Интерпретируемые тематические модели для анализа новостных потоков

Косарев Евгений Александрович

Студент

Факультет ВМК МГУ им. М.В.Ломоносва, Москва, Россия

E-mail: evgenijkkk@yandex.ru

Научный руководитель *— Воронцов Константин Вячеславович*

Определения:

- **Событие** - нечто произошедшее, что нашло отголосок в СМИ.
- **Новостная цепочка** - упорядоченный набор новостей из новостного потока, объединенный одним событием [1].
- **Тема** - композиция событий. Обычно сообщения из новостного потока, объединенные общей темой, содержат значительное количество общей лексики.

На выход алгоритму подается список новостей, собранных со следующих новостных ресурсов: *currenttime, rbc, ria news, tass, meduza и другие*. Всего 19 источников.

Способ построения цепочек

Алгоритм итеративный и очень похож на кластеризацию [2–3]. Очередная поступающая новость присоединяется либо к одной из уже образованных цепочек, либо создает новую цепочку. Особенность состоит в том, что цепочки содержат упорядоченные новости, и решение о присоединении новой цепочки выносится на основании некоторого подмножества новостей из цепочки.

Формально

Пусть после шага j построено множество цепочек C, где $C = \{C_1, C_2, \ldots, C_m\}$. Каждая из цепочек состоит из своего набора документов (новостей): $C_i = \{d_{i_1}, d_{i_2}, \ldots, d_{i_{n(i)}}\}$. На j + 1 итерации новость d нужно либо отнести к очередной цепочке, либо составить новую.

Произведем расчет расстояния от новости d до всех уже построенных цепочек по следующему правилу:

$$\rho(d, C) = \min_{C_i \in C} \rho(d, C_i),$$

где расстояние от новости до цепочки высчитывается по правилу:

$$\rho(d, C_i) = \max_{d_i \in C_i^*} \rho(d, d_i)$$

C_i^* - подмножество C_i, в которое включаются последние k новостей из цепочки C_i.

Расстояние от одной новости до другой будем считать по следующей формуле:

$$\rho(x, y) = 1 - cosine(x, y) = 1 - \frac{\sum_{j=1}^{|W|} x_j * y_j}{\sqrt{\sum_{j=1}^{|W|} x_j^2} \times \sqrt{\sum_{j=1}^{|W|} y_j^2}}$$

Если $\rho(d, C) \leq threshold$, то новость присоединяется в ближайшей цепочке, иначе создается новая цепочка C_{m+1}, где $C_{m+1} = \{d\}$. Тогда $C_{new} = C \bigcup C_{m+1}$

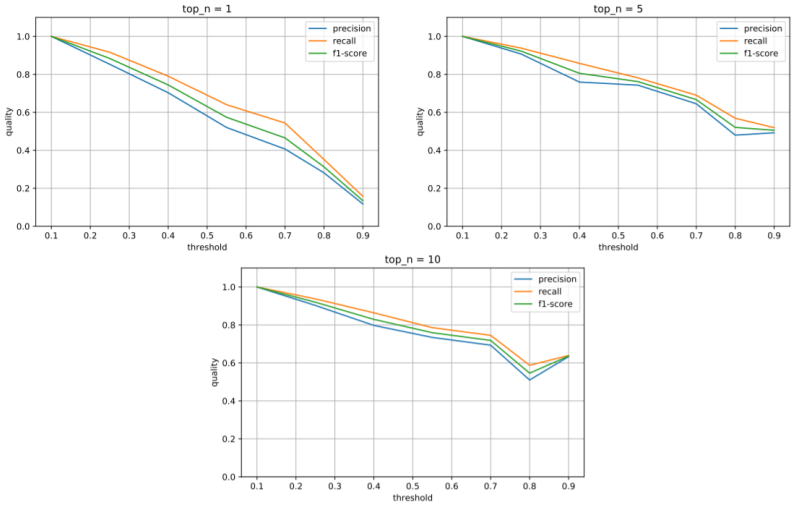
Устойчивость алгоритма

Считаются меры precision, recall и f1 меру по способам построения цепочек. За верный вариант берется хронологическое по времени построение [4].

Эксперименты

Проведем эксперименты по изучению устойчивости алгоритма к порядку входных данных. Разберем различные стратегии подачи данных. В каждой стратегии рассмотрим несколько способов поиска расстояния от цепочки до новости и выберем лучший результат. Устоцчивость измеряется на top@k документах.

Лучшая устойчивость достигается на топ 5 и топ 10, метрики качества не сильно отличаются друг от друга, а это значит, что алгоритм вполне устойчив. Благодаря полученным результатам можно оценить степень устойчивости алгоритма в зависимости от размера порога присоединения документа к цепочке новостей.



Исследование устойчивости

Литература

1. Zou X., Zhu Y., Feng J., Lu J., Li X. A novel hierarchical topic model for horizontal topic expansion with observed label information // IEEE Access. — 2019. — Vol. 7. — Pp. 184242–184253.
2. Hu L., Li J.-Z., Zhang J., Shao C. o-hetm: An online hierarchical entity topic model for news streams // PAKDD. — 2015.
3. Xi Y., Li B., , Tang Y. Topic model based new event detection within topics // Journal of Advanced Computational Intelligence and Intelligent Informatics. — 2016. — Vol. 20, no. 3. — Pp. 467–476.
4. Allan J., Carbonell J., Doddington G., Yamron J., Yang Y. Topic detection and tracking pilot study: Final report // Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop. — 1998. — Pp. 194–218.