Алгоритм нечеткого текстового поиска в виртуальных социальных сетях

Ю. В. Давыдова

***Аннотация*—Поиск по ключевым словам в задаче мониторинга виртуальных социальных сетей существенно осложняется наличием ошибок, опечаток, сленга в текстах сообщений пользователей. Для снижения чувствительности поиска к ошибкам и повышения полноты поисковой выдачи предлагается использовать нечеткий поиск с фильтрацией. В данной статье представлен алгоритм, состоящий из двух этапов – сканирования и верификации. На этапе сканирования осуществляется фильтрация текста с целью исключения из рассмотрения сообщений, точно не содержащих искомые слова. Оставшиеся сообщения проверяются на этапе верификации. Интеграция в поиск лингвистических правил и статистики ошибок позволяет при этом сохранить достаточную точность. В статье приведены оценки эффективности нечеткого поиска в целом, а также использующегося в алгоритме классификатора. Тестирование проводилось на выборке текстов сообщений Генерального интернет-корпуса русского языка.**

***Ключевые слова*—виртуальные социальные сети, классификация словоформ, модель ошибок, нечеткий текстовый поиск.**

I. Введение

В связи с высокой степенью вовлеченности населения в виртуальные социальные сети (ВСС), мониторинг ВСС становится все более популярным. Существуют различные системы поиска упоминаний и анализа контента [1], [2], выявляющие отношение пользователей к объекту мониторинга и частоту его упоминаний. При этом одним из этапов мониторинга является поиск по набору ключевых слов, характеризующих тот или иной объект мониторинга.

Особенности языка коммуникаций в сети Интернет, такие как прерывистость синтаксических конструкций, использование неофициального стиля, языковые игры, употребление сленга, позволяют исследователям рассматривать его как отдельное явление [3]. Таким образом, обработка языка в ВСС требует новых подходов.

Важной проблемой при организации поиска по ключевым словам является большое количество грамматических ошибок, опечаток, сокращений и сленга в сообщениях пользователей [4], поэтому необходим такой алгоритм поиска, который обеспечивал бы низкую чувствительность к различного рода ошибкам.

Статья получена 23 марта 2018.
Юлия Витальевна Давыдова, Орловский государственный университет имени И.С. Тургенева, (e-mail: j.davydova@ ostu.ru).

Большинство исследований посвящено этапам анализа ВСС более высокого уровня, как правило, рассматриваются алгоритмы классификации и кластеризации сообщений и пользователей, методы сетевого анализа [5], [6]. Проблеме эффективного поиска по ключевым словам не уделяется достаточного внимания, однако, ввиду вышеперечисленных особенностей, поиск является нетривиальной задачей. Для ее решения предлагается использовать алгоритм нечеткого поиска без предварительной индексации, осуществляющий поиск непосредственно в текстовых массивах данных. Алгоритм состоит из двух этапов – сканирования текста и последующей верификации предварительно найденных результатов. Данная статья является продолжением работы [7], где была описана разработанная автором модель ошибок, интегрированная в фазу верификации, для повышения точности поиска. В настоящей статье рассматривается весь алгоритм нечеткого текстового поиска, в частности разработанный этап сканирования и классификатор для обработки результатов поиска. Эффективность алгоритма поиска была оценена на подкорпусе текстов Генерального интернет-корпуса русского языка [8], [9].

II. Нечеткий онлайн поиск

Традиционно информационно-поисковые системы осуществляют поиск в индексе – структуре данных, содержащей информацию о документах [10]. Скачанные краулером страницы проходят этап токенизации, полученные словоформы подвергаются нормализации путем стемминга или лемматизации. После этого сведения о документе добавляются в индекс, который обновляется с заданной периодичностью. Как правило, обработка ошибок в поисковых системах проводится на этапе первичного анализа запроса пользователя – осуществляется коррекция ошибок для слов из запроса с учетом или без учета контекста [11]. Далее организуется поиск в соответствии с исправленным запросом.

Одна из ключевых особенностей ВСС – наличие большого количества текстовых сообщений. Согласно исследованию Brand Analytics только в ВСС Вконтакте за май 2017 года было опубликовано более 300 миллионов сообщений [12]. Среди них содержится большое количество дубликатов в виде репостов, сообщения в среднем имеют небольшую длину. В связи с этим формирование и обновление индекса потребовало бы очень больших затрат. Кроме того, необходимо увеличить полноту поиска, предоставляя в

поисковой выдаче варианты упоминаний об объекте мониторинга с грамматическими ошибками и опечатками. С учетом данных особенностей предлагается рассматривать алгоритмы нечеткого поиска без предварительной индексации, так называемый нечеткий онлайн поиск [13].

Задачей нечеткого поиска является нахождение всех возможных подстрок текста, отличающихся от искомого паттерна не более чем на заданное число ошибок определенного типа. В контексте мониторинга паттерн – искомое слово, характеризующее объект мониторинга. Тип возможных ошибок определяется расстоянием редактирования – минимальным количеством операций редактирования, необходимых для преобразования одной строки в другую. Для задачи нечеткого поиска по ключевым словам было выбрано взвешенное расстояние Дамерау-Левенштейна, поскольку оно является наиболее подходящим для обработки естественного языка. Данное расстояние учитывает следующие операции:

– вставка символа в строку;
– удаление символа из строки;
– замена символа строки на другой символ;
– транспозиция смежных символов.

Для организации поиска в процессе мониторинга ВСС из нескольких категорий алгоритмов нечеткого поиска был выбран подход с фильтрацией как удовлетворяющий особенностям задачи наилучшим образом. Обзор подходов в нечетком онлайн поиске, обоснование выбора расстояния редактирования и категории алгоритмов с фильтрацией было приведено автором в [7].

Нечеткий текстовый поиск является ресурсоемкой задачей. Для повышения скорости работы в алгоритмах с фильтрацией вводится этап сканирования, который позволяет отфильтровать фрагменты текста, не содержащие искомые слова. Это является особенно актуальным при обработке больших объемов текста. Традиционно искомое слово разбивается на $k+1$ подстрок, где k – количество допустимых ошибок. Затем организуется поиск данных подстрок в тексте. Идея фазы сканирования заключается в том, что хотя бы один из $k+1$ отрезков без изменений встретится в подстроке, отвечающей условиям нечеткого поиска. Существуют различные методы организации поиска подстрок паттерна в тексте [14]. Однако такой подход к фильтрации имеют высокую чувствительность к количеству ошибок в словах, показывая плохие результаты при множественных ошибках. Особенно заметно это проявляется при обработке сленга и языковых игр в виде намеренного искажения написания слов, например, «остается» - «астаецца».

Подстроки, отобранные на этапе сканирования, проходят этап верификации, при этом часто применяется динамическое программирование. Однако классическое решение задачи нечеткого поиска методом динамического программирования представляется довольно грубым для задачи обработки естественного языка, поскольку не учитывает частотность различных ошибок. Например, замена буквы «о» на «а» является распространенной ошибкой, а замена «я» на «л» - нет. В работе [7] автором была приведена модель ошибок, интегрируемая в фазу верификации для сохранения точности поиска. Используемый алгоритм динамического программирования и модель ошибок остаются без изменений, но общий подход к верификации был модифицирован – добавлен морфологический анализатор и классификация результатов нечеткого поиска.

С учетом вышеизложенных особенностей обработки текстов в ВСС автором предлагается алгоритм нечеткого текстового онлайн поиска на основе фильтрации.

III. Предлагаемый алгоритм нечеткого текстового поиска в ВСС

A. Фаза сканирования

Текстовые сообщения пользователей, прошедшие первичную обработку (в том числе проверку сегментации слов), хранятся в базе данных сообщений системы мониторинга. Слова для поиска, характеризующие объект мониторинга, и соответствующая лингвистическая информация содержатся в отдельной базе данных. Под лингвистической информацией понимаются парадигмы словоизменения и семантика сленговых слов, если таковые имеются [15].

Фаза сканирования реализована посредством трех фильтров (рис. 1). Тексты сообщений из базы данных сообщений поступают на обработку первому фильтру стоп-слов, при этом единицей анализа является словоформа. Словоформа – это слово в любой его форме согласно словоизменительной парадигме. Стоп-слова были сформированы на основе частотного распределения словоформ по материалам Национального корпуса русского языка [16]. Сюда, как правило, относятся предлоги, местоимения. Также в список стоп-слов был добавлен частотный сленг, например «ваще», «собсна», сформированный на основе анализа подкорпуса ГИКРЯ [9]. Проверка на вхождение в сформированное множество стоп-слов позволяет исключить из рассмотрения слова, несущие незначительную информацию.

Далее осуществляется фильтрация словоформ по длине в соответствии с набором искомых слов, загружаемых из базы данных. Данный фильтр удаляет из рассмотрения словоформы, длины которых более чем в два раза превышают длины искомых слов.

Третий фильтр реализует фильтрацию словоформ на основе меры сходства с искомыми словами. Существуют различные подходы к измерению степени близости между строками [17], [18]:

– меры сходства на основе множества общих символов;
– фонетические меры сходства;
– меры сходства на основе редакционного расстояния.

Фаза сканирования должна быстро отфильтровывать тексты сообщений, точно не содержащие ключевые

слова, поэтому третий фильтр должен вычислять минимально необходимую степень сходства между словами и не быть ресурсоемким. В связи с этим предлагается использовать меры на основе множества общих символов как наиболее простые в реализации.

Коэффициент Жаккара (1) не учитывает порядок букв в словах, поэтому количество ложноположительных результатов будет велико.

$$J = \frac{A \cap B}{A \cup B}, \quad (1)$$

где A, B – множество символов первой и второй строк, участвующих в сравнении, соответственно.

Мера сходства на основе n-грамм фактически является коэффициентом Жаккара на множестве n-грамм. N-грамма – последовательность символов длиной n. При этом повышается точность сравнения слов, однако n-граммы плохо справляются с ошибкой транспозиции и сленгом (например, «астаецца»). Для того чтобы увеличить полноту поиска, но минимизировать количество ложноположительных срабатываний, предлагается использовать совокупность коэффициента Жаккара на основе букв и n-грамм в виде усредненного показателя меры сходства.

Пусть даны две строки: S_1 и S_2 длиной m и l соответственно. $X = \{x_1x_2, x_2x_3, ..., x_{m-1}x_m\}$ – множество биграмм строки S_1. $Y = \{y_1y_2, y_2y_3, ..., y_{l-1}y_l\}$ - множество биграмм строки S_2. Тогда с учетом (1) усредненный коэффициент сходства примет вид:

$$J_{avg} = \frac{1}{2}\left(\frac{A \cap B}{A \cup B} + \frac{X \cap Y}{X \cup Y}\right). \quad (2)$$

Словоформы из сообщений пользователей, удовлетворяющие минимальному порогу сходства с искомыми словами, формируют список предварительно найденных результатов и подвергаются дальнейшей верификации. Следует отметить, что в процессе сравнения используются начальные формы искомых слов. При этом данное решение не сказывается на результатах сканирования – словоформы в рамках парадигмы словоизменения всегда достигают порога минимально необходимого сходства.

B. Фаза верификации

На этапе верификации (рис. 1) рассчитывается взвешенное расстояние Дамерау-Левенштейна между каждой словоформой из списка предварительно найденных результатов и соответствующим искомым словом. Расстояние вычисляется с использованием весов из статистической компоненты модели ошибок. Данная компонента хранит информацию о статистике ошибок по типам операций редактирования, например статистику ошибочного удаления буквы «ь» после «ш». Более подробно статистическая компонента модели ошибок и ее интеграция в вычисление взвешенного расстояния Дамерау-Левенштейна методом динамического программирования описана автором в работе [7]. В случае значения расстояния, превышающего единицу, рассматриваемая словоформа удаляется из списка потенциально найденных лексем. При нулевом значении расстояния словоформа добавляется в поисковую выдачу первой очередности. Первая очередность предполагает точное совпадение с искомым словом.

Если значение расстояния находится в диапазоне от нуля до единицы, то рассматриваемая словоформа подается на вход морфологическому анализатору [19]. Если словоформа распознается, то слово считается найденным и добавляется в поисковую выдачу, в противном случае для обработки словоформы используется алгоритм машинного обучения. Анализатор работает на основе словарей, таким образом, если словоформа является какой-либо формой искомого слова, но отсутствует в словаре (особенно это касается неологизмов и сленга) или написана с ошибками, анализатор не сможет ее корректно распознать. Для того, чтобы классифицировать словоформу как являющуюся или не являющуюся искомым словом используется шесть признаков.

В процессе расчета расстояния Дамерау-Левенштейна формируется выравнивание между словами – каждый символ одного слова ставится в соответствие символу другого слова. Таким образом, легко фиксируются места ошибок в слове. Данные ошибки проверятся на соответствие фонетическим правилам русского языка, при этом рассматриваются ошибки, состоящие как из одной буквы, так и из нескольких букв. Такой подход обусловлен необходимостью обработки ошибок по принципу «как слышится, так и пишется». Например, «сонце» – ошибка непроизносимого согласного, «жолтый» – гласная после шипящего звука. Кроме того, в модель добавлены типовые правила для проверки сленга, например, «ца», «цца», «цо», «ццо» для «тся» и «ться». Принадлежность ошибок множеству фонетических правил является одним из признаков для классификации.

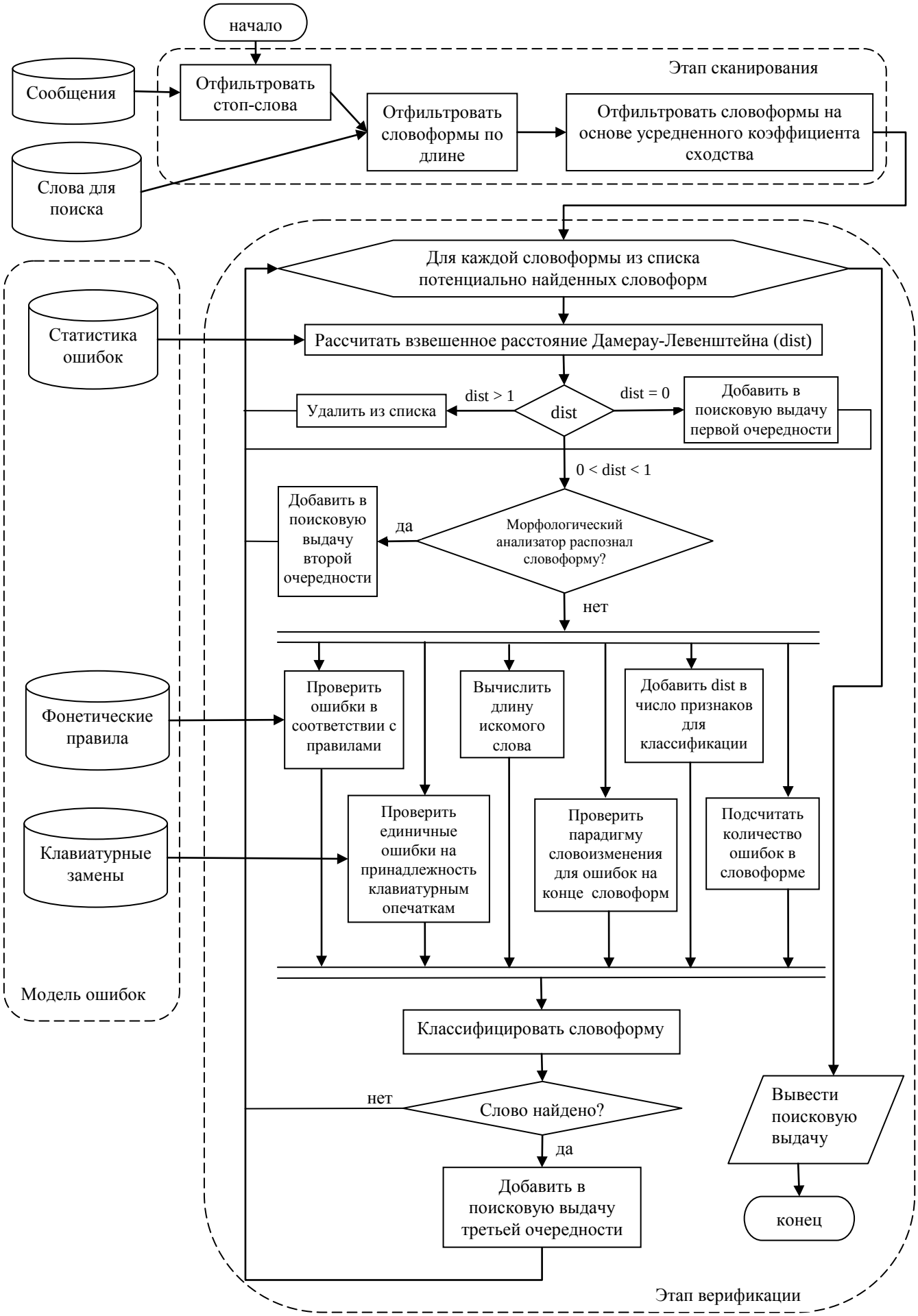

Рис. 1. Алгоритм нечеткого текстового онлайн поиска

Другой признак для классификации - принадлежность множеству клавиатурных замен. Данная проверка необходима для выявления возможных клавиатурных опечаток. При этом для текущей клавиши множество замен формируется только из двух смежных с ней клавиш, находящихся в том же клавиатурном ряду [7].

В случае наличия ошибки на конце словоформы необходимо проверить ошибочные буквы на принадлежность словоизменительной парадигме искомого слова, поскольку при расчете расстояния Дамерау-Левенштейна также используется начальная форма искомого слова.

Количество возможных ошибок имеет прямую зависимость от длины слова, поэтому вычисляется длина искомого слова и количество ошибок в словоформе согласно выравниванию. Также во множество признаков добавляется значение взвешенного расстояния Дамерау-Левенштейна.

Вышеописанные признаки подаются на вход классификатору, который определяет, является ли рассматриваемая словоформа искомым словом, написанным с ошибкой (а также сленгом, неологизмом, не распознанным анализатором), или не является. Классификация осуществляется на основе дерева решений, обучение классификатора проводилось на подвыборке корпуса ГИКРЯ. При выводе результатов поиска ранжирование осуществляется в соответствии со степенью точности совпадения с искомыми словами, поэтому в случае положительной классификации словоформа добавляется в поисковую выдачу третьей очередности.

IV. Оценка качества поиска

Эффективность разработанного алгоритма нечеткого поиска оценивалась на основе показателей полноты и точности [10]. Единицей анализа в данном случае является не документ, а слово, поскольку оценивается способность рассматриваемого алгоритма находить слова с ошибками. При этом информация о принадлежности словоформы тому или иному сообщению сохраняется. Тематическая релевантность выходит за рамки данной статьи. Полнота (3) показывает долю правильно найденных слов среди всех релевантных (т.е. искомых). Точность (4) означает долю правильно найденных слов среди всех найденных.

$$R = \frac{W_f \cap W_{rel}}{W_{rel}}, \quad (3)$$

где W_f – количество найденных слов, W_{rel} – количество искомых слов в тестовой выборке.

$$P = \frac{W_{rel} \cap W_f}{W_f}. \quad (4)$$

Оценка эффективности поиска проводилась на подвыборке корпуса ГИКРЯ, результаты представлены в таблице 1.

Таблица 1. Оценка эффективности поиска по показателям точности и полноты

Объем выборки, слов	Полнота, %	Точность, %
300 000	96.62	96.88

В процессе тестирования алгоритма поиска были проанализированы результаты работы классификатора. Оценка качества классификации также проводилась на основе показателей точности и полноты. Для задачи классификации показатели будут выглядеть следующим образом (5), (6):

$$P_{cl} = \frac{TP}{TP + FP}, \quad (5)$$

где TP – количество истинно положительных результатов классификации. FP – количество ложно положительных результатов классификации.

$$R_{cl} = \frac{TP}{TP + FN}, \quad (6)$$

где FN – количество ложно отрицательных результатов классификации.

Значения показателей точности и полноты классификатора представлены в таблице 2.

Таблица 2. Оценка качества классификатора

Полнота, %	Точность, %
98.05	81.55

Следует отметить, что выборки для обучения классификатора и для тестирования алгоритма поиска не пересекались, так же как и поисковые запросы.

V. Заключение

Поиск по ключевым словам является неотъемлемой частью процесса мониторинга виртуальных социальных сетей. В данной статье предложен алгоритм двухэтапного нечеткого текстового поиска, который существенно снижает чувствительность поиска к ошибкам и опечаткам, тем самым увеличивая полноту поисковой выдачи, при этом используемая модель ошибок позволяет сохранять достаточную точность. На этапе сканирования осуществляется поиск словоформ, удовлетворяющих минимально необходимому порогу сходства. Этап верификации позволяется подтвердить или опровергнуть предварительно найденные результаты. Для этого используется взвешенное расстояние Дамерау-Левенштейна, рассчитываемое на основе динамического программирования, модель ошибок [7], морфологический анализатор и классификатор.

Проведенное тестирование на подвыборке ГИКРЯ позволило оценить эффективность поиска по показателям полноты и точности, значения которых оказались достаточно высокими. В дальнейшем планируется улучшать точность классификатора в процессе поиска за счет обратной связи, предоставляемой пользователем при просмотре результатов поисковой выдачи. Повышение точности классификации позволит сделать алгоритм нечеткого поиска более эффективным.

Библиография

[1] Brand Analytics – система мониторинга социальных медиа и СМИ [электронный ресурс] // URL: https://br-analytics.ru/ (дата обращения 20.02.2018).

[2] Youscan – система мониторинга социальных медиа и социальных сетей [электронный ресурс] // URL: https://youscan.io/ (дата обращения 20.02.2018).

[3] Современный русский язык в Интернете / под ред. Я. Э. Ахапкина, Е.В. Рахилина. – М.: Языки славянской культуры, 2014. – 328 с.

[4] Давыдова Ю. В. Проблема обработки ошибок в текстах сообщений пользователей в задаче мониторинга виртуальных социальных сетей // Новые информационные технологии и системы: материалы XIV Международной научно-технической конференции. – Пенза, 2017. – С. 342-345.

[5] Губанов Д. А., Чхартишвили А. Г. Концептуальный подход к анализу социальных сетей // Управление большими системами: сборник трудов. – 2013. – № 45. – С. 226-236

[6] Batrinca B., Treleaven P. C. Social media analytics: a survey of techniques, tools and platforms // AI & Society. – 2015. – Vol. 30, No. 1. – pp. 89-116.

[7] Давыдова Ю. В. Модель ошибок для нечеткого текстового поиска в задаче мониторинга виртуальных социальных сетей для обеспечения информационно-психологической безопасности личности // Современные информационные технологии и ИТ-образование. – 2017. – Т. 13, № 3. – С. 72-82.

[8] Belikov V., Kopylov N., Piperski A., Selegey V., Sharoff S. Big and diverse is beautiful: A large corpus of Russian to study linguistic variation // Proceedings of the 8th Web as Corpus Workshop (WaC-8), 2013. – pp. 24-28.

[9] Генеральный Интернет-корпус русского языка [электронный ресурс] // URL: http://www.webcorpora.ru/ (дата обращения 25.02.2018).

[10] Manning C. D., Raghavan P., Schutze H. Introduction to information retrieval. – Cambridge: Cambridge University Press, 2008. – 496 p.

[11] Панина М. Ф., Байтин А. В., Галинская И. Е. Автоматическое исправление опечаток в поисковых запросах без учета контекста // Компьютерная лингвистика и интеллектуальные технологии: материалы ежегодной Международной конференции «Диалог», 2013. – С. 568-579.

[12] Социальные сети в России, лето 2017: цифры и тренды [электронный ресурс] // URL: http://blog.br-analytics.ru/sotsialnye-seti-v-rossii-leto-2017-tsifry-i-trendy/ (дата обращения 22.02.2018).

[13] Navarro G., Raffinot M. Flexible pattern matching in strings: practical on-line search algorithms for texts and biological sequences. – Cambridge: Cambridge University Press, 2007. – p. 232.

[14] Navarro G. A guided tour to approximate string matching // ACM Surveys. – 2001. – Vol. 33, No. 1. – pp. 33-88.

[15] Savva Yu. B., Davydova Yu. V. Linguistic database for monitoring system of online social networks in providing information and psychological security // European integration: justice, freedom and security: proceedings of VII scientific and professional conference with international participation: in 3 volumes. – Belgrade: “Criminalistic-Police Academy” Publisher, 2016. – Vol. 1. – P. 145-154.

[16] Частоты словоформ и словосочетаний [электронный ресурс] // URL: http://www.ruscorpora.ru/corpora-freq.html (дата обращения 2.03.2018).

[17] Ingersoll G.S., Morton T.S., Farris L.A. Taming text. How to find, organize and manipulate it. – NY: Manning Publications Co., 2013. – 320 p.

[18] Ульман Дж., Раджараман А, Лесковец Ю. Анализ больших наборов данных. – М.: ДМК Пресс, 2016. – 498 с.

[19] Korobov M. Morphological analyzer and generator for Russian and Ukrainian languages // Analysis of Images, Social Networks and Texts: proceedings of International conference, 2015. – pp. 320-332

Algorithm of fuzzy text search in online social networks

Yulia Davydova

***Abstract*—In the task of online social networks monitoring search with keywords is complicated by misspellings, typos, slang in users' posts. To reduce search sensitivity to misspellings and improve the completeness of search results it is proposed to use fuzzy search with filtration. This article presents the algorithm consisting of two stages – scanning and verification. On the scanning stage, text is being filtered with the aim to exclude posts, which definitely do not contain keywords from consideration. Remaining post are checked on the verification stage. Integration of linguistic rules and misspellings statistics in text search allows to preserve its accuracy. The article presents estimation of effectiveness of the whole algorithm of fuzzy search and of the classifier used in it in particularly. Testing was done on the sample of posts from The General Internet-Corpus of Russian.**

***Keywords*—online social networks, word forms classification, misspellings model, fuzzy text search.**