Разработка метода анализа тенденций развития Интернет
Elaboration of a method for analyzing Internet evolution

Куликовская А. А.
НИЦ Курчатовский институт, ИТЭФ, Москва

Аннотация

В данной статье рассматривается алгоритм, определяющий тематику статьи на основе ключевых слов. Был проведен анализ всех статей на выбранном сайте, и на основе собранной статистики были выбраны наиболее отличные друг от друга категории, по которым можно однозначно распределить все статьи на данном сайте. Учитывая даты написания статей и изменения их количества можно сделать выводы о том, в каких категориях на данный момент ведутся активные разработки, а какие темы сейчас являются менее обсуждаемыми.

This article is devoted to the created algorithm which determines the theme of the article based on keywords. An analysis of all the articles on the saturn.itep.ru website was carried out, and on the basis of the received statistics the most distinct categories were selected, by which all articles on this site can be uniquely assigned. Considering the dates of the articles and changes in their quantity, the conclusions can be made about the most developing categories and which themes are now less discussed.

Введение

В настоящее время темп роста объемов данных стал таким, что человеку уже стало невозможно все прочесть и проанализировать (эффект big data) всю информацию по интересующей теме, это делает компьютерную аналитику найденных материалов безальтернативной. Сейчас по каждому запросу к поисковому сервису находится как правило до нескольких тысяч статей. При всем желании мы сможем изучить только малую часть этого количества. И не обязательно это будет наиболее значимая часть, по которой можно будет сделать однозначный вывод обо всех остальных статьях.

Один из возможных вариантов решения этой задачи - это анализатор, который позволяет фильтровать и предлагать пользователям только те статьи, которые удовлетворяют заданным критериям.

На основе большого количества статистических данных предполагается делать некоторые аналитические выводы. В данной работе это определение наиболее актуальных тем среди исследуемых.

Цель исследования

Создание фильтра для отбора по ключевым словам или словосочетаниям в названиях статей. Распределение статей по категориям для поиска статей, аналогичных ранее найденным. Получение вывода о развитии технологий Интернет на основе количества статей, изданных для каждой из выбранных тематик.

Работа состоит из следующих частей:

- в первом разделе описывается создание программных модулей, выполняющих поиск и сохранение статей по ключевым словам, поиск даты написания статьи и создание программного модуля, распределяющего все статьи на категории, согласно заданным критериям.
- во втором разделе описывается метод оценки количества статей в каждой категории, написанных за все годы и за каждый год в отдельности.
- в третьем разделе производится сравнение результатов статистического анализа количества статей и определение основных направлений развития технологий Интернет и тенденций последних лет.
- в заключении подводятся итоги сравнения и намечаются задачи для дальнейших исследований.

Программное обеспечение и банк данных

В качестве банка данных использовались статьи RFC на сервере http://saturn.itep.ru/rfc_docs0.htm. На данный момент количество статей составляет 8496. Общее направление статей – технологии работы и дальнейшего развития Интернет.

Созданные модули – это программы на языке Python 3.7, которые работают последовательно, независимо друг от друга.

Создание программных модулей

Отбор и загрузка статей осуществляется с помощью составления регулярных выражений, для поиска ссылок на статьи и заданных ключевых слов в названиях статей. Далее в отобранных статьях выполняется поиск даты написания статьи. Это делаетя также с помощью регулярного выражения. После этого все статьи с найденным ключевым словом в названии сортируются по годам. Время работы алгоритма зависит от количества найденных статей и скорости копирования. В среднем на поиск статей по одному ключевому слову уходит не более минуты.

На рисунке 1 представлены результаты поиска статей по ключевому слову “DNS”. Столбцы обозначают количество вышедших статей с данным словом в каждый год, начиная с 1989 года. На диаграмме можно заметить, что количество статей с каждым годом в среднем за каждые 3-4 года увеличивается, что говорит об активных разработках в данном направлении.

На рисунке 2 представлены результатыпоиска статей по ключевому слову “IPv6”.На этой диаграмме можно заметить, что пик выхода статей данной направленности пришелся на 2011 год и в данный момент идет на спад.

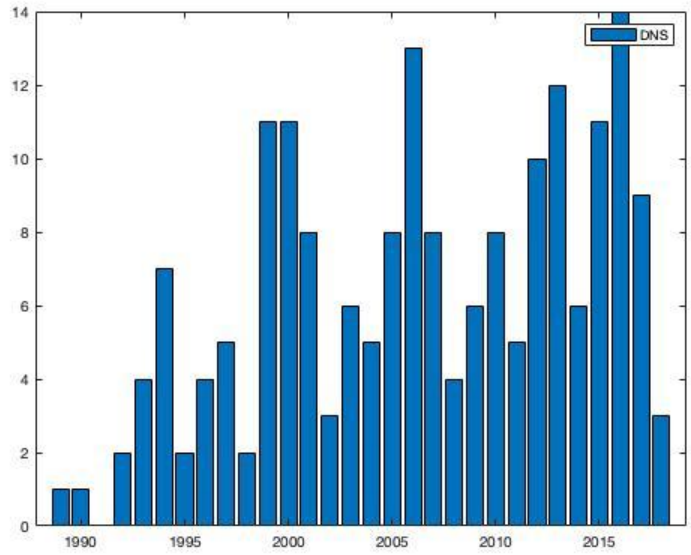

Рис. 1. Количество вышедших статей со словом “DNS” в названии

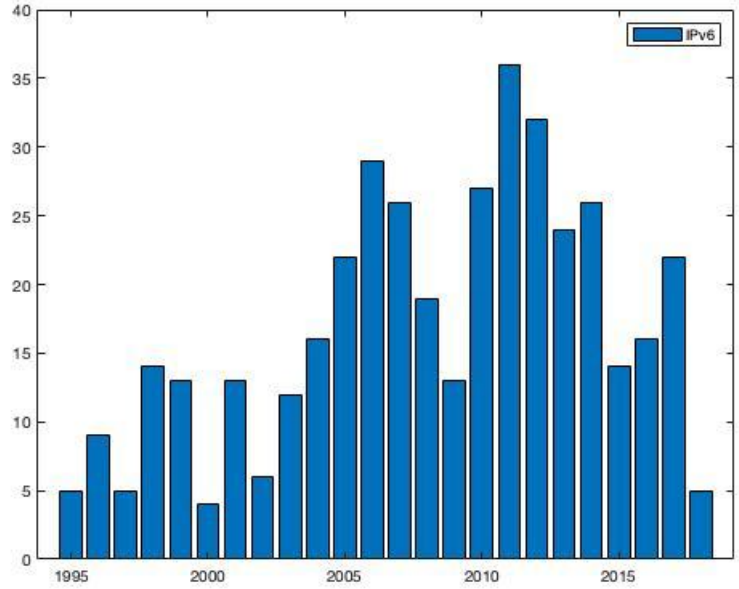

Рис. 2. Количество вышедших статей со словом “IPv6” в названии

Метод оценки количества статей в каждой категории

Все ключевые слова, по которым производился поиск делятся на несколько максимально непересевающихся множеств. Эти множества соответствуют категориям, к которым относятся отобранные статьи. Статей, не попавших ни в одну категорию и попавших в несколько категорий сразу, нет.

На рисунке 3 приведен пример категоий и количества статей, вышедших за все годы в каждой категории. По оси уотображено количество статей, вышедших за все рассматриваемое время (1969-2018гг).

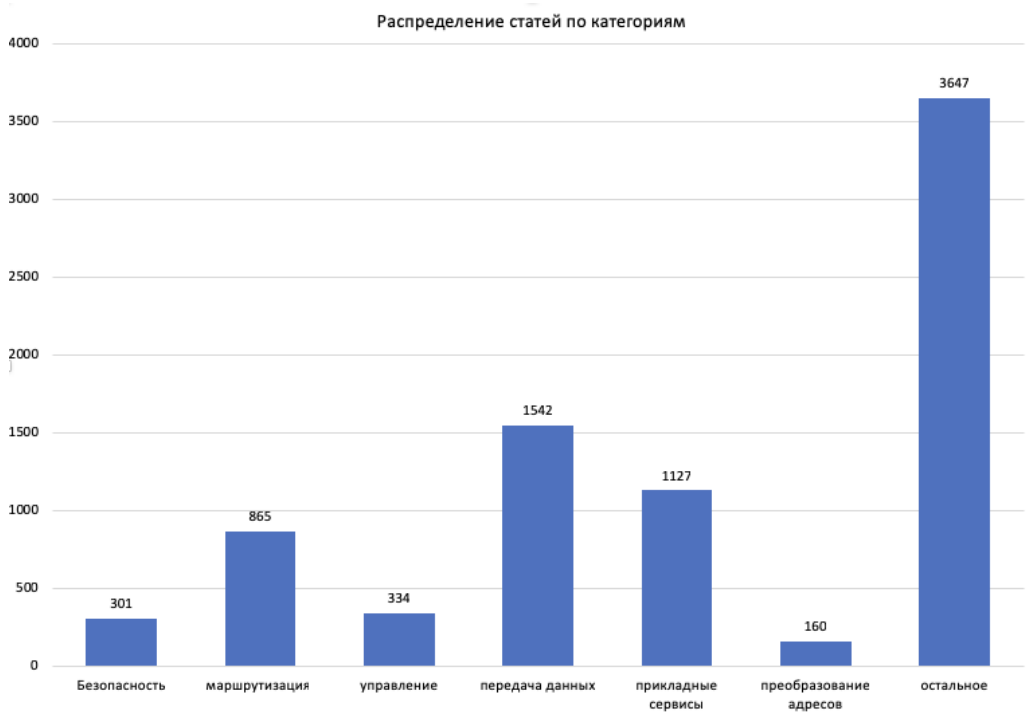

Рис. 3. Количество статей, вышедших за все годы в каждой категории

Сравнение результатов статистического анализа

На основе полученных данных о количестве статей в каждой категории, за каждый год в отдельности, можно оценить динамику изменения процентного соотношения количества статей.

На рисунке 4 представленно изменение процентного соотношения статей каждой категории ко всем статьям, вышедшим за год. Здесь можно заметить как какие-то категории со всеменем становятся более популярными, например “передача данных”, а категоия “преобразование адресов” становится менее актуальной.

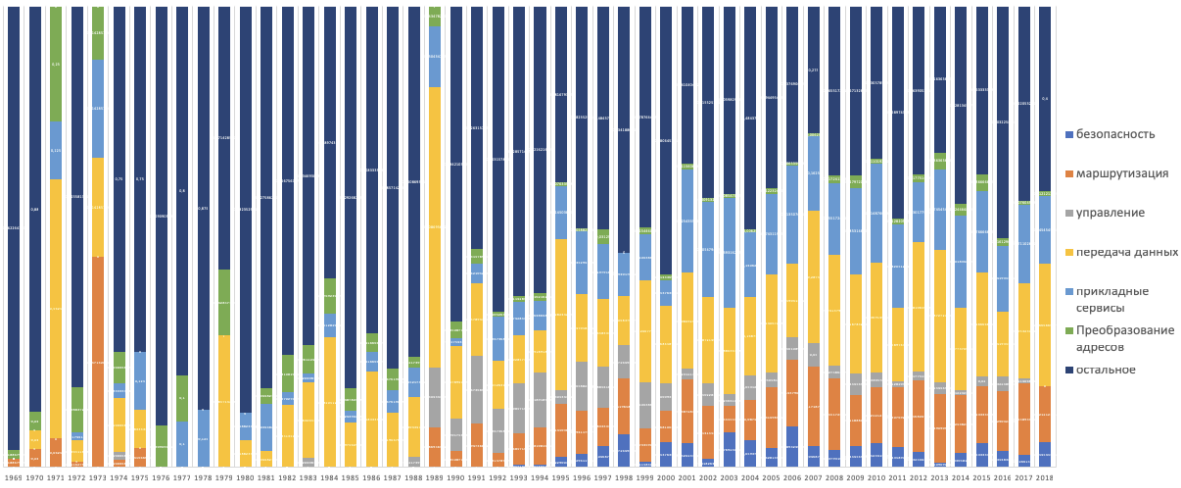

Рис. 4. Процентное соотношение статей каждой категории ко всем статьям, вышедшим за год.

Заключение

В автоматическом режиме, без участия человека, определены наиболее развивающиеся направления технологий Интернет и те направления, которые уже достигли своего решения.

Данный метод анализа развития можно применять не только на выбранной тематике, но и других направлениях развития науки и технологии. Комбинируя банки данных, категории и критерии отбора можно получить наиболее актуальные на данный момент направления в любой отрасли.

В данной работе рассматривались только статьи с выбранного сайта, поэтому информация, полученная в данном исследовании, может отличаться от других источников. Однако при дальнейшем масштабировании исследования данные будут более точными.