"ЧЕРНЫЕ ЛЕБЕДИ": ИЗВЛЕЧЕНИЕ РЕДКИХ СОБЫТИЙ ИЗ ТЕКСТА

А.М. Попов, Ю.В. Адаскина
InfoQubes
Санкт-Петербургский государственный университет
Москва, Санкт-Петербург

Задача поиска аномалий в текстах становится все более востребованной в области анализа клиентского опыта. Это связано как с наработанными за последние годы знаниями об отзывах клиентов, так и со сдвигом интереса от статистически значимого к статистически незначимому в других научных дисциплинах. Статистический анализ частотных и легко прогнозируемых причин обращения клиентов в службу контроля качества и клиентского негативного опыта — задача не новая и до известной степени решенная. Противоположная же ей задача — поиск и идентификация редких и нечастотных аспектов негативного опыта клиента — задача весьма новая и не имеющая на сегодняшний день стандартных подходов к решению. Извлечение аномалий (anomaly detection) существует как самостоятельная задача во многих областях, где применяется машинное обучение и анализ большого объема данных, а также при создании различных систем мониторинга. В ходе нашего исследования мы разрабатываем инструментарий для анализа таких случаев для коммерческого проекта, а сами аномалии получили метафорическое название «черные лебеди» вслед за известной работой экономиста Н. Талеба «Черный лебедь. Под знаком непредсказуемости». Обычно считается, что именно эта книга и предложенная в ней терминология привлекла внимание к исследованиям редких непредсказуемых событий, имеющих значительные последствия.

На этом этапе работы нам не хотелось бы ограничивать понятие «черный лебедь» или «аномалия» строгим определением, так как в него входит экспертная логика. В нашем исследовании мы ограничиваемся только формальными статистическими критериями, пытаясь предложить эксперту упорядоченный список документов, отсортированных по убыванию степени аномальности. Таким образом, результатом работы нашего алгоритма становится не список аномальных документов, а некоторый аналог поисковой выдачи, который пользователю необходимо просмотреть и выбрать из него действительно релевантные для него документы. В частности, по этой причине оценка качества в нашем случае возможна скорее не в терминах полноты-точности, а в терминах релевантности: какое количество релевантных документов, т.е. «черных лебедей», содержится в некотором топ-списке выдачи алгоритма.

Исследователи, которые решают похожие задачи, зачастую используют такие подходы как «мешок слов» или различные методы машинного обучения (классификация, кластеризация). На наш взгляд, использование «мешка слов» затрудняет поиск собственно аномальных фрагментов текста, т.к. без учета информации о сочетаемости основным критерием определения аномальности слова становятся собственно его частотные характеристики в корпусе. Более того, есть серьезное опасение, что использование единичных слов как маркеров аномальности приведет к конфликту с релевантностью, т.к. наличие большого числа нечастотных слов в документе может свидетельствовать как о его аномальности, так и об отсутствии релевантности для исследуемой предметной области. По этой причине мы используем в качестве маркеров аномальности пары синтаксически связанных слов, а критерий аномальности опирается на их лексическую сочетаемость. Таким образом, мы опираемся и на информацию о синтагматических свойствах лексем, и на синтаксические связи, что позволяет использовать преимущества обоих методов. В нашем исследовании мы не стали применять традиционные методы машинного обучения, т.к. наряду с определением того, является ли документ «черным лебедем» нас интересуют и те фрагменты текстов, которые описывают аномальные ситуации. Многие существующие реализации классификаторов и кластеризаторов на машинном обучении не предоставляют возможности выяснить, какие признаки стали ключевыми для идентификации конкретного документа как аномального.

Наш метод во многом опирается на идею, предложенную в [1], где для извлечения «черных лебедей» используется семантическая информация о словах, полученная путем анализа «независимых» корпусов, например WordNet. Кроме того, хорошие результаты дает использование информации о семантической близости слов [2, 3]. Другие подходы основаны на представлении документов из корпуса как многомерных таблиц совместной встречаемости слов, для анализа которых применяются нейронные сети [4] или SVM-классификатор [5].

Стандартные методики поиска главных проблем, или негативных тем, обычно представляют собой частотный анализ различных лингвистических объектов, извлекаемых из исследуемых текстов: слов и их нормальных форм, словосочетаний (n-грамм), разрывных конструкций, синтаксических связей или фрагментов синтаксических структур и т.д. Этот метод достаточно эффективен, так как для сортированного частотного списка хорошо работает принцип Парето: небольшая выборка из верхней части частотного

списка покрывает большинство случаев, однако при этом остается «хвост» нечастотных элементов огромной длины. Очевидно, что поиск аномалий в инвертированном частотном списке неэффективен из-за большого количества «мусора», ведь далеко не все низкочастотные объекты являются аномальными.

Поэтому при разработке нашего инструмента мы решили вместо частотного критерия использовать критерий лексической сочетаемости. В некотором смысле наш подход напоминает метод извлечения коллокаций из корпуса текстов: если пара слов встречается вместе часто, а по отдельности эти слова встречаются редко, то вес такого сочетания как устойчивого будет высоким. При этом в нашем эксперименте мы использовали не биграммы, а пары синтаксически связанных слов, полученные в ходе синтаксического анализа. За основу метрики аномальности синтаксической связи мы взяли фрагмент формулы PMI:

$$\mathbf{w_r = f_r / (f_s + f_t),}$$

где f_r — это частота совместной встречаемости двух лемм в рамках синтаксического отношения, а f_s и f_t — соответственно самостоятельные частоты леммы-вершины и леммы-зависимого. Чем выше значение w_r, тем выше аномальность сочетания этих двух лемм.

Легко заметить, что в чистом виде эта формула будет завышать в том числе и те устойчивые сочетания, которые часто встречаются в рассматриваемой коллекции и не являются аномальными, например «станция метро», «водительское удостоверение» и т.д. Очевидно, что данный эффект необходимо как-то компенсировать. Для этого мы ввели в формулу дополнительный коэффициент, основанный на встречаемости сочетания в документах, формула приобрела вид:

$$\mathbf{w_r = f_r \times (- \log(DF)) / (f_s + f_t),}$$

где DF – это доля документов, в которых рассматриваемое сочетание встретилось как минимум по одному разу, логарифм здесь добавляет нелинейность, сильнее завышая вес сочетаний, которые встречаются в небольшом количестве документов.

Таким образом, на подготовительном этапе мы провели полный лингвистический анализ всего корпуса текстов, накопленного за несколько лет сотрудничества с одним из наших заказчиков, и построили по нему два частотных списка: список единичных лемм и список синтаксически связанных пар лемм. В статистику не включались слова и сочетания, которые встретились в корпусе только один раз.

Этап собственно поиска аномалий в нашем понимании состоит из двух частей:

– Поиск аномальных документов из предложенной выборки;
– Поиск ключевых слов, маркирующих аномальность, в этих документах.

В качестве критерия аномальности документа мы взяли собственную метрику, производную от весов синтаксических связей, входящих в этот документ. Наша гипотеза, позднее подтвержденная экспериментально, состояла в следующем: аномальные документы имеют некоторое количество связей с высоким весом, в то время как для нормальных документов распределение весов связей более равномерное. Это позволило нам взять за вес документа отношение суммы весов n наиболее весомых связей к n наименее весомым, соответственно, чем выше это соотношение — тем выше аномальность документа. Формально вес документа можно описать так:

$$\mathbf{w_d = \sum a_1 \ldots a_n / \sum a_{m-n} \ldots a_m}$$

где a_i — это показатель аномальности связи, m — это число связей в документе, а n — размер выборки связей для вычисления веса (в наших экспериментах n = 10). Данная формула имеет тенденцию к завышению аномальности более длинных документов, т.к. в них обычно разрыв между полюсами сортированного списка связей выше, поэтому в некоторых случаях полученный результат следует поделить на длину документа или нормировать его каким-то иным образом.

Теперь у нас есть вся необходимая информация для анализа корпуса на предмет выявления аномалий. Для решения задачи поиска «черных лебедей», или аномальных документов в коллекции текстов мы используем сортировку документов по критерию аномальности. Для решения задачи маркирования аномальных слов в отобранном подмножестве текстов мы подсвечиваем в документах слова, входящие в наиболее весомые синтаксические связи. Первичная верификация методики происходила на основе экспертной разметки корпуса документов, результат подтвердил работоспособность предложенного алгоритма: тексты, размеченные вручную как аномальные, получили значительно более высокое значение показателя аномальности.

Следует отметить, что первые результаты продемонстрировали высокую аномальность таких документов, которые обычно считаются «мусорными» или «спамом». Поэтому одной из основных задач дальнейшего исследования становится вычисление некоторой меры релевантности документа данной

предметной области в целом, который можно было бы использовать для исключения из выдачи тех документов, которые представляют собой информационный шум.

ЛИТЕРАТУРА

1. Mahapatra A., Srivastava N., Srivastava J. Contextual Anomaly Detection in Text Data // Algorithms. 2012. №5.
2. Cilibrasi R., Vitanyi P. The google similarity distance // IEEE Transactions on Knowledge and Data Engeneering. 2007. №19.
3. Lin D. An Information-Theoretic Definition of Similarity // Proceedings of the 15th International Conference on Machine Learning (Madison, WI, USA, 24–27 July 1998). 1998. P. 296–304.
4. Manevitz L. Yousef M. Document Classification on Neural Networks Using Only Positive Examples // Proceedings of the 23rd Annual International ACM SIGIR Conference Research and Development in Information Retrieval (New Orleans, USA, 24–28 July 2000). 2000. Vol. 34. P. 304–306.
5. Manevitz L. Yousef M. One-class SVMs for document classification // Journal of Machine Learning Research. 2002. №2.