Научная статья
УДК 004.622; DOI: 10.61260/2218-13X-2023-3-118-128

АЛГОРИТМ МНОГОКРИТЕРИАЛЬНОГО АНАЛИЗА ТЕ КСТОВОЙ ИНФОРМАЦИИ

✉Вострых Алексей Владимирович;
Медведев Дмитрий Валерьевич.
Санкт-Петербургский университет ГПС МЧС России, Санкт-Петербург, Россия
✉a.vostrykh@list.ru

Аннотация. Представлен авторский алгоритм многокритериального анализа текстовой информации, основная цель которого заключается в повышении эффективности работы экстренных служб, в том числе и МЧС России. Данный алгоритм основан на всесторонней оценке информации, поступающей из социальных сетей с помощью спектра различного вида исследований гетерогенных данных.

Разработанный алгоритм позволит ускорить сбор данных, оптимизировать их анализ, разработать сценарии превентивного характера для возможных происшествий, а также сократить расходы ресурсов, оптимизировать управленческие предложения и обосновывать принятые решения.

Ключевые слова: социальная сеть, прогнозирование происшествий, анализ данных, алгоритм, эффективность, текстовая информация

Для цитирования: Вострых А.В., Медведев Д.В. Алгоритм многокритериального анализа текстовой информации // Науч.-аналит. журн. «Вестник С.-Петерб. ун-та ГПС МЧС России». 2023. № 3. С. 118–128. DOI: 10.61260/2218-13X-2023-3-118-128.

Scientific article

ALGORITHM FOR MULTI-CRITERIA ANALYSIS OF TEXT INFORMATION

✉Vostrykh Alexey V.;
Medvedev Dmitry V.
Saint-Petersburg university of State fire service of EMERCOM of Russia, Saint-Petersburg, Russia
✉a.vostrykh@list.ru

Abstract. The article presents the author's algorithm for multi-criteria analysis of textual information, the main purpose of which is to increase the efficiency of emergency services, including the Ministry of Emergency Situations of Russia. This algorithm is based on a comprehensive assessment of information coming from social networks using a range of different types of heterogeneous data studies.

The developed algorithm will speed up data collection, optimize their analysis, develop preventive scenarios for possible incidents, as well as reduce resource costs, optimize management proposals and justify decisions made.

Keywords: social network, accident forecasting, data analysis, algorithm, efficiency, text information

For citation: Vostrykh A.V., Medvedev D.V. Algorithm for multi-criteria analysis of text information // Scientific and analytical journal «Vestnik Saint-Petersburg university of State fire service of EMERCOM of Russia». 2023. № 3. P. 118–128. DOI: 10.61260/2218-13X-2023-3-118-128.

Введение

Современные мегаполисы ежедневно сталкиваются с широким спектром происшествий различного характера и масштаба. От скорости реагирования на данные вызовы зависит как

степень их развития, так и ущерб во всех его аспектах [1, 2]. Одним из ведомств, занимающихся реагированием на чрезвычайные ситуации, является МЧС России. От технического и информационного обеспечения деятельности министерства зависит успех ликвидации происшествий, своевременное реагирование, а также выполнение превентивных мероприятий [3, 4].

В сегодняшнем мире цифровых технологий огромную роль играет информация, её точность и скорость получения, а также обработки. В настоящее время система МЧС России обладает множеством различных баз данных, получение, обработка и объединение которых могло бы составить единую централизованную сеть, способную анализировать, измерять разнообразную информацию о происшествиях [5, 6], в том числе об её концентрации и зависимостях, например, от территории поступления сигнала.

Обладая достоверными источниками данных, необходимо также иметь средство получения и анализа информации из них, которое позволит обрабатывать весь огромный массив, выделяя только тематически релевантную информацию и предоставляя её в удобном для анализа виде оператору.

В этих целях в настоящей статье представлен авторский алгоритм многокритериального анализа текстовой информации, применение которого в работе сотрудников МЧС России позволит значительно облегчить сбор большого количества данных, оптимизировать их анализ, разработать сценарии превентивного характера для возможных происшествий, а также сократить расходы ресурсов (временных, финансовых и др.), оптимизировать управленческие предложения, обосновывать принятые решения, основываясь на конъюнктуре большого числа параметров и математического аппарата.

Методы исследования

Сложность анализа естественного языка заключается в том, что данные, подающиеся на вход, являются не структурированными, поэтому при анализе текстов возникают следующие проблемы:

– неоднозначность на уровне слов (некоторые слова имеют несколько значений);

– неоднозначность на уровне предложений (одно и то же слово может быть разными частями речи);

– флективность (изменение формы слова с помощью приставок, суффиксов, окончаний, также к этому относится падеж, число, род, время и вид);

– свободный порядок слов в предложении.

Для исключения выявленных проблем в предлагаемом алгоритме используются механизмы обработки естественного языка, которые состоят из графематического, морфологического, семантического, синтаксического и прагматического анализа текстовой информации.

Графематический анализ выделяет элементы структуры текста и является первым шагом при обработке естественного языка, в результате которого происходит декомпозиция текста на токены. Этапами проведения графематического анализа являются:

– сегментация текста на абзацы;

– сегментация каждого абзаца на предложения (основной сложностью данного этапа является то, что точки не всегда являются свидетельством об окончании предложения, например, инициалы, сокращение имени организации, название продукта через точку);

– сегментация предложений на слова. Для осуществления данного процесса существует несколько подходов: декомпозиция по пробелам, декомпозиция по знакам препинания, разделение на слова с полным удалением знаков препинания, деление на символы, разделение предложения на части слов (наиболее популярными алгоритмами являются Byte-Pair Encoding, WordPiece, Unigram и SentencePiece);

– приведение слов к нижнему регистру;

– удаление знаков препинания;

– удаление стоп-слов (как правило, это предлоги, союзы и т.д.).

Морфологический анализ представляет собой второй этап в обработке текста. Он позволяет определить морфологические характеристики для каждого из выделенных слов текста и определяет их нормальную форму (лемму).

Морфологический анализ обладает следующими ключевыми понятиями:

– словоформа (слово в тексте);

– лемма (словарная или каноническая форма слова);

– основа слова (грамматическая характеристика слова: время, род, часть речи, время и т.д.);

– лексема (набор всех форм одного слова).

Также на этапе морфологического анализа используется стемминг – каждое слово в тексте заменяется на основу слова. В некоторых случаях стемминг показывает лучшие результаты, чем при лемматизации, но на практике чаще всего применяется последнияя, так как она сохраняет больше информации из текста. Процессу лемматизации присуща лексическая и морфологическая неоднозначность, которая выражается в том, что одному слову могут соответствовать несколько лемм. Данная задача решается на следующем этапе – при синтаксическом анализе. В программной реализации данного этапа на языке программирования «Python» чаще всего используются библиотеки «pymorphy2» и «mystem».

Синтаксический анализ позволяет снять неоднозначность, которая возникает на морфологическом уровне, дополняя его результаты и формируя «дерево разбора предложения» (структура, элементы которой связаны синтаксическими правилами). Основными способами построения дерева синтаксического разбора являются: деревья зависимостей и деревья составляющих. Вершинами дерева зависимостей являются слова, а дугами – связи (зависимости) между словами, корневой вершиной является глагол. Наиболее эффективными библиотеками, используемыми для синтаксического анализа, являются: «spaCy» и «Natasha».

Следующим уровнем анализа текста является семантический анализ, который служит переходом от структуры синтаксических связей к ее смысловой интерпретации. На вход данного этапа подается синтаксическая структура текста, представленная в виде деревьев разбора, а на выходе формируется множество семантических структур, построенных в соответствии с принятой формальной нотацией (семантической моделью).

Завершающим этапом анализа текстов является прагматический анализ, который по своей структуре и методам похож на семантический. Принципиальной разницей между ними является то, что семантический анализ представляет собой процесс извлечения смысла текста на основе некоторой модели знаний, а прагматический анализ выходит за рамки моделей о предметной области и зачастую опирается на экстралингвистические факторы, такие как намерения автора, социальный контекст высказывания и т.д.

Предлагаемый в настоящем параграфе алгоритм включает в себя все рассмотренные виды анализа текста и состоит из трех этапов:

Этап 1 – Фильтрация данных.

Этап 2 – Смысловой анализ данных (включает в себя графематический, морфологический и синтаксический анализ текста).

Этап 3 – Анализ уровня эмоциональной напряженности (включает в себя семантический и прагматический анализ текста).

На первом этапе производится фильтрация поступающих данных из социальных сетей по определенным параметрам, таким как: временной интервал публикации сообщений, геопозиция автора, геопозиция места происшествия, фильтрация по тегам и т.д.

На втором этапе отобранные данные с первого этапа анализируются по смысловой нагрузке – производится поиск сообщений, по ключевым словам и словосочетаниям с целью составления выборки, например, по определенному происшествию.

На третьем этапе собранные данные анализируются по уровню эмоциональной нагрузки, выделяются сообщения с повышенной эмоциональной насыщенностью.

Рассмотрим более подробно способы получения данных и информационные процессы на каждом этапе.

На первом этапе, при получении оператором МЧС России списка сообщений из подтвержденных источников социальных сетей о происшествиях, они отправляются в систему фильтрации по заданным критериям (время публикации $\{P_{time}\}$, геопозиции $\{P_{lok}\}$, тегам $\{P_{teg}\}$ и т.д.):

$$\theta_{\{F\}}\left(\{P_{time}\},\{P_{lok}\},\{P_{teg}\},\dots,\{P_n\}\right)=\begin{cases}\text{A, если использован 1 критерий;}\\ B\text{, если использован 2 критерий;}\\ C\text{, если использован 3 критерий;}\\ \cdot\\ \cdot\\ D\text{ другие варианты 4.}\end{cases} \quad (1)$$

После проведения фильтрации переходим на второй этап, где происходит сегментация текста на абзацы, затем на предложения и слова (токены). Полученные токены приводятся к нижнему регистру, удаляются знаки препинания и «стоп-слова». Затем слова приводятся к нормальной форме или к основе (лемматизация). Далее вычисляется ряд индикаторов, позволяющих ранжировать выделенные слова и словосочетания, а вместе с ними и сообщения:

1) индикатор дивергенции Кульбака-Лейблера [7] позволяет провести сравнение распределения терминов (реальное и теоретическое). Рассчитывается с помощью формулы:

$$Kl(z)=\sum_{x\epsilon D} p_t(z,d)*ln\left(\frac{p_t(z,d)}{p_n(d)}\right), \quad (2)$$

где z – реальное распределение термина; p_n – вероятность найти термин z во множестве исследуемых сообщений относительно длины определенного сообщения *d,* вычисляется с помощью формулы:

$$p_n(d)=\frac{N(d)}{\sum_{x\epsilon D} N(x)},$$

где $N(d)$ – сумма терминов в сообщении *d*; $\sum_{x\epsilon D} N(x)$ – сумма терминов *x* во всем массиве сообщений *D*; $p_t(z,d)$ – вероятность появления термина *z* в документе *d,* вычисляется с помощью формулы:

$$p_t(z,d)=\frac{tf(z,d)}{F(z)},$$

где $tf(z,d)$ – вероятность употребления термина *z* в сообщении *d*; $F(z)$ – вероятность употребления термина *z* в массиве сообщений *D*;

2) индикатор информационной энтропии характеризует равномерность распределения термина по сообщению, вычисляется с помощью формулы:

$$Ie(z)=\sum_{d\epsilon D} p_t(z,d)*ln\left(\frac{1}{p_t(z,d)}\right). \quad (3)$$

Если данный показатель $Ie(z)>0$, то термин равномерно представлен в коллекции документов, если $Ie(z)=0$, то термин *z* встречается только в одном тексте или сообщении;

3) индикатор выделения общеупотребительных слов [8] демонстрирует отличие распределения терминов *z* в эталонном массиве. Индикатор вычисляется по формуле:

$$R(z)=\frac{p_e(z)}{p_k(z)}, \quad (4)$$

где p_e – относительная частота встречаемости термина в эталонном массиве; p_k – относительная частота встречаемости термина в Национальном корпусе русского языка.

Данный индикатор позволяет выделить большую часть общих слов, если они представлены на портале. Для общеупотребительных терминов индикатор имеет значение около 1, а для малоупотребительных >> 1;

4) индикаторы, основанные на распределении Бернулли [9], вычисляются с помощью сравнения реального распределения терминов в массиве с теоретическим распределением Бернулли по формулам:

$$\begin{cases} Z_1(z) = \sum_{x \in D} Z_{risk,1}(z,x) \\ Z_2(z) = \sum_{x \in D} Z_{risk,2}(z,x) \\ Z_{risk,1}(z,d) = \frac{-\log_2 Prob_{norm}(z,d)}{tf(z,d)+1} \\ Z_{risk,2}(z,d) = \frac{F(z)(-\log_2 Prob_{norm}(z,d))}{df(z)(tf(z,d)+1)} \end{cases}, \quad (5)$$

где $df(z)$ – количество документов в массиве, содержащих термин z; $Prob_{norm}$ рассчитывается по формулам:

$$\begin{cases} Prob_{norm}(w,d) = \frac{Prob(w,d)}{\sum_{x \in D} Prob(w,d)} \\ Prob(w,d) = 2^{-\log_2 Prob_1(w,d)} \\ Prob_1(w,d) = B(N,F,X) = \left(\frac{F(w)}{tf(w,d)}\right) p^{tf(w,d)} q^{F(w)-tf(w,d)} \end{cases};$$

5) индикатор Флеша-Кинсайда [10] оценивает читабельность текста с помощью формулы:

$$FRE = 206{,}835 - 1{,}3\left(\frac{a}{b}\right) - 60{,}1 * \left(\frac{c}{a}\right), \quad (6)$$

где a – количество слов в документе; b – количество предложений в документе; c – количество слогов в документе.

Чем выше полученное значение, тем легче восприятие информации;

6) индикатор семантического алгоритма Гинзбурга [11] определяет семантическую близость двух терминов относительно их окружения в пределах сообщения. Индикатор вычисляется с помощью формулы:

$$ind(z|c) = \frac{N_{zc} N_t}{N_{tc} N_z}, \quad (7)$$

где N_{zc} – встречаемость термина z со словом c; N_t – общее число терминов в массиве сообщений; N_{tc} – сумма терминов в окружении термина c; N_z – встречаемость термина z в массиве сообщений.

Индекс значимости рассчитывается для всех терминов Z, встречающихся в определенном предложении C. Если $ind(z|c) > 1$ – то данный показатель значим при расчёте. Индикатор связанности по Гинзбургу [11] определяет силу семантической связи между двумя словами. Рассчитывается на основе индексов значимости по формуле:

$$ginz(Z,C) = 1 - \frac{sum(Z)+sum_{razn}+sum(C)}{sum_{all}}, \quad (8)$$

где $sum(Z)$ – сумма индексов значимости с термином Z >1 | $Z \notin C$; sum_{razn} – сумма абсолютных значений разностей индексов значимости в общей части; $sum(C)$ – сумма

индексов значимости с термином $C > 1 \mid C \notin Z$; sum_{all} – сумма всех индексов значимости больших 1.

Значения индикатора связанности по Гинзбургу лежат в интервале $0 < ginz(Z, C) < 1$ (0 – слова не связаны, 1 – слова связаны максимально).

Далее, с помощью полученных значений производится отбор сообщений и текстов. Оператору представляется отфильтрованный список по требуемой теме. Результаты тематического поиска могут содержать большое число сообщений и текстов, с целью их сокращения и конкретизации под установленные задачи (например, срочность, неотложность) запускается третий модуль, основанный на оценке эмоциональности сообщений. Данный модуль позволяет выделить эмоционально насыщенные тексты и сообщения, которые могут свидетельствовать о критической ситуации, в которой находится автор. Полученные данные после ранжирования передаются оператору.

Для оценки уровня эмоционального напряжения (степени возбуждения автора в момент написания текста) используется набор маркеров $\{M\} = \{m_1, \dots \dots, m_n\}$, m_i, $i = \overline{1, n}$ – отдельный маркер, n – мощность множества M) [12–14]:

– соотношение количества глаголов к количеству существительных в единице текста M_{ver};

– соотношение количества глаголов к количеству прилагательных в единице текста M_{ve};

– соотношение суммы существительных и глаголов к сумме прилагательных и наречий в единице текста M_{vz};

– соотношение предлогов к общему количеству слов в единице текста M_p;

– соотношение существительных и прилагательных к количеству глаголов и причастий в единице текста M_{pz};

– соотношение предлогов к общему количеству предложений в единице текста M_{pc};

– наличие ненормативных слов M_{an};

– количество слов в тексте M_q;

– средний размер предложений M_s;

– количество знаков восклицания в документе M_{at};

– наличие иконок с эмоциями M_{em}.

Также имеется список коэффициентов, которые наиболее сильно отражают эмоциональную возбужденность [15–17]:

– коэффициент агрессивности вычисляется с помощью следующей формулы (нормальное значение не превышает 0,6):

$$K_a = \frac{V + Vf}{\sum w}, \tag{9}$$

где V – количество глаголов; Vf – количество глагольных форм (причастий и деепричастий); $\sum w$ – общее количество всех слов;

– коэффициент Трейгера вычисляется с помощью следующей формулы (оптимальное значение, близкое к единице):

$$K_t = \frac{V}{A}, \tag{10}$$

где A – количество прилагательных;

– коэффициент определенности действия, вычисляется с помощью следующей формулы (оптимальное значение, близкое к единице):

$$K_{oa} = \frac{V}{N_n}, \tag{11}$$

где N_n– количество существительных.

Данные маркеры и коэффициенты имеют общие характеристики при диагностировании, а именно: если их значения завышены, то у автора имеется эмоциональное беспокойство или волнение.

Результаты исследования и их обсуждение

Перейдём к пошаговому описанию алгоритма:

Шаг 1 – Проведение фильтрации всего массива поступающих данных по времени публикаций, геопозиции, тегам и т.д. с помощью формулы (1).

Шаг 2 – Выборка кандидатов *N* сообщений для анализа.

Шаг 3 – Сегментация текста на абзацы S_a.

Шаг 4 – Сегментация текста на предложения S_s.

Шаг 5 – Сегментация текста на слова S_w.

Шаг 6 – Приведение слов к нижнему регистру W_{lov}.

Шаг 7 – Удаление знаков препинания и стоп-слов W_{st}.

Шаг 8 – Проведение лемматизации W_{lem}.

Шаг 9 – Вычисление индикаторов.

Шаг 9.1 – Проведение сравнения распределения терминов в тексте с помощью дивергенции Кульбака–Лейблера $Kl(z)$ по формуле (2).

Шаг 9.2 – Оценка равномерности распределения термина по сообщению с помощью информационной энтропии $Ie(z)$ по формуле (3).

Шаг 9.3 – Выделение общеупотребительных слов $R(z)$ с помощью формулы (4).

Шаг 9.4 – Вычисление ключевых терминов с помощью распределения Бернулли $Z_n(z)$ по формуле (5).

Шаг 9.5 – Оценка читабельности текста с помощью индикатор Флеша-Кинсайда FRE по формуле (6).

Шаг 10 – Если проверены все сообщения переход на шаг 11, если нет, возврат к шагу 2.

Шаг 11 – Нормализация полученных значений индикаторов для вычисления единого значения по каждому кандидату.

Шаг 12 – Ранжирование полученных значений по каждому кандидату.

Шаг 13 – Инициализация терминов наивысшего ранга, как ключевые термины.

Шаг 14 – Формирование на основе ключевых терминов биграмм и триграмм.

Шаг 15 – Вычисление биграмм и триграмм.

Шаг 15.1 – Определение силы семантической связи между словами с помощью семантического алгоритма Гинзбурга $ind(z|c)$ по формуле (7, 8).

Шаг 16 – Ранжирование полученных значений по каждому словосочетанию.

Шаг 17 – Выбор биграмм и триграмм с наивысшим значением ранга, присвоение им значения – ключевые.

Шаг 18 – Фильтрация полученных результатов с применением «стоп-слов».

Шаг 19 – Вывод ключевых слов и словосочетаний.

Шаг 20 – Оценка эмоциональной нагрузки сообщения (расчет значений маркеров).

Шаг 20.1 – Расчет соотношения количества глаголов к количеству существительных в единице текста M_{ver}.

Шаг 20.2 – Расчет соотношения количества глаголов к количеству прилагательных в единице текста M_{ve}.

Шаг 20.3 – Расчет соотношения суммы существительных и глаголов к сумме прилагательных и наречий в единице текста M_{vz}.

Шаг 20.4 – Расчет соотношения предлогов к общему количеству слов в единице текста M_p.

Шаг 20.5 – Расчет соотношения существительных и прилагательных к количеству глаголов и причастий в единице текста M_{pz}.

Шаг 20.6 – Расчет соотношения предлогов к общему количеству предложений в единице текста M_{pc}.

Шаг 20.7 – Проверка на наличие ненормативных слов M_{an}.

Шаг 20.8 – Вычисление количества слов в тексте M_q.

Шаг 20.9 – Вычисление среднего размера предложений M_s.

Шаг 20.10 – Вычисление количества знаков восклицания в документе M_{at}.

Шаг 20.11 – Определение наличия иконок с эмоциями M_{em}.

Шаг 20.12 – Определение коэффициента агрессивности K_a по формуле (9).

Шаг 20.13 – Определение коэффициента Трейгера K_t по формуле (10).

Шаг 20.14 – Определение коэффициента определенности действия K_{oa} по формуле (11).

Шаг 21 – Если проверены все тексты переход к шагу 22, если нет – возврат на шаг 19.

Шаг 22 – Нормализация полученных с помощью маркеров значений.

Шаг 23 – Ранжирование текстов и сообщений по эмоциональной нагрузке.

Шаг 24 – Вывод результатов оператору.

Конец алгоритма.

Схема алгоритма семантического, синтаксического и прагматического анализа текстовой информации представлена на рисунке.

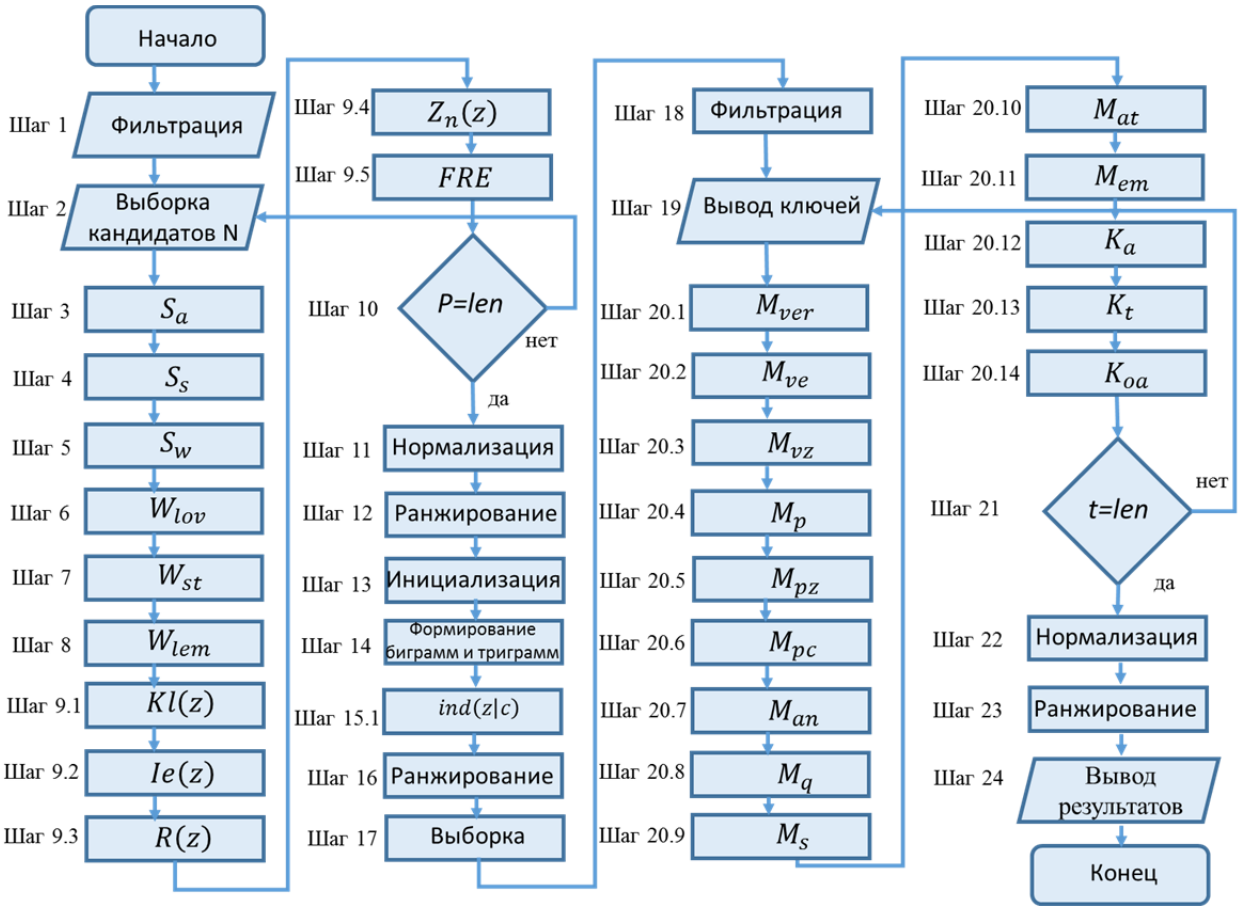

Рис. **Схема алгоритма семантического, синтаксического и прагматического анализа текстовой информации**

Таким образом, представленный в настоящей статье алгоритм многокритериального анализа текстовой информации позволяет провести оценку данных с помощью графематического, морфологического, синтаксического, семантического и прагматического анализа, что дает возможность облегчить сбор большого количества данных, оптимизировать их анализ, разработать сценарии превентивного характера для возможных происшествий природного и техногенного типа, а также сократить расходы ресурсов (временных, финансовых и др.), оптимизировать управленческие предложения, обосновывать принятые решения, основываясь на конъюнктуре большого числа параметров и математического аппарата.

Заключение

Разработанный алгоритм многокритериального анализа текстовой информации позволит повысить эффективность работы экстренных служб, в том числе и МЧС России за счёт всесторонней оценки данных с помощью спектра механизмов, в которые входят графематический, морфологический, синтаксический, семантический и прагматический анализ.

В дальнейшем планируется разработка программного продукта на основе представленного в настоящей статье алгоритма, что позволит автоматизировать процесс обработки информации.

Список источников

1. Вострых А.В., Шуракова Д.Г. Компоненты специальной информационной технологии построения оптимальных маршрутов // Актуальные проблемы инфотелекоммуникаций в науке и образовании (АПИНО 2018): сб. науч. статей VII Междунар. науч.-техн. и науч.-метод. конф. 2018. С. 213–218.

2. Воднев С.А., Матвеев А.В. Оценка эффективности реагирования аварийно-спасательных служб на чрезвычайные ситуации на транспорте // Проблемы управления рисками в техносфере. 2019. № 2 (50). С. 110–117. EDN XDDTYZ.

3. Матвеев А.В., Максимов А.В., Попивчак И.И. Перспективные направления информацинно-аналитической деятельности в области обеспечения пожарной безопасности // Геополитика и безопасность. 2015. № 2 (30). С. 113–117. EDN VMLYLY.

4. Matveev A., Maksimov A., Vodnev S. Methods improving the availability of emergency-rescue services for emergency response to transport accidents // Transportation Research Procedia. SPb.: Elsevier, 2018. P. 507–513. DOI: 10.1016/j.trpro.2018.12.137. EDN: AWTRJK.

5. Вострых А.В. Модель описания пользователей социальных сетей // Актуальные проблемы математики и информационных технологий: сб. материалов IV Всерос. конф. с междунар. участием. 2023. С. 48–51.

6. Вострых А.В. Анализ эффективности информационных систем, используемых сотрудниками МЧС России // Актуальные проблемы обеспечения безопасности в Российской Федерации: сб. материалов Дней науки с междунар. участием, посвящ. 90-летию Гражданской обороны России. Екатеринбург, 2022. С. 62–66.

7. Rabinovich A.E., August A.V. Application of Big Data technology in the field of railway communication // Original research. 2021. Vol. 11. P. 155–161.

8. Taduri A. Railway assets: a potential area for big data analysis // Procedia Computer Science. 2015. Vol. 53. P. 457–467.

9. Eremenko K. Working with data in any field. How to reach a new level using analytics. M.: Alpina Publisher, 2019. 304 p.

10. Stevens-Davidovits,S. Everyone lies. Search engines, big data and the Internet know everything about you. M.: Eksmo, 2018. 384 p.

11. Gudovskikh D.V., Moloshnikov I.A., Rybka R.B. Analysis of emotivity of texts based on psycholinguistic markers with determination of morphological properties // Bulletin of the Vsu. series: linguistics and intercultural communication. 2015. № 3. P. 92–97.

12. Pang B., Lee L. Collecting opinions and analyzing moods // Fundamentals and trends in the search for information. 2008. Vol. 2. № 1/2. P. 543–561.

13. Leontiev A.A. Fundamentals of psycholinguistics. SENSE. 1997. 287 p.

14. Sadegh M. Collecting Opinions and sentiment Analysis: A Survey // In: International Journal of Computers and Technologies. 2012. P. 171–178.

15. Beigi G. Overview of sentiment analysis in social networks and its application in disaster relief // In: Sentiment analysis and ontology development. Springer. 2016. P. 313–340.

16. Mika V.M., Graziotin D., Kuutila M. The evolution of sentiment analysis – an overview of research topics, venues and the most cited articles // In: Computer Science Review 27. 2018. P. 16–32.

17. Mikolov T. Distributed representations of words and phrases and their compositionality // In: Achievements in the field of neural information processing systems. 2013. P. 3111–3119.

References

1. Vostryh A.V., Shurakova D.G. Komponenty special'noj informacionnoj tekhnologii postroeniya optimal'nyh marshrutov // Aktual'nye problemy infotelekommunikacij v nauke i obrazovanii (APINO 2018): sb. nauch. statej VII Mezhdunar. nauch.-tekhn. i nauch.-metod. konf. 2018. S. 213–218.

2. Vodnev S.A., Matveev A.V. Ocenka effektivnosti reagirovaniya avarijno-spasatel'nyh sluzhb na chrezvychajnye situacii na transporte // Problemy upravleniya riskami v tekhnosfere. 2019. № 2 (50). S. 110–117. EDN XDDTYZ.

3. Matveev A.V., Maksimov A.V., Popivchak I.I. Perspektivnye napravleniya informacinno-analiticheskoj deyatel'nosti v oblasti obespecheniya pozharnoj bezopasnosti // Geopolitika i bezopasnost'. 2015. № 2 (30). S. 113–117. EDN VMLYLY.

4. Matveev A., Maksimov A., Vodnev S. Methods improving the availability of emergency-rescue services for emergency response to transport accidents // Transportation Research Procedia. SPb.: Elsevier, 2018. P. 507–513. DOI: 10.1016/j.trpro.2018.12.137 EDN: AWTRJK.

5. Vostryh A.V. Model' opisaniya pol'zovatelej social'nyh setej // Aktual'nye problemy matematiki i informacionnyh tekhnologij: sb. materialov IV Vseros. konf. s mezhdunar. uchastiem. 2023. S. 48–51.

6. Vostryh A.V. Analiz effektivnosti informacionnyh sistem, ispol'zuemyh sotrudnikami MCHS Rossii // Aktual'nye problemy obespecheniya bezopasnosti v Rossijskoj Federacii: sb. materialov Dnej nauki s mezhdunar. uchastiem, posvyashch. 90-letiyu Grazhdanskoj oborony Rossii. Ekaterinburg, 2022. S. 62–66.

7. Rabinovich A.E., August A.V. Application of Big Data technology in the field of railway communication // Original research. 2021. Vol. 11. P. 155–161.

8. Taduri A. Railway assets: a potential area for big data analysis // Procedia Computer Science. 2015. Vol. 53. P. 457–467.

9. Eremenko K. Working with data in any field. How to reach a new level using analytics. M.: Alpina Publisher, 2019. 304 p.

10. Stevens-Davidovits,S. Everyone lies. Search engines, big data and the Internet know everything about you. M.: Eksmo, 2018. 384 p.

11. Gudovskikh D.V., Moloshnikov I.A., Rybka R.B. Analysis of emotivity of texts based on psycholinguistic markers with determination of morphological properties // Bulletin of the Vsu. series: linguistics and intercultural communication. 2015. № 3. P. 92–97.

12. Pang B., Lee L. Collecting opinions and analyzing moods // Fundamentals and trends in the search for information. 2008. Vol. 2. № 1/2. P. 543–561.

13. Leontiev A.A. Fundamentals of psycholinguistics. SENSE. 1997. 287 p.

14. Sadegh M. Collecting Opinions and sentiment Analysis: A Survey // In: International Journal of Computers and Technologies. 2012. P. 171–178.

15. Beigi G. Overview of sentiment analysis in social networks and its application in disaster relief // In: Sentiment analysis and ontology development. Springer. 2016. P. 313–340.

16. Mika V.M., Graziotin D., Kuutila M. The evolution of sentiment analysis – an overview of research topics, venues and the most cited articles // In: Computer Science Review 27. 2018. P. 16–32.

17. Mikolov T. Distributed representations of words and phrases and their compositionality // In: Achievements in the field of neural information processing systems. 2013. P. 3111–3119.

Информация о статье:
Статья поступила в редакцию: 29.06.2023; одобрена после рецензирования: 22.07.2023;
принята к публикации: 24.07.2023
Information about the article:
The article was submitted to the editorial office: 29.06.2023; approved after review: 22.07.2023;
accepted for publication: 24.07.2023

Сведения об авторах:
Вострых Алексей Владимирович, преподаватель кафедры прикладной математики и информационных технологий Санкт-Петербургского университета ГПС МЧС России (196105, Санкт-Петербург, Московский пр., д. 149), кандидат технических наук, e-mail: a.vostrykh@list.ru, https://orcid.org/0000-0002-8261-0712, SPIN-код: 4788-4683
Медведев Дмитрий Валерьевич, адъюнкт Санкт-Петербургского университета ГПС МЧС России (196105, Санкт-Петербург, Московский пр., д. 149), e-mail: meedvedevdv@mail.ru, https://orcid.org/0009-0002-9436-4376

Information about authors:
Vostrykh Aleksey V., lecturer, department of applied mathematics and information technology of Saint-Petersburg university of State fire service of EMERCOM of Russia (196105, Saint-Petersburg, Moskovsky ave., 149), candidate of technical sciences, e-mail: a.vostrykh@list.ru, https://orcid.org/0000-0002-8261-0712, SPIN: 4788-4683
Medvedev Dmitry V., associate professor of Saint-Petersburg university of State fire service of EMERCOM of Russia (196105, Saint-Petersburg, Moskovsky ave., 149), e-mail: meedvedevdv@mail.ru, https://orcid.org/0009-0002-9436-4376