ТЕМА

ПРЕДСТАВЛЕНИЕ ЗНАНИЙ В ИНТЕЛЛЕКТУАЛЬНЫХ СИСТЕМАХ

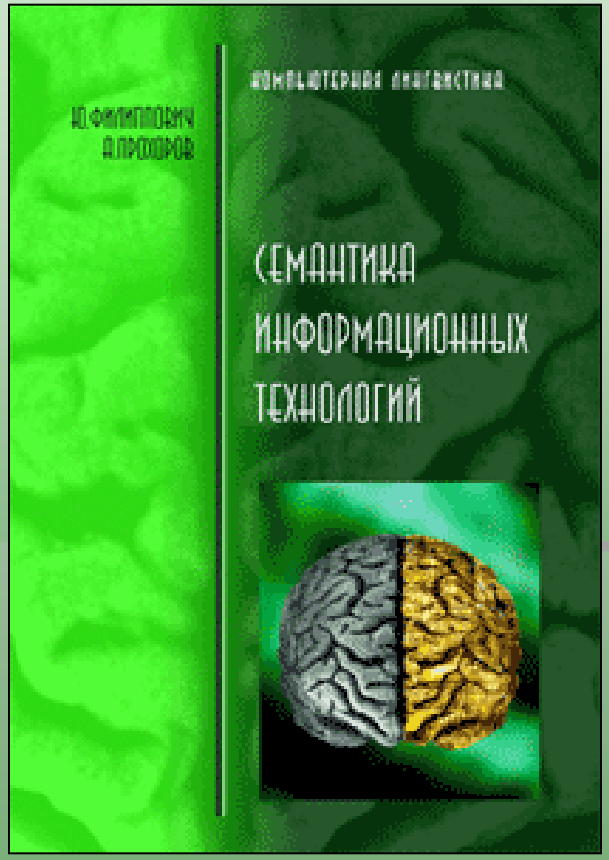
ОСНОВНЫЕ РАЗДЕЛЫ ТЕМЫ

1. ***Формализация знаний в интеллектуальных системах.***
2. ***Количественная спецификация ЕЯ систем.***
3. ***Логико-статистические методы извлечения знаний.***
4. ***Формально-логические модели .***
5. ***Продукционные модели .***
6. ***Сетевые модели***

2. КОЛИЧЕСТВЕННАЯ СПЕЦИФИКАЦИЯ ЕСТЕСТВЕННО-ЯЗЫКОВЫХ СИСТЕМ

- Статистический анализ ЕЯ описания.
- Модель «ранг-частота».
- Закон Ципфа.
- Формула Мандельброта.
- Построение ядра ЕЯ описания.

Литература



Материал лекции представлен в книге:

Ю.Н.Филиппович, А.В.Прохоров.
Семантика информационных технологий: опыты словарно-тезаурусного описания. /

Серия «Компьютерная лингвистика».
Вступ. Статья А.И.Новикова.
М.: МГУП, 2002.
— книга в комплекте с CD ROM
— С. 34–45.

СТАТИСТИЧЕСКИЙ АНАЛИЗ ЕСТЕСТВЕННО-ЯЗЫКОВОГО ОПИСАНИЯ

Лингвистическая статистика, лингвостатистика

— раздел языкознания, занимающийся исследованиями статистическими методами количественных закономерностей в языке и речи.

↓

Энциклопедия «Русский язык»

(1) в широком смысле — область применения статистических методов в языкознании (то есть опирающаяся на математическую статистику подсчетов и измерений при изучении языка и речи);

(2) в узком смысле — изучение некоторых математических проблем, связанных с лингвистическим материалом, главным образом с типами статистических распределений языковых единиц в тексте.

ПОНЯТИЯ ЛИНГВИСТИЧЕСКОЙ СТАТИСТИКИ

ТЕКСТ

↓

последовательность лингвистических единиц:

букв, морфем, словоформ, словосочетаний, предложений и др.

количественные характеристики лингвистических форм: употребительность, совместная встречаемость, законы распределения в тексте, их физические размеры.

↓

ОСНОВНЫЕ ПОНЯТИЯ И КАТЕГОРИИ ЛИНГВОСТАТИСТИКИ:

генеральная совокупность, выборки, частоты и вероятности, вероятностные распределения и статистические оценки.

↓

ВИДЫ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ:

- Тексты (корпусы текстов).
- Языковые единицы лингвистического уровня.

ТЕОРИЯ ЧАСТОТ СЛОВ

Предпосылки теории частот слов

А. Методологические (концептуальные) суждения о мире

Б. Эмпирические наблюдения

В. Измерения

Г. Практические приложения

МЕТОДОЛОГИЧЕСКИЕ ПРЕДПОСЫЛКИ ТЕОРИИ ЧАСТОТ СЛОВ

РЕАЛЬНЫЙ МИР (ПОРМ)

↓

<ЗНАНИЯ СУБЪЕКТА О ПОРМ> ←→ <ЕЯ ОПИСАНИЕ>

↓

МЕТАЗНАНИЕ

↓

ЗНАНИЕ ОБ ОРГАНИЗАЦИИ ЯЗЫКОВЫХ ЕДИНИЦ (СИНТАГМ):

…,
морфемы,
словоформы,
словосочетания,
предложения,
…,
тексты,
…

ЭМПИРИЧЕСКИЕ НАБЛЮДЕНИЯ ТЕОРИИ ЧАСТОТ СЛОВ

- Значительные количественные (номенклатурные) и комбинаторные ограничения на использование языковых единиц.
- Существенная избыточность некоторых типов языковых единиц.
- Сложная иерархическая структура синтагм.
- Последовательная во времени организация языковых единиц.

ИЗМЕРЕНИЯ ТЕОРИИ ЧАСТОТ СЛОВ

- **Исследования произведений А.С.Пушкина:**

словоупотреблений – 545 000; разных слов – 21 000.

- **Исследования языковой деятельности школьников:**

корпус текстов(писем, сочинений, заданий и т.п.) – 100 000; словоупотреблений – 6 000 000; разных словоформ – 25 000; разных слов – 2 500.

- **Исследования современных английских текстов:**

словоупотреблений – 250 000; разных словоформ в книжных текстах – 24 000, в разговорных – 10 000.

- **Исследования французской разговорной речи:**

50% словоупотреблений – это 37 слов, 75% – 120 слов, 90% – 887 слов;

95% словоупотреблений языка телефонных разговоров – 737 слов.

ПРАКТИЧЕСКИЕ ПРИЛОЖЕНИЯ ТЕОРИИ ЧАСТОТ СЛОВ

- Криптография
- Стенографирование
- Полиграфия
- Редакционно-издательская подготовка рукописей
- Распознавание текстов (печатных и рукописных)
- Распознавание аудиовизуальной речи
- Автоматизированное создание баз данных
- Автоматический перевод
- Сжатие данных
- Информационный поиск
- Автоматическое индексирование и реферирование

МОДЕЛЬ «РАНГ-ЧАСТОТА»

Жан.-Батист Эступ (Jean Baptiste Estoup).

Джордж Кингсли Зипф (George Kingsley Zipf),

ОПРЕДЕЛЕНИЯ МОДЕЛИ «РАНГ-ЧАСТОТА»

<ТЕКСТ>

↓

<ЧАСТОТНЫЙ СЛОВНИК>

↓

Ранг r	**Слово W(r)**	**Частота f(r)**
1	W(1)	f(1)
2	W(2)	f(2)
…		
r	W(r)	f(r)

Пример:

Ранг r	**Слово W(r)**	**Частота f(r)**
1	the	245
2	of	136
3	terms	98
4	to	81
5	a	65
6	and	61
7	in	55
8	we	52
...	…	…

ЗАКОН ЧАСТОТ СЛОВ ЦИПФА

$$i(k, r)/k = 0.1 * r^{-1} = 1/(10 * r), \quad (1.0)$$

где: $i(k,r)/k$ – относительная частота слова в тексте,

k – общее число слов в тексте,

r – ранг слова, т.е. его порядковый номер в упорядоченном по убыванию частотной функции словнике.

ЛИТЕРАТУРА

Дж. Солтон.
Динамические библиотечные информационные системы.
М.: Наука, 1979.

Б.Мандельброт.
Теория информации и психолингвистика: теория частот слов // Математические методы в социальных науках /
Сб. статей под ред. П.Лазарсфельда и Н.Генри.
М.: Прогресс, 1973. – С. 316–337.

«ВЫВОД» ЗАКОНА ЧАСТОТ СЛОВ (1)

Текст — случайная последовательность символов (букв и пробелов). Пробелы обозначают границы между словами.

Обозначим:

W(r) — слово; **r** — ранг слова; ***k*** — количество слов; ***i* (r, *k*)/*k*** — относительная частота слова; ***p* (r)** — вероятность слова; p_0 — вероятность пробела; ***M*** — количество типов букв, ***M*>1**, $(1 - p_o)/M$ — вероятность буквы в тексте; **m** — количество букв в слове.

Вероятность слова, состоящего из *m* букв:

$$p_0\left[\frac{1-p_0}{M}\right]^m = P_0 \exp\{-m \log[M/(1-p_0)]\}.$$

Это может быть записано как

$p_0 exp\{-\beta m\}$, где

$\beta = log\ (M/(1-p_0))$ — положительная величина, зависящая от p_o и ***M***.

«ВЫВОД» ЗАКОНА ЧАСТОТ СЛОВ (2)

Зависимость между числом букв *m* и рангом слова *r*

Букв в слове	Типов слов	***Пример:***		
		пробел: _; буквы: a,b,c; M=3.	*Типов слов*	*Вер-ть слова*
0	1	_	1	0.2500
1	M	a,b,c	3	0.0625
2	M^2	aa, ab, ac, ba, bb, bc, ca, cb, cc	9	0.0153
3	M3	aaa, aab, aac, aba, abb, abc,…	27	0.0038

Пример: *объем текста 1000 символов;*
пробелов — ~ *250,* {a,b,c} — ~ 62; {aa,…cc} — ~ 15; {aaa,…ccc} — ~ 3.

Ранг	1	2	3	4	5	6	…	12	…
Частота	~ 62	~ 62	~ 62	~ 15	~ 15	~ 15	…	~ 15	…
Вер-ть	0.0625	0.0625	0.0625	0.0153	0.0153	0.0153	…	0.0153	…

«ВЫВОД»

ЗАКОНА ЧАСТОТ СЛОВ (3)

$$r > 1 + M + M^2 + \ldots + M^{m-1} = \frac{M^m - 1}{M - 1},$$

или, наоборот,

$$m < \frac{\log [(M-1)\, r + 1]}{\log M};$$

$$r \leqslant 1 + M + M^2 + \ldots + M^m = \frac{M^{m+1} - 1}{M - 1},$$

или, наоборот,

$$m \geqslant -1 + \frac{\log [(M-1)\, r + 1}{\log M}.$$

«ВЫВОД» ЗАКОНА ЧАСТОТ СЛОВ (4)

Зависимость между рангом и вероятностью:
ступенчатая функция, которая постоянна,
когда *r* изменяется между двумя $(M^m - 1)/(M - 1)$,
соответствуя последовательным значениям *m.*

Если *m* велико, то:

$$r > (M^m - 1)/(M - 1), \quad r < (M^m - 1)/(M - 1);$$

$$r \cong (Mm - 1)/(M - 1),$$

или

$$m \sim \frac{\log [(M-1)\, r]}{\log M}.$$

«ВЫВОД»
ЗАКОНА ЧАСТОТ СЛОВ (5)

Вероятность слова из *m* букв

$$p(r) = p_0 \exp\left\{-\frac{\beta}{\log M} \cdot \log\left[(M-1)\,r\right]\right\} = p_0 (M-1)^{-B} r^{-B} = \\ = P r^{-B}.$$

Где:

B = β/log M ; β=log (M/(1-p0)); P = p_0 (M—1):

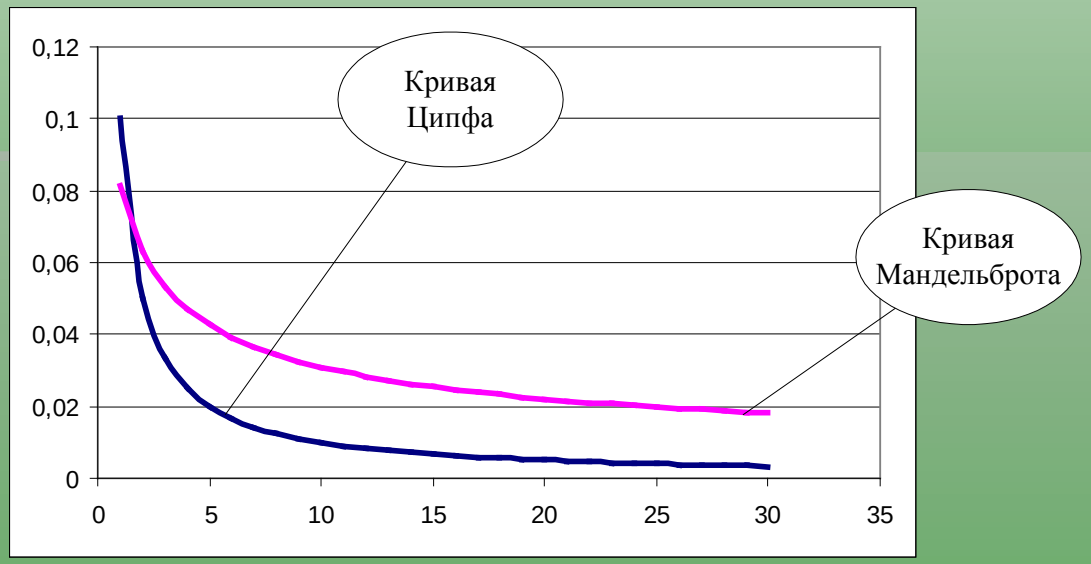
Связь между вероятностью слова и его рангом почти идентична закону Ципфа при значениях B = -1 и P = 0.1

ФОРМУЛА МАНДЕЛЬБРОТА

«Поведение» наиболее часто употребляющихся слов, а также редких, которые характеризуют «богатство словарного состава» текста не соответствует закону Ципфа.

Формула Б.Мандельброта (Benoit Mandelbrot)

$$i(k,r) = pk\,(r+v)^{-b},$$ где: b, k, v – const (1.1)



ВОПРОСЫ ДЛЯ САМОКОНТРОЛЯ

- Опишите модель текста «ранг-частота».
- Сформулируйте закон частот слов Ципфа.
- «Выведите» закон Ципфа.
- Напишите формулу Мандельброта для закона частот слов.