DOI: 10.25728/mlsd.2022.1276

МОНИТОРИНГ ЭНТРОПИЙНОГО АНАЛИЗА ЛИНГВИСТИЧЕСКОЙ ИНФОРМАЦИИ В КОМПЬЮТЕРНОЙ СЕТИ

Полтавский А.В., Русяева Е.Ю.
Институт проблем управления им. В.А. Трапезникова РАН
Россия, г. Москва ул. Профсоюзная д.65
1779624@mail.ru, rusyaeva@ipu.ru, avp57avp@yandex.ru

Аннотация. Представлены энтропийный анализ знаковых систем для формирования информационно-аналитических проектов и формализованная модель лингвистического анализа текстов. Приводится алгоритм по оценке качества преобразования информации в компьютерной сети. Получен технологический и технический результат на основе вычислительного эксперимента, включения в информационный процесс сетевых вычислительных ресурсов.

Ключевые слова: конструкция, язык, вычислительная система, алгоритмы, информационная модель, связь, энтропия, символ, лингвистика, идентификация.

Введение

В процессе эволюционного развития не просто повышается уровень информатизации человеческого общества, а создается как бы иная виртуальная (цифровая) реальность. В сложной нынешней ситуации в России и мире, в условиях нового геополитического передела, смены расстановки сил в мировом противостоянии систем, внимание к совершенствованию информационно-управленческих многократно увеличивается. Теме развития цифровых технологий уделяется большое внимание в управлении крупномасштабными системами. Востребованы в России именно отечественные (импортозамещение) теоретические и прикладные программно-аппаратные средства для сетевых вычислительных систем (ВС), телекоммуникационных и информационных технологий. Они создаются в том числе и в целях обеспечения требуемых (заданных или желаемых) траекторий образования для нового поколения работников ИТ-сферы.

Цифровые инновационные проекты и технологии, направленные на создание автоматизированных процедур с разработкой новых адаптивных алгоритмов для сетевых информационных систем (ИС) и информационно-аналитических систем (ИАС), которые, как правило, должны учитывать неопределенность - энтропийные составляющие для передачи и обработки информации по различным каналам связи. Данное обстоятельство, а также и сам такой случайный (стохастический) процесс непосредственно связаны с потерей качества передачи информации по каналу связи и управления в сетевых ИС и ИАС, которая может быть утрачена, частично утеряна или преобразована (получила недопустимое ее искажение) с большими ошибками [1,2]. Проблема, которая требует решения, кроется, на наш взгляд, в нарастающей актуальности анализа лингвистической информации, идентификации перевода оригиналу, мониторингу подобия и соответствия, адекватности и полноте данных. На поиск решения этих задач и направлено данное исследование. Выработан подход к энтропийному анализу знаковых систем и символьной информации в сетевых структурах для формирования современных информационно-аналитических проектов. Создан алгоритм по оценке качества преобразования информации в компьютерной сети на основе основных положений и понятия энтропии по теоремам Кл. Шеннона [2] и формализованная модель энтропийного анализа. Для примера приводится информационное моделирование и анализ знаковой системы в компьютерной сети были следующие: получить характеристики компьютерного контент-анализа из известных песен китайского исполнителя Lu Han [3-8], раскрыть семантику для контента произведений на примере анализа иероглифов и получить сравнительные оценки характеристик данных произведений по признакам.

1 Краткий экскурс в историю проблемы

Большинство разрабатываемых и вновь создаваемых сетевых объектов в ИС и ИАС, исторически и изначально ориентированы на математическую обработку информации в символьной форме. Особенным в данном процессе является то, что для обеспечения желаемого уровня качества обработки информации и идентификации текстовых сообщений, является преобразование данных с помощью программ текстовых переводчиков с одного языка на другой. Методы и модели, алгоритмы и программы такого преобразования являются достижением для всей вычислительной техники. В процессе эволюции они впервые появились в Нью-Йорке (в 1954 г.), когда была проведена первая публичная демонстрация перевода с русского языка на английский с помощью первых

вычислительных систем, в частности, с помощью цифровой электронной вычислительной машины ИБМ-701. Для проведения испытаний (постановки вычислительного эксперимента) автоматического перевода текста был подготовлен словарь объемом в 250 русских слов, часто применяемых в области политики, юстиции и математики, записанных латинскими буквами. Слова для их математической обработки в ЭВМ были подобраны так, чтобы каждое русское слово имело один или два английских эквивалента. Каждому слову из разработанной программы были приписаны в электронном словаре три цифровых дополнительных кода. Программа машинного перевода символов и слов подготовленного текста, для первых цифровых ЭВМ ИБМ-701 содержала всего 2400 одноадресных команд, была затем введена непосредственно в электронную цифровую вычислительную машину [9-11]. Далее, после этого, в цифровую электронную вычислительную машину вводились слова выбранных русских предложений, непосредственно подлежащие переводу в ЭВМ. Предложения из текста пробивались непосредственно на перфокартах стандартным для машины цифровым кодом, а для ввода переводимого текста в цифровую вычислительную машину каждая буква из латинского алфавита заменялась определенным набором цифр. При этом каждое английское слово заменялось соответствующим ему условным числом (цифровым кодом). Следует отметить, что и аналогичные работы в области машинных переводов прикладной математической лингвистики велись и в нашей стране. В те, условно далекие, годы в Институте точной механики и вычислительной техники АН СССР при составлении алгоритма и программы для перевода с английского языка на русский, произведенного на вычислительной машине БЭСМ-1, был избран иной путь, состоящий в воспроизведении работы, выполняемой непосредственно самим переводчиком текста. Эта «лингвистическая» работа по переводу текстовой информации с помощью компьютера составила следующие этапы:

1. чтение английской фразы, подлежащей машинному переводу в цифровой ЭВМ;
2. выявление тех слов переводимой фразы, которые были знакомы переводчику.

Выяснение некоторых грамматических признаков этих слов как по их окончаниям, так и путем сопоставления их друг с другом и с остальными словами фразы.

Для опытов автоматического перевода на вычислительной машине БЭСМ-1 был составлен словарь из 952 английских и 1073 русских слов. В общую программу входили подпрограммы «синтаксиса» и «изменение порядка слов». Первая из них расставляла знаки препинания, а вторая изменяла в русской фразе расположение слов непосредственно по правилам русской грамматики. Работа цифровых вычислительных систем и ЭВМ первых поколений в условиях неопределенности внешней среды, а также вероятностного характера математической обработки информации (сбои, внутренние и внешние помехи ЭВМ влияют на точность проводимых вычислений) порождала сложности. С развитием средств телекоммуникаций ВС и алгоритмов программных средств идентификации текстовых сообщений преобразуются и методы их создания. Современные программы-переводчики глобальных, региональных и локальных ИС имеют широкие возможности в оценках показателей качества передаваемой информации, в том числе, и с учетом ее энтропийного анализа. Методы теории вероятностей и формулы математической статистики все чаще применяется в компьютерной лингвистике. Один из подходов для системного анализа был основан на общем понятии об энтропии информационного процесса и теореме Кл. Шеннона [2]. Аналогичный подход в оценках вероятностных характеристик и получения количественных значений энтропии для анализа потоков (из корпусов) символьной информации приводится и нашими отечественными учеными: академиками В.С. Пугачевым [1], А.А. Дородницыным, Л.Н. Столяровым и др.

2 Основы информационной модели и энтропийного анализа корпусов текста в сети

Ранее известно, что основоположник теории информации Клод Шеннон определил синтаксическую меру информации. Им было показано, что объем данных $V_д$ в сообщении измеряется количеством символов в текстах. В различных системах счисления один разряд имеет свой различный вес и, соответственно, меняется и сама единица измерения информации [1, 2, 11]:

- в двоичной системе счисления единица измерения – бит (*bit – binary digit* – двоичный разряд (код одного символа в памяти машины занимает 1 байт);
- в десятичной СС единица измерения–дит (*dit – decimal digit* – дес. разряд).

Количество информации (энтропию) на синтаксическом уровне невозможно определить без рассмотрения понятия о неопределенности состояния системы. Сам термин «энтропия» используется Кл. Шенноном по совету фон Неймана. Получение информации о какой-либо системе

(или процессе) всегда будет связано с изменением степени неосведомленности получателя о состоянии этой системы (и/или протекающего информационного процесса). Покажем данное направление и задачи информационной технологии обработки данных. Пусть до получения информации ее потребитель имеет предварительные (как априорные) сведения о системе (или процессе) α. Мерой его неосведомленности о системе [2] является некоторая функция H(α), которая служит мерой о неопределенности состояния системы. После получения некоторого сообщения β его получатель приобрел дополнительную информацию Hβ(α), уменьшившую априорную неопределенность так, что апостериорная (после получения сообщения β) неопределенность состояния стала Iβ(α). Тогда количество информации Iβ(α) о системе, полученной в сообщении β, определится из формулы для разности $H_\beta(\alpha) = H(\alpha) - H_\beta(\alpha)$, т. е. количество информации измеряется уменьшением неопределенности о состоянии системы (информационного процесса). Информация – это противоположность неопределенности. Энтропия (как неопределенность) системы H(α), имеющая N состояний, может рассматриваться как мера недостающей информации по формуле Кл. Шеннона. Теоретические основы по разработке алгоритмов в идентификации текстовых сообщений для сетевой информационной системы базируются на их моделях [1,2]. Подробная формализация разработанной информационной модели и используемого интегративного метода представлена в [11, 12 и 13]. Здесь укажем лишь параметры вычисления динамики, влияния скорости передачи сообщений на адекватную идентификацию символов.

Энтропия Hy(x) в ИС характеризует среднюю неопределенность принимаемых сообщений или потерю информации, вызванную наличием ошибок (учитываемые помехи сетевой ИС). При полном отсутствии ошибок передачи сообщения из множества символов вероятность $P(i,j)= 0$,
тогда

$$H_\gamma(x) = -\sum_{(i,j)} P(i,j)\log_2 P_i(i) = 0$$

и энтропия характеризует, что соблюдается условное «идеальное равновесие» и равенство как H(x,y) = H(y) = H(x). Источники текстовых сообщений, у которых отсутствует коррелятивная связь, называют эргодическими, а выдаваемые ими последовательности (в виде множества) символов называют эргодическими последовательностями. Для эргодического источника сообщений существует конечное число состояний, в которых он может находится, причем условная вероятность появления очередного символа зависит от того, в каком состоянии находится в этот момент источник. Кроме понятия энтропии на символ текста в ИС имеет место и понятие поток информации – это скорость сообщений как энтропия источника, приходящаяся на единицу времени [14]

$$H'(x) = \frac{H(x)}{\bar{\tau}}, \quad (1)$$

где $\bar{\tau}$ - средняя длительность символа в секундах.

3 Информационное моделирование задач энтропийного анализа текстов

В сформированной компьютеризированной экспертной (информационной) системе проводился сравнительный контент-анализ известных песен ***Lu Han*** [3-8]:

- Lu Han - Dream With A Childlike Heart (追梦赤子心);
- Lu Han - Promises (諾言);
- Lu Han - 勋章 (Medals);
- Lu Han - That Good Good (有点儿意思);
- Lu Han - On Fire (零界點).

Задачами информационного моделирования и анализа знаковой системы в компьютерной сети были следующие: получить характеристики компьютерного контент-анализа из известных песен исполнителя Lu Han, раскрыть семантику для контента произведений на примере анализа иероглифов и получить сравнительные оценки характеристик данных произведений по признакам (местоимения 我 你; глагол 是; служебные частицы 不, 的).

Далее представлены в табличном виде результаты информационного моделирования данных энтропийного и контент-анализа решаемых задач в созданной ИС.

Таблица 1. Энтропийные характеристики из анализа песни Lu Han On Fire (零界點)

	Иероглиф	Кол-во	Pi	Pi*log(Pi;2)
1	我	9	0,01848	0,106408
2	的	10	0,020534	0,11511
3	不	3	0,00616	0,045233
4	是	7	0,014374	0,087973
5	你	1	0,002053	0,018332

Таблица 2. Энтропийные характеристики из анализа песни Lu Han That Good Good (有点儿意思)

	Иероглиф	Кол-во	Pi	Pi*log(Pi;2)
1	我	16	0,023739	0,128109
2	的	16	0,023739	0,128109
3	不	10	0,014837	0,090129
4	是	5	0,007418	0,052483
5	你	17	0,025223	0,13391

Таблица 3. Энтропийные характеристики из анализа песни Lu Han 勋章 (Medals)

	Иероглиф	Кол-во	Pi	Pi*log(Pi;2)
1	我	27	0,074176	0,278375
2	的	27	0,074176	0,278375
3	不	3	0,008242	0,057056
4	是	14	0,038462	0,180786
5	你	1	0,002747	0,023373

Таблица 4. Энтропийные характеристики из анализа песни Lu Han Promises (諾言)

	Иероглиф	Кол-во	Pi	Pi*log(Pi;2)
1	我	16	0,046784	0,206683
2	的	17	0,049708	0,215253
3	不	7	0,020468	0,114835
4	是	1	0,002924	0,024614
5	你	16	0,046784	0,206683

Таблица 5. Энтропийные характеристики песни Lu Han Dream With A Childlike Heart (追梦赤子心)

	Иероглиф	Кол-во	Pi	Pi*log(Pi;2)
1	我	14	0,031461	0,156998
2	的	17	0,038202	0,17994
3	不	14	0,031461	0,156998
4	是	3	0,006742	0,048625
5	你	0	0	0

Так выглядит сравнение спектральных характеристик песен между собой в процентном соотношении.

Заключение

Итак, задачи нынешнего этапа развития отечественных информационных систем обращены на новый геополитический контекст, ориентированы на евразийский ареал и знаменуют новый этап развития. В свете этого анализ текстовой переводной информации китайского языка, энтропийный анализ китайских иероглифов как символьной информации актуален как никогда ранее. Сейчас есть множество отечественных прикладных программ и определенный арсенал аппаратно-технических

средств для реализации инновационных замыслов по установлению актуальных лингвистических классификаций [15], мультимедийных связей, создания новых информационно-лингвистических [15], логистических и пр. цепочек коммуницирования. В этом ключе особенно важно наладить адекватное взаимодействие с восточно-азиатского ареала, для чего необходимо наладить канал лингвистического, языкового взаимодействия, текстового понимания.

Естественно, что такой информационный аспект и процесс, не мог не коснутся образовательной среды. Информатизация образования – это непрерывный управляемый процесс обеспечения системы образования методами, моделями и средствами современных информационных технологий. Технологии и эволюция [9, 10] применяемых в данными процессе различных компьютеризированных ИС (мониторинговых, фактографических, документальных, экспертных, информационно-аналитических и др.) направлена на широкий охват средств и методов информационного управления в обществе. При этом, защита информации в компьютерной сети ИС и ИАС – это вынужденные меры, направленные против несанкционированного доступа, прежде всего, к данным, хранящейся в памяти компьютера. Одним из основных способов защиты данных, хранящихся в сетевых ВС, является использование символов для паролей, которые в итоге, имеют цифровой код. Эффективными методами защиты информации в сетевых ВС являются методы, основанные на различных подходах криптографии, они включают комплекс алгоритмов преобразования информации, обеспечивающие скрытность из смыслового содержания данных. К защите информации можно также отнести организацию учета потери информации в процессе ее преобразования и передачи по каналам сетевой ВС. Современные телекоммуникационные и программные средства ИС и ИАС для текстовых переводов с одного языка на другой имеют различный уровень по точности [9]. Работы в этом направлении и над «приемлемыми» в сети методами, моделями и алгоритмами для их идентификации – актуальная проблема информационного обеспечения в достижении желаемого уровня и траектории для решения поставленных задач по обработке потоков из корпусов лингвистической информации.

Литература

1. *Пугачев В.С.* Теория случайных функций и ее применение к задачам автоматического управления. М., «Наука», 1962.
2. *Шеннон К.Э.* Работы по теории информации и кибернетике (с предисловием академика А.Н.Колмогорова). Издательство иностранной литературы, 1963 г. – М.: – С. 824.
3. Текст песни Dream With A Childlike Heart (追梦赤子心) Источник: https://lyricstranslate.com/ru/luhan-追梦赤子心-sky-hunter-ost-lyrics.html
4. Текст песни Promises (諾言) Источник: https://lyricstranslate.com/ru/luhan-promises-諾言-lyrics.html
5. Текст песни 勋章 (Medals) Источник: https://lyricstranslate.com/ru/lu-han-鹿晗-lyrics.html-1
6. Текст песни That Good Good (有点儿意思) Источник: https://lyricstranslate.com/ru/ lu-han-good-good-lyrics.html-0
7. Текст песни On Fire (零界點) Источник: https://lyricstranslate.com/ru/luhan-fire-零界點-lyrics.html
8. Лу Хань Источник: https://ru.wikipedia.org/wiki/Лу_Хань
9. *Полтавский А.В.* Программные средства вычислительных систем. Часть I. ЭВМ первых поколений. Учебное пособие. – М.: МГПУ, 2014 – 87 с.
10. *Полтавский А.В.* Программные средства вычислительных систем. Часть II. ЭВМ третьего и четвертого поколений. Учебное пособие. – М.: МГПУ, 2016 – 96 с.
11. *Полтавский А.В.* Основы математической обработки информации вычислительных систем. Учебное пособие. – М.: МГПУ, 2017 – 97 с.
12. *Полтавский А.В., Русяева Е.Ю.* Интегративный подход к построению информационной системы мониторинга для комплексов беспилотных летательных аппаратов / Труды 14-й Международной конференции «Управление развитием крупномасштабных систем» (MLSD-2021). М.: ИПУ РАН, 2021. С. 1670-1677
13. *Полтавский А.В., Русяева Е.Ю., Бурба А.А.* Устройство для содержательного анализа текстовой информации: Патент на изобретение № 2568272 РФ; Дата публикации 27.10.2015. Бюл. №30
14. *Полтавский А.В.* Информационная модель случайного процесса // Информационные войны. 2018. №3 (47). С. 98-101.
15. *Батура. Т.В.* Методы автоматической классификации текстов // Программные продукты и системы / Software & Systems.2017. Т. 30. № 1. С. 85–99.