УДК 621.391 : 004.052.3 : 004.052.4

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

МЕТОДЫ АВТОМАТИЧЕСКОЙ КОРРЕКЦИИ ОШИБОК АДРЕСАЦИИ

МЕТОДИ АВТОМАТИЧНОЇ КОРЕКЦІЇ ПОМИЛОК АДРЕСАЦІЇ

METHODS OF AUTOMATIC ADDRESSING–ERRORS' CORRECTION

Аннотация. Рассмотрены требования, предъявляемые к качеству предоставления услуг в сфере телекоммуникаций, в частности службами электронной почты. Проанализированы отказы, возникающие вследствие ошибок адресации, допускаемых абонентами e-mail. Предложены методы, способные в определенных условиях осуществлять автоматическую коррекцию ошибок в электронных адресах. Указанные методы базируются на математической теории подобия конечных последовательностей.

Анотація. Розглянуто вимоги, що висуваються до якості надання послуг у сфері телекомунікацій, зокрема службами електронної пошти. Виконано аналіз відмов, що трапляються внаслідок помилок адресації, які допускають абоненти e-mail. Запропоновано методи, здатні за певних умов виконати автоматичну корекцію помилок в електронних адресах. Наведені методи ґрунтуються на математичній теорії подібності скінченних послідовностей.

Summary. The requirements to the quality of telecommunication services, specifically to electronic mail, are considered. The failures appearing due to errors in electronic addresses made by e-mail consumers are analyzed. Some methods that can automatically correct electronic addresses under certain conditions are proposed. These methods are based on the mathematical theory of finite sequences similarity.

Обеспечение качества услуг (QoS – Quality of Service) – наиболее важная и сложная проблема в сфере телекоммуникаций. Именно по показателям качества телекоммуникационных услуг опосредованно оценивается степень удовлетворения потребителей предоставляемым сервисом [1].

В концепции развития телекоммуникаций в Украине [2], [3] определены требования, предъявляемые к качеству предоставления телекоммуникационных услуг, ориентированные на международные стандарты. Как отмечено в [4], стандартные меры оценки надёжности электронного оборудования и систем основаны на классических моделях отказов, таких как прогнозирование *среднего времени между отказами* (*СВМО*) и *среднего времени между перерывами в обслуживании* (*СВМПО*). Ведущие разработчики телекоммуникационного оборудования направляют свои усилия именно на удовлетворение данных требований, и уделяют сравнительно мало внимания другим причинам сетевых отказов.

В то же время, в предложенной Киасом классификации отказов [5] в особую категорию выделена *ошибка оператора*, которая определяется как отказ, вызванный непосредственно действиями человека. Такие ошибки, при расчете показателей надёжности и отказов телекоммуникационных сетей, учитываются как отказы по вине телекоммуникационного оборудования и, по данным Киаса [5], влекут за собой свыше 5% всех системных отказов.

Одним из видов предоставления услуг в телекоммуникациях является пересылка почтовых отправлений. С развитием информационных сетей появились службы доставки таких сообщений через электронные средства телекоммуникации, эта услуга получила название «электронная почта». Требования к пересылке почтовых сообщений описаны в правилах предоставления услуг почтовой связи [6], а стандарты работы электронной почты в [7], [8]. В частности, [6] содержит описание причин, по которым сообщение не может быть доставлено адресату традиционной почты, и рекомендации по дальнейшим действиям с такой почтой.

Однако в литературе не уделено должного внимания вопросам, связанным с устранением отказов в электронной почте за счет возможности корректировки ошибочной записи адреса, допущенной по вине абонента e-mail.

Целью статьи является предложение методов, позволяющих автоматически анализировать и в ряде случаев исправлять некоторые типы *ошибкок оператора*, в частности ошибок в e-mail-адресах, допускаемых абонентами электронной почты.

1. «Ошибки оператора» в записях электронных адресов. При формировании электронных адресов к ним не предъявляются требования «осмысленности» имени адреса. Зачастую электронный адрес содержит ряд трудно запоминаемых неосмысленных символов. Такой подход в составлении адреса приводит к тому, что пользователи иногда сталкиваются с проблемой незнания точного адреса (возможно, помнят его приблизительно), и в результате допускают ошибки в записи адреса электронного ресурса или адреса получателя e-mail. Кроме того, ясно, что даже зная точный (и даже «осмысленный») адрес, пользователь может допустить случайную описку в его записи при передаче сообщения.

Подобные ошибки влекут сбой в предоставлении соответствующей услуги, и по классификации Киаса должны быть отнесены к «ошибкам оператора», хотя их виновниками являются конечные пользователи. Последние, тем не менее, часто предъявляет претензии к качеству услуги, считая, по-видимому, что их ошибки должны были быть исправлены оператором или системой.

В случае e-mail ошибка в адресе приводит к отказу – невозможности доставки письма, что снижает качественные показатели системы.

С другой стороны, ошибки в адресах могут быть намеренно использованы, например, когда для «раскручивания» сайта он регистрируется под именем, мало отличающимся от имени популярного ресурса (выбираются имена, содержащие наиболее вероятные ошибки/описки при «ручном» наборе адреса) [9].

В целом, подобные ситуации приводят к снижению показателей надежности сети, и к заявлениям со стороны потребителя о некачественном предоставлении ему услуг.

Если реализовать возможность *автоматизированной коррекции* электронного адреса в случае его незначительного искажения, это могло бы повысить качество услуг, предоставляемых потребителю. Следует отметить, что типами ошибок, допускаемых при «ручном» наборе адреса, чаще всего бывают:

– искажения символа (нескольких символов);

– пропуск символа(ов);

– вставка «лишнего» символа(ов);

– перестановки нескольких (часто – рядом стоящих) символов.

Существуют математические методы, позволяющие оценивать «подобие» произвольных символьных последовательностей, для которых характерны искажения указанных типов [10], [11]. В случаях, когда указанных искажений немного, эти методы оценят степень «подобия» правильного и искаженного электронного адреса как весьма высокую.

Таким образом, возникает возможность повышения качества услуг, предоставляемых телекоммуникационными службами (в частности, электронной почтой), с помощью автоматической коррекции некоторых (не очень грубых) ошибок, допускаемых операторами или пользователями сетей телекоммуникации. Рост качества будет связан с устранением/уменьшением числа отказов (для e-mail – количества писем, возвращенных отправителю по причине допущения ошибок в записи адреса).

2. Основные понятия теории подобия конечных последовательностей (ТПКП). Эти понятия будут здесь изложены максимально кратко в той мере, в какой знакомство с ними необходимо, чтобы оценить возможности ТПКП для коррекции электронных адресов (подробное изложение ТПКП см. в [10], [11]).

Пусть M_0 – множество символов, называемое далее «множеством нулевого уровня». Это может быть, например, множество A всех символов ANSI-таблицы; или его подмножество A_E, используемое при формировании электронных адресов; или множество букв какого-либо (естественного или искусственного) языка.

Объектами 1-го уровня называются конечные последовательности элементов из M_0, они составляют множество M_1. Объекты 2-го уровня – это конечные последовательности элементов из M_1, они составляют множество M_2, и т.д. Так, если $M_0 = A_E$, то в M_1 войдут всевозможные (в том числе неправильно сформированные) электронные адреса; а для примера с языком объекты 1-го уровня – это *слова* этого языка, а объекты 2-го уровня – его *предложения*.

Считается, что для элементов базового множества M_0 имеет место рефлексивное и симметричное отношение «подобия». Например, если $M_0 = A_E$, «подобием» можно считать обычное равенство символов. В случае натурального языка «подобие» может задаваться более сложным образом – скажем, можно считать «подобными» символы е и ё.

Вводятся два типа «подобия» для объектов *a* и *b*, принадлежащих уровню M_i ($i > 0$). Именно, если *a* и *b* – два объекта одного и того же ненулевого уровня, то задаются две следующие меры:

$$Fw(a,b) = dw(a,b) / \max(|a|,|b|); \quad (1)$$

$$Fs(a,b) = ds(a,b) / \max(|a|,|b|), \quad (2)$$

где $dw(a,b)$ – число подобных суб-объектов в *a* и *b*;

$ds(a,b)$ – длина длиннейшей подпоследовательности подобных суб-объектов в *a* и *b*;

$|a|$ – длина (число суб-объектов) объекта *a*.

Говорят, что объекты *a* и *b* подобны в широком смысле на уровне подобия δw ($0 < \delta w \le 1$), если $Fw(a,b) \ge \delta w$; т.е. число подобных суб-объектов в *a* и *b* достаточно велико.

Говорят, что *a* и *b* подобны в узком смысле на уровне подобия δs ($0 < \delta s \le 1$), если $Fs(a,b) \ge \delta s$; т.е. длиннейшая подпоследовательность их подобных суб-объектов, сохраняющая порядок следования этих суб-объектов в составе как *a*, так и *b*, достаточно длинная.

Например, пусть $M_0 = A_E$. Рассмотрим следующие объекты 1-го уровня:

a = n_severin, b = m_severin, c = m_severni.

Нетрудно видеть, что

$$Fw(a,b) = Fs(a,b) = 8/9 \approx 0{,}889;$$

$$Fw(a,c) = 8/9 \approx 0{,}889;$$

$$Fs(a,c) = 7/9 \approx 0{,}778.$$

В действительности последовательность символов n_severin является частью адреса e-mail, принадлежащего автору: n_severin@ukr.net. Если оценивать степени подобия «полных» электронных адресов:

a = n_severin@ukr.net,
b = m_severin@ukr.net,
c = m_severni@ukr.net,

то получим:

$$Fw(a,b) = Fs(a,b) = 16/17 \approx 0{,}941;$$

$$Fw(a,c) = 16/17 \approx 0{,}941;$$

$$Fs(a,c) = 15/17 \approx 0{,}882.$$

Из этих примеров видно, что при сравнении электронных адресов целесообразно применять узкую меру подобия (2) *Fs*: перемешивание символов в адресе является существенным. Что касается широкой меры (1) *Fw*, то в работах [10], [11] она чаще всего применялась для оценки подобия объектов 2-го уровня – предложений натуральных языков, где порядок слов менее значим, чем порядок букв в слове.

Доказано [11], что каждая из функций $1–Fw$ и $1–Fs$ является метрикой на множестве M_i объектов *i*-го уровня, т.е. задает «расстояние» между ними. Предложены также алгоритмы рассчета мер *Fw* и *Fs*, и доказано, что они имеют полиномиальную сложность, т.е. вполне эффективны [10], [11].

Обобщением мер *Fw* и *Fs* являются меры подобия, учитывающие (кроме подобия и порядка суб-объектов) *относительную значимость* различных суб-объектов объекта *a* заданного уровня M_i [11]. Например, можно считать, что пропуск/замена символа @ в адресе e-mail является более «грубой» ошибкой, чем пропуск/замена какого-либо иного символа. Это означает, что относительный «вес» @ в адресе выше, чем у других символов. Если, допустим, приписать @ вес 5, а всем остальным символам адреса – вес 1, мера узкого подобия $Gs(a,b)$ из [11] дает следующие значения подобия адресов:

b = n_severi@ukr.net;
c = n_severin_ukr.net

относительно правильного адреса a = n_severin@ukr.net:

$$Gs(a,b) \approx 0{,}952;$$

$$Gs(a,c) \approx 0{,}762.$$

Отсюда видно, что 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.

3. Сравнение методов ТПКП и методов оценки подобия символьных последовательностей по расстоянию Хэмминга. В математике, логике и технических дисциплинах предлагались различные меры, оценивающие «близость» или «расстояние» между последовательностями того или иного вида. В этом разделе методы ТПКП будут сравнены с методами, базирующимися на расстоянии Хэмминга, применяемом в теории кодирования. Это, как рассчитывает автор, способствует прояснению некоторых специфических черт методов ТПКП.

Расстояние Хэмминга используется для оценки обнаруживающей и исправляющей способности кода [12]. Кодовое расстояние между двумя последовательностями символов («словами») численно равно минимальному числу ошибок, исправление которых может превратить одно слово в другое. Декодер, который декодирует каждую принятую последовательность в ближайшее к ней по расстоянию Хэмминга кодовое слово, выбирает то кодовое слово, условная вероятность передачи которого максимальна, и потому называется декодером максимального правдоподобия.

Приведем максимально простой пример, показывающий как близость, так и различия метода Хэмминга с методами, предложенными в [10], [11].

Рассмотрим три слова:

$$a = 0111100, \qquad b = 0100101, \qquad c = 0010110,$$

и сравним меру Хэмминга с описанной выше мерой узкого подобия символьных последовательностей *Fs*. Для их сравнения нужно, во-первых, учесть не различие, а подобие кодов по Хэммингу (число совпавших разрядов). Далее это число следует нормализовать, разделив на длину кодового слова, т.е. в данном случае на 7. Тогда получим меру χ «нормализованного подобия по Хэммингу». При этом:

$$\chi(a, b) = 4/7 \approx 0{,}571;$$
$$\chi(a, c) = 4/7 \approx 0{,}571;$$
$$\chi(b, c) = 3/7 \approx 0{,}429.$$

Что касается меры *Fs*, то будем иметь:

$Fs(a, b) \approx 0{,}571$ – поскольку max длинная общая подпоследовательность у этих слов – это 0111;

$Fs(a, c) \approx 0{,}714$ – (max общая подпоследовательность: 01110);

$Fs(b, c) \approx 0{,}714$ – (max общая подпоследовательность: 01010).

Представим теперь, что получено слово $p = 0011100$, т.е. отличающееся от *a* в единственном – 2-м разряде. Код Хэмминга в сочетании с методом наибольшего правдоподобия сочтет такое слово «искаженным *a*». А вот что дает применение меры *Fs*:

$Fs(p, a) \approx 0{,}857$ – max длинная общая подпоследовательность у этих слов – это 011100;

$Fs(p, b) \approx 0{,}571$ – (max общая подпоследовательность: 0011);

$Fs(p, c) \approx 0{,}857$ – (max общая подпоследовательность: 001110).

Таким образом, с точки зрения меры *Fs*, слова *a* и *c* «равно близки» – и весьма близки, поскольку каждая из двух общих подпоследовательностей имеет длину 6 – полученному слову *p*.

Это говорит приблизительно вот о чем. Искажения, учитываемые расстоянием Хэмминга – это искажения разрядов (*символов*). В дополнение к искажениям этого типа мера *Fs* учитывает возможности пропуска, вставки «лишней» буквы, сдвига и/или перестановок букв (с возможностью изменения *длины* кодового слова). В тех задачах, где такое следует учитывать, мера *Fs* будет лучше меры Хэмминга. Задачи оценки близости текстов (в частности, электронных адресов), вводимых человеком, как раз таковы.

4. Методы анализа/коррекции адресов e-mail. В настоящее время в электронной почте применяют следующие приемы обработки «неправильных» адресов. Агент передачи сообщений MTA (Mail Transfer Agent) при отсутствии указанного отправителем адресата формирует возвращенное письмо (bounce message). Это сообщение отсылается обратно отправителю, когда почтовый ящик получателя не существует или недоступен. (Существуют различные шаблоны таких писем, и содержимое возвращенного письма может отличаться в зависимости от настроек почтового сервера). В тексте возвращенного письма указывается текст ошибки, адрес почтового ящика, список кодов ошибок, и причина, по которой письмо не могло быть доставлено. Также к bounce message может быть добавлено исходное письмо или некоторая его часть (на это обычно влияют ограничения по размеру возвращенного письма).

Таким образом, отправленная пользователем срочная и важная корреспонденция может не достичь адресата из-за ошибки (опечатки) в записи электронного адреса. А в случае несохранения им отправленной копии – к ее потере. При возможности повторной отправки письма его отправителю необходимо будет выяснить правильный адрес, а в случае невозможности этого он может пытаться подобрать его методом перебора. Это приводит как к увеличению нагрузки на сеть, так и снижению качества предоставляемых услуг.

В традиционной почтовой связи существует требование о принятии всех мер по доставке письма адресату, в том числе и по выяснению нового места пребывания последнего. Электронная почта является альтернативой традиционной, и качество предоставляемого ею сервиса должно быть

по крайней мере «не хуже». Учитывая опыт, полученный в традиционной почте, прежде чем отправить e-mail-сообщение о невозможности доставки письма адресату, следует предпринять все меры по «отысканию» адресата.

Широкое распространение в сети серверов, построенных на Linux/Unix-платформах с использованием свободного программного обеспечения [13], дает возможность разработчикам не только изучать исходные коды, но и выполнять их модификацию. С учетом этого, можно предложить – в качестве одного из методов борьбы с ошибками адресации – внедрение в MTA модуля, позволяющего выполнять коррекцию ошибочного или же поиск «близкого» адреса. В случае возможности однозначного определения адресата, модуль позволит выполнить корректировку адреса и доставку письма. При невозможности однозначно определить адресата в возвращенное письмо может быть включен перечень «близких» адресов.

Рассмотрим пример оценок близости электронных адресов, полученных при использовании методов ТПКП. В качестве примера возьмем электронный адрес автора: n_severin@ukr.net. Оценка близости выполнена 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. Пример оценок приведен в табл. 1.

Таблица 1 – Оценка близости электронных адресов, при использовании методов ТПКП

Искаженный адрес	Оценка близости к «правильному адресу»
nseverin	0,889
m_severin	0,889
n-severin	0,889
nv_severin	0,900
n_sveerin	0,889
n_siverin	0,889
n_sivirin	0,778
n_civerin	0,778

Из примера видно, что незначительные искажения (одиночные ошибки, перестановки символов) «мало» искажают электронный адрес. Установив порог меры близости, при котором «близкие» к правильному адреса допустимо считать результатами «опечаток», равным 0,88, первые шесть 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 шесть адресов можно включить в bounce message.

Таким образом, даже в случаях невозможности или недопустимости автоматической коррекции адреса, предложенные методы делают bounce message более информативным. В итоге применение методов ТПКП позволит как повысить качество работы сервиса, так и снизить «ошибочный» трафик сети. Ведь пользователь, не зная точного адреса, может пытаться выполнить отправку путем его неоднократного «угадывания». Такое поведение неоправданно расходует ресурсы, как почтового сервера, так и сети, поскольку влечет неоднократную передачу по сети невостребованной информации (иногда больших объемов, например, фотографий, звуковых файлов и т.п.).

В заключение работы отметим следующее. Предложенные методы позволяют оценивать степень искажения электронного адреса, формирования множества «близких» к искаженному действительных адресов, а в случае малых искажений – автоматическую корректировку адреса. Для служб электронной почты это позволяет ускорить процесс нахождения правильной записи адреса, и тем самым выполнить необходимую доставку e-mail (устранить отказ в услуге); а также уменьшить нагрузку на сеть,

Помимо систем электронной почты, методы ТПКП представляются перспективными для устранения ошибок:

– при наборе номера; в телефонных сетях использующих IP-телефонию;

– при вводе адреса Интернет ресурса: ввод ошибочного имени домена;

– на сервере мобильного оператора: обработка ошибочных номеров при отправке sms или mms сообщения.

Литература

1. *Свиридов Е.* Градации качества [*Электронный ресурс*] / Е. Свиридов – Сети и телекоммуникации, 2006 – №11. – Режим доступа: http://www.seti-ua.com/?in=seti_show_article&seti_art_ID=256&_by_id=1&_CATEGORY=35
2. *Верховна Рада України*, Закон України про телекомунікації від 18.11.2003 № 1280-IV [*Электронный ресурс*]. – Режим доступу: http://zakon.rada.gov.ua/cgi-bin/laws/main.cgi?page=1&nreg=1280-15
3. *Кабінет Міністрів України*, Розпорядження від 07.06.2006 № 316-р Про схвалення концепції розвитку телекомунікацій в Україні [*Электронный ресурс*]. – Режим доступу: http://zakon.rada.gov.ua/cgi-bin/laws/main.cgi?nreg=316-2006-%F0
4. *Hudyma R.* Causes of failure in IT telecommunications networks. [*Электронный ресурс*] / R. Hudyma, Deborah I. Fels – Proceedings of SCI – 2004. Florida. – С. 35-38. – Режим доступа: http://www.ryerson.ca/clt/publications/papers/ITFailure.doc
5. *Kyas O.* Network Troubleshooting. / O. Kyas // Palo Alto California, Agilent Technologies. – 2001.
6. *Постанова Кабінету Міністрів України* від 5 березня 2009 р. N 270 «Про затвердження Правил надання послуг поштового зв'язку» [*Электронный ресурс*]. – Режим доступу: http://www.ukrposhta.com/www/upost.nsf/(documents)/4DA6FED7F3BD25BDC225746E002F0456
7. *Enhanced Mail System Status Codes*: RFC 3463 – Internet Engineering Task Force (IETF), 2003. – 16 с. – (Международный стандарт).
8. *Internet Message Format*: RFC 2822 – Internet Engineering Task Force (IETF), 2001. – 51 с. – (Международный стандарт).
9. *Кокшаров С.* Роль опечаток в SEO [*Электронный ресурс*]. – Режим доступа: http://devaka.ru/articles/seo-misprints
10. *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.
11. *Leonenko L.* Analogies between Texts: Mathematical Models and Applications in Computer-assisted Knowledge Testing / L. Leonenko // Information Models of Knowledge. – Kiev, Ukraine – Sofia, Bulgaria: ITHEA, 2010. – P. 128 – 134.
12. *Хэмминг Р.В.* Теория кодирования и теория информации: пер. с англ. / Хэмминг Р.В. – М. : Радио и связь, 1983. – 176 с.
13. *Security Lab. И*сследование: «*И*деальный почтовый сервер» [*Электронный ресурс*]. – Режим доступа: http://www.securitylab.ru/analytics/309889.php.