Метод прогнозирования развития ситуаций на основе обнаружения событий в потоке текстовых документов

© А.М. Андреев © Д.В. Березкин © И.А. Козлов

Московский государственный технический университет им. Н. Э. Баумана, Москва

arkandreev@gmail.com berezkind@bmstu.ru kozlovilya89@gmail.com

Аннотация. Рассмотрен метод автоматизированного прогнозирования развития ситуаций на основе обнаружения событий в потоке текстовых документов. Описаны существующие подходы к анализу ситуаций, выявлены их преимущества и недостатки с точки зрения специфики решаемой задачи. Предложен метод формирования сценариев развития ситуаций на основе принципа исторической аналогии, учитывающий динамику развития ситуаций. Этот метод позволяет оценивать вероятность реализации сформированных сценариев с помощью логистической регрессии. Представлен метод выделения оптимистического и пессимистического сценариев на основе метода анализа иерархий. Описан способ снабжения сценариев предложениями для лиц, принимающих решения. Представлены результаты экспериментальной оценки качества разработанного метода.

Ключевые слова: ситуационный анализ, прогнозирование, сценарный анализ, система поддержки принятия решений, аналогия, анализ текстового потока.

Method for Forecasting of Situations Development Based on Event Detection in Text Stream

© Ark Andreev © Dmitry Berezkin © Ilya Kozlov

Bauman Moscow State Technical University, Moscow

arkandreev@gmail.com berezkind@bmstu.ru kozlovilya89@gmail.com

Abstract. The article deals with the problem of automated forecasting of situations development based on event detection in a stream of text documents. Existing methods of situational analysis are analyzed and their advantages and disadvantages in view of the specifics of the task are determined. A method for generation of possible scenarios of situations development is described. The method generates scenarios on the principle of historical analogy, taking into account the dynamics of situation development. The probability of the generated scenarios' implementation is estimated via logistic regression. A method for the optimistic and the pessimistic scenario identification based on analytic hierarchy process is proposed. A way to supplement scenarios with recommendations for decision-makers is described. The results of experimental evaluation of the developed method's quality are presented.

Keywords: situational analysis, forecasting, scenario analysis, decision support system, analogy, text stream analysis.

1 Введение

В настоящий момент большое количество данных, обрабатываемых современными информационными системами (ИС), имеет форму информационных потоков: новые информационные сообщения постоянно поступают из источников и должны обрабатываться ИС с минимальной задержкой. Как правило, информация в потоке представлена в неструктурированном виде, в частности, в форме текста. Так, форму текстовых потоков имеют сообщения пользователей в социальных сетях, новости СМИ, официальные заявления органов власти.

Динамический характер текстовых потоков делает их важным средством информационной поддержки для людей, которым требуется принимать управленческие решения в режиме реального времени в условиях меняющейся обстановки. Задачи своевременного обнаружения проблемной ситуации, отслеживания её развития и оперативного принятия решений по управлению развитием ситуации возникают в различных сферах – политической, социальной, военной, экономической.

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL'2017), Москва, Россия, 10–13 октября 2017 года

Анализ текстового потока позволяет осуществлять мониторинг интересующих пользователей тем, т. е. обнаруживать возникновение важных событий, относящихся к тем или иным явлениям или объектам [5]. Обнаруживаемые события отражают развитие различных ситуаций с течением времени. Однако для принятия наилучших решений необходимо также определять возможные варианты дальнейшего развития этих ситуаций – это позволяет на основе полученного прогноза предпринимать определенные шаги, направленные на изменение ситуации в нужную сторону.

В статье предложено решение задачи автоматизированного прогнозирования развития ситуаций на основе анализа потока текстовых сообщений.

2 Постановка задачи

2.1 Функционирование системы мониторинга развития ситуаций

В [5] предложено решение задачи мониторинга тем на основе обнаружения событий, релевантных заданным темам, в потоке текстовых документов. Под событием понимается некоторое изменение, произошедшее в реальном мире и отраженное в текстовом потоке. Обнаружение событий рассматривается как задача кластеризации, заключающаяся в разбиении текстового потока на группы документов, описывающих различные события. Для этого каждый документ представляется многокомпонентной моделью, компоненты которой описывают содержание, структуру и метаданные документа: $d_i = (d_i^w, d_i^{tw}, d_i^c, d_i^p, d_i^n, d_i^{dt}, d_i^e, d_i^g, d_i^t)$. В частности, текстовое содержание документа представлено вектором $d_i^w = (w_i^1, w_i^2, \dots, w_i^{N^w})$, каждый элемент которого w_i^k отражает значимость k-го терма в контексте документа и рассчитывается с помощью метода TF-IDF. Каждое событие также описано многокритериальной моделью, компоненты которой формируются на основе документов, относящихся к событию: $\varepsilon_j = (\varepsilon_j^w, \varepsilon_j^{tw}, \varepsilon_j^c, \varepsilon_j^p, \varepsilon_j^n, \varepsilon_j^{dt}, \varepsilon_j^e, \varepsilon_j^g, \varepsilon_j^t)$. Объединение документов в группы выполняется с помощью алгоритма инкрементальной кластеризации, в основе которого лежит покомпонентное сопоставление каждого документа с ранее обнаруженными событиями. Более подробно модели документа и события, а также метод обнаружения событий описаны в [5].

Метод позволяет работать с документами на произвольном языке при наличии подготовленных экспертами тематических запросов, а также словарей имен персон, названий организаций и географических наименований на соответствующем языке. Для повышения качества обнаружения событий могут быть использованы наработки авторов в области морфологического [6], синтаксического [4] и семантического [7] анализа текстов. Для отслеживания изменения обстановки с течением времени необходимо формировать ситуации – цепочки взаимосвязанных событий, отражающие развитие тех или иных процессов. Для этого из множества обнаруженных событий выделяют пары взаимосвязанных событий $p_{ij} = (\varepsilon_i, \varepsilon_j)$, потенциально принадлежащих одной ситуации. На основе формирования таких пар выполняется построение ситуационного графа $G = (E, P)$. В этом графе узлы $E = \{\varepsilon_i\}$ соответствуют событиям, а ребра $P = \{p_{ij}\}$ – выделенным парам (каждое ребро является ориентированным и направлено к более позднему событию пары). Любой путь в этом графе является потенциальной ситуацией $s = (\varepsilon_s^1, \varepsilon_s^2, \dots, \varepsilon_s^n)$.

На Рис. 1 представлен пример ситуации, представляющей собой последовательность из четырех взаимосвязанных событий.

2.2 Особенности решаемой задачи

Прогнозирование заключается в построении возможных сценариев развития ситуации. Каждый сценарий представляет собой потенциальное продолжение текущей ситуации, т. е. цепочку событий, которые могут наступить в будущем. Для эффективного использования результатов прогнозирования из множества сформированных сценариев необходимо выделить три варианта, представляющих наибольший интерес для лиц, принимающих решения (ЛПР), – пессимистический, оптимистический и наиболее вероятный. На основе результатов прогнозирования должны приниматься решения по управлению ситуацией. Поэтому помимо сформированных сценариев пользователю должны предлагаться предложения по действиям, которые необходимо предпринять для способствования развитию ситуации по наиболее благоприятному сценарию.

3 Обзор существующих подходов к анализу ситуаций

В некоторых работах, посвященных ситуационному анализу, ситуации описываются совокупностями определенных числовых показателей [13]. Для прогнозирования в этом случае могут использоваться методы анализа временных рядов и методы регрессионного анализа. Такие подходы не могут быть использованы для анализа развития ситуаций на основе текстового потока, поскольку требуемый результат прогнозирования имеет качественный, а не количественный характер.

В ряде работ предложены подходы к формированию сценариев на основе когнитивных карт и знаковых орграфов [14, 16]. В них ситуация представляется как граф, узлы которого соответствуют факторам ситуации, а ребра отражают влияние факторов друг на друга. Прогнозирование заключается в оценке будущих значений факторов путем моделирования изменения ситуации с учетом различных управляющих воздействий. Построение описания ситуации в виде когнитивной карты выполняется экспертом, поэтому такие подходы неприменимы для автоматического формирования сценариев развития ситуаций.

Ситуация: Тестирование беспилотных такси Uber

Компания Uber запустила беспилотное такси в США

Имя документа	Дата публикации	Время публикации	Источник
Компания Uber запустила беспилотное такси в США	14.09.2016	14:41:39	ТАСС
Uber запустил первые беспилотные такси в США	14.09.2016	17:19:37	РБК
Беспилотные такси выехали на дороги	14.09.2016	19:40:00	Комсомольская Правда

Власти Калифорнии требуют от Uber прекратить использование беспилотных такси

Имя документа	Дата публикации	Время публикации	Источник
Власти Калифорнии требуют, чтобы Uber свернул сервис беспилотного такси в Сан-Франциско	15.12.2016	07:27:16	ТАСС
Власти Калифорнии требуют от Uber прекратить использование беспилотных такси	15.12.2016	07:55:00	Коммерсант
Власти Калифорнии потребовали прекратить эксперимент Uber с беспилотными такси	15.12.2016	09:15:00	Интерфакс
В Калифорнии потребовали ликвидировать сервис беспилотного такси Uber	15.12.2016	09:28:00	Комсомольская Правда

Власти вынудили Uber свернуть онлайн-вызов такси с автопилотом в Сан-Франциско

Имя документа	Дата публикации	Время публикации	Источник
Власти вынудили Uber свернуть онлайн-вызов такси с автопилотом в Сан-Франциско	22.12.2016	06:01:00	ТАСС
Uber приостановил испытания беспилотных такси в Калифорнии	22.12.2016	10:59:00	Интерфакс

Uber перенесла беспилотные такси в Аризону

Имя документа	Дата публикации	Время публикации	Источник
Uber перенес испытания беспилотных такси в Аризону	23.12.2016	19:51:00	Интерфакс
Uber перенесла беспилотные такси в Аризону	23.12.2016	20:10:00	Вести Экономика
Uber после неудачи в Калифорнии протестирует сервис беспилотного такси в Аризоне	24.12.2016	07:01:18	ТАСС

Рисунок 1 Пример выявления событий и формирования ситуации

Многие работы используют принцип аналогии – прогнозирование дальнейшего развития ситуации и формирование предложений по управляющим действиям основано на поиске аналогичных ситуаций, имевших место в прошлом. В работах, базирующихся на принципе аналогии, используются различные подходы к представлению ситуаций.

В [10] ситуация представляется фрагментом семантической сети, содержащим объекты и их отношения в рамках ситуации. Получить такое представление автоматически возможно лишь для определенных предметных областей, поэтому такой подход нельзя использовать для прогнозирования развития произвольных ситуаций.

В ряде работ предложено описание ситуаций в виде набора или вектора параметров с определенными значениями [1, 12]. Для сравнения ситуаций с целью определения аналогии используются евклидово расстояние, манхэттенская метрика, расстояние Чебышева, мера Хэмминга, косинусная мера и другие меры близости. Недостаток данных подходов заключается в статическом описании ситуаций – при определении близости между ситуациями не учитывается сходство динамики их развития.

Для учета динамики можно использовать описание эталонной ситуации в виде графа или автомата [3, 9, 11, 15, 18], пути в котором отражают различные варианты развития ситуации. Все эти подходы позволяют применять лишь принцип строгой аналогии: анализируемая текущая ситуация должна точно соответствовать некоторому пути в графе, построенном экспертом. Однако цепочка событий, автоматически построенная при анализе текстового потока, не всегда точно соответствует эталону – в ней могут содержаться дополнительные события или, напротив, отсутствовать какие-либо события из графа.

Подход на основе нестрогой аналогии предложен в [8]. Ситуации представляются цепочками событий, близость между ними определяется с помощью модифицированного расстояния Левенштейна. Но этот подход требует выделения для каждого события объекта и субъекта, что не может быть сделано автоматически для произвольных текстовых сообщений. Кроме того, результат определения аналогов текущей ситуации в названной работе используется лишь для отнесения этой ситуации к одному из заданных классов.

4 Предлагаемый метод прогнозирования развития ситуаций

Обнаружение для текущей последовательности s_c цепочки-аналога s_e позволяет не только определить вероятный итог развития ситуации (как предлагается в [8]), но и объяснить, какие события могут привести к этому итогу. Такой прогноз можно получить, если обнаружено сходство всей текущей последовательности с начальной частью $st(s_e, s_c)$ цепочки-аналога. В этом случае можно предположить, что в будущем наступят события, аналогичные тем, которые составляют заключительную часть цепочки-аналога $fin(s_e, s_c)$. Таким образом, эту заключительную часть можно рассматривать как возможный сценарий дальнейшего развития текущей ситуации.

Для выполнения сопоставления необходимо наличие базы ситуаций-эталонов $S_e = \{s_e^i\}$. Такие эталоны отбираются экспертами в зависимости от

задачи, для которой используется система мониторинга. Так, для анализа ситуации, связанной с тестированием беспилотных такси (рис. 1), использовалась база эталонных ситуаций, отражающих развитие различных технологий в прошлом.

Поскольку текущие ситуации представляют собой пути в ситуационном графе, процесс прогнозирования состоит из следующих этапов:

1. При появлении в ситуационном графе нового события ε_c (либо при изменении существующего события) осуществляется поиск аналогичных ему событий, принадлежащих эталонным ситуациям.
2. При нахождении эталонного события $\varepsilon_e \in s_e$, аналогичного событию ε_c, выполняется попытка выделить в графе цепочку событий s_c (текущую ситуацию), которая содержит событие ε_c и имеет максимальное сходство с начальной частью $st(s_e, s_c)$ последовательности s_e. Если s_c является аналогом s_e, то заключительная часть эталонной ситуации $fin(s_e, s_c)$ признается возможным сценарием развития текущей ситуации.
3. Сценарии, сформированные для текущей ситуации, ранжируются по приоритетности. Наиболее приоритетный сценарий считается оптимистическим, наименее приоритетный – пессимистическим.
4. Формируются предложения по действиям, которые необходимо предпринять для способствования развитию текущей ситуации по благоприятным сценариям.

4.1 Обнаружение аналогичных событий

Событие представляет собой некоторое изменение ситуации в реальном мире. Однако текстовое описание события характеризует не только само изменение, но и его контекст, т. е. содержит информацию о ситуации в целом. Например, в сообщении о завершении тушения пожара содержится некоторая общая информация о чрезвычайной ситуации – место и время возникновения пожара, причина и условия протекания. Аналогичными будем считать события, соответствующие схожим изменениям ситуаций без учета контекста.

Для определения аналогичности события ε_i, принадлежащего ситуационному графу, и события $\varepsilon_{s_e}^j$, принадлежащего эталонной ситуации s_e, определим расстояние $\gamma_{an}(\varepsilon_i, \varepsilon_{s_e}^j)$ между ними. Функция $\gamma_{an}(\varepsilon_i, \varepsilon_j)$ принимает неотрицательные значения, причем $\gamma_{an}(\varepsilon_i, \varepsilon_j) = 0$, если события ε_i и ε_j описывают полностью идентичные изменения, произошедшие в рамках соответствующих ситуаций. Если расстояние $\gamma_{an}(\varepsilon_i, \varepsilon_{s_e}^j)$ меньше порогового значения Th_{an}, делается вывод о том, что текущее событие ε_i аналогично эталону $\varepsilon_{s_e}^j$.

При определении аналогичности событий учитываются их названия, текстовые описания и тематический состав. Текстовое описание события ε_i задается вектором $\varepsilon_i^w = (w_i^1, w_i^2, \dots, w_i^{N^w})$, где N^w – количество различных слов, встречающихся в описаниях событий, w_i^l – вес -го слова в описании i-го события, который находится методом TF-IDF.

Для того чтобы наиболее важную роль при определении аналогичности играли термы, характерные для конкретного события, а не ситуации в целом, было решено умножать вес каждого терма w_i^l в ε_i^w на коэффициент k_i^l, отражающий соотношение значимости терма для события и для ситуации s, к которой относится это событие: $k_i^l = w_i^l len(s) / \sum_{\varepsilon_j \epsilon s} w_j^l$, где $len(s)$– количество событий в ситуации s:

$$\varepsilon_i^{w\prime} = (w_i^1 k_i^1, w_i^2 k_i^2, \dots, w_i^{N^w} k_i^{N^w}).$$

Расстояние между событиями с точки зрения текста рассчитывается на основе косинусной меры: $\gamma_{i,j}^w = 1 - sim_{\cos}(\varepsilon_i^{w\prime}, \varepsilon_j^{w\prime})$. Представление слов названия события $\varepsilon_i^{tw\prime}$ и расчёт расстояния между событиями с точки зрения названий $\gamma_{i,j}^{tw}$ выполняется аналогично.

Тематический состав события характеризует вектор $\varepsilon_i^t = (t_i^1, t_i^2, \dots, t_i^{N^t})$, где N^t – количество анализируемых тем, а t_i^l – значение, отражающее релевантность -го события l-ой теме. Темы задаются экспертами в виде формализованных поисковых запросов, а значения t_i^l рассчитываются на основе модифицированного метода Okapi BM25 с помощью поисковой машины Sphinx [2]. Расстояние между событиями с точки зрения тематического состава $\gamma_{i,j}^t$ также определяется на основе косинусной меры близости векторов: $\gamma_{i,j}^t = 1 - sim_{\cos}(\varepsilon_i^t, \varepsilon_i^t)$.

Расстояние между событиями с точки зрения аналогичности может быть представлено как взвешенная сумма расстояний по различным критериям:

$$\gamma_{an}(\varepsilon_i, \varepsilon_j) = \lambda^w \gamma_{i,j}^w + \lambda^{tw} \gamma_{i,j}^{tw} + \lambda^t \gamma_{i,j}^t.$$

Нахождение значений коэффициентов λ^w, λ^{tw}, λ^t и порогового значения Th_{an} может быть выполнено путем решения задачи линейной бинарной классификации, состоящей в отнесении векторов $\gamma_{i,j} = (\gamma_{i,j}^w, \gamma_{i,j}^{tw}, \gamma_{i,j}^t)$ к одному из двух классов: один означает аналогичность сравниваемых событий, а второй – её отсутствие. Решение задачи заключается в построении разделяющей плоскости:

$$\lambda^w \gamma_{i,j}^w + \lambda^{tw} \gamma_{i,j}^{tw} + \lambda^t \gamma_{i,j}^t - Th_{an} = 0.$$

Анализируемый вектор $\gamma_{i,j}$ относится к одному из классов, исходя из его расположения относительно плоскости.

Для решения задачи может быть использована машина опорных векторов (SVM). Чтобы обеспечить возможность нахождения расстояния $\gamma_{an}(\varepsilon_i, \varepsilon_j)$ как взвешенной суммы значений $\gamma_{i,j}^w$, $\gamma_{i,j}^{tw}$ и $\gamma_{i,j}^t$, необходимо использовать SVM с линейным ядром. Для обучения машины используется набор векторов

обоих классов, подготовленный экспертами.

Описанный способ обнаружения аналогов позволяет находить для текущих событий схожие события, происходившие в прошлом. Так, для события «Власти вынудили Uber свернуть онлайн-вызов такси с автопилотом в Сан-Франциско» (Рис. 1) такими аналогами являются другие случаи запрета использования тех или иных технологий органами власти по соображениям безопасности, в частности, событие «США официально запретили продажу Samsung Galaxy Note 7». После обнаружения события-аналога выполняется попытка выделения в ситуационном графе цепочки, аналогичной соответствующей эталонной ситуации (в данном случае – ситуации, касающейся проблем Samsung, связанных со смартфоном Galaxy Note 7).

4.2 Определение близости между ситуациями

На формируемую текущую ситуацию накладывается следующее ограничение: события, аналогичные событиям из эталонной цепочки, должны следовать друг за другом в том же порядке, что и соответствующие события в эталонной ситуации. Это связано с тем, что последовательность событий в эталонной цепочке отражает их причинно-следственную связь и логику развития ситуации. Если в текущей и эталонной последовательностях события располагаются в разном порядке, значит, логика их развития различна, и они не могут быть признаны аналогами.

Таким образом, при определении близости между ситуациями необходимо учитывать, что цепочки содержат ряд попарно аналогичных событий, располагающихся в цепочках в одинаковом порядке (на Рис. 2 они выделены серым цветом, пунктирной линией соединены события-аналоги). Кроме того, каждая из ситуаций может содержать события, аналоги которых отсутствуют в другой цепочке. На Рис. 2 эталонные события, аналоги которых отсутствуют в текущей ситуации, выделены вертикальной штриховкой, «лишние» события текущей ситуации – горизонтальной. Также необходимо помнить о том, что при сравнении учитывается лишь начальная часть эталонной ситуации $st(s_e, s_c)$ – от её первого события ($\varepsilon_{s_e}^1$ на рис. 2) до последнего события, имеющего аналог в текущей ситуации ($\varepsilon_{s_e}^6$ на рис. 2). События, составляющие заключительную часть эталонной ситуации $fin(s_e, s_c)$, не влияют на значение близости.

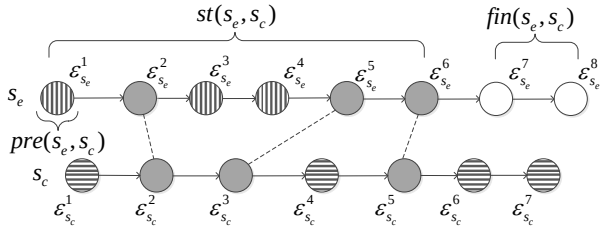

Рисунок 2 Сопоставление цепочек событий

В Таблице 1 представлен пример сравнения текущей ситуации с эталонной. В данном случае пары событий (ε_1^1,ε_2^1), (ε_1^2,ε_2^2) и (ε_1^3,ε_2^3) являются аналогами, событие ε_1^4 является «лишним» событием текущей ситуации, а событие ε_2^4 является заключительной частью эталонной ситуации.

Таблица 1 Сопоставление текущей и эталонной ситуаций

Текущая ситуация	**Эталонная ситуация**
ε_1^1: Компания Uber запустила беспилотное такси в США	ε_2^1: Выпущен Samsung Galaxy Note 7
ε_1^2: Власти Калифорнии требуют от Uber прекратить использование беспилотных такси	ε_2^2: Власти США призвали отказаться от использования Samsung Galaxy Note 7
ε_1^3: Власти вынудили Uber свернуть онлайн-вызов такси с автопилотом в Сан-Франциско	ε_2^3: США официально запретили продажу Samsung Galaxy Note 7
ε_1^4: Uber перенесла беспилотные такси в Аризону	
	ε_2^4: Samsung объявил о прекращении производства Galaxy Note 7

Для измерения близости ситуаций используется метод, представляющий собой модификацию расстояния Левенштейна: расстояние между цепочками определяется суммарным весом операций, необходимых для превращения одной цепочки в другую. Рассмотрим операции, которые требуются для превращения начальной части эталонной ситуации $st(s_e, s_c)$ в текущую ситуацию s_c, а также способы измерения веса этих операций.

- Удаление событий $\varepsilon_{s_e}^i$, аналоги которых отсутствуют в текущей ситуации. В качестве веса операции $w_{del}(\varepsilon_{s_e}^i)$ может использоваться значимость удаляемого события – показатель, учитывающий количество документов, описывающих событие, и авторитетность источников, опубликовавших эти документы. Множество удаляемых событий обозначим E_{del}. Суммарный вес таких операций: $W_{del} = \sum_{\varepsilon_{s_e} \epsilon E_{del}} w_{del}(\varepsilon_{s_e})$.
- Добавление событий $\varepsilon_{s_c}^i$, аналоги которых отсутствуют в эталонной ситуации. Вес операции $w_{add}(\varepsilon_{s_c}^i)$ вычисляется аналогично. Множество добавляемых событий обозначим E_{add}. Суммарный вес операций добавления: $W_{add} = \sum_{\varepsilon_{s_c} \epsilon E_{add}} w_{add}(\varepsilon_{s_c})$.
- Замена события из эталонной цепочки $\varepsilon_{s_e}^i$ на его аналог $\varepsilon_{s_c}^j$. Вес этой операции $w_{rep}(\varepsilon_{s_e}^i, \varepsilon_{s_c}^j)$ определяется расстоянием $\gamma_{an}(\varepsilon_{s_e}^i, \varepsilon_{s_c}^j)$ между событиями $\varepsilon_{s_e}^i$ и $\varepsilon_{s_c}^j$ с точки зрения аналогичности.

Множество пар $(\varepsilon_{s_e}^i, \varepsilon_{s_c}^j)$ обозначим P_{rep}. Суммарный вес операций этого вида: $W_{rep} = \sum_{(\varepsilon_{s_e}^i, \varepsilon_{s_c}^j)\epsilon P_{rep}} w_{rep}(\varepsilon_{s_e}^i, \varepsilon_{s_c}^j)$.

- Изменение (сокращение или удлинение) временного интервала $t_{s_e}^{i,j}$ между событиями. Интервалу $t_{s_e}^{i,j}$ в эталонной последовательности соответствует интервал $t_{s_c}^{k,l}$ в текущей ситуации, где $\varepsilon_{s_c}^k$ – аналог $\varepsilon_{s_e}^i$, а $\varepsilon_{s_c}^l$ – аналог $\varepsilon_{s_e}^j$. Вес этой операции $w_{trep}(t_{s_e}^{i,j}, t_{s_c}^{k,l})$ определяется относительной разностью между величинами интервалов: $w_{trep}(t_{s_e}^{i,j}, t_{s_c}^{k,l}) = |t_{s_e}^{i,j} - t_{s_c}^{k,l}|/t_{s_e}^{i,j}$. Множество пар $(t_{s_e}^{i,j}, t_{s_c}^{k,l})$ обозначим T_{trep}. Суммарный вес таких операций: $W_{trep} = \sum_{(t_{s_e}^{i,j}, t_{s_c}^{k,l})\epsilon T_{trep}} w_{trep}(t_{s_e}^{i,j}, t_{s_c}^{k,l})$.

Ввиду различия способов расчёта веса операций разных типов, при вычислении расстояния между цепочками они должны учитываться с различными коэффициентами. Кроме того, расстояние необходимо нормировать, поскольку, чем короче учитываемая при сравнении часть эталонной цепочки, тем меньше модифицирующих операций с ней можно произвести с сохранением аналогичности полученной последовательности оригиналу. Таким образом, расстояние между эталонной и текущей цепочкой

$$\rho(s_e, s_c) = \frac{\theta^T W}{len\big(st(s_e, s_c)\big)} = \\ = \frac{(\theta_{del} W_{del} + \theta_{add} W_{add} + \theta_{rep} W_{rep} + \theta_{trep} W_{trep})}{len(st(s_e, s_c))},$$

где $len(st(s_e, s_c))$ – количество событий в начальной части $st(s_e, s_c)$ эталонной ситуации s_e, а θ_{del}, θ_{add}, θ_{rep} и θ_{trep} – коэффициенты, определяющие вклад операций различных типов в значение расстояния.

4.3 Определение вероятности аналогичности ситуаций

На основе расстояния $\rho(s_e, s_c)$ необходимо определить, является ли текущая ситуация аналогом эталонной и какова вероятность того, что текущая ситуация будет развиваться по сценарию, определяемому эталонной ситуацией. С этой целью было принято решение рассмотреть сравнение цепочек как задачу логистической регрессии. Для этого введем переменную y, принимающую одно из двух возможных значений:

$$y = \begin{cases} 1, \text{если цепочки не являются аналогами,} \\ 0, \text{если цепочки являются аналогами.} \end{cases}$$

Предположим, что вероятность наступления события $y = 0$ (т. е. вероятность того, что текущая ситуация является аналогом эталонной) задана функцией:

$$\mathrm{P}(y = 0|s_e, s_c) = 1 - \frac{1}{1 + \exp\left(-\frac{\theta^T W}{len\big(st(s_e, s_c)\big)}\right)}.$$

Значения параметров θ подбираем методом максимального правдоподобия на основе обучающей выборки, состоящей из множества пар аналогичных и неаналогичных ситуаций.

Логистическая регрессия позволяет также выполнить бинарную классификацию пар ситуаций: цепочки s_e и s_c считаются потенциальными аналогами при $\mathrm{P}(y = 0|s_e, s_c) > 0.5$.

4.4 Формирование сценария

Построение текущей ситуации начинается с нового или измененного события ситуационного графа ε_c, которое обязательно должно ей принадлежать. Далее на каждом шаге выполняется попытка дополнить ситуацию путем присоединения к цепочке одного из соседей события, которое на данный момент является первым или последним в цепочке. При этом необходимо рассмотреть различные варианты интерпретации добавляемого в цепочку события. Оно может интерпретироваться и как аналог некоторого события из s_e, и как «лишнее» событие, не имеющее аналогов в эталонной цепочке. Путем выбора на каждом шаге одного из возможных событий, добавляемых в цепочку, а также одного из возможных вариантов его интерпретации формируется дерево возможных вариантов построения текущей ситуации. Из всех вариантов текущей ситуации, рассмотренных в процессе построения, выбирается цепочка s_c^{max}, имеющая максимальную близость к эталону. Эта последовательность считается завершенной текущей ситуацией.

Если $\mathrm{P}(y = 0|s_e, s_c^{max}) > 0.5$, полученная текущая ситуация признается аналогом s_e. В этом случае $fin(s_e, s_c^{max})$ считается возможным сценарием дальнейшего развития текущей ситуации, а значение $\mathrm{P}(y = 0|s_e, s_c^{max})$ – вероятностью того, что текущая ситуация будет развиваться в соответствии с этим сценарием. На основе всех эталонных ситуаций, аналогичных текущей, формируется множество возможных сценариев её дальнейшего развития. Заключительная часть цепочки, для которой вероятность аналогичности текущей ситуации максимальна ($s_e^{prob} = \underset{s_e}{\mathrm{argmax}}[\mathrm{P}(y = 0|s_e, s_c^{max})]$), является наиболее вероятным сценарием.

4.5 Выделение оптимистического и пессимистического сценариев

Для выделения оптимистического и пессимистического сценариев необходимо определить оптимальность каждого из них. Для этого используется метод анализа иерархий (МАИ), позволяющий определить приоритет различных альтернатив с точки зрения цели с учетом различных критериев [17]. Целью в данном случае является выбор оптимального сценария, альтернативами – сформированные сценарии, а в качестве критериев могут использоваться такие характеристики сценариев, как длительность, экономическая эффективность и другие. Выбор критериев определяется предметной областью, в рамках которой используется прогнозирование развития ситуаций.

Значения критериев для эталонных ситуаций определяются экспертами на этапе подготовки базы эталонов S_e. Также эксперты путем попарных сравнений определяют приоритетность критериев относительно цели. Приоритетность сценариев относительно каждого из критериев может быть определена автоматически при анализе ситуационного графа на основе сравнения характеристик соответствующих эталонных ситуаций. Это позволяет автоматически определить приоритет относительно цели для каждого из сценариев, сформированных для текущей ситуации. Сценарий с максимальным приоритетом считается оптимистическим, сценарий с минимальным приоритетом – пессимистическим.

На Рис. 3 показаны оптимистический и пессимистический сценарии, сформированные для ситуации, связанной с тестированием беспилотных такси. Для определения приоритетности сценариев использовались такие критерии, как «безопасность», «длительность» и «экономическая эффективность». Также на рисунке представлен наиболее вероятный сценарий, определенный с помощью логистической регрессии.

4.6 Формирование предложений для лиц, принимающих решения

С целью последующего формирования предложений эксперты должны снабжать каждое событие ε_e каждой эталонной ситуации s_e рекомендациями по действиям, которые должны предприниматься при наступлении аналогичного события в будущем. Рекомендация $rec_{\varepsilon_e} = \langle action_{\varepsilon_e}, actor_{\varepsilon_e}, period_{\varepsilon_e} \rangle$ содержит информацию о действиях $action_{\varepsilon_e}$, которые должны быть предприняты лицом $actor_{\varepsilon_e}$ в срок $period_{\varepsilon_e}$ для содействия или противодействия развитию текущей ситуации по сценарию, сформированному на основе s_e.

На Рис. 3 показаны рекомендации для ЛПР с учетом сценариев, сформированных для ситуации с тестированием беспилотных такси.

Рисунок 3 Пример формирования оптимистического, наиболее вероятного и пессимистического сценариев развития ситуации

5 Экспериментальная проверка метода

На основе предложенного метода разработана система автоматизированного мониторинга и прогнозирования развития ситуаций. Обучение системы выполняется экспертами на основе эталонных событий и ситуаций. Обученная система автоматически обрабатывает текстовый поток, обнаруживает события и формирует ситуации, а также определяет вероятные сценарии их дальнейшего развития и вырабатывает рекомендации.

Результаты качества работы подсистемы обнаружения событий приведены в [5]. Эксперименты показали, что при использовании для обучения 1300 пар документов и событий достигается значение точности 85,2%, полноты – 76% и F-меры – 79,8%.

Для анализа качества работы подсистемы формирования сценариев был проведен эксперимент с целью определения зависимости точности, полноты и F-меры выявления аналогичных ситуаций от мощности обучающей выборки. Полученные зависимости приведены на Рис. 4. В результате проведения эксперимента оказалось, что для обучения системы достаточно 90 пар ситуаций. При таком количестве обучающих примеров достигается значение F-меры около 0.8, с дальнейшим увеличением обучающей выборки качество работы метода не улучшается.

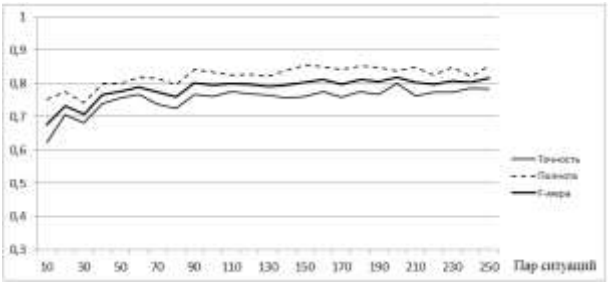

Рисунок 4 Зависимость точности (тонкая сплошная линия), полноты (пунктирная линия) и F-меры (жирная линия) от мощности обучающей выборки

6 Направления дальнейших исследований

Предложенный метод прогнозирования развития ситуаций предоставляет пользователю сценарии дальнейшего развития ситуации и рекомендации по действиям, необходимым для их реализации, но не позволяет осуществлять управление развитием ситуации по оптимальному сценарию. Пользователю требуется определять, соответствует ли развитие ситуации сформированному ранее сценарию, и получать рекомендации в случае необходимости корректировки намеченного плана мероприятий. В связи с этим дальнейшим направлением развития метода является разработка более сложных сетевых моделей эталонных ситуаций, способных отражать различные варианты возможного развития текущей ситуации в зависимости от действий ЛПР на каждом этапе управления ситуацией.

Выше описан эксперимент по оценке качества обнаружения аналогичных ситуаций, однако необходимо также оценивать качество прогнозирования. В связи с этим в рамках дальнейших исследований планируется выработать критерий качества ситуационного прогноза и выполнить оценку результатов прогноза по этому критерию.

7 Заключение

Предложен метод прогнозирования развития ситуаций на основе обнаружения событий в потоке текстовых документов. Прогнозирование состоит в формировании сценариев дальнейшего развития ситуаций по принципу исторической аналогии: выполняется построение текущей ситуации, для которой существует аналог в базе эталонных ситуаций. Этот аналог считается возможным сценарием развития текущей ситуации. Предложенный метод формирования сценариев учитывает динамику развития ситуаций и нестрогий характер аналогии между ситуациями. Из множества сформированных сценариев выделены оптимистический и пессимистический, для этого использован метод анализа иерархий. Также предложен способ подготовки предложений по действиям, которые необходимо предпринять для способствования или препятствования развитию ситуации по построенным сценариям.

Литература

[1] Aggarwal, C.C., Subbian, K.: Event Detection in Social Streams. Proc. of the 2012 SIAM Int. Conf. on Data Mining, pp. 624-635. Society for Industrial and Applied Mathematics, Philadelphia (2012). doi: 10.1137/1.9781611972825.54

[2] How Sphinx Relevance Ranking Works. http://sphinxsearch.com/blog/2010/08/17/how-sphinx-relevance-ranking-works/

[3] van der Aalst, W.M.P.: Process Mining: Data Science in Action. Springer, Heidelberg (2016). doi: 10.1007/978-3-662-49851-4

[4] Андреев, А.М., Березкин, Д.В., Брик, А.В., Смирнов, Ю.М.: Вероятностный синтаксический анализатор для информационно-поисковых систем. Вестник МГТУ. Сер. Приборостроение, 2, сс. 34-53 (2000)

[5] Андреев, А.М., Березкин, Д.В., Козлов, И.А.: Подход к автоматизированному мониторингу тем на основе обнаружения событий в потоке текстовых документов. Информационно-измерительные и управляющие системы, 15 (3), сс. 49-60 (2017)

[6] Андреев, А.М., Березкин, Д.В., Симаков, К.В.: Обучение морфологического анализатора на большой электронной коллекции текстовых документов. Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды Седьмой Всерос. науч. конф. (RCDL–2005), сс. 173-181 (2005)

[7] Андреев, А.М., Березкин, Д.В., Симаков, К.В.: Особенности проектирования модели и онтологии предметной области для поиска противоречий в правовых электронных библиотеках. Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды Шестой Всерос. науч. конф. RCDL, сс. 93-102 (2004)

[8] Ахременко, А.С.: Политический анализ и прогнозирование: Учеб. пособие. М.: Гардарики (2006)

[9] Борисов, В.В., Зернов, М.М.: Реализация ситуационного подхода на основе нечеткой иерархической ситуационно-событийной сети. Искусственный интеллект и принятие решений, 1, сс. 18-30 (2009)

[10] Варшавский, П.Р.: Методы и программные средства поиска решения на основе аналогий в интеллектуальных системах поддержки принятия решения. Дисс. … канд. техн. наук, Московский энергетический институт (2005)

[11] Волгин, Н.С.: Исследование операций, ч. 1. С-Пб.: ВМА им. Н. Г. Кузнецова (1999)

[12] Еремеев, А.П., Варшавский, П.Р.: Моделирование рассуждений на основе прецедентов в интеллектуальных системах поддержки принятия решений. Искусственный интеллект и принятие решений, 2, сс. 45-57 (2009)

[13] Зацаринный, А.А., Сучков, А.П.: Некоторые подходы к ситуационному анализу потоков событий. Открытое образование, 1, сс. 39-46 (2012)

[14] Кононов, Д.А., Косяченко, С.А., Кульба, В.В.: Формирование и анализ сценариев развития социально-экономических систем с использованием аппарата операторных графов. Автоматика и телемеханика, 68 (1), сс. 121-136 (2007)

[15] Косяченко, С.А., и др.: Модели, методы и автоматизация управления в условиях чрезвычайных ситуаций. Автоматика и телемеханика, 59 (6), сс. 3-66 (1998)

[16] Кулинич, А.А.: Компьютерные системы моделирования когнитивных карт: подходы и методы. Проблемы управления, 3, сс. 2-16 (2010)

[17] Саати, Т.: Методы анализа иерархий. М.: Радио и связь (1993)

[18] Ситчихин, А.Н.: Иерархические ситуационные модели с предысторией для автоматизированной поддержки решений в сложных системах. Дисс. … канд. техн. наук, Уфимский гос. авиационный технический университет (2002)