**УДК 681.511.3**

# ДИАГНОСТИКА И АНАЛИЗ ПОТОКОВ ASCII-ДАННЫХ

***Александр Манухин***
*Военный институт Национального технического университета Украины "КПИ"*

*Анотація*: **Діагностика й аналіз потоків ASCII-даних. Розглянуто побудову пошукового серверу з позиції аналізу структури потоків і з наступної їхньої кластеризації.**
*Summary:* **Diagnostics and stream analysis of the ASCII-data. The common principle of build-up of the retrieval server from a position of the analysis of contents of a treated stream and its consequent clusterization surveyed.**
*Ключові слова:* **a stream, quantum of a stream, attribute, diagnostics, analysis.**

## Вступление

Политика взаимоотношений "клиент-сервер" предполагает предоставление определенного рода услуг пользователю, в том числе и поиск информации по заданному критерию. В качестве такового обычно выступают связки ключевых слов. В результате предварительного отбора инициатору выдается вал информации, сортировкой которой он будет занят время, зависящее от собственного терпения. Полнота данных в случае подобного поиска превосходит точность более чем в десятки раз.

Рассмотрение подобной проблемы отталкивается от механизма сбора информации провайдером. Программа-робот случайным образом выбирает информацию (сайт) в сети и строит реестр, в котором буква сайта индексируется в зависимости от местоположения в слове. Ключевое слово клиента включает механизм обработки реестра, в результате чего образуются ссылки на тот или иной сайт сети, что будет со временем и выдано пользователю.

Предлагается вместо реестра в поисковом сервере использовать образы потоков, в которых заложена информация об их предметных характеристиках. Задачей пользователя с этого момента становится подбор профиля требуемой предметной области.

## I Общие определения

Рассмотрение любой проблематики предполагает наличие единой терминологии.

*Определение 1.* Потоком называется множество данных, поступающих на мнимое устройство их обработки.

Поток делится на кванты – составляющие потока, объединенные определенной логической завершенностью. Под квантом потока в применении к ПЭВМ будем понимать отдельные файлы. Квант потока имеет собственные атрибуты, определяющие его характеристику дальнейшей обработки (допустим, расширение имени файла). Поток данных представляет собой структуру из элементарных символов, совокупность которых является алфавитом потока. Поскольку информационный поток ПЭВМ состоит из ASCII-символов, назовем его ASCII-потоком (данные – ASCII-данными).

Исходя из критерия органов восприятия человека, весь ASCII-поток разбит на литерный (L), звуковой (V) и поток изображений (S). Происхождение любого потока связывается с определенным источником.

*Определение 2.* Диагностикой потока называется его аутентификация по композиционным и структурным особенностям алфавита с последующей идентификацией (анализ потока).

В качестве объекта диагностики и анализа выступает спектральный образ потока [1-2].

## II Классификация потоков данных

Поскольку образ является продуктом восприятия, целесообразно провести классификацию образов в соответствии с критериями $(K_j)$, перечень которых может выглядеть следующим образом:

I. По отношению к средствам восприятия;

II. По отношению к форме отображения потока в независимости от средств восприятия;

III. По отношению к средствам передачи образов.

Рассмотрим типы возможных образов в соответствии с *первым* критерием $(K_1)$, ограничиваясь спецификой их восприятия, характерной для технических систем:

1. Образы могут представлять собой временные сечения некоторых фрагментов изображения;
2. Образы могут быть сформированы таким способом, что отображают фрагменты динамики или динамические процессы в целом;
3. Образы могут отображать фрагменты действительности в координатах параметров, восприятие которых возможно только специальными техническими средствами;

4. Образы могут отображать фрагменты предметной области посредством параметров, которые не имеют целостной физической интерпретации в предметной среде, а являются лишь допустимым способом абстрагирования в описании законов функционирования (*псевдообразы* потока).

Классификация возможных образов в соответствии со *вторым* критерием *($K_{II}$)* может состоять в следующем:

1. Внешнее отображение факта существования некоторой физической или технической сущности;
2. Отображение факта взаимодействия некоторой сущности с близким или удаленным окружением;
3. Отображение законов и принципов существования некоторой сущности как таковой.

Классификация образов по отношению к *третьему* критерию *($K_{III}$)* может иметь следующий вид:

1. Электронные каналы передачи образов;
2. Оптические средства связи;
3. Акустические каналы (дефектоскопия, передача звукового потока);
4. Другие каналы передачи образов (воздействие на иные типы сенсоров или органов восприятия).

Поскольку мы провели некоторую систематизацию образов, то целесообразно увязать соответствующие типы образов с известными подходами формального описания потока. Прежде всего, рассмотрим, как соотносятся между собой различные типы образов в рамках приведенной классификации. Для удобства в дальнейшем типы будем нумеровать следующим способом. Первой цифрой будем обозначать номер критерия классификации*($K_j$)*, а второй – номер типа образа в рамках этой классификации. К примеру, запись *Z(II.2)* будет обозначать образ второго типа по второму критерию. Поскольку образы могут представлять собой сложные типы изображений, будем применять составную запись, например:

$$Z_i = Z(II.3)\&Z(III.1).$$

Пусть эти критерии и параметры представляют собой иерархическую структуру. Параметры $P_i$ будем использовать для определения уровней иерархии, а критерии $K_j$ – для разделения образов в пределах одного *j*–уровня. Кроме того, введем представление о целях $C=\{c_y\}$, которые будем различать в соответствии с целевыми признаками $c_1, c_2,.., c_b$. Примем, что мы располагаем некоторым набором образов *{I(T)}*, которые в соответствии с *{K}* и *{P}* сформированы в некоторую структуру *G=Q(P, K)*. Естественно, что каждый образ *I(T)* нумеруется двумя индексами. Первый выбирается из *{P}*, второй – из *{K}*. Те или иные взаимоотношения в рамках структуры *G* определяются некоторой целью $c_y \in C$. Определенная совокупность *{P}*, *{K}* и *{C}* основывается на некоторой интерпретации, которая для квантов потока $T_q$ и их образов $I(T_q) \in G$ предопределяет некоторые зависимости между различными типами $\hat{I}_i(p, k, c)$, составляющие систему аксиом *U*:

$$U=L[\ddot{I}_i(p, k, c)].$$

В дальнейшем, параметрами классификации образов выберем тип, форму, вид, класс и подкласс. Классификация образов конкретного типа *($P_1$)* рассмотрена выше (критерии 1 – 3). Классификация по параметру формы *($P_2$)* структуры *G* имеет критерии органов восприятия потока:

1. Литерный поток;
2. Звуковой поток;
3. Поток изображений;
4. Иной поток.

Классификация образов по параметру вида *($P_3$)* структуры *G* должна учитывать идентификационные свойства источника потока в понятии его языка (языковая модель источника). Параметры класса и подкласса *($P_4$* и *$P_5$)* определяют свойства источника с необходимой точностью (модель композиции лексикона источника). Поскольку классификации образов по параметрам $P_3 - P_5$ имеет привязку к конкретной предметной области, обобщенная запись будет иметь для них вид ".~".

Обобщенная запись образа имеет параметрический вид критериев, разделенных точками.

В контексте установленной классификации, постановка задачи исследования будет выглядеть следующим образом. Пусть имеем квант входного символьного потока в рамках рассмотренной классификации:

$$T_q \in "Z(I.1)\&Z(II.1)\&Z(III.1).Z(1-4).\sim".$$

Необходимо квант выходного образного потока преобразовать к виду $I(T_q) \in "Z(I.4)\&Z(II.2)\&Z(III.3).Z(3).\sim"$ через промежуточную форму *Z(1)*.

## III Метод диагностики и анализа ASCII-данных

Решение этой задачи следует искать в структуре исследуемого кванта потока и основных семантических особенностях языка, представляющих собой множество грамматических правил источника потока.

Обобщенная методика построения образного представления символьного потока выглядит следующим образом:

1. Алфавитный состав источника разбивается на литерные группы согласно фонетическому наполнению (алфавитные неоднородности).
2. К группам литер применяется система правил языкового словообразования.

3. Словарный состав исследуемого потока сортируется согласно композиционным правилам построения сообщения.

Пример образа кванта потока (конкретно – данной работы) приведен на рисунке 1. В таком виде текстовый анализ далее именуется спектральным.

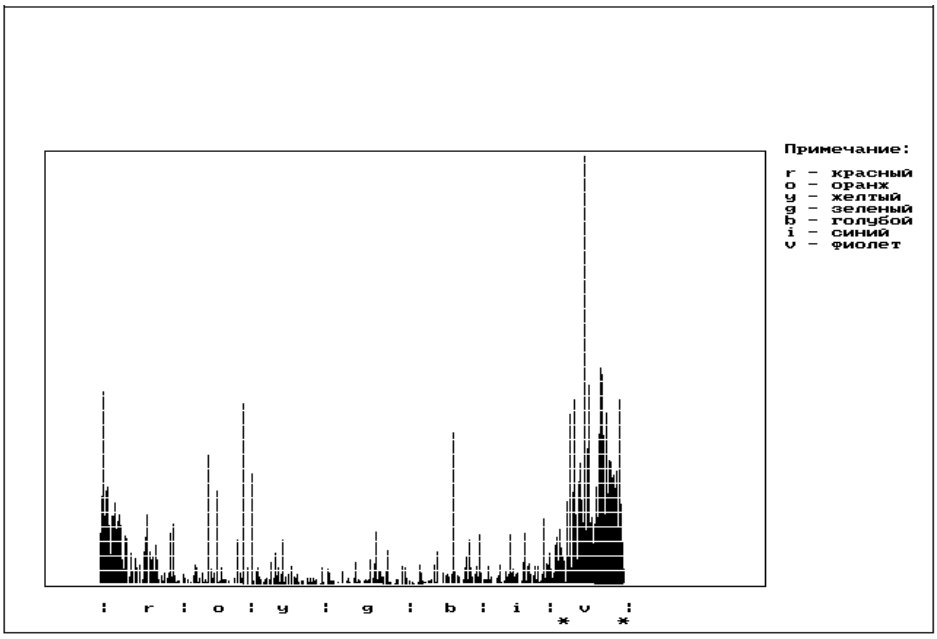


**Рисунок 1 Спектральный образ кванта потока**

Обучающая информация $I_0(K_1,..,K_L)$ в задаче кластеризации строится по принципу “обучения с учителем” и определяется введением шаблона *l*–стилистической направленности $\hat{I}_l$, $\hat{I}_l=I_0(K_l)$.

Пусть задана совокупность допустимых квантов $T_l^1,..,T_l^N$ одного *l*–класса стилистической направленности и их спектры $\{I(T_l^n)\}$, $n \in N$.

*Определение 3.* Шаблоном *l*–стилистической направленности называется спектр $\hat{I}_l$,, гармоники $\hat{g}_h$ которого являются результатом среднестатистического наложения гармоник $g_h^n$ квантов *l*–стилистической направленности:

$$\widetilde{I}_l = \bigcup_H \ddot{g}.$$

Закономерности исследуемого потока $T_q$ возможно охарактеризовать огибающей гармоник $f(T_q)$, применяя предметную область теории вероятностей и математической статистики. Закономерности изменения параметров огибающей $f(T_q)$ объективно отражают лингвистические закономерности трансформации потока $T_q$. Точная обработка символьного потока заключается в анализе огибающей спектра потока $f(T_q)$.

Степень детализации табличного спектрального прототипа $\hat{I}_l$ определяет параметр кластеризации потока *T* согласно классификации, предложенной в разделе II.

Процесс диагностики заключается в том, что задан способ определения величины близости образа и соответствующих шаблонов стилистик в виде следующего функционала, далее называемого “функционалом близости”, $B(I(T),\hat{I}_l)$:

$$B(I(T),\widetilde{I}_l) = \sum_{h=1}^{H} b_h / H,$$

$$b_h = \begin{cases} 1, & \text{если } \widetilde{g}_h * 1.682 \geq g_h \geq \widetilde{g}_h * 0.682 \\ 0, & \text{иначе.} \end{cases} \qquad (1)$$

Здесь $b_h$ – близость элементарной гармонической составляющей, представляющей результат стробирования гармоники исследуемого потока к гармонике шаблона.

Величина $B(I(T), \hat{I}_l)$ называется также оценкой по шаблону или степенью близости спектра сообщения $I(T)$ к классу *l*. Выражается в процентах.

Процесс анализа потока использует метод наибольшего правдоподобия. Для имеющегося спектра, представляющего собой выборку $I(T)=(g_1, g_2,.., g_h)$, значение функции правдоподобия $B(I(T), \hat{I}_l)$ для искомого класса $l$ выбирается таким, при котором она достигает своего максимума. Это значение является функцией $B_l$:

$$B_l = \max\left(B(I(T), \tilde{I}_l)\right) \tag{2}$$

## IV Алгоритм работы поискового сервера

В контексте введенных определений и формул (1) - (2) работа поискового сервера заключается в получении и использовании шаблонов потоков различных типовых форм согласно классификации, предложенной в разделе 2.

Такая работа была проделана и результаты сведены в таблицы 1 и 2, характеризующие формульные огибающие потоков по параметру формы и вида потока $(P_2$ и $P_3)$.

*Таблица 1*

**Формульное выражение огибающих кластеризации ASCII-потока по параметру $P_2$**

| Форма потока | Период | Число отрезков | Функция 1, 3 отрезков | Функция 2 отрезка | Интервал 2 отрезка |
|---|---|---|---|---|---|
| S | 3π | 2 | $1.3r * \sin(\frac{\pi}{5} - g_h) + 1.6r$ | $\frac{r}{2} * \sin(g_h - 0.7\pi) + 1.15r$ | >150 |
| L | 6π | 3 | $\frac{r}{30}$ | $10r * \sin(g_h - 1.8\pi)$ | 230-285 |
| V | 5π | 3 | $2r * \sin(\frac{\pi}{1.6} - g_h) + 2.6r$ | r/4 | 80-170 |

*Таблица 2*

**Формульное выражение огибающих кластеризации ASCII-потока по параметру $P_3$**

| Модель языка | Период | Число отрезков | Функция 1, 3 отрезков | Функция 2 отрезка | Интервал 2 отрезка |
|---|---|---|---|---|---|
| Русский | 6π | 3 | $1.4r * \sin(\frac{\pi}{3} - g_h) + 1.9r$ | r/2 | 50..290 |
| Украинский | 6π | 3 | $1.3r * \sin(\frac{\pi}{18} - g_h) + 1.3r$ | r/150 | 33..275 |
| Английский | 6π | 2 | $\frac{r}{1.5} * \sin(\frac{\pi}{4} - g_h) + r$ | a*x+b, a>0 | 200..360 |

Алгоритм работы поискового сервера представляет собой последовательное прохождение шагов 1-4.

Шаг 1. Формирование образа потока.

Шаг 2. Анализ близости (1) – (2) образа потока по параметру $P_2$.

Шаг 3. Анализ близости (1) – (2) образа потока по параметру $P_3$.

Шаг 4. Анализ близости (1) – (2) образа потока по параметрам $P_4$, $P_5$.

## Выводы

Поток источника сообщения характеризуется собственным лексиконом, композиционными и семантическими особенностями, накладывающими индивидуальный отпечаток на соответствующий им спектральный образ. По спектру сообщения возможно идентифицировать как модель источника (с вероятностью до 90%), так и индивидуальную принадлежность потока к источнику (до 83%).

С позиции спектрального анализа возможно ввести иерархию объектов, предсказывать развитие их характеристик, а также оценивать целевые свойства объектов на базе лингвистических особенностей языка источника потока и его преобразующих особенностей.

*Литература: 1. Осинский Л.М., Манухин А.В. Обработка текстовой информации. Портретный метод // Материалы международной научно–технической конференции "Повышение эффективности систем защиты информации". "Защита–97". – Киев: КМУГА, 1997. – с. 51–53. 2. Четвериков И.А., Манухин А.В. Спектральный анализ текстовой информации и ее кластеризация // Защита информации: Сборник научных трудов. – Киев: КМУГА, 1998. – с. 224–229.*

**УДК 621.391**

# ПРОБЛЕМИ РАХУВАННЯ КІЛЬКОСТІ ТОЧОК ЕЛІПТИЧНИХ КРИВИХ НАД ПОЛЯМИ ХАРАКТЕРИСТИКИ 2

***Олександр Теліженко***
*Київський військовий інститут управління і зв'язку*

*Анотація:* **Аналізується одна з важливих проблем створення ефективних криптографічних систем на еліптичних кривих.**
*Summary:* **Analyses one of important creation problems of effective cryptographic systems on elliptic curves.**
*Ключові слова:* **Еліптичні криві, відкриті ключі, несимитрічні криптосистеми.**

Застосування еліптичних кривих для проблем захисту інформації стає все більш поширеним у всьому світі. Це обумовлено тим, що криптографічна система, яка побудована на еліптичних кривих є системою з відкритими ключами, дає не меньшу стійкість ніж інші несиметричні криптосистеми, але має довжину ключів, що використовуються, в кілька разів менше. Найбільш ефективне використання мають криптосистеми, які створюються за допомогою еліптичних кривих над полями характеристики 2. В цьому випадку такі системи технічно реалізуються з меншими складнощами та дозволяють вирішувати питання не тільки шифрування, а й знаходження та виправлення помилок після передачі повідомлення каналом зв'язку. Для практичного застосування рекомендують не суперсингулярні криві, які мають наступний вигляд:

$$y^2+xy=x^3+a_2x^2+a_6, \quad \text{де } a_2,a_6 \in GF(2^n).$$

При цьому n обирається більш ніж 160. Звідти виникає питання підрахування кількості точок еліптичної кривої, тому як кількість елементів поля дуже велика. Відома теорема Хассе, в якій дана оцінка кількості точок $N$:

$$q + 1 - 2\sqrt{q} \le N \le q + 1 + 2\sqrt{q}\ , \text{ де } q=/GF(2^n)/.$$

Однак для максимальної стійкості криптосистеми необхідно обирати криві, для яких число, яке дорівнює кількості точок, було б максимальним, а крім того мало б великий простий дільник. Для знаходження кривих з такими властивостями необхідно знати точну кількість точок еліптичної кривої, яку ми обрали. Це дуже складна задача, однак в окремому випадку вона може бути вирішена саме для кінцевих полів.

Якщо число $n$ має вигляд $n=h*m$, тобто має два прості множники, існує рекурсивна формула, за допомогою якої можна підрахувати кількість точок еліптичної кривої над полем $GF(2^{kh})$. Для цього вводиться спеціальна функція $L(p,Z,K)=V(K)$, яка має початкові дані:

$$V(0)=2,\ V(1)=p,\ V(K)=pV(k-1)-ZV(k-2).$$

Кількість точок $R$ еліптичної кривої в цьому випадку буде дорівнювати:

$$R=2^{hm}+1-L(2^h-(N-1),2^h,K).$$

Значення $Z$ залежить від кількості точок еліптичної кривої над полем $GF(2^h)$ [1].

Вибір значення $n$ повинен задовольняти вимозі існування в кінцевому полі оптимального нормального базису для спрощення арифметичних операцій [2]. Мінімальне таке значення, яке задовольняє всім вимогам, буде 183=3*61. Крім того, існують складнощі в випадковому виборі самої еліптичної кривої, тобто вибір коефіцієнтів $a_2$ та $a_6$ таким чином, щоб кількість точок еліптичної кривої була максимальною, а для ефективного використання криптосистеми для полів характеристики 2 потрібно щоб:

$$R=2^s qb,$$

де $b$ - велике просте число, $q$ - невеликий множник.