*УДК 621.391 : 004.052.4 : 004.056.53*

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

# МЕТОД АВТОМАТИЧЕСКОЙ ФИЛЬТРАЦИИ ЗАПРЕЩЕННОГО КОНТЕНТА В ЗАПРОСАХ ИНТЕРНЕТ-РЕСУРСОВ

# МЕТОД АВТОМАТИЧНОЇ ФІЛЬТРАЦІЇ ЗАБОРОНЕНОГО КОНТЕНТУ У ЗАПИТАХ ІНТЕРНЕТ-РЕСУРСІВ

# A METHOD FOR AUTOMATIC FORBIDDEN CONTENT FILTRATION IN INTERNET RESOURCES' QUERIES

***Аннотация.*** Рассмотрены вопросы фильтрации нецелевого (в частности, запрещенного) контента в сети Интернет. Проанализирован ряд систем безопасного поиска и пути их обхода, в частности, посредством ввода «искаженного» запроса. Предложено использование метода, базирующегося на математической теории подобия конечных последовательностей, для автоматической оценки контента ресурсов, предполагаемых запросом.

***Анотація.*** Розглянуто питання фільтрації нецільового (зокрема, забороненого) контенту в мережі Інтернет. Проаналізовано низку систем безпечного пошуку і шляхи їх обходу, зокрема, через введення «спотвореного» запиту. Запропоновано використання методу, що ґрунтується на математичній теорії подібності скінченних послідовностей, для автоматичної оцінки контенту ресурсів, передбачених у запиті.

***Summary.*** I discuss some questions concerning the filtration of inappropriate (in particular, forbidden) content in Internet. A group of secure search systems were analyzed, as well as certain ways of their evasion, in particular by means of entering a "garbled" query. I propose a method, basing on the mathematical theory of finite sequences' similarity, for automatic estimation of content of resources that are guessed in a query.

Защита пользователей Интернет от агрессивного контента (порнография, пропаганда и распространение наркотиков, экстремизм и т.д.) является чрезвычайно важной проблемой, которой уделяется внимание во всем мире. Особенно остро стоит этот вопрос по отношению к несовершеннолетним пользователям, поскольку неконтролируемый доступ к Интернет-ресурсам данного характера может вызвать не только «заболевание» Интернет-зависимостью, но и оказать пагубное воздействие на развитие и психику ребенка. По инициативе Международного Союза Электросвязи защита детей в онлайн пространстве названа приоритетом № 1 в 2009 г. [1 … 4]. В Украине принят ряд законопроектов, направленных на борьбу с распространением противоправного контента. Они предусматривают, в т.ч., и уголовную ответственность [5, 6]. Ведущие разработчики программного обеспечения и операторы телекоммуникаций активно развивают программное обеспечение, направленное на защиту от агрессивного контента в сети Интернет. Был разработан ряд программных продуктов, а также стандартов, основанных на наиболее распространенных механизмах фильтрации web-контента в сети Интернет [1 … 4, 7]. В частности, в [1] предложена обобщенная модель фильтрации web-контента в сети Интернет, позволяющая выполнить квалификацию существующих методов и подходов к фильтрации.

Однако нельзя говорить о каком-либо «общем» решении данной проблемы, способном удовлетворить все предъявляемые к задачам фильтрации требования. Вопрос развития существующих систем и методов, а также поиска новых решений остается открытым. В частности, существующие системы фильтрации «плохо справляются» с идентификацией контента, содержащего запрещенные слова, умышленно представленные в искаженном, но вполне понятном человеку виде.

**Целью статьи** является разработка метода автоматического анализа содержимого запрошенного Интернет-ресурса на наличие нецелевого (в частности, запрещенного) контента, в том числе когда запрос ресурса представлен в «искаженном» виде. Этот метод базируется на математической теории подобия конечных последовательностей.

**1. Анализ систем и средств ограничения доступа к нецелевому контенту.** Широко используются следующие методы доступа к ресурсам Интернет, в том числе содержащим запрещенный контент: 1) ввод адреса сайта (URL) в строке браузера (*type-in*); 2) переход по ссылке (рекламному баннеру) с другого сайта; 3) переход по результату запроса в поисковой системе [8].

Первые два способа можно считать несколько менее значимыми в плане обсуждаемой проблемы. В первом случае человеку довольно сложно запомнить и безошибочно ввести адрес запрещенного ресурса. Второй способ, хотя и более популярен (весь Интернет пестрит рекламой аморального содержания), все же предоставляет доступ только к определенному ограниченному ряду рекламируемых ресурсов. Первоочередное внимание, на наш взгляд, следует уделить третьему способу доступа, так как именно поисковым системам выпала роль основного информатора пользователей Интернет о существующих и вновь появляющихся ресурсах агрессивного контента (см. [8]). Следует отметить, что ведущие разработчики поисковых систем занимаются проблемой безопасного Интернета и предоставляют ряд проектов, направленных на обеспечение безопасного поиска [4]. К ним можно отнести: *сервис безопасного поиска от Google (http://www.google.ru/familysafety/)*, *семейный фильтр от Yandex* (*http://yandex.ua/familysearch*) и ряд других, в том числе специализированных, систем – как, например, *безопасная поисковая система интернет Цензор (http://search.icensor.ru)*. Однако даже поверхностное исследование данных систем показывает их уязвимость и необходимость совершенствования. Так, например, в системе безопасного поиска от *Google*, введя сообщение *поститутка* (т.е. допустив банальную опечатку – пропуск «*р*» в запрещенном слове), на момент написания статьи мы получили ссылки на 43200 ресурсов, где в первой же десятке находятся ресурсы откровенного порнографического характера. Более детально о результатах исследований автора, касающихся подобных «обманов» безопасных систем, можно узнать из доклада [9].

Конечно же, системы безопасного поиска – это всего лишь один из инструментов фильтрации трафика, содержащего нецензурный контент. Поскольку контроль работы на стороне пользователя в большинстве случаев не представляется возможным, не стоит полагаться на обязательное использование им только безопасного поиска. Но заметим, что даже в последнем случае «искажение» запроса может выступать инструментом для обхода защиты системы. Например, работая в созданном специально для детей *обозревателе Интернет-ресурсов Buddy Browser*, который по утверждению разработчиков гарантирует 100%-ю защиту от неуместного контента [10], используя искаженные слова в запросах поисковых систем, можно получить доступ к web-ресурсам порнографического характера.

В любой системе ограничения доступа могут использоваться такие виды фильтрации, как фильтрация по адресу, фильтрация по содержимому ресурса (контенту), либо их комбинация. В современном программном обеспечении широкое распространение получил подход применения «белых» и «черных» списков. Такие списки 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 [1]. Использование «черных» списков URL-адресов имеет очевидный недостаток в скорости реагирования системы на возникновение новых ресурсов. При современном интенсивном развитии Интернет 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 [1], [2].

Что касается запрещенных слов, которые используются в качестве эталонов сравнения при *динамической фильтрации* [2] запрошенного контента, то и здесь имеются свои трудности. Помимо запрещенных слов, представленных в словарях либо используемых в обиходе, на практике встречаются их искаженные («мутированные») формы. Т. е. в результате случайной, а зачастую и умышленной, ошибки происходит изменение слова, но при этом оно остается легко узнаваемым – для человека, но не для системы фильтрации.

Такой подход к «обману» систем фильтрации не является новым. Так, на электронных досках объявлений (Bulletin Board System - BBS) с 1990 г. применялся язык Leet (1337). Суть его – замена латинских букв на похожие цифры и символы, имитация и пародирование ошибок, характерных для быстрого набора текста, имитация жаргона и т.д. [11]. В результате пользователи могли получать доступ к хранящимся на BBS файлам и папкам, играм, специальным чатам. Часто в дополнение к этому можно было скачать архивы с пиратскими программами, порнографией, и файлы с текстами противоправного содержания, например, как сделать бомбу или как изготовить наркотик в домашних условиях. Язык Leet (1337) был создан, чтобы обходить запреты и языковые фильтры, разработанные для BBS, а также чтобы администратор не понял, о чём идёт речь, и не закрыл тему.

Для исследования популярности использования искаженных слов можно воспользоваться представленными в сети Интернет сервисами статистики ключевых запросов в поисковых системах: *http://wordstat.yandex.ru, http://adwords.google.com, http://adstat.rambler.ru* и др. В [9] представлены результаты таких исследований для искаженных слов («слов-мутантов»), принадлежащих к

запрещенному контенту. В качестве примера рассмотрим одну из таблиц, содержащую результаты всего по одному слову (*порнография*) из обширного множества запросов. В табл. 1 наряду с результатами обработки таких запросов системами безопасного поиска показано (в скобках) количество ответов на запрос при обычном поиске.

Таблица 1 – Выборочная статистика запросов поисковых систем

| Ключевое слово | Количество запросов в месяц http://wordstat.yandex.ru | Количество найденных ответов | | |
|---|---|---|---|---|
| | | Семейный фильтр *Yandex* (обычный) | Безопасная поисковая система *Цензор* | Безопасный поиск *Google* (обычный) |
| ***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 (изм. на ***мафия***) (19 300) |
| порноргафия | 21 | 24 (94) | 0 | 268 (5 900) |

Представленные искажения слова могут быть следствиями ошибок человека при работе с клавиатурой. Например, «п***ро***ногр***фа***ия» могла возникнуть из-за того, что: (1) ***ро*** – соседние символы на клавиатуре; (2) ***фа*** – латинское ***a*** и русское ***ф*** размещены на одной клавише клавиатуры, и пользователь, полагая, что нажимает ***a***, получил ***ф***.

Однако указанные в табл. 1 искажения могут быть – и часто бывают – намеренными. Анализ статистики из [9] говорит о том, что запросы, содержащие «*очепятки*» (грамматические ошибки), довольно широко распространены. 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.

Так, 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 работы SEO-технологий (Search Engine Optimization – поисковая оптимизация) [12]. В последнем случае при формировании *семантического ядра сайта* [8] в тексты его страниц умышленно внедряются определенные «искаженные» слова и словосочетания, которые впоследствии выступают «ключевыми» словами для поисковых систем [12] и предоставляют пользователю возможность получить доступ к запрещенному контенту в обход некоторых систем безопасного доступа.

Здесь становится ясным, что предусмотреть все возможные варианты ошибок в записи запрещенных слов и внести их в «черный» список на практике невозможно. Одним из примеров, подтверждающим это, может послужить анализ Интернет-фильтра «БлокПрограмма» (*http://blokprogramma.ru*) основанного на идентификации запрещенного контента по ключевым словам и фразам. База этого фильтра превосходит 8600 слов ненормативной лексики и 3100 слов порнографического содержания, но тем не менее программа не во всех случаях адекватно реагирует на слова порнографического характера, в особенности представленные в искаженном виде. Следовательно, современные системы фильтрации должны уметь распознавать «подлоги» и выполнять обработку естественного языка с учетом возможных намеренных искажений.

**2. Методы распознавания «искаженных» запросов нецелевого контента**. В качестве базиса автоматизированной системы оценки контента на наличие в нем «слов-мутантов» могла бы, по нашему мнению, выступать теория подобия конечных последовательностей (**ТПКП**). Краткий обзор ТПКП можно найти в статьях [13, 14], а более подробно она описана в работах [15, 16, 17]. Мы приведем максимально краткое изложение идей ТПКП, демонстрируя применение ее методов на примерах, связанных с темой статьи.

Пусть $M_0$ – произвольное множество, элементы которого будем называть объектами нулевого уровня. Объектами 1-го уровня называются конечные последовательности элементов из $M_0$ (обозначим их множество через $M_1$); объектами 2-го – конечные последовательности элементов из $M_1$ (множество $M_2$) и т.д.

В ТПКП для объектов одного и того же уровня вводится несколько видов подобия, мы рассмотрим здесь только два из них. Первый – это так называемое *невзвешенное подобие в узком смысле* (*Fs*). Если *a* и *b* – два объекта одного уровня, то они считаются *Fs*-подобными, если длиннейшая подпоследовательность их подобных суб-объектов, сохраняющая порядок следования этих суб-объектов в составе как *a*, так и *b*, достаточно длинна [15]. Численная мера для данного подобия проста:

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

где *ds(a,b)* – длина длиннейшей подпоследовательности сходных суб-объектов в *a* и *b*; |*a*| – длина (число суб-объектов) объекта *a*.

Следующие далее примеры будут «привязаны» к применению мер ТПКП в задачах фильтрации текстового контента в системах безопасного доступа. Эти системы работают в основном с естественными языками. Оценивая подобие текстовых сообщений натурального языка методами ТПКП, к объектам нулевого уровня целесообразно отнести *символы*, используемые при построении сообщений. Объектами первого уровня будем считать *слова*, формирующие контент web-ресурсов (причем допустимы «искаженные», грамматически ошибочные слова). К объектам второго уровня можно отнести комбинации ключевых слов, по которым поисковые системы производят индексацию и поиск web-ресурсов, а объектами третьего уровня будут предложения.

Рассмотрим сообщение 1-го уровня *a = порнография*, относящееся к запрещенному контенту, и его «искаженные» формы – *b = пронография* и *c = п0рн0графия*, а также близкое к *a* по звучанию *d = монография*. Применение меры подобия *Fs* дает для всех приведенных «искажений» *a* одинаковые результаты:

$$Fs(a,b) = 9/11 \approx 0{,}818;\ Fs(a,c) = 9/11 \approx 0{,}818;\ Fs(a,d) = 9/11 \approx 0{,}818.$$

Заметим, что сообщение *b* могло быть и результатом непреднамеренной опечатки, тогда как *c* – часто встречающаяся умышленная подмена буквы о на цифру 0.

Мы видим, что мера *Fs* не вполне адекватна задаче отделения запрещенных и разрешенных слов, поскольку безобидное слово *d* оказывается столь же подобным запрещенному *a*, что и его умышленное искажение *c*.

Ввиду этого здесь уместнее использовать другие меры подобия ТПКП, а именно позволяющие учитывать *относительную значимость* различных суб-объектов объекта *a*. В частности, мера *взвешенного* узкого подобия *Gs* позволяет приписать суб-объектам исходного сообщения *a* различные *веса*. Численная мера для подобия *Gs* выражается не так просто, как для *Fs*, поскольку использует ряд дополнительных математических понятий (см. [16]). Не приводя ее здесь, покажем на примерах, какой «выигрыш» она дает в плане рассматриваемой задачи.

Естественно считать, что для *a = **порно**графия* подчеркнутая часть слова более значима (при оценке его сходства с «искажениями»), чем остальные символы. Учитывая, что ошибки в записи согласных встречаются реже, чем для гласных (например, п***а***рнография), для них можно установить более высокий вес. Зададим, например, для согласных «п», «р» и «н» в выделенной части слова *a* вес 2, а для всех остальных его символов – вес 1. Даже при такой сравнительно небольшой разнице в весе, применяя меру *Gs* к рассмотренным выше сообщениям *b*, *c* и *d*, получим следующие оценки их подобия *a*:

$$Gs(a,b) \approx 0{,}857; \qquad Gs(a,c) \approx 0{,}825; \qquad Gs(a,d) \approx 0{,}714.$$

Если применить более «тонкое» ранжирование весов в слове *a* (считая, например, что корень слова более важен при его распознавании, чем, допустим, окончание и т.п.), отделимость «искаженных» форм *a* от других слов, близких к *a* по буквенному составу, возрастет. Например, если распределить веса так:

**пор но графия**
**212 11 000000,**

получим значения подобия:

$$Gs(a,b) \approx 0{,}857; \qquad Gs(a,c) \approx 0{,}714; \qquad Gs(a,d) \approx 0{,}429.$$

Итак, использование взвешенного подобия *Gs* улучшает результаты оценки, поскольку значения меры сходства с *a* для запрещенных слов контента (сообщений *b* и *c*) оказываются существенно выше, чем сходство *a* со сторонним сообщением *d*.

Приведем пример оценки запросов из табл. 1 с помощью ТПКП, используя меру подобия *Gs* с приведенным выше распределением весов слова *a*. В следующей таблице первый столбец содержит искажения слова *a = порнография,* встречающиеся в запросах поисковых систем, а второй – численные значения *Gs*-подобия этих искажений слову *a*.

Из табл. 2 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 высоко-подобные ему слова.

Таблица 2 – Оценки подобия искажений слова *a*

| Сообщение *b* | *Gs(a,b)* |
|---|---|
| п***а***рнография | 0,857 |
| понография | 0,714 |
| порногафия | 1,000 |
| п***ро***нография | 0,857 |
| порномафия | 0,917 |
| порно***рг***афия | 0,917 |
| п***ро***ногр***фа***ия | 0,825 |

При анализе «черных» списков запрещенных слов систем безопасного доступа можно обнаружить, что там довольно часто присутствуют как запрещенное слово, так и слова, полученные в результате его склонения. Такой подход приводит к неоправданному наращиванию объемов «черных» списков, тем более, что всеравно невозможно предусмотреть все возможные варианты записи слова, хотя бы потому, что в нем может присутствовать ошибка. Применение методов оценки ТПКП устраняет этот недостаток. Достаточно ввести одно слово, а близкие слова (полученные в результате 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. Эти слова, в отличие от слов из табл. 1 и табл. 2, отобраны *не* по результатам статистики запросов из сети Интернет (хотя вполне могут в них появиться). Цель табл. 3 – продемонстрировать отделимость посредством меры *Gs* «искажений» *a* от незапрещенных слов:

Таблица 3 – Искажения слова *a* в сравнении с «разрешенными» словами

| Сообщение *b* | *Gs(a,b)* |
|---|---|
| порно | 1,000 |
| порно***индустрия*** | 0,647 |
| порн***уха*** | 0,825 |
| ***ф***онография | 0,429 |
| ***г***о***л***ография | 0,429 |
| ***о***порно-графи***ческий*** | 0,579 |
| ***оп***пор***ту***ни***зм*** | 0,611 |

Табл. 2 и 3 демонстрируют достоинства предложенного метода оценки слов на их «запрещенность». Так, в рассмотренных примерах слова *монография*, *фонография*, *голография* (казалось бы, 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 запрещенных слов, наряду с *порнографией*.

Рассмотрим еще один пример в котором выполним оценку грамматически правильного *длинного составного* слова: *a = порнозвезда*. При распределении весов (аналогичном выбранному выше для *порнография*):

**порнозвезда**
**21211000000**

для слов *b= порно-звезда*, *c= порно-звездища* и *d= порно-звездочка* будем иметь:

$$Gs(a, b) = 0{,}917; \quad Gs(a, c) = 0{,}786; \quad Gs(a, d) = 0{,}733.$$

Полученные результаты показывают, что наличие *ожидаемых* в ответе (запросе) символов, даже при их *нулевых весах*, приводит к существенному *росту* подобия.

В заключение отметим: в данной работе предложен метод автоматического анализа содержимого запрошенных пользователями Интернет-ресурсов на наличие запрещенного контента; предложенный метод базируется на математической теории подобия конечных последовательностей.

Автор планирует произвести практический эксперимент по оценке эффективности предложенного метода, как в плане качества распознавания искажений запрещенных слов, так и по скорости работы алгоритма.

***Литература***


1. *Каптур В.А.* Узагальнена класифікаційна модель фільтрації контенту в мережі Інтернет / В.А. Каптур // Збірник наукових праць ВІТІ НТУУ „КПІ”. – 2011. – №1.
2. *Моисеев К.В.* Динамический метод фильтрации Интернет сайтов с агрессивным содержанием [Электронный ресурс]. – Режим доступа: http://www.controlchaostech.com/demo/Public_FilterDinamik.pdf.
3. Всеукраїнська система обмеження доступу до нецільових ресурсів мережі Інтернет [*Электронный ресурс*]. – Режим доступа: http://copworldwide.org.
4. *Прохоров А.* «Приличный» Интернет в школе и дома / А. Прохоров // Компьютер Пресс, 2007 – № 2 [*Электронный ресурс*]. – Режим доступа: http://www.compress.ru/article.aspx?id=17262&iid=799.
5. *Верховна Рада України.* Закон про внесення змін до деяких законодавчих актів України щодо протидії розповсюдженню дитячої порнографії від 20.01.2010 № 1819-VI [Электронный ресурс]. – Режим доступа: http://zakon.rada.gov.ua/cgi-bin/laws/main.cgi?nreg=1819-17.
6. *Верховна Рада України*. Закон України про телекомунікації від 18.11.2003 № 1280-IV [*Электронный ресурс*]. – Режим доступа: http://zakon.rada.gov.ua/cgi-bin/laws/main.cgi?page=1&nreg=1280-15.
7. *Сидельников А.Б.* Организация фильтрации контента в школе и дома [*Электронный ресурс*]. – Режим доступа: http://sidelnikov.wordpress.com/2010/12/03/контент-фильтрация/.
8. *Ашманов И.* Оптимизация и продвижение сайтов в поисковых системах / И. Ашманов, А. Иванов. – СПб. : Питер, 2008. – 400 с. – ISBN 978-5-388-00008-8.
9. *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. [*Электронный ресурс*]. – Режим доступа: *http://seminar.onat.edu.ua/public/doc/presentations/child_protect/Day2/Severin.ppt*.
10. *Boddy Browser* [Електронний ресурс]. – Режим доступу: http://www.buddybrowser.com/Free-Parental-Controls.cfm.
11. *Википедия – свободная энциклопедия* Leet [*Электронный ресурс*]. – Режим доступа: *http://ru.wikipedia.org/wiki/Leet.*
12. *Кокшаров С.* Роль опечаток в SEO [*Электронный ресурс*]. – Режим доступа: http://devaka.ru/articles/seo-misprints.
13. *Леоненко Л.Л.* Идентификация сообщений, характеризуемых неустранимой неоднозначностью, в системах компьютерного тестирования знаний / Л.Л. Леоненко, В.Ю. Баранов, Н.В. Северин // Вісник Східноукраїнського національного університету ім. В. Даля. – 2010. – № 10 (152). – Ч. 2. – С. 134 – 142.
14. Северин Н.В. Методы автоматической коррекции ошибок адресации / Н.В. Северин // Наукові праці ОНАЗ ім. О. С. Попова. – 2011. – № 1. – С. 127 – 132.
15. *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.
16. *Leonenko L*. Analogical inferences in computer assisted knowledge testing systems / *L. Leonenko* // 6-th Multi-Conference on Systemics, Cybernetics and Informatics. Proceedings. – 2002. – XVIII. – P. 371 – 376.

17. *Леоненко Л. Л.* Алгоритмы оценки аналогичности текстов и их применение в компьютерном тестировании / Л.Л. Леоненко // Сб. трудов VII междунар. конф. "Интеллектуальный анализ информации". – К.: Просвіта, 2007. – С. 210 – 220.