Синтактически инвариантный метод идентификации семантики информации

Syntactically the invariant method of identification of semantics of the information

Потапов М. В. (potapov_mv@rgrta.ryazan.ru)

Государственное образовательное учреждение высшего профессионального образования «Рязанский государственный радиотехнический университет», Рязань, Россия

Содержится описание практически апробированного метода оценки смыслового содержания информационных потоков, основанного на статистико-лингвистическом способе их представления и обработки с использованием подходов теории распознавания образов при анализе многомерных признаков.

Для оценки состояния контролируемых на удалении сложных технических комплексов актуальна задача распознавания и идентификации генерируемых информационных сообщений. Для этого предлагается использовать текстовое представление информационных потоков:

$$\vartheta = (A_X, \Theta), \quad (1)$$

где A_X — алфавит сообщений, Θ — отношения между элементами текста. Текст Θ — это множество элементарных знаков, между которыми установлены отношения $\Theta=\{\Theta_1, \Theta_2, \Theta_3\}$, определяемые правилами функционирования объектов, генерирующих информацию, принятой системой интерпретации и целями исследований. Синтактика Θ_1 связывает с текстом некоторую структуру отношений между знаками независимо от их содержания. Отношения между объектами и их обозначениями рассматривают семантика Θ_2 и прагматика Θ_3. Каждое из слов текста ϑ состоит из символов A_X. Пусть все слова текста ϑ образуют множество:

$$X_\Gamma(t) = \{x_1, x_2, ..., x_{N_\Gamma}\}, \quad (2)$$

где $N_\Gamma=\mathrm{Card}(X_\Gamma)$, а различные словоформы (2) образуют множество (алфавит):

$$Ax = \{a_1, a_2, ..., a_n\},\ a_i \neq a_j,\ i, j \in \{1, 2, 3, ..., n\}, \quad (3)$$

где $n = \mathrm{Card}(A_X)$ — мощность алфавита, причём $A_X \subset X_\Gamma(t)$.

Выбор типа словаря (3) зависит от особенностей формирования и дальнейшего использования информационного процесса (2). Место каждого знака a_i внутри каждого слова определяется смысловым содержанием и моделью функционирования объекта, порождающего эту информацию. Когда известна структура (синтактика), режимы и программы функционирования объекта (семантика и прагматика), анализ смыслового содержания состоит в оперативной оценке Θ знаковой системы (1). Это возможно при отсутствии помех естественного и искусственного характера, кодирующих преобразований, которые, разрушая структуру передаваемых данных, приводят к неоднозначности (1) — к потере их смыслового содержания. В этих условиях качественное и эффективное решение рассматриваемой задачи обеспечивает статистико-лингвистический подход [1, 2]. В качестве элементов (словоформ) алфавита (3) выступают слова сообщений:

$$A_X = \{a_1=x_1, a_2=x_2, a_3=x_3, \dots\} \quad (4)$$

В условиях неизвестной структуры сообщений, в качестве словоформ предлагается использовать длины блоков одноименных, подряд следующих одноимённых кортежей “0” и “1”:

$$A_X = \{a_1=\{0, 1\}; a_2=\{00, 11\}; \\ a_3 = \{000, 111\}; \dots; a_n=\{0 \dots 0,1 \dots 1\}\} \quad (5)$$

Установлено, что осмысленные, информационно наполненные сообщения, содержат достаточно длинные кортежи $n>>30$. Это позволяет идентифицировать эти сообщения на фоне помех.

Предлагаемый метод сводится к последовательности процедур. Вначале, по словарю A_X (4) или

(5) строится эмпирический ряд частот появления словоформ:

$$f(x) = \{f(x)_i = Q^a_i / N_\Gamma,\ i=1, 2, 3, \ldots, n\}, \quad (6)$$

где Q^a_i — число вхождений слов $x \in A_X$ в $X_\Gamma(t)$; N_Γ — объём выборки.

Отметим, что для неискажённых, семантически нагруженных данных распределение (6) является негауссовым. Ранговым распределением назовём функцию $\Phi(r)$, которая ставит в соответствие номеру (рангу) r слова $x \in A_X$ частоту его появления f(x):

$$\Phi(r) = \{\Phi(r)_i \leftrightarrow f(x)_i,\ \Phi(r)_i > \Phi(r)_i + 1,\ i = 1, 2, 3, \ldots, n-1\} \quad (7)$$

Эмпирический ряд (7) прологарифмируем и аппроксимируем модифицированной зависимостью Ципфа-Мандельброта [1, 2] вида:

$$\Phi(r) = C_n r^{-\gamma_o \mathrm{Exp}(dr)}, \quad (8)$$

где C_n — константа, зависящая от n, γ_0 — начальный показатель рангового распределения $\gamma(r)=\gamma_o \mathrm{Exp}(dr)$, d — коэффициент его прироста. Их начальные значения определим как:

$$\begin{aligned} &C_n = \mathrm{MAX}\,\Phi(r) = \Phi(0);\\ &\gamma_r = (\mathrm{Lg}(\Phi_{max}) - \mathrm{Lg}(\Phi_r)) / \mathrm{Lg}(r)\ ;\\ &\gamma_0 = \gamma_r / \mathrm{Exp}(dr)\ ;\\ &d_{ij} = (\mathrm{Ln}(\gamma_i) - \mathrm{Ln}(\gamma_j)) / (r_i - r_j);\ i \neq j. \end{aligned} \quad (9)$$

Первичная аппроксимация (8–9) в последующем улучшается последовательно-итерационным подбором коэффициентов C_n, γ_0 , d . Критерием наилучшего приближения является минимум суммы квадратов разностей эмпирического $\Phi^Э(r)$ и аппроксимирующей зависимости теоретического $\Phi^Т(r)$ законов распределения:

$$S_\phi = 1/n \sum_{i=1}^{n} (\Phi^Э(r)_j - \Phi^Т(r)_j)^2 \quad (10)$$

При достижении Sф заданного порогового уровня $S^П_ф$ по критерию

$$\lim_{n\to\infty} S_\phi \to S^П_\phi \quad (11)$$

фиксируются — записываются в базу данных коэффициенты модели (8–9) и формируется точка в признаковом пространстве Далее проводится проверка гипотезы о принадлежности анализируемой информации к одной из эталонных $U_{\gamma d,\, i}$ с использованием критерия «минимаксного» расстояния, по результату которого либо дополняется эталонная база, либо отвергается выдвинутая гипотеза. В первом случае по расстоянию между полученными координатами (C_n, γ_0, d) и центром «тяжести» $Z_{u,\, i}$ наиболее близкого кластера $U_{\gamma d,\, i}$ оценивается мера качества, например отношение сигнал/шум:

$$\mathrm{MIN}_i(R_i = |(\gamma_0, d) - Z_{u,\, i}|). \quad (12)$$

Проведённые исследования показали, что модель (8–9) является точной и хорошо отражает закономерности эмпирического ряда в выборках различного объёма (рис. 1) разнородных информационных сообщений. Выпуклые области изменений коэффициентов γ_0 и d, построенные по интервальным оценкам при $P_Д = 0{,}95$ и фиксированном объёме выборки 20 Кбайт. Области возможных изменений γ_0 и d представляют статистические образы различных информационных процессов и могут использоваться для их идентификации.

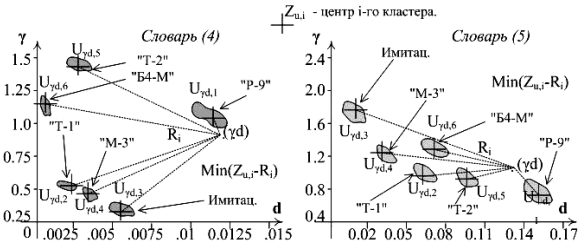

Рис. 1. Области коэффициентов аппроксимации ранговых распределений

Искажение шумами, шифрование, изменение смыслового содержания в целом или какой-либо его части приводят к изменению соответствий (8–9) и, как следствие, к трансформированию частотно-рангового распределения и увеличению размеров областей изменения коэффициентов γ_0 и d, что подтверждается результатами проведённого моделирования (рис. 2). Таким образом, можно выделить области $U_{\gamma d}=\{\gamma_{min}<\gamma_0<\gamma_{max};\ d_{min}<d<d_{max}\}$ для сообщений с нарушенной структурой сообщений $U^Ш_{\gamma d}$ и области $U^Э_{\gamma d}$, характерные для смысловой информации. Используя эти свойства и зная $U_{\gamma d,\, i} = \{U^Ш_{\gamma d}, i, U^Э_{\gamma d}, i\}$ словарей (4), (5) для каждого i-го типа данных можно строить алгоритмы оценки качества, отбраковки недостоверных участков, оценки отношения сигнал/шум идентификации сообщений и др.

Для сопоставления (идентификации) процессов предлагается использовать меры, предоставляемые теорией распознавания образов, рассматривая в качестве признаков коэффициенты γ и d. Статистико-лингвистический способ оценки достоверности и смыслового содержания информации с использованием геометрической интерпретации, кластеризации объектов, получения решающих функций с учётом весов и мер расстояний, устранения незначимых параметров и уменьшения размерности признакового пространства практически апробирован на разнообразных данных. Результаты быстрой и надёжной кластериза-

ции при наполнении эталонной базы и последующее гарантированное распознавание информации позволяют в качестве меры качества использовать расстояние между оценками (γ_0, d) и центрами «тяжести» $Z_{u,i}$ кластеров $U_{\gamma d,i}$ в признаковом пространстве γ, d.

При аппроксимации частотно ранговых распределений некоторых информационно-выраженных процессов зависимостью (8) был выявлен эффект получения неудовлетворительной оценки ($R \to max$) при $\Phi(r \to max) = 0$, $n_i < n_{max}$. Положительный эффект даёт исключения из анализа элементов $\Phi(r) = 0$, а также более точный подбор коэффициентов апроксимирующей зависимости γ_0 и d. Отмечено также, что для выполнения надёжного оценивания, необходимо выбирать одинаковые тип словаря (4) или (5), $N_Г$, $R_{i\,зад}$, $S^{П}_{ф} < 1\%$, $P_Д$.

Для гарантированной идентификации в алгоритм включены процедуры получения оценок тяжести «хвостов» [3], а также анализа законов распределения и статистических характеристик коэффициентов частотно-ранговых распределения (8). Кроме этого, дополнительными идентифицирующими критериями могут быть использованы факты выхода элементов частотно-рангового распределения анализируемого процесса из статистически накопленных эталонных допусковых коридоров разброса (6–12). Статистико-лингвистический алгоритм позволяет получать оценки качества при анализе как перекрывающихся, так и не перекрывающихся выборок из информационного потока (2).

Частотно-ранговые распределения (6–13) являются характеристиками неискажённого сигнала и сигнала с разрушенной информационной структурой кодирующим псевдослучайным преобразованием или шумами (рис. 2). Эти распределения и отмеченные закономерности их параметров являются важной характеристикой, отражающей статистические и структурные свойства анализируемой информации.

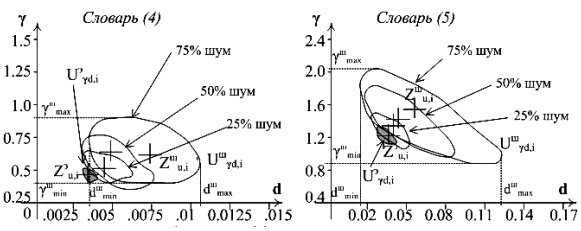

Рис. 2. Изменения областей коэффициентов аппроксимации ранговых распределений при зашумлении

На рис. 3 приведены результаты обработки, представляющие собой аппроксимацию на основе зависимости (8) эмпирического ряда распределений для различных по информативности источников информации.

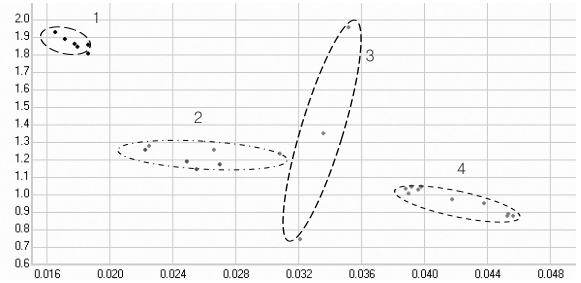

Рис. 3. Области возможных вариаций коэффициентов аппроксимации γ_0 и d

Области возможных вариаций коэффициентов аппроксимации γ_0 и *d* на рис. 3 подразделяются на четыре группы систем: 1 — малоинформативные (50 кбит/с), 2 — среднеинформативные (220 кбит/с), 3 и 4 — высокоинформативные (750, 2300 кбит/с). Решение о принадлежности значений {γ_0, d}тому или иному типу систем принимается на основании критерия качества

$$F = \ln (d / (c \cdot \lambda)), \qquad (14)$$

где c — расстояние между значениями {γ_0 , d} внутри группы, d — расстояние между значениями {γ_0, d} разных групп, λ — мера «одинаковости структуры» групп.

Точка со значениями {γ_0, d} относится к тому типу систем, где максимально значение критерия качества F. Частоты встречаемости словоформ для информации каждой из групп занимают устойчивые положения. Их можно использовать в качестве статистико-лингвистических образов систем, порождающих эту информацию, они характеризуют качество проектирования структур передаваемых данных телеизмерений. По тенденциям изменения коэффициентов γ_0 и d, используя свойство стабильности проявления закона распределения словоформ, можно идентифицировать характерные участки жизненного цикла одной и той же системы (рис. 4).

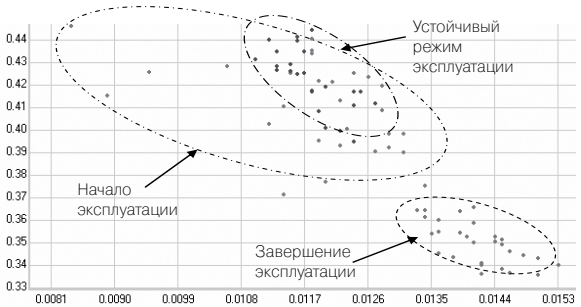

Рис. 4. Результаты анализа словоформ {0–1} словарь (5)

Использование предлагаемого подхода позволяет решить задачу классификации без выделения конкретных элементов (параметров), идентифицировать информационные фрагменты. Перспек-

тивным прикладным направлением полученных результатов является оценивание достоверности измерительной информации в нештатных, аварийных ситуациях и при наличии шумов и сбоев. Достоинства предлагаемого метода: работоспособность в условиях семантической неопределённости; синтаксическая инвариантность; снижение размерности анализируемых признаков; высокое быстродействие; возможность автоматизации процесса анализа; инвариантность к фазовым характеристикам анализируемых данных; высокая чувствительность; эффективная идентификация аномальных явлений в потоке данных; точность анализа вплоть до бита.

Предлагаемый метод статистико-лингвистического и графического анализа приводит к порождению новой информации о характере анализируемых данных — знаниепорождающей (когнитивной) графики [4]. Он практически апробирован на реальной разнородной информации и дал положительные результаты. Этот метод использован для построения высокоразвитых комплексов средств автоматизации распределённой автоматизированной системы, реализующей сквозную технологию регистрации, обработки, анализа, представления, хранения, тиражирования и архивирования потоковой циклической информации [5, 6].

Литература

1. *Мандельброт Б.* Теория информации и психолингвистика// Математические модели в социальных науках. — М.: Наука, 1973. С. 316–322.
2. *Орлов Ю. К.* Обобщенный закон Ципфа—Мандельброта и частотные структуры информационных единиц различных уровней// Вычислительная лингвистика. — М.: Наука, С. 179–194.
3. *Кукушкин С. С., Потапов М. В.* Статистико-лингвистические методы оценки смыслового содержания информации. / Проектирование ЭВМ. Межвузовский сборник научных трудов. Рязань, 1994. 112 с.
4. *Потапов М. В.* Когнитивная графика в задачах анализа телеметрической информации. 12-я международная научно-техническая конференция «Проблемы передачи и обработки информации в сетях и системах телекоммуникаций»: Тез. докл. /Рязан. гос. радиотехн. акад. Рязань, 2004. -180 с. ISBN 5-7722-0209-X.
5. *Везенов В. И., Гусев И. А., Потапов М. В., Селецкий О. Б., Юдин В. И.* Комплекс средств автоматизации ввода, регистрации, обработки, анализа и представления телеметрической информации. 3-я международная научно-техническая конференция «Космонавтика. Радиоэлектроника. Геоинформатика»: Тез. докл. /Рязан. гос. радиотехн. акад. Рязань, 2000. — 358с. ISBN 5-7722-0147-6.
6. *Потапов М. В., Кузин В. А.* Анализ характеристик телеметрической информации в условиях недостаточных априорных сведений. 4-я международная научно-техническая конференция «Космонавтика. Радиоэлектроника. Геоинформатика»: Тез. докл. /Рязан. гос. радиотехн. акад. Рязань, 2003. — 360с.