

ОБНАРУЖЕНИЕ НОВЫХ СОБЫТИЙ В НОВОСТНОМ ПОТОКЕ

Амелин Владислав Владимирович

Студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: amelin408@gmail.com

Развитие Интернета привело к всеобщему увеличению текстовых данных из электронных писем, новостных источников, форумов и т.д. В последние годы набирают популярность сетевые новостные агентства, во многом заменившие традиционные средства информации. Для отдельно взятого человека стало, в большинстве случаев, невозможно отслеживать все релевантные текстовые данные и, кроме того, находить изменения в тенденциях или темах. В результате, появилась необходимость в разработке систем автоматизированного анализа новостного потока. Задачи обнаружения, отслеживания и группировки событий на основе анализа новостей актуальны как с научной, так и с практической точки зрения. Необходимость журналистам и компаниям по мониторингу СМИ обрабатывать большое количество новостных статей, приходящих ежедневно, для выявления интересных тем и событий, подчеркивает важность этих задач.

В задаче по выявлению новых событий из потока новостей (New Event Detection) [1] предполагается, что на вход соответствующего алгоритма поступают новые документы. В результате работы алгоритма каждому возникающему кластеру соответствует новое событие, отражаемое в документах данного кластера. Наибольший интерес представляет именно онлайн выявление событий, т.к. это позволяет предупреждать о новых и интересных событиях, происходящих в мире в реальном времени, а также помогает в принятии важных решений. Таким образом, алгоритм New Event Detection должен быть точным, масштабируемым и обрабатывать поток новостных статей за один проход.

В работе исследованы подходы, в основе которых лежит вычисление меры сходства между новостными статьями с использованием векторного представления слов (word embedding) [2]. Наиболее известной моделью такого типа является модель word2vec [3], в которой слова из одного контекста имеют семантически близкие значения. Использование word2vec позволяет решать проблему синонимии, когда события повествуют об одном и том же разными словами.

Для ускорения времени работы алгоритмов рассматривается при-

менение метода понижения размерности — Locality-Sensitive Hashing, чтобы быстро получить наилучшее соответствие для каждого нового документа. Также рассматривается введение скользящего временного окна: попарно сравниваются не все документы, а только N последних документов.

Для оценки качества работы реализованных методов применяется DET-кривая (Decision Error Tradeoff curve), которая иллюстрирует компромисс между пропущенными обнаружениями нового события (missed detection) и ложными обнаружениями (false alarm). Эксперименты проводятся на размеченной по событиям новостной коллекции.

В исследовании представлены результаты работы реализованных алгоритмов, приводится сравнение качества их работы между собой, а также с традиционными методами, основанными на применении статистической меры TF-IDF, используемой для оценки важности слов в контексте документа.

Литература

1. Allan J. Topic detection and tracking: Event-based Information Organization. Springer Science+Business Media, LLC, 2002.
2. Kusner M., Sun Y., Kolkin N., Weinberger K. Weinberger K. From word embeddings to document distances // ICML'15 Proceedings of the 32nd International Conference on International Conference on Machine Learning — Volume 37, 2015, P. 957–966.
3. Mikolov T., Sutskever I., Chen K., Corrado G., Dean J. Distributed representations of words and phrases and their compositionality // Advances in neural information processing systems, Lake Tahoe, Nevada, USA, 2013, P. 3111–3119.