ОЦЕНКА РЕЛЕВАНТНОСТИ КОНТЕНТА В РЕЖИМЕ РЕАЛЬНОГО ВРЕМЕНИ

Шинкевич Н. Н.
Кафедра программного обеспечения информационных технологий, Белорусский государственный университет информатики и радиоэлектороники
Минск, Республика Беларусь
E-mail: sn0wf1llin@gmail.com

В данной работе рассматривается построение и применение дерева синонимов для определения релевантности контента. В качестве примеров актуальных запросов (актуального контента) выступает информация с Google Trends [1]. Описывается алгоритм нахождения минимального расстояния во взвешенном графе, представляющем текст, которое является мерой релевантности контента примерам запросов Google Trends.

Введение

Актуальность – один из важных факторов релевантности документа. Если документ содержит устаревшую, неактуальную информацию, его нельзя считать хорошим ответом и не стоит оценивать положительно. Конечно, существуют документы, информация в которых остается актуальной еще долгое время. Например, биография известного художника или писателя. Но многие документы имеют некоторый «срок жизни», в течение которого содержащаяся в них информация актуальна.

I. Содержание документа

Пусть имеется документ, содержащий некоторую информацию и обладающий определенными признаками, такими как время создания автором и геопозиция. Автора интересует, насколько актуальной является его информация с учетом рассматриваемой темы, а так же перечисленных выше признаков.

Таким образом, необходимо построить систему, позволяющую определить меру релевантности имеющегося документа пользователя актуальным поисковым запросам Google Trends. Мерой релевантности назовем число, в соответствии со значением которого можно проводить сортировку документов по релевантности. Это число должно обладать следующими свойствами:

1. мера релевантности - неотрицательное вещественное число;
2. чем выше релевантность контента, тем выше мера релевантности;
3. мера релевантности должна быть ограничена сверху.

Последнее условие крайне важно с точки зрения удобства пользователя, потому что так он сможет оценить абсолютную степень релевантности контента. Меру релевантности предоставим в виде целого числа от 0 % до 100 %. Для этого мы округлим и нормируем меру релевантности, подразумевая при этом, что 100 % - релевантный документ представляет собой контент, содержащий более 20 % слов или более 30 % синонимов запросов самой актуальной выдачи Google Trends.

Для получения актуальных поисковых запросов используются инструменты, называемые web-crawlers (web-spiders), позволяющие собирать и обрабатывать данные автоматически в режиме реального времени. Процесс повторяется каждый час с целью получения самой актуальной информации. В результате работы краулеров в базе данных хранятся записи, каждая из которых представляет собой поисковый запрос в определенный момент времени в том виде, в котором оно представлено в Google Trends.

Каждый полученный поисковый запрос содержит текстовое описание, количество раз ввода поискового запроса в строку поиска Google, график популярности (набор точек, где осью абсцисс представлено время, а ординат - количество раз ввода запроса) и картинку. Вся информация сохраняется в базу данных с целью дальнейшего изучения зависимости популярности контента от степени его соответствия актуальным поисковым запросам в определенный момент времени.

Поисковые запросы разделяются по категориям, каждая категория представлена в Google Trends специфическими для конкретного языка наборами слов по теме поисковых запросов, отсортированных по убыванию количества переходов пользователей. Так же полученные наборы имеют краткие описания (пояснения) к уже приведенным словам.

Реализованная система после поступления запроса пользователя относительно оценки релевантности контента обращается к базе данных, содержащей уже обработанную самую актуальную информацию. Для каждого набора каждой категории строится взвешенный граф, содержащий синонимы данных слов. Для этого используется пакет для анализа естественных языков NLTK Wordnet Synsets [2] на языке программирования Python. Вес ребра графа в данном случае представляет степень близости того или иного синонима. Максимальное число синонимов

для каждого слова было выбрано равным трем, так как выбор большего числа резко увеличивает вычислительную сложность алгоритма. Стоит отметить, что на точность результатов негативно влияет специфика языка, а так же большое количество имен и названий в поисковых запросах. Информация такого рода может не присутствовать в словаре nltk.wordnet, или присутствовать с исправлениями, что существенно усложняет поиск вплость до невозможности его осуществления. В ходе разработки системы было принято решение игнорировать такие слова и не учитывать их в процессах дальнейшего анализа и представления результатов.

Далее при помощи алгоритма Rake [3] из исходного текста извлекаются ключевые слова. Метод Rapid Automatic Keyword Extraction (RAKE) основывается на предположении, что почти всегда ключевыми будут именно те слова, которые содержат значимые слова, но редко включают стоп-слова, местоимения или дру- гие слова с минимальным лексическим значением. Извлечение ключевых слов происходит следующим образом:

1. текст разбивается на слова по позициям стоп-слов и знаков препинания - разделителей;
2. полученные последовательности слов формируют список “кандидатов” в ключевые слова;
3. строится граф встреч данных кандидатов друг с другом в тексте документа;
4. вес каждого слова вычисляется по формуле (1);
5. N слов с наибольшим весом выбираются в качестве ключевых.

$$W(w_i) = \frac{D(w_i)}{F(w_i)}. \quad (1)$$

где (w_i) - некоторое слово, $W(w_i)$ - word weight, $D(w_i)$ - word degree, $F(w_i)$ - word frequency.

После определения ключевых слов для них составляется граф синонимов с глубиной равной двум. При помощи встроенных алгоритмов nltk.wordnet оценивается близость контента к определенному набору слов, представляющему поисковый запрос. Полученное число является усредненным показателем для всех слов, входящих в запрос, по отношению к каждому слову исходного текста.

Тот же процесс повторяется для каждого набора каждой категории, в результате пользователю предоставляется набор поисковых запросов каждой категории, отсортированных по мере релевантности и для удобства просмотра усеченный до трех единиц в каждом наборе.

II. Заключение

Таким образом, суть описанного выше подхода к анализу релевантности контента состоит в использовании взвешенного графа синонимов и поисковых запросов Google Trends. Данная методика призвана помочь пользователю в создании актуального контента. Предлагаемый подход успешно применен в рамках практической реализации системы оценки релевантности контента для проекта Exponenta.io, и доступен по ссылке `https://app.exponenta.io`. Дальнейшее развитие применяемой методики может заключаться в разработке системы определения популярности контента на основе вычисленной динамики его соответствия изменяющимся во времени поисковым запросам.

III. Список литературы

1. Google Trends [Electronic resource] / – Mode of access: https://ru.wikipedia.org/wiki/Google_Trends. – Дата доступа: 08.09.2018.
2. NLTK Wordnet Synsets [Electronic resource] / – Mode of access: http://www.nltk.org/howto/wordnet.html. – Date of access: 09.09.2018.
3. Automatic Keyword Extraction from Individual Documents / S. Rose, D. Engel, N. Kramer, W. Cowley // M. W. Berry & J. Kogan (Eds.), Text Mining: Theory and Applications: John Wiley & Sons. – 2010. – Vol.18, № 2. – P. 37–46.