УДК 621.391 : 004.052.3 : 004.052.4

Северин Н.В.
Северин М.В.
Severin N.V.

ПРИМЕНЕНИЕ МЕТОДОВ ТЕОРИИ ПОДОБИЯ КОНЕЧНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ ДЛЯ ИДЕНТИФИКАЦИИ СООБЩЕНИЙ В СИСТЕМАХ ТЕЛЕКОММУНИКАЦИЙ

ЗАСТОСУВАННЯ МЕТОДІВ ТЕОРІЇ ПОДІБНОСТІ СКІНЧЕННИХ ПОСЛІДОВНОСТЕЙ ДЛЯ ІДЕНТИФІКАЦІЇ ПОВІДОМЛЕНЬ У СИСТЕМАХ ТЕЛЕКОМУНІКАЦІЙ

APPLYING THE METHODS OF FINITE SEQUENCES' SIMILARITY THEORY TO MESSAGES' IDENTIFICATION IN TELECOMMUNICATION SYSTEMS

Аннотация. Исследуются проблемы внедрения в сети телекоммуникаций средств распознавания «искаженных» сообщений, в том числе с неустранимыми искажениями, обусловленными представлением одной и той же информации различными кодовыми последовательностями. Показано, что распознать некоторые типы таких сообщений возможно с применением методов теории подобия конечных последовательностей.

Анотація. Досліджуються проблеми впровадження у мережі телекомунікацій засобів розпізнавання "спотворених" повідомлень, зокрема з неусувними спотвореннями, які обумовлені зображенням однієї і тієї ж інформації різними кодовими послідовностями. Показано, що розпізнати деякі типи таких повідомлень можливо шляхом застосування методів теорії подібності скінченних послідовностей.

Summary. I investigate some problems of implementing in telecommunication networks facilities able to recognize "corrupted" messages, specifically with unavoidable deformations that appear since one and the same information can be presented in different sequences of codes. It is shown that some types of such messaged can be recognized using the methods of finite sequences' similarity theory.

В сетях телекоммуникаций при передаче сообщений к системе выдвигается ряд требований, одним из которых является обеспечение заданного порога вероятности ошибки [1]. Для этого обычно используются различные методы кодирования информации. Они позволяют определить, претерпели ли данные в процессе передачи изменения, и – при наличии таковых – принять решение о дальнейшем использовании данных.

Однако методы кодирования способны обеспечить достоверность только в плане идентичности полученного сообщения относительно переданного. Они не подходят для задач, когда ошибка в сообщении может быть допущена еще на входе в систему.

Современные сети телекоммуникаций поддерживают работу многочисленных информационных систем (ИС), где нередки ситуации, когда на вход системы подается сообщение, уже содержащее «искажения» [2; 3]. Эти искажения могут быть ошибками, но могут также быть связанными с представлением одной и той же информации различными синтаксическими конструкциями. Такие ситуации распространены в системах, где источником формирования сообщений является человек. Актуальной является проблема идентификации сообщений с искажениями указанного типа. Предлагаются различные подходы и методы, способствующие ее решению [4; 5; 6]. Их, однако, нельзя признать «универсально» применимыми. Указанная область исследования остается открытой.

Целью статьи является исследование применимости теории подобия конечных последовательностей [7; 8] для распознавания «искаженных» сообщений в сетях телекоммуникаций, поддерживающих функционирование ИС.

Будут рассмотрены следующие примеры «нечеткости» сообщений, вносимой пользователями ИС на входе системы:

– случайные искажения адресов e-mail, нуждающиеся в исправлении;

– намеренные искажения электронных адресов в коммерческих целях;

– искажения в запросах интернет-ресурсов, вносимые с целью обхода систем фильтрации;

– поиск информации в сетях и базах данных по «нечеткому» запросу.

Как увидим далее, общим для всех перечисленных примеров является то, что искажения сообщений, критичные для их идентификации ИС, не являются таковыми для распознавания этих же сообщений человеком.

Мы предлагаем использовать для идентификации сообщений, «нечеткость» которых связана с вносимой пользователями ИС «неустранимой неоднозначностью», методы теории подобия конечных последовательностей (ТПКП).

1. Кратко о теории ТПКП. Не описывая здесь подробно понятия и методы ТПКП, отметим, что в ней предлагаются меры, оценивающие подобие конечных иерархических систем с линейным порядком на каждом уровне иерархии [7; 8]. Таковыми являются, в частности, натуральные или искусственные языки, исходящие из некоторого алфавита M_0. Из элементов M_0 («букв») строятся конечные последовательности, ограниченные некоторыми символами из выделенного подмножества M_0. Их естественно назвать буквосочетаниями или словами. Далее строятся «предложения» как «слова второго уровня», состоящие из «букв-слов»; и «тексты», буквами которых выступают предложения.

ТПКП исходит из допущения о возможности представления одной и той же информации разными последовательностями слов или кодов, и предлагает алгоритмы, позволяющие выявлять подобие данных последовательностей, т. е. давать количественную оценку их «сходства» или «близости».

Принимается, что на M_0 определена двухместная функция $\approx (x, y)$ со значениями в [0,1], которая называется «функцией подобия (сходства)». При этом $\approx (x, x) = 1$, т.е. подобие тождественных символов равно 1. Но допустимо и $\approx (x, y) > 0$ при $x \neq y$, так, для русских букв *е* и *ё* можно задать ненулевой уровень подобия, скажем $\approx$ (е, ё) = 0,8. Аналогично можно положить, скажем $\approx$ (в, ф) = 0,4.

Функция $\approx (x, y)$ служит базисом для введения различных численных мер подобия для объектов следующих уровней иерархии («предложений», «текстов» и т.д.). Некоторые из этих мер используют дополнительное предположение о различной значимости или «весе», букв в слове либо слов в тексте. Такова, в частности, так называемая мера *взвешенного узкого подобия Gs*. Она оценивает сходство слов, отправляясь от их *максимально весомой* (значимой) *подпоследовательности общих вхождений символов* (сравн. [9]).

В дополнение к мерам подобия, базирующимся на функции $\approx (x, y)$, в ТПКП вводятся численные меры сравнения кодов по их *структуре*. Предполагается, что в последовательности слов по некоторым основаниям выделяются особые ***именные группы*** (ИГ). Например, если речь идет о текстах русского языка, это могут быть группа подлежащего и группа сказуемого в предложении и т.п. Принимается следующий *«принцип сплоченности»* для групп: *слова, образующие группу, обычно соседствуют в тексте; и перестановки слов внутри групп «разрушают» структуру текста в меньшей степени, чем чередование слов, принадлежащих разным группам* (сравн. [10]). Предложены меры оценки сходства текстов, комбинирующие меры их лексического «*взвешенного*» подобия с мерами *сплоченности* входящих в текст ИГ. Вводится понятие «средней разобщенности» слов заданной ИГ (упрощенно говоря, сколько «в среднем» чужеродных слов расположено между двумя словами данной ИГ). Средняя разобщенность служит аргументом функции сплоченности φ, оценивающей сходство исходной и «засоренной» ИГ. В качестве φ можно выбирать различные функции – выбор зависит от принимаемой гипотезы, насколько узнаваемость ИГ искажается от ее «засорения». Если, допустим, принимается, что даже одно «разобщающее» слово сильно портит узнаваемость ИГ, рационально выбрать $\varphi = e^{-x}$.

Доказано, что предлагаемые ТПКП алгоритмы вычисляют степени подобия за полиномиальное время. Таким образом, они достаточно быстро оценивают сходство «длинных» текстов. Сравнение этих мер с другими известными мерами оценивания «схожести» текстов выявило ряд их преимуществ [11]. Теоремы и формулы, характеризующие математические свойства упомянутых здесь мер ТПКП см., например в [12].

2. Ошибки адресации в сетях телекоммуникаций и методы их корректировки. В сети Интернет довольно распространены случаи ошибочного набора электронного адреса. Опечатки такого рода при записи URL приводят к потере потенциальных клиентов сайта, а при пересылке писем (e-mail) – к недоставке письма адресату.

Исследование "проблемы опечаток" в сети может представлять коммерческий интерес. Так, в [13] рассматриваются вопросы, связанные с привлечением посетителей на сайт за счет использования опечаток при ручном наборе URL адреса. Для этого регистрируется домен, близкий по названию с популярным доменом (его адрес содержит наиболее вероятные ошибки при наборе). Такой метод получил название "type-in-traffic".

Не менее "популярны" и ошибки запросов в поисковых системах. Они также используются для привлечения клиентов на свой сайт [14]. Статистика поисковых систем говорит о большом количестве орфографических ошибок или опечаток вида: A) искажения символа (нескольких символов); B) пропуска символа(ов);C) вставки "лишнего" символа(ов);D) перестановки нескольких символов. В поисковых системах применяются меры, направленные на устранение таких ошибок. При вводе "искаженного" текста система предлагает пользователю возможные варианты "правильных" слов из словаря. Это снижает нагрузку на поисковый сервер и улучшает качество предоставляемых им услуг.

Ниже предлагаются методы распознавания искаженных сообщений, базирующиеся на мерах подобия "текстов", предлагаемых ТПКП.

3. Коррекция адресов электронной почты. Сейчас здесь применяют следующие приемы обработки "неправильных" адресов. Агент передачи сообщений MTA (Mail Transfer Agent) при отсутствии указанного отправителем адресата формирует возвращаемое письмо (bounce message) с описанием причины, по которой письмо не могло быть доставлено. Bounce message отсылается обратно отправителю.

В традиционной почтовой связи существует требование принятия всех мер по доставке письма адресату, в том числе и по выяснению его нового места пребывания. Качество сервиса электронной почты должно быть, по крайней мере, "не хуже". Ввиду этого можно предложить внедрение в MTA модуля, выполняющего коррекцию ошибочного или же поиск "близкого" адреса. При возможности однозначного определения адресата модуль позволит выполнить корректировку адреса и доставку письма. При невозможности однозначно определить адресата в bounce message может быть включен перечень "близких" адресов.

Рассмотрим пример оценок близости адресов e-mail, полученных при применении упомянутой выше меры подобия $Gs(a, b)$ ТПКП. В качестве a возьмем n_severin@ukr.net. В качестве b берем различные искаженные адреса. Оценку близости выполним только для первой части адреса (до символа @), поскольку здесь подразумевается, что сообщение находится на сервере адресуемого домена, и на почтового агента возложена задача только поиска адресата в своем домене. Веса всех символов в сравниваемых адресах вначале примем равными 1.

Таблица 1 – Оценка близости электронных адресов мерой подобия *Gs*

Искаженный адрес *b*	*Gs*(*a*, *b*)	Искаженный адрес *b*	*Gs*(*a*, *b*)
nseverin	0,889	n_sveerin	0,889
m_severin	0,889	n_siverin	0,889
n-severin	0,889	n_sivirin	0,778
nv_severin	0,900	n_civerin	0,778

Из табл. 1 видно, что единичные ошибки и перестановки символов "мало" искажают электронный адрес. Установив порог меры близости, при котором "близкие" к правильному адресу допустимо считать результатами "опечаток", равным 0,88, первые 6 из приведенных адресов можно автоматически "исправить". Конечно, на практике все адреса в приведенном списке могли бы принадлежать другим пользователям сети. Если такая ситуация считается высоковероятной, первые 6 адресов можно включить в bounce message. Таким образом, bounce message становится более информативным. В итоге применение методов ТПКП позволит как повысить качество работы сервиса, так и снизить "ошибочный" трафик сети.

Как отмечалось выше, мера $Gs(a, b)$ позволяет учесть различную *относительную значимость* различных частей электронного адреса. Например, можно считать, что пропуск/замена символа @ в адресе e-mail является более "грубой" ошибкой, чем пропуск/замена какого-либо иного символа. Если, допустим, приписать @ вес 5, а всем остальным символам адреса – вес 1, мера узкого подобия $Gs(a, b)$ даст следующие значения подобия адресов b = n_severi@ukr.net и c = n_severin_ukr.net относительно правильного адреса a = n_severin@ukr.net: $Gs(a, b) \approx 0{,}952$; $Gs(a, c) \approx 0{,}762$. Итак, адрес з заменой @ знаком подчеркивания, с "точки зрения" меры $Gs(a, b)$, намного менее похож на правильный адрес, чем адрес с пропуском отличной от @ буквы.

4. Распознавание «злонамеренных искажений» в запросах интернет-ресурсов. Усилия разработчиков программного обеспечения и операторов телекоммуникаций направлены на защиту пользователей Интернет от агрессивного контента (порнография, наркотики, экстремизм и т.д.). Но

существующие системы фильтрации «плохо справляются» с идентификацией контента, содержащего запрещенные слова, *умышленно* представленные в искаженном, хотя вполне понятном человеку виде.

Ведущие разработчики поисковых систем предоставляют ряд проектов, направленных на обеспечение безопасного поиска [15]. Но исследование данных систем выявляет их уязвимость и необходимость совершенствования. Так, в системе безопасного поиска от Google, введя запрос *поститутка* (пропуск «р» в запрещенном слове), на момент написания этого текста были получены ссылки на 43200 ресурса, где в первом же десятке – ресурсы откровенного порнографического характера. Детальнее результаты исследований автора, касающихся подобных «обманов» безопасных систем см. в [16].

Системы ограничения доступа используют фильтрацию по адресу, по содержимому ресурса (контенту), либо их комбинации. Широкое распространение получил подход применения «белых» и «черных» списков. Они могут содержать как URL-адреса разрешенных или запрещенных ресурсов, так и слова, отнесенные к запрещенному контенту. В большинстве случаев формирование списков происходит в ручном режиме [17], и имеет очевидный недостаток в скорости реагирования системы на возникновение новых ресурсов. Сейчас каждый день появляется масса новых ресурсов, еще не зарегистрированных в системах безопасного доступа, а URL-адрес ресурса попадает в черный список уже после посещения его пользователем [17], [18].

В отношении запрещенных слов, используемых как эталоны сравнения при динамической фильтрации [18] запрещенного контента, то и здесь имеются проблемы. Кроме запрещенных слов, представленных в словарях либо используемых в обиходе, на практике встречаются их «мутированные» формы: в результате случайной или намеренной ошибки слово изменяется, но остается легко узнаваемым – для человека, но не для системы фильтрации.

Степень популярности использования искаженных слов можно оценить с помощью Интернет-сервисов статистики запросов в поисковых системах. В [16] представлены результаты таких исследований для «слов-мутантов», принадлежащих к запрещенному контенту. Приведенная ниже таблица содержит результаты всего по одному слову *порнография* из обширного множества запросов. Наряду с результатами обработки таких запросов системами безопасного поиска указано (после знака/) и количество ответов на запрос при обычном поиске.

Таблица 2 – Выборочная статистика запросов поисковых систем

Ключевое слово	Количество запросов в месяц http://wordstat.yandex.ru	Количество найденных ответов		
		Семейный фильтр *Yandex* /обычный поиск	Безопасный поиск: система *Цензор*	Безопасный поиск *Google* /обычный поиск
порнография	***320 328***	***0 / 3 000 000***	***0***	***0 / 3 880 000***
парнография	6 981	8 654 / 8 673	0	5 950 / 34 300
понография	8 432	2 038 / 2 040	47	2 110 / 15 600
порногафия	1 079	3 050 / 3 074	0	2 090 / 19 100
пронография	711	2 304 / 2 306	90	2 070 / 12 900
порномафия	27	3 978 / 3 978	0	46 800 000
порнргафия	21	24 /94	0	268 / 5 900

«Мутации» слова могут быть следствиями ошибок человека при работе с клавиатурой. Однако указанные в табл. 2 искажения часто бывают намеренными. Анализ статистики говорит о том, что запросы, содержащие «*очепятки*», довольно широко распространены [16]. Поскольку спрос формирует предложение, то растет и количество нецелевых ресурсов сети, которые, подстраиваясь под спрос, используют подобные слова в своем контенте. При этом поисковые системы выполняют индексацию web-ресурсов, содержащих в своем контенте данные опечатки, и, как видно из результатов поиска, их количество довольно велико.

Например, несмотря на то, что «искаженное» слово *порно**рг**афия* было запрошено всего 21 раз в месяц, в сети по данным статистики от *Google* уже имеется как минимум 5900 источников, содержащих его в своем контенте. Наличие web-ресурсов, содержащих опечатки слов, отнесенных к запрещенному контенту – не случайность. Это результат как умышленного использования этих слов пользователями сети для «маскировки» запрещенного контента, так и SEO-технологий (поисковой оптимизации). В последнем случае при формировании семантического ядра сайта в тексты его

страниц умышленно внедряются определенные «искаженные» слова и словосочетания. Они впоследствии выступают «ключевыми» словами для поисковых систем и предоставляют возможность получить доступ к запрещенному контенту в обход систем безопасного доступа.

Отсюда ясно: предусмотреть все возможные варианты ошибок в записи запрещенных слов и внести их в «черный» список на практике невозможно. Поэтому современные системы фильтрации должны уметь распознавать «подлоги» и выполнять обработку естественного языка с учетом возможных намеренных искажений.

Следующие примеры показывают, что может дать применение меры подобия *Gs(a, b)* для фильтрации текстового контента в системах безопасного доступа. Эти системы работают в основном с естественными языками. Вначале примем, что запрос, «запрещенность» которого следует распознать системе фильтрации, есть отдельное *слово*. Рассмотрим слово *a=порнография*, и его «искаженные» формы – *b=пронография* и *c=n0рн0графия*, а также близкое к *a* по звучанию *d = монография.* Заметим: слово *b* может быть результатом опечатки; тогда как *c* – часто встречающаяся умышленная подмена буквы *о* на цифру 0.

Уместно считать, что для *a=**порно**графия* подчеркнутая часть слова более значима (при оценке его сходства с «искажениями»), чем остальные символы. Если задать следующие веса вхождений букв слова *a* (указаны как верхние индексы): $п^{2}о^{1}р^{2}н^{1}о^{1}г^{0}р^{0}а^{0}ф^{0}и^{0}я^{0}$, мы получим такие значения его подобия «искажениям» *b* и *c*: *Gs(a, b) ≈ 0,857; Gs(a, c) ≈ 0,714; Gs(a, d) ≈ 0,429.* Итак, при использовании меры *Gs* сходство *a* с запрещенными словами (*b* и *c*) оказываются существенно выше, чем сходство *a* со сторонним словом *d*.

В следующей таблице слева приведены некоторые другие искажения слова *a=порнография*, встречающиеся в запросах поисковых систем, а справа – численные значения *Gs*-подобия этих искажений слову *a*.

Таблица 3 – Оценки подобия $Gs(a, b_i)$ слова *a* его искажениям b_i

«Искажение» b_i	$Gs(a, b_i)$	«Искажение» b_i	$Gs(a, b_i)$
п***а***рнография	0,857	п***ро***нография	0,857
понография	0,714	порномафия	0,917
порногафия	1,000	порно***рг***афия	0,917

Мы видим, что все приведенные искажения с высокой степенью подобны слову *a = порнография*. Следовательно, возникает возможность включить *только это слово* в перечень запрещенных слов, если система анализа контента будет способна реагировать не только на слово *a*, но и на высокоподобные ему слова.

Анализ «черных» списков систем безопасного доступа показывает, что там часто присутствуют как запрещенное слово, так и слова-результаты его склонения. Это приводит к неоправданному наращиванию объемов черных списков: невозможно предусмотреть все возможные варианты записи слова, хотя бы потому, что в нем может присутствовать ошибка. Применение методов ТПКП устраняет этот недостаток. Достаточно ввести одно слово, а близкие слова (результаты склонения или ошибок) будут определены системой анализа.

В следующей таблице приведем оценки подобия слова *a* некоторым словам, близким к *a* по буквенному составу и/или по звучанию. Цель – продемонстрировать отделимость посредством меры *Gs* «искажений» *a* от ***не***-запрещенных слов:

Таблица 4 – Искажения слова *a* в сравнении с «разрешенными» словами

«Искажение» b_i	$Gs(a, b_i)$	«Не-запрещенное» b_i	$Gs(a, b_i)$
порно	1,000	***ф***онография	0,429
порно***индустрия***	0,647	***г***о***л***ография	0,429
порн***уха***	0,825	***о***порно-граф***ический***	0,579
п***ро***ногр***фа***ия	0,825	***оп***пор***ту***низ***м***	0,611

Табл. 3 и 4 демонстрируют достоинства предложенного метода оценки слов на «запрещенность». Слова *монография, фонография, голография* (близкие по написанию и звучанию с исходным словом *a*) получили низкую степень подобия. В то же время слово *порнуха* (кажущееся, наоборот, менее схожим по написанию и звучанию с *a*) высоко-подобно *a* и потому может быть отнесено к запрещенным словам. Заметим, что слово *порно**индустрия*** оказалось сравнительно мало-

подобным *a*. Причина – в большом количестве букв (***индуст***), отсутствующих в слове-эталоне *a*. Но *порноиндустрия* – грамматически правильное *длинное составное* слово. Это как раз тот случай, когда его можно рекомендовать включить в «черный» список, наряду с *порнографией*.

Таким образом, можно сделать вывод о перспективности применения *метода взвешенного подобия* (базирующегося на мере *Gs*) в задачах фильтрации нецелевого контента. Этот вывод подкрепляется статистическими исследованиями, выполненными автором с привлечением ряда реальных сайтов, содержащих запрещенный контент, – их результаты представлены в [19].

Дополнительные проблемы возникают, когда запрещенными являются не отдельные слова, а *словосочетания* из *разрешенных вне данного контекста* слов. Пример: ***девочки по вызову***. Доступ к Интернет-ресурсу уместно ограничить, только если данная группа слов входит в его контент «*сплоченно*», *не слишком перемежаясь* другими словами. В подобных случаях можно применить меры ТПКП, оценивающие степень «засоренности» именных групп чужеродными словами (см. выше). В следующей таблице приведены оценки сходства ИГ ***девочки по вызову*** с текстами, в которых между словом ***девочки*** и словами ***по вызову*** размещены какие-нибудь «разобщающие» слова. В качестве функций сплоченности φ выбраны: $\varphi_1 = 1/(1 + x)$; $\varphi_2 = e^{-x}$.

Таблица 5 – Убывание сходства в результате «засорения» ИГ

Количество «разобщающих» слов	Сходство $\varphi_1(x)$	Сходство $\varphi_2(x)$
0	1,000	1,000
1	0,693	0,641
2	0,529	0,411
3	0,429	0,264
10	0,184	0,012

Из табл. 5 видно, что учет сплоченности ИГ позволяет отделять тексты с запрещенными и не запрещенными словосочетаниями.

Заметим, что методы ТПКП позволяют приписать различным словам различные «потенциалы разобщения», и если задать эти потенциалы равными 0, скажем, для слов, подобных словам ***любой*** и ***ваш***, то сходство ИГ из вышеприведенного примера с текстом "***Дэвочки! – лубые по Вашему вызыву!***" будет равно 1.

5. Поиск информации по «нечётким» запросам. Средства нечеткого поиска становится все более востребованными в системах телекоммуникации [20 – 22]. В современных ИС часто возникает задача определения схожести полученного сообщения с некоторым имеющимся набором сообщений. Так, по статистике Яндекса, около 12 % поисковых запросов содержат ошибки и опечатки [23]. "Умная" система могла бы предложить пользователю коррекцию его нечеткого запроса.

Задачи, предполагающие оценку подобия сообщений некоторому "образцу", ставятся и решаются в рамках различных ИС. Пример: браузер Chrome от корпорации Google Inc обрабатывает информацию, вводимую пользователем в адресную строку, и, используя асинхронный теневой режим, делает запрос к серверу, получая в ответ набор предполагаемых адресов, в том числе варианты коррекции исходного адреса [24]. Такой подход уменьшает число ошибочных запросов и тем самым снижает нагрузку как линии связи, так и серверного оборудования.

Для оценки сходства текстовых сообщений применяются различные методы, основанные обычно на тех или иных функциях расстояния (метриках). В зависимости от области использования выбирается тот или иной метод, либо используется комбинация некоторых методов.

Рассмотрим результаты применения ряда методов оценки подобия сообщений в некоторых задачах нечеткого поиска ИС, и сравним их с оценками, полученными посредством использования меры *Gs* ТПКП.

Целью системы нечеткого поиска может являться как предоставление пользователю доступа к ресурсам, содержащим "похожие" на его запрос сообщения, так и *запрет* такого доступа – когда задачей системы контроля является сравнение запроса с запрещенными текстами или адресами. Ниже рассматривается второй случай. Для соблюдения приличий приводимые примеры не относятся к запрещенному контенту, но они достаточно наглядно показывают характер проблем, возникающих в данной области. Для сравнения возьмем три простых сообщения:

a = fastmail, *b = emailing,* *c = fastfood.*

Мы будем считать *mail* "запрещенным" буквосочетанием и, следовательно, *a* – запрещенным сообщением. Предположим, что *b* и *c* – сообщения, поступившие на вход некой системы нечеткого поиска.

Расчет подобия и сравнение полученных результатов выполним для метрик Хемминга [25], Левенштейна [9], функций Soundex и Metaphone (они оценивают *фонетическое* сходство звучания слов – см. [26]), а также для меры *Gs*. Поскольку ТПКП предлагает меры не для расстояния, а для подобия сообщений, ниже на базе перечисленных функций расстояния вводятся соответствующие им функции подобия $H(a, b)$, $L(a, b)$, $S(a, b)$ и $M(a, b)$.

Учитывая, что в сообщении $a = fastmail$ подстрока *mail* считается запрещенной, т.е. более значимой, установим веса (они отображены верхними индексами) следующим образом: $f^0a^0s^0t^0m^1a^1i^1l^1$. Результаты оценок подобия будут следующими:

Таблица 6 – Оценки подобия для сравниваемых методов

подобие	*H*	*L*	*S*	*M*	*Gs*
a и *b*	0,000	0,125	0,250	0,400	0,667
a и *c*	0,500	0,500	0,750	0,600	0,250
b и *c*	0,000	0,000	0,000	0,000	0,222

Таким образом, сравнительный анализ показывает перспективность применения метода взвешенного подобия *Gs*: его оценки оказались наиболее приближены к "человеческим".

В заключение можно сделать следующие выводы. В статье выполнен анализ функционирования ряда информационных систем, для которых характерны искажения входных сообщений, вносимые пользователями ИС. Проведено исследование применимости ТПКП для идентификации таких сообщений. Результаты исследования показывают, что применение методов ТПКП открывает возможности адекватной обработки сообщений, характеризуемых «неустранимой неоднозначностью», заключающейся в случайных или преднамеренных искажениях адресов электронной почты и сетевых ресурсов, и использовании «нечетких» запросов в сетях и базах данных.

Кроме этого, представляется, что методы ТПКП могут оказаться полезными:

– в системах национальной безопасности, где их можно использовать для организации дополнительной автоматизированной проверки и сортировки sms-сообщений в сети мобильных операторов, организации контроля электронной переписки (e-mail), проверки электронных сообщений и электронных ресурсов на наличие запрещенного контента;

– в поисковых системах сети Интернет;

– для автоматизации обработки регистрационных данных в сети Интернет, так как, начиная от сайтов интернет магазинов и заканчивая сайтами государственного значения, требуется заполнение анкет регистрации.

Представляется также, что методы ТПКП могут оказаться перспективными в таких предметных областях, как распознавание образов, диагностика и т.п.

Литература

1. *Стеклов В.К.* Теорія електричного зв'язку / В.К. Стеклов, Л.Н. Беркман. – К.: Техніка, 2006. – 552 с.
2. *Умаров А. С.* Некоторые аспекты создания информационных систем для сбора и хранения научной и наукометрической информации / А. С. Умаров, Н. В. Попова, В. А. Зелепухина // Прикаспийский журнал: управление и высокие технологии. – 2013. – № 3 (23). – С. 111 – 118.
3. *Осипов Г.С.* Контентная фильтрация в Интернете: современный уровень и перспективы развития. / Г.С. Осипов, И.В. Смирнов, И.В. Соченков, И.А. Тихомиров // Труды 4-й Международной конференции «Системный анализ и информационные технологии», (17-23 августа 2011г.; Абзаково, Россия). – Челябинск: Издательство Челябинского гос. ун-та, 2011. – С. 103 – 109.
4. *Мироненко А.Н.* Модель фильтрации спам-сообщений в потоке электронной почты / А.Н. Мироненко, С.В. Белим // Вестник компьютерных и информационных технологий. – 2011. – № 11. – С. 34 – 36.

5. *Бойцов Л.М.* Классификация и экспериментальное исследование современных алгоритмов нечеткого словарного поиска / Л.М. Бойцов // Труды 6-ой Всероссийской научной конференции “Электронные библиотеки: перспективные методы и технологии, электронные коллекции” – RCDL2004, Пущино, Россия – 2004 / [Электронный ресурс]. – Режим доступа: http://rcdl.ru/doc/2004/paper27.pdf.
6. *Ставровиецкий В. В.* Алгоритм нечеткого фонетического поиска на основе простых чисел / В. В. Ставровиецкий, Ю.Е. Гапанюк, В.А. Галкин // Молодежный научно-технический вестник. – М.: МГТУ им. Н.Э. Баумана. – 2012. – № 7 / [Электронный ресурс]. – Режим доступа: http://sntbul.bmstu.ru/doc/457989.html.
7. *Леоненко Л.Л.* Теория подобия конечных последовательностей и ее приложения к распознаванию образов / Л.Л. Леоненко, Г.В. Поддубный // Автоматика и телемеханика. – 1996. – № 8. – С. 119 – 131.
8. *Leonenko L.* Analogies Between Texts: Mathematical Models and Applications in Computer-assisted Knowledge Testing // Information Models of Knowledge. – Kiev, Ukraine – Sofia, Bulgaria: ITHEA, 2010. – P. 128 – 134.
9. *Смит У.* Методы и алгоритмы вычислений на строках / Смит У. – М.: "И.Д. Вильямс", 2006. – 496 с.
10. *Гладкий А.В.* Математические методы изучения естественных языков / А.В. Гладкий // Труды МИАН им. В.А.Стеклова. – 1973. – Т. 133. – С. 95 – 108.
11. *Северин Н.В.* Методы нечеткого поиска в системах контроля нецелевого контента / Н.В. Северин // Вісник Східноукраїнського національного університету ім. В. Даля. – 2012. – № 8 (179). – Ч. 2. – С. 199 – 205.
12. *Леоненко Л.Л.* Выводы по аналогии и методы идентификации сообщений / Л.Л. Леоненко, Н.В. Северин // Наукові праці ОНАЗ ім. О.С. Попова. – 2007. – Вип. 1. – с. 57 – 63.
13. *Жарков С.* Кривые пальцы, приносящие трафик / С. Жарков / [Электронный ресурс]. – Режим доступа: http://www.linkz.ru/promotion/krivuee_palmztcue_prinosyascie_trafik.html.
14. *Кокшаров С.* Роль опечаток в SEO / С. Кокшаров / [Электронный ресурс]. – Режим доступа: http://devaka.ru/articles/seo-misprints.
15. *Прохоров А.* «Приличный» Интернет в школе и дома / А. Прохоров // Компьютер Пресс. – 2007. – №2. – http://www.compress.ru/article.aspx?id=17262&iid=799.
16. *Северин Н.В.* Теория подобия конечных последовательностей в задачах идентификации сообщений / Н.В.Северин // Материалы семинара МСЭ/БРЭ «Комплексные аспекты защиты детей в сети Интернет» – ОНАС им. А.С. Попова, 2011. – http://www.itu.int/ITU-D/cyb/events/2011/Odessa/docs/day2/Severin.ppt.
17. *Каптур В.А.* Узагальнена класифікаційна модель фільтрації контенту в мережі інтернет / В.А. Каптур // Збірник наукових праць ВІТІ НТУУ „КПІ”. – 2011. – №1. – С. 65 – 70.
18. *Моисеев К.В.* Динамический метод фильтрации интернет сайтов с агрессивным содержанием / К.В. Моисеев / [Электронный ресурс]. – Режим доступа: http://www.controlchaostech.com/demo/Public_FilterDinamik.pdf.
19. *Северин Н.В.* Анализ эффективности одного метода автоматической фильтрации запрещенного контента в интернет-запросах / Н.В. Северин // 66 наук.-техн. конф. професорсько-викладацького складу, науковців, аспірантів та студентів. – Одеса, 2011. – С. 58 – 60. – http://onat.edu.ua/dif_files/66_konferencia/seminar_3.doc.
20. *Тодорико О.А.* Оценка сигнатурных алгоритмов поиска по сходству в словаре / О.А. Тодорико, Г.А. Добровольский // Вестник ХНТУ. – 2011. – №2(41). – С. 250 – 254.
21. *Ломакин А.А.* Совершенствование методов идентификации персональных данных в автоматизированной информационной системе «Электронный социальный регистр населения» Санкт-Петербурга / А.А. Ломакин // Электронный научный журнал «Исследовано в России». – 2011. – С. 156 – 163. – http://zhurnal.ape.relarn.ru/articles/2011/015.pdf.
22. *Кулай А.Ю.* О статистических методах идентификации языка искаженных текстовых и речевых сообщений / А.Ю. Кулай, Д.А. Леднов, С.Ю. Мельников // Известия ЮФУ.Технические науки. – 2008. – № 8. – С. 177 – 183.
23. Пресс-релизы Яндекса за 2011 год: Яндекс отвечает сразу на два вопроса / [Электронный ресурс]. – Режим доступа: http://company.yandex.ru/press_releases/2011/0902/index.xml.
24. Google Chrome / [Электронный ресурс]. – Режим доступа: https://www.google.com/chrome?hl=ru.
25. *Харитоненков А. В.* Поиск на неточное соответствие: коды Хемминга / А. В. Харитоненков // Журнал научных публикаций аспирантов и докторантов. – ISSN 1991-3087. – 2009. – Режим доступа: http://jurnal.org/articles/2009/inf32.html.
26. Фонетические алгоритмы. [Электронный ресурс]. – Режим доступа: http://habrahabr.ru/post/114947/.