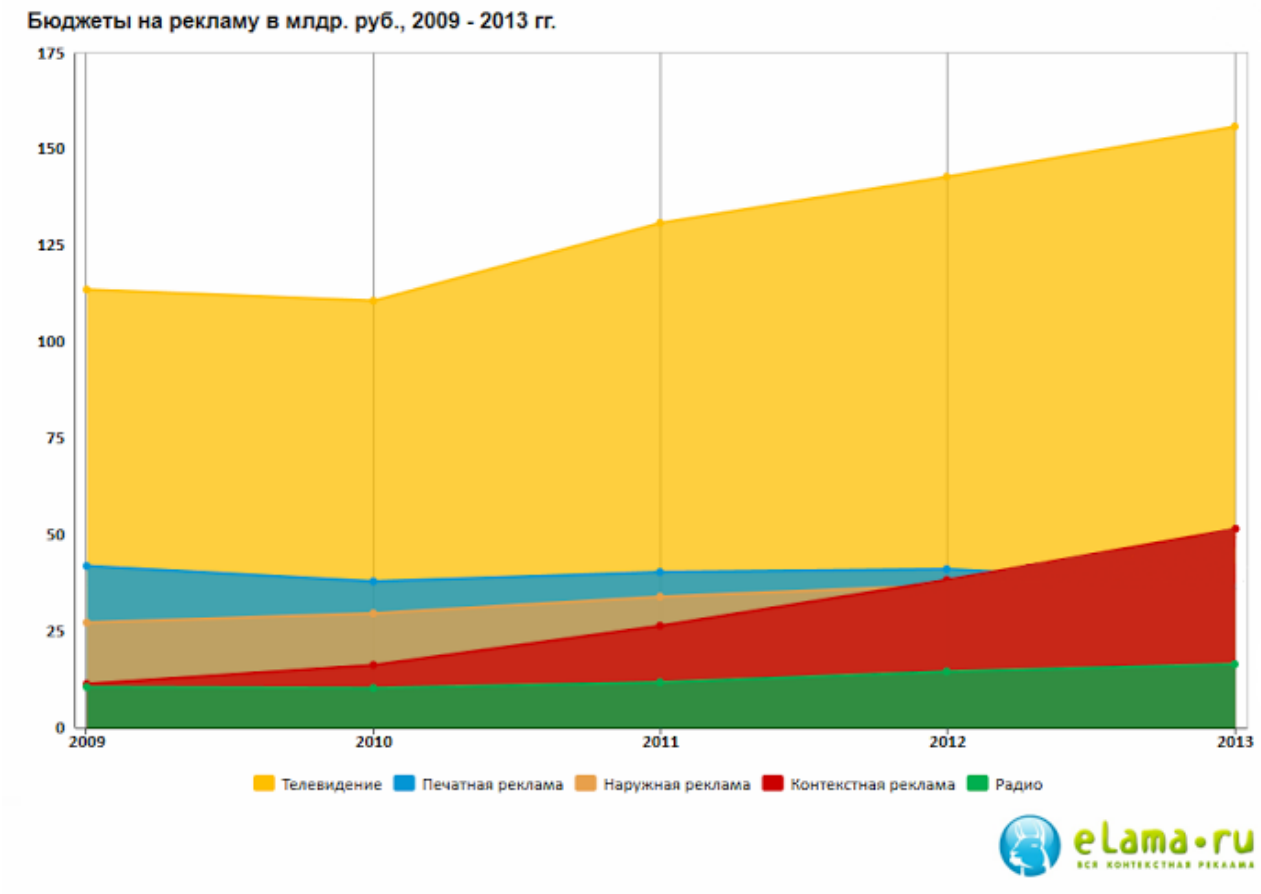

Рисунок 1 – График распределения маркетинговых бюджетов по различным рекламным каналам

Как видно из графика, к началу 2014 года контекстная реклама уступает только бюджетам на телевидение.

Размещение контекстной рекламы отличается от традиционных методов маркетинга. Платное объявление вместе со ссылкой на сайт рекламодателя показывается только по тем запросам, которые были заранее указаны. Эти запросы принято называть ключевыми словами. Именно ключевое слово определяет ту цену, которую заплатит рекламодатель за переход пользователя по ссылке. Поэтому важно иметь список поисковых запросов, по которым компании выгодно показывать свои объявления. В рамках данной статьи оценка качества подобранных ключевых слов будет исключительно экспертной.

Интернет-магазины и другие интернет-компании сталкиваются с проблемами при размещении контекстной рекламы, связанными с крайне большим количеством товарных позиций. Например, OZON.ru имеет более 2 млн различных позиций [2]. Для каждого товара нужно иметь отдельный список ключевых слов. Специалисты по контекстной рекламе не могут обработать такое количество веб-страниц вручную. Таким образом становится актуальной задача автоматизированной генерации ключевых слов на основании имеющихся веб-страниц (для интернет-магазинов – это карточки товаров).

Существует два известных подхода к генерации ключевых слов. Первый основан на глубоком анализе структуры документа, частности каждого слова и составлении фраз состоящих из нескольких слов. Второй подход базируется на анализе группы документов. При этом помимо частности слова рассчитывается TF-IDF – статистическая мера значимости слова внутри группы документов. После их практической реализации оказалось, что оба метода в отдельности имеют свои существенные недостатки. Однако удалось найти способ объединить два подхода в один алгоритм так, что новый подход показывает более качественный результат, чем каждый из подходов в отдельности.

1 Подход на основе глубокого анализа документа

Группа исследователей [3] предлагает подход, основанный на глубоком анализе документа. Алгоритм построения набора ключевых слов состоит из двух независимых модулей:

- построение списка слов и соответствующих весов, где вес – действительное число, которое характеризует значимость слова внутри документа;
- построение фраз (сочетаний из 2 и более слов) на основе списка слов и их весов.

Алгоритм первого модуля, строящего список слов с весами, основан на определенной структуре HTML-документа и частности слов.

Язык разметки HTML предусматривает опциональное наличие различных тегов. Чтобы веб-страница была легкочитаема и понятна для пользователя, определенные части документа принято выделять соответствующими тегами. Например, название документа обычно заключают внутри тэга <title></title>, заголовок большого параграфа выделяют с помощью <h1></h1>. Также внутри веб-страницы могут быть мета-тэги, которые не видны пользователю и обычно служат дополнительной информацией для поисковых систем.

Авторы статьи сопоставляют каждому тегу действительное число, которое характеризует значимость данного тэга, что показано в таблице 1.

Таблица 1

Вес различных тэгов

Element	Assigned Weight
<title>	50
meta keywords	40
meta description	40
anchor text	30
<h1>	30
<b>	10
other	1

Данные числа были получены на основе анализа большой выборки веб-страниц. Наибольший вес получили мета-тэги и тэг <title></title>.

Далее мы будем рассматривать в документе не слова, а лексемы. Например, два слова «окно» и «окна» относятся к одной и той же лексеме. При этом у каждой лексемы есть основная словоформа, в данном случае «окно».

Для каждой лексемы внутри документа подсчитывается его частотность внутри каждого тэга и на основе этой частности и веса тэга рассчитывается действительное число внутри промежутка [0;1]. Математически это выглядит следующим образом:

$$special\_weight_j = \sum w_{jtag}$$

$$relevance\_score_j = \frac{special\_weight_j}{MAX\_WIEGHT}$$

Это число автор называет весом данной лексемы внутри документа. Если полученное число оказывается меньше порогового значения, то слово не включается в результирующий список. Автор статьи предлагает использовать следующую формулу для фильтрации:

$$r = 0.001 \cdot relevance_{\max}$$

.

Полученный вектор пар («основная_словоформа», k), где k – действительное число в [0;1], подается во второй модуль программы.

Второй модуль, который предназначен для построения фраз, получает в качестве исходной информации список слов с соответствующими весами. Алгоритм построен на последовательности построения сначала фраз из двух слов, а затем из трех.

На первом этапе строится квадратная матрица, с количеством строк (и столбцов) равным количеству слов в списке, полученным из первого модуля программы. Матрица заполнена нулями, как это показано в таблице 2. Затем последовательно просматривается каждый тэг документа, и если два слова встретились внутри одного тэга, то программа добавляет 1 к соответствующей ячейке внутри матрицы.

Таблица 2

Первоначальное состояние матрицы

	$Word_1$	$Word_2$	...	$Word_N$
$Word_1$	0	0	0	0
$Word_2$	0	0	0	0
...	0	0	0	0
$Word_N$	0	0	0	0

При этом автор статьи не приводит конкретные формулы для фильтрации полученных пар и 3-х словных комбинаций, а высказывают лишь общую идею.

2 Подход при анализе коллекции документов

Классическим подходом к анализу группы документов является вычисление статической меры TF-IDF. Алгоритм программы разбивается на следующие этапы:

- формирование коллекции документов для анализа;
- составление списка слов и их частности для каждого документа в коллекции;
- расчет меры TF-IDF;
- фильтрация слов;
- составление 2- и 3-словных комбинаций.

Сначала происходит формирование коллекции документов. Если веб-страницы загружаются с сайта, то на этом этапе происходит скачивание документов на локальный компьютер.

Далее каждый документ рассматривается как набор слов. Тэги и знаки препинания не рассматриваются. Как и в предыдущем методе, необходимо учитывать русскую морфологию. Например, слова «окно» и «окна» являются различными словоформами, но принадлежат одной и той же лексеме. Для каждой лексемы подсчитывается ее частота в пределах данного документа. В результате данного анализа определена функция:

$$tf(t,d) = \frac{n_i}{\sum_{i=1}^{n} n_i},$$

где t – лексема; d – документ.

На следующем этапе происходит расчет функции инверсии частоты:

$$idf(t,D) = \log \frac{|D|}{n(t)},$$

где $n(t)$ – количество документов, в которых содержится лексема t;
$|D|$ – общее количество документов в коллекции.

Часто используется модифицированная формула, которая была впервые реализована в 1980-ых годах в Лондонском городском университете в рамках проекта поисковой системы «Okapi BM25»:

$$idf(t,D) = \log \frac{|D| - n(t) + 0.5}{n(t) + 0.5},$$

где $n(t)$ – количество документов, в которых содержится лексема t;
$|D|$ – общее количество документов в коллекции.

Далее рассчитывается функция TF-IDF как простое произведение функции частоты и инверсии частоты по данной коллекции документов:

$$tfidf(t,d,D) = tf(t,d) \cdot idf(t,D),$$

где t − лексема; d − документ; D − коллекция документов.

Затем происходит нормализация значений функции TF-IDF для всех лексем в пределах одного документа, и таким образом для каждой лексемы в пределах документах получается действительное число [0;1]. Далее происходит фильтрация, и отбрасываются все слова, которым сопоставлено число меньше порогового.

Составление 2- и 3-словных комбинаций происходит таким же образом, как и в предыдущем методе.

3 Достоинства и недостатки двух подходов. Формирование нового алгоритма для генерации ключевых слов

Два метода были реализованы на языке Java. Были проанализированы коллекции документов из веб-страниц.

Содержимое файла links1.txt (сайт компании по пластиковым окнам):
http://www.barinoff.ru/rehau.html
http://www.barinoff.ru/kbe-windows.html
http://www.barinoff.ru/proplex-windows.html
http://www.barinoff.ru/roto.html
http://www.barinoff.ru/wooden-windows.html
http://www.barinoff.ru/balconies.html
http://www.barinoff.ru/glass-replace.html
http://www.barinoff.ru/repair.html
http://www.barinoff.ru/delivery.html
http://www.barinoff.ru/montage.html
http://www.barinoff.ru/installment.html

Содержимое файла links2.txt (интернет-магазин электроники):
http://apples-msk.ru/iPhone/Apple-iPhone-5-16Gb-Black
http://apples-msk.ru/iPhone/Apple-iPhone-5-16Gb-White
http://apples-msk.ru/iPad/Apple-iPad-4-WiFi-16Gb-Black
http://apples-msk.ru/Mac/Apple-MacBook-Air-Z0NB-002
http://apples-msk.ru/iPod/Apple-iPod-touch-5-32Gb-Slate
http://apples-msk.ru/Accessories/Yoobao-iSmart-Leather-Case-for-iPad-2

Содержимое файла links2.txt (сайт компании по строительству):
http://kinghouse.ru/kingcatalog/pechi_dlya_ban_i_domov_kaminy_dymohody/
http://kinghouse.ru/kingcatalog/stroitel_stvo_doma/
http://kinghouse.ru/kingcatalog/stroitel_stvo_bani/
http://kinghouse.ru/kingcatalog/okna/

В результате экспертной оценки были определены особенности двух подходов.

При глубоком анализе документов в результирующем списке было слишком большое количество «лишних» слов, которые прямо не соответствуют тематике страницы. Главным минусом второго подхода оказалось отсечение важных слов. Например, при анализе документов с сайта по реализации продукции Apple, слово «apple» есть на каждой странице и его TF-IDF = 0.

Таким образом, очевидна актуальность совмещения двух данных подходов и построение другого алгоритма, который учитывает как тэги внутри документа, так и значение TF-IDF внутри всей коллекции документов.

В результате разработанного подхода, каждое слово характеризуется тремя значениями – частотность, вес тэга и TF-IDF (рисунок 2). Веса тэгов взяты из оригинальной статьи в первом подходе.

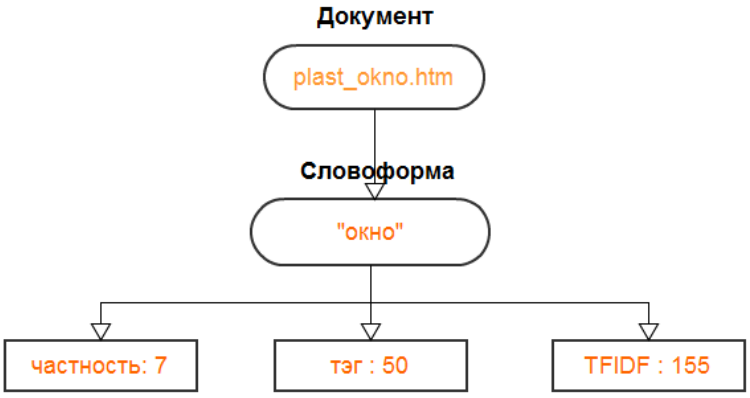

Рисунок 2 − Статистика по словам внутри коллекции документов

После получения такой статистики по каждому слову в коллекции документов происходит фильтрация (рисунок 3).

Как видно из картинки, фильтрация устроена таким образом, что если слово входит в важный тэг (с весом 50 – главные тэги в документе, см. статью), то оно попадает в результирующий массив. Однако если слово находится в тэге с низким весом, то мы проверяем его по второму критерию – превышение по статистической мере TF-IDF определенной величины. При этом сравнение происходит с максимальным значением TF-IDF (в рамках коллекции документов) умноженным на коэффициент.

По той же выборке сайтов были проведены эксперименты, и эксперты однозначно согласились с двумя утверждениями:

- при новом подходе доля «лишних» слов меньше, чем в каждом из двух первоначальных подходах;

- количество правильных слов не меньше, чем в каждом из двух первоначальных подходах.

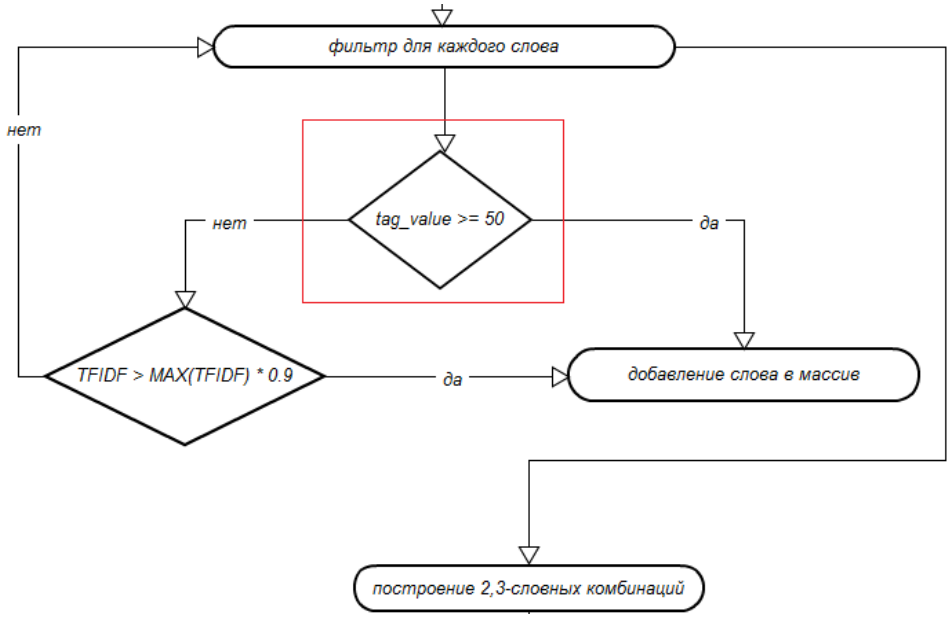

Рисунок 3 − Новый подход к фильтрации слов

Далее была разработана следующая методология для автоматизации процесса:

- на вход программе поступает коллекция документов (локальные или удаленные адреса веб-страницы);
- программа генерирует список ключевых слов и производит фильтрацию по новому подходу;
- оператор производит дополнительную фильтрацию или, в особых случаях, добавляет новые слова в список;
- программа автоматически генерирует 2- и 3-словные сочетания, и производит запись в файл в необходимом формате.

В ходе проведения эксперимента, такой порядок действий показал себя как эффективный способ автоматизированной генерации ключевых слов.

4 Заключение

Автор статьи считает, что в данной работе новыми являются алгоритм фильтрации ключевых слов и методология автоматизированной генерации ключевых слов.

Безусловная важность данной разработки была обоснована потребностями рынка. С другой стороны актуальность данной проблемы подтверждена появлением множества автоматизированных и автоматических сервисов, таких как eLama.ru, r-broker.ru и другие.

Разработанная методология позволяет эффективно расходовать время сотрудника. В то же время сотрудник будет заниматься только творческой работой, которую в данный момент сделать полностью автоматически не представляется возможным.

Литература

1. http://blog.elama.ru/post.php?id=16661248

2. http://www.shopolog.ru/metodichka/analytics/top-10-internet-magazinov-runeta-itogi-2012-goda

3. Amruta, J. & Rajeev, M. (2006). Keyword Generation for Search Engine Advertising. Sixth IEEE International Conference on Data Mining, 2006, 490-94.

4. Brian Lott (2012). Survey of Keyword Extraction Techniques.

5. *Алексеева Т.В., Дик В.В., Кокорева Л.А.* Оперативный анализ данных в электронном бизнесе // Славянский форум. 2014. № 6 (2). С.6–12.

6. *Бондаренко Т.Е.* Психология рекламы // Славянский форум. 2014. № 6 (2). С.13–23.

Program of automated keyword generation based on a collection of web documents

***Alexander Sergeevich Zhuravel**, Developer, Wikimart Technologies Ltd.*

The paper is about two approaches to the problem of keyword generation based on a collection of web documents. A new approach is developed on the basis of the analysis.

Keywords: Context advertising, keyword generation, TF-IDF, analysis of a web document.

УДК 378.1

ГЕОСТАТИСТИЧЕСКИЙ АНАЛИЗ В ОБРАЗОВАНИИ

***Ольга Викторовна Зайцева**, канд. техн. наук, зав. отделом статистики*
Центра мониторинга и статистики образования,
E-mail: cvdisser@list.ru,
Федеральный институт развития образования,
http://www.firo.ru

Статья описывает новое научное направление – геостатистику. Показана связь геостатистики с научной картиной мира. Показано, что как прикладное направление геостатистика создает механизм исследования и управления системой образования. Исследован аспект применения геостатистики в образовании. Показано, что геостатистика создает информационные поля, которые связывают воедино разрозненные пространственные объекты одного качества, особенности геостатистики. Показано, что геостатистика в образовании отражает тенденцию информатизации образования.

Ключевые слова: образование, геоинформатика, статистика, геостатистика, моделирование, визуализация, геоданные, принципы геостатистики, управление образованием.

Введение

Научная картина мира − одно из основополагающих понятий современной науки, которое включает обобщение и синтез различных научных теорий [1]. Научная картина мира включает общенаучную картину мира и картины мира отдельных наук. Картины мира отдельных наук включают в себя определённые способы понимания и трактовки предметов, явлений и процессов объективного мира, существующие в каждой отдельной науке. Научная картина мира выступает как специфическая форма систематизации научного знания, задающая видение предметного мира науки соответственно определенному этапу её функционирования и развития [2].


О.В. Зайцева

Наряду с понятием «научная картина мира» и «картины мира отдельных наук» применяют термин «картина мира». В философско-методологической литературе термин «картина мира» применяется как для обозначения мировоззрения, так и в более узком смысле, когда речь заходит о таких представлениях об окружающей действительности, которые являются особым типом теоретического или практического знания. Картина мира в рамках какой-либо отрасли [3] включает понятие информационного пространства этой отрасли. Информационное пространство является отражением информационной сферы Земли [4]. Для построения такого пространства, в частности в сфере образования, необходимо использовать статистические методы. Таким образом, для построения описания [5] отраслевой картины мира, необходимо применять статистические методы