УДК 007.681.5:519.714:519.766

АТРИБУЦИЯ АНОНИМНЫХ И ПСЕВДОНИМНЫХ ТЕКСТОВ В СИСТЕМНО-КОГНИТИВНОМ АНАЛИЗЕ

Луценко Е.В. – д. э. н., профессор
Кубанский государственный аграрный университет

В данной статье исследуется возможность атрибуции текстов с применением технологии и инструментария системно-когнитивного анализа. Приведен подробный численный пример реализации всех этапов СК-анализа при атрибуции текстов, т. е. когнитивной структуризации и формализации предметной области; формирования обучающей выборки; синтеза семантической информационной модели; ее оптимизации и измерения адекватности; адаптации и пересинтеза; а также типологического и кластерно-конструктивного анализа. Для специалистов по атрибуции и контент-анализу текстов на естественном языке. Материал может быть использован в качестве руководства к лабораторной работе по дисциплине: "Интеллектуальные информационные системы".

Краткая теория

Под атрибуцией анонимных и псевдонимных текстов понимается установление их вероятного авторства [1–5].

Анонимные тексты – это тексты вообще без подписи автора, а псевдонимные – подписанные под псевдонимом.

Задача идентификации текстов на основе анализа предложений является тривиальной из-за уникальности предложений. Поэтому больший интерес представляет задача идентификации текстов на основе анализа слов, т. е. задача атрибуции текстов, имеющая большое научное и практическое значение. К этой задаче сводится:

- определение вероятного авторства текстов в случае, когда автор не указан (анонимный текст) или указан его псевдоним (псевдонимный текст);
- датировка текста.
- ***задачи идентификации, прогнозирования, сравнения и классификации объектов, описанных на естественном языке (причем не важно, на каком именно).***

С ней связаны также задачи автоматического выделения дескрипторов и задачи нечеткого поиска и идентификации.

Все эти задачи имеют практическое значение для специалистов по прикладной информатике в экономике и юриспруденции, которых готовят в Кубанском государственном аграрном университете.

Одному из вариантов решения этих задач с применением интеллектуальной технологии "Эйдос" посвящена данная статья.

Задания

Согласно логике системно-когнитивного анализа, выполнить следующие работы.

1. Осуществить когнитивную структуризацию предметной области.
2. Выполнить формализацию предметной области.
3. Сформировать обучающую выборку.
4. Осуществить синтез семантической информационной модели.
5. Оптимизировать семантическую информационную модель.
6. Проверить семантическую информационную модель на адекватность, измерить внутреннюю и внешнюю, дифференциальную и интегральную валидность.
7. Выполнить адаптацию модели и измерить, как изменилась ее адекватность.
8. Осуществить пересинтез модели и измерить, как изменилась ее адекватность.
9. Вывести информационные портреты текстов и дать их интерпретацию.
10. Провести кластерно-конструктивный анализ модели.

Пример решения

1. Осуществить когнитивную структуризацию предметной области

Под когнитивной структуризацией в СК-анализе понимается определение причин и следствий, факторов и состояний объекта управления, исходной информации и того, на что она влияет.

В данной лабораторной работе необходимо решить задачу идентификации текстов по входящим в них словам. Следовательно, необходимо будет сформировать обобщенные образы текстов, соответствующих определенной тематике или автору (будем считать, что сочинение принадлежит тому писателю, творчеству которого оно посвящено). Для этого в качестве объектов обучающей выборки использовались фрагменты текстов школьных сочинений, взятые из Internet, а в качестве признаков текстов – входящие в них слова.

Каждое сочинение разобьем случайным образом на примерно равные по объему небольшие фрагменты, которые используем в качестве объектов обучающей выборки.

2. Выполнить формализацию предметной области

Под формализацией предметной области понимается разработка классификационных и описательных шкал и градаций и ввод их в программную систему "Эйдос", являющуюся инструментарием СК-анализа.

2.1. Формирование классификационных шкал и градаций

В подсистеме "Классификационные шкалы и градации" введем классы, соответствующие следующим писателям: Ф.М. Достоевский; Н.В. Гоголь; А.С. Грибоедов; М.Ю. Лермонтов; А.С. Пушкин; Л.Н. Толстой; И.С. Тургенев (рис. 1).

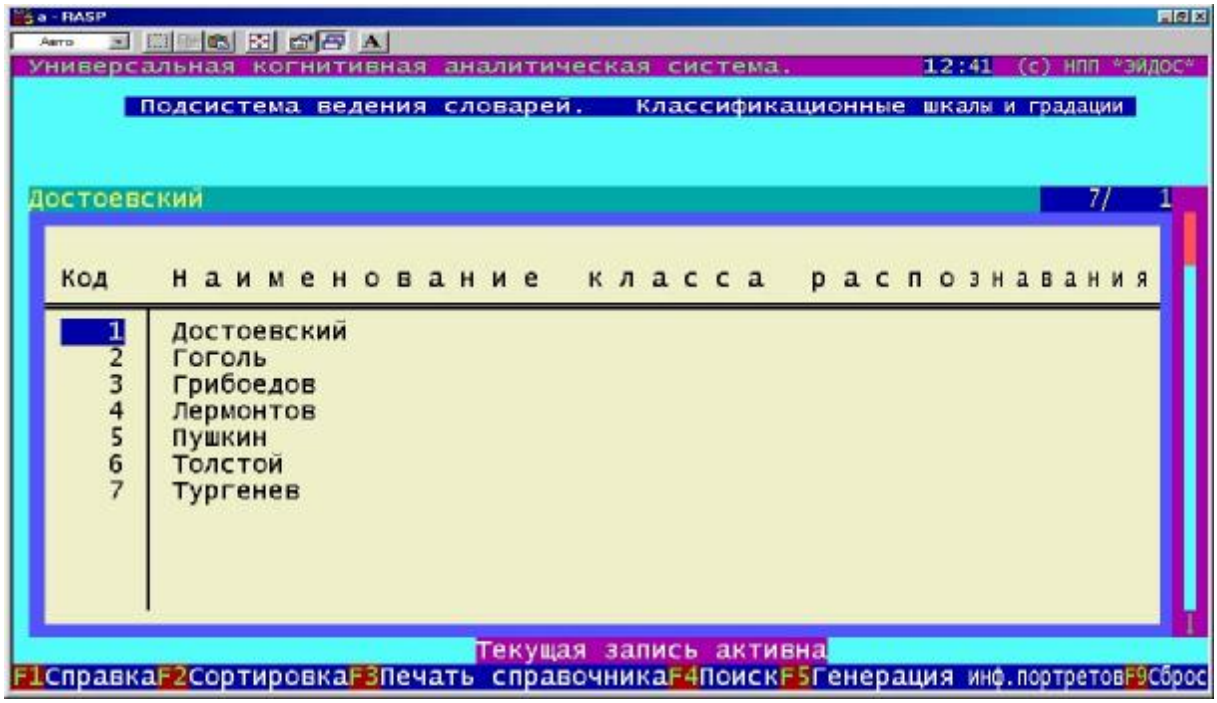


Рис. 1. Ввод классов

2.1. Формирование описательных шкал и градаций

С этой целью исходные файлы для формирования объектов обучающей выборки должны быть средствами Word и представлены в виде текстовых файлов, стандарта "Текст DOS" (без разбиения на строки).

Затем каждый из этих файлов разбивается на столько файлов, сколько в нем строк, причем имена этих файлов должны иметь вид: ####SUBSTR(File_name,4).TXT, где #### – сквозной номер файлов, соответствующий будущему номеру анкеты обучающей выборки, SUBSTR(File_name,4) – первые 4 символа имени исходного файла.

Полученные файлы должны быть помещены в поддиректорию DOB системы "Эйдос", а исходные – удалены из нее.

Это осуществляется одним из трех способов:

1. Вручную.

2. С использованием специальной программы, текст которой приводится ниже (язык программирования xBase).

3. В режиме: "Словари – Программные интерфейсы для импорта данных – Импорт данных из TXT-файлов стандарта "Текст DOS" формируем описательные шкалы и градации (рис. 2), причем в качестве признаков выбираем слова**.**

Исходный текст программы записи TXT-файлов с данными по строкам

```
*************************************************************************
********** Разбиение текстовых файлов DOS на нумерованные файлы по строкам
********** Луценко Е.В., 03/31/04 04:24pm
*************************************************************************
scr_start=SAVESCREEN(0,0,24,79)
SHOWTIME(0,58,.T.,"rb/n")
FOR j=0 TO 24
    @j,0 SAY SPACE(80) COLOR "n/n"
NEXT
********** Удаление TXT-файлов, имена которых начинаются на 0
FILEDELETE("0*.TXT")
***** РЕКОГНОСЦИРОВКА
Count = ADIR("*.TXT")             && Кол-во TXT-файлов
IF Count = 0
   Mess = "В текущей директории TXT-файлов не обнаружено !!!"
   @15,40-LEN(Mess)/2 SAY Mess COLOR "gr+/n"
   INKEY(0)
   RESTSCREEN(0,0,24,79,scr_start)
   SHOWTIME()
   QUIT
ENDIF
PRIVATE Name[Count],Size[Count]   && Имена и размеры файлов
Count = ADIR("*.txt",Name,Size)
SortData(Name,Size,LEN(Name),1)   && Сортировка файлов по алфавиту
CrLf = CHR(13)+CHR(10)            && Конец строки (абзаца) (CrLf)
*** Загрузка TXT-файлов
Num_pp = 0                        && Номера выходных файлов
FOR f = 1 TO Count                && Начало цикла по TXT-файлам
    ****** Загрузка файла
    Buffer = FILESTR(Name[f],.T.)
    Buffer = CHARONE(" ",Buffer)  && Удаление повторяющихся пробелов
    Buffer = Buffer + CrLf
    Len = AT(CrLf,Buffer)
    DO WHILE Len > 0 .AND. LASTKEY() <> 27     && Цикл по строкам
       Len = AT(CrLf,Buffer)
       IF Len > 0
          ****** Запись фрагмента файла
          Str_pr = ALLTRIM(SUBSTR(Buffer,1,Len-1))
          Fn_out = STRTRAN(STR(++Num_pp,4)," ","0")+SUBSTR(Name[f],1,4)+".TXT"
          STRFILE(Str_pr,Fn_out)
          ****** Исключение из буфера записанной строки
          Buffer = ALLTRIM(SUBSTR(Buffer,Len+1))
       ENDIF
    ENDDO
NEXT
*** Удаление исходных TXT-файлов
FOR f=1 TO Count
    FILEDELETE(Name[f])
NEXT
RESTSCREEN(0,0,24,79,scr_start)
SHOWTIME()
```

```
QUIT
```

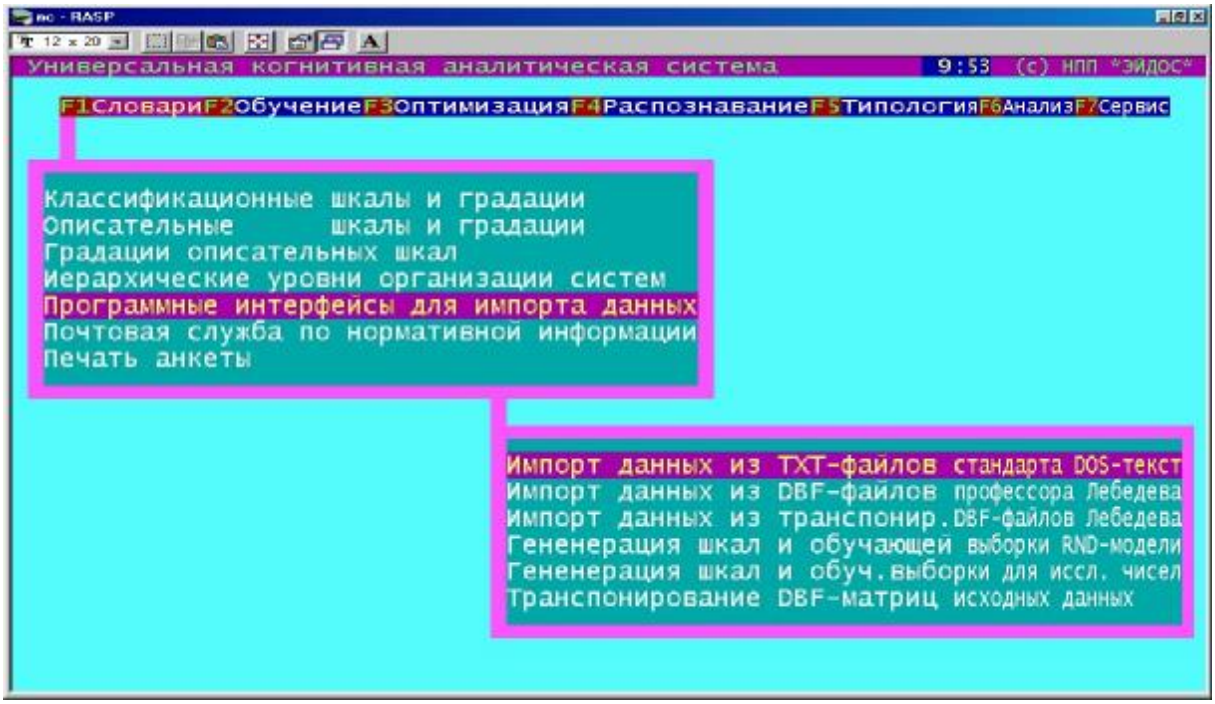


Рис. 2. Выход на режим генерации справочников на основе текстовых файлов

На рисунке 3 приведена экранная форма, отображающая ход процесса генерации описательных шкал и градаций и TXT-файлов, содержащих примеры текстов с разбивкой по строкам.

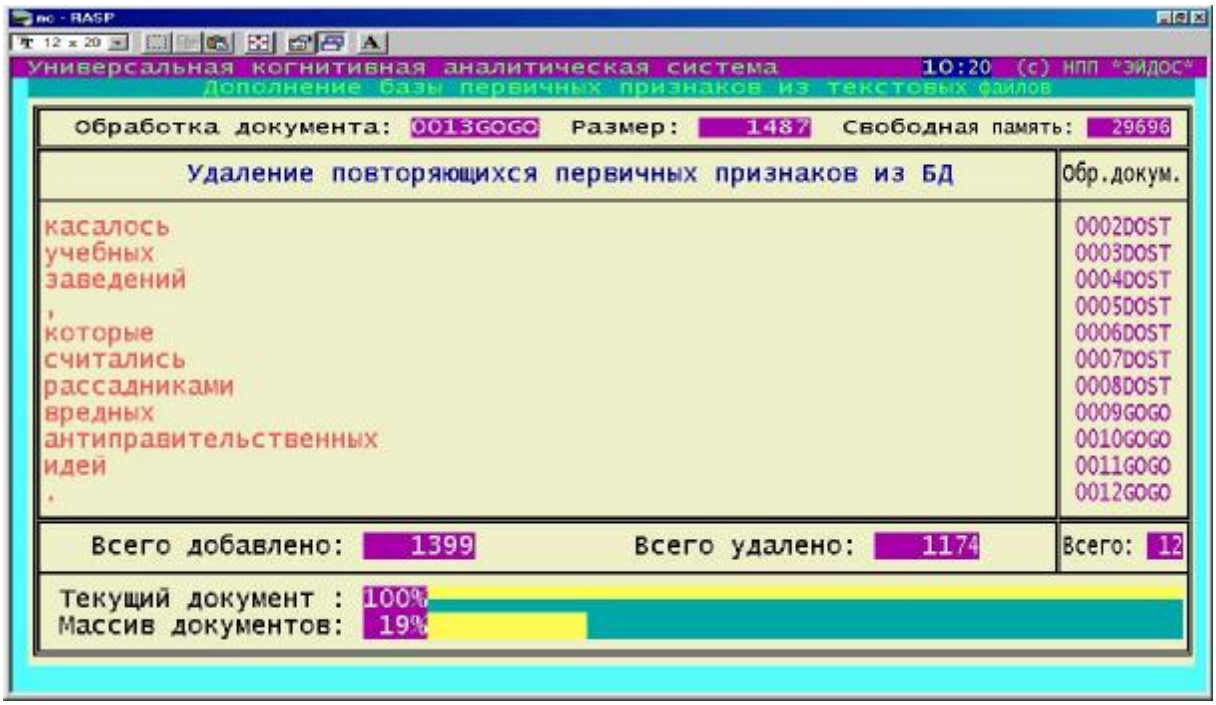


Рис. 3. Генерация описательных шкал и градаций на основе TXT-файлов

В результате получаем классификационные и описательные шкалы и градации, приведенные в таблицах 1 и 2. В таблице 2 отражена лишь часть справочника, т. к. его размерность составляет 3522 градации (т. е. слова).

Таблица 1. Классификационные шкалы и градации

Код	Наименование
1	Достоевский
2	Гоголь
3	Грибоедов
4	Лермонтов
5	Пушкин
6	Толстой
7	Тургенев

Таблица 2. Описательные шкалы и градации (фрагмент)

Код	Наименование	Код	Наименование	Код	Наименование
1	!	41	Бедные	81	Все
2	(	42	Без	82	Вспомним
3	(основной	43	Бездушных	83	Встреча
4	)	44	Безумным	84	Всюду
5	,	45	Безумных	85	Вы
6	-	46	Безухов	86	Вызывают
7	.	47	Безухову	87	Высокие
8	1812	48	Белинский	88	Высокопарные
9	20-	49	Бессильной	89	Г
10	30-е	50	Бог	90	Герой
11	30-х	51	Боже	91	Главная
12	60-х	52	Болконский	92	Глухость
13	:	53	Болконскому	93	Говоря
14	;	54	Бордо	94	Гоголь
15	?	55	Борис	95	Гоголя
16	Встает	56	Бориса	96	Годунов
17	XIX	57	Бородинским	97	Горе
18	А	58	Бородинского	98	Гости
19	Автор	59	Буянов	99	Грибоедов
20	Авторский	60	Была	100	Грибоедова
21	Агрессивная	61	В	101	Гулливера
22	Адама	62	Ведь	102	Да
23	Александр	63	Везде	103	Даже
24	Александра	64	Век	104	Дворянин-аристократ
25	Алексеевна	65	Великий	105	Действительно
26	Алексеевна	66	Великолепная	106	Дельвигу
27	Аммоса	67	Вернулся	107	Денисова
28	Андреевич	68	Взволнованный	108	Дидло
29	Андрей	69	Взгляды	109	Для
30	Андрею	70	Власы	110	Дмитриевна
31	Анной	71	Вместе	111	Добролюбова
32	Архивам	72	Внешней	112	Достоевского
33	Афанасьевича	73	Внешние	113	Драматична
34	Ах	74	Воды	114	Друбецкого
35	Базаров	75	Возникает	115	Другое
36	Базарова	76	Война	116	Думы
37	Базаровым	77	Вообще	117	Дуни
38	Балы	78	Вопрос	118	Дуня
39	Бегущим	79	Вот	119	Душа
40	Бедность	80	Время	120	Евгений

3. Сформировать обучающую выборку

Обучающая выборка представляет собой фрагменты текстов различных авторов, используемые в качестве примеров для формирования семантической информационной модели. На основе анализа этих примеров выявляются взаимосвязи между теми или иными словами и принадлежностью текстов разным авторам.

Для генерации обучающей выборки используется 1-й режим 2-й подсистемы, функция F7InpTXT – F6Ввод из всех файлов. При этом в качестве признаков, также как при формировании описательных шкал и градаций, выбираются слова (рис. 4).

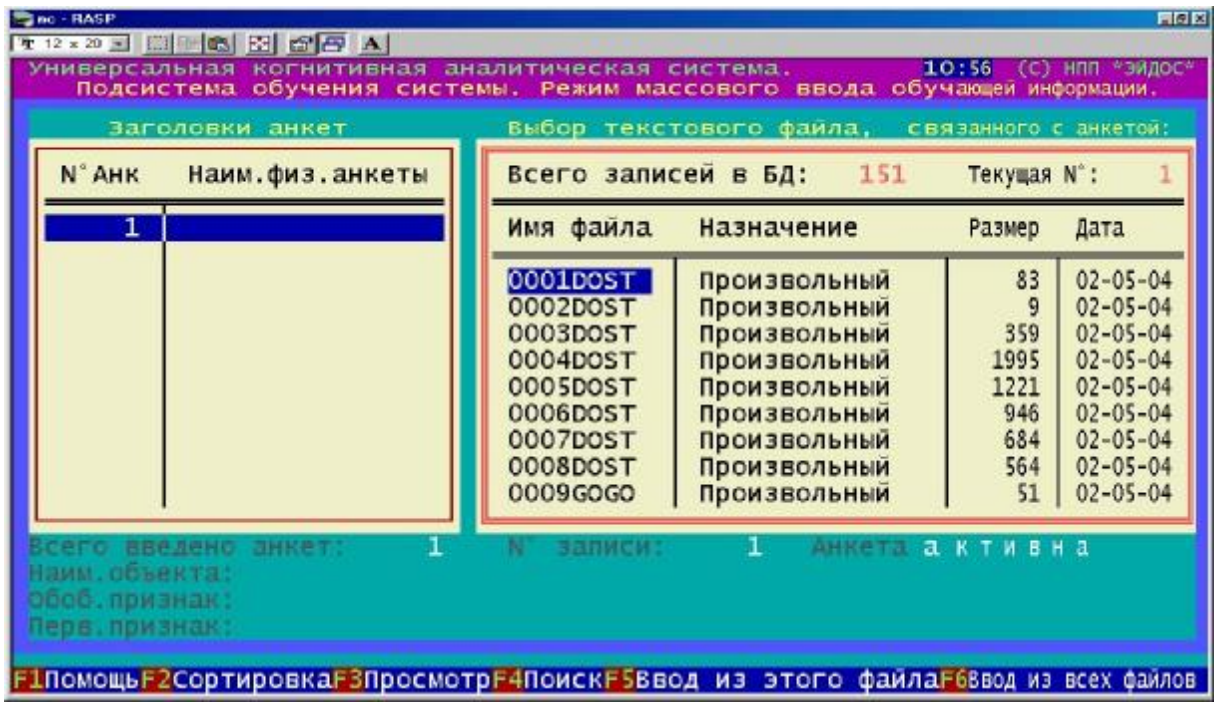


Рис. 4. Генерация обучающей выборки из TXT-файлов

В результате формируется обучающая выборка, состоящая из 151 примера фрагментов текстов различных авторов. Остается лишь проставить в каждом примере (анкете) код писателя, о котором данный текст, т. е. код класса (в левом окне).

4. Осуществить синтез семантической информационной модели

Синтез модели осуществляется во 2-й подсистеме, 4-м режиме, 5-й функции (рис. 5).

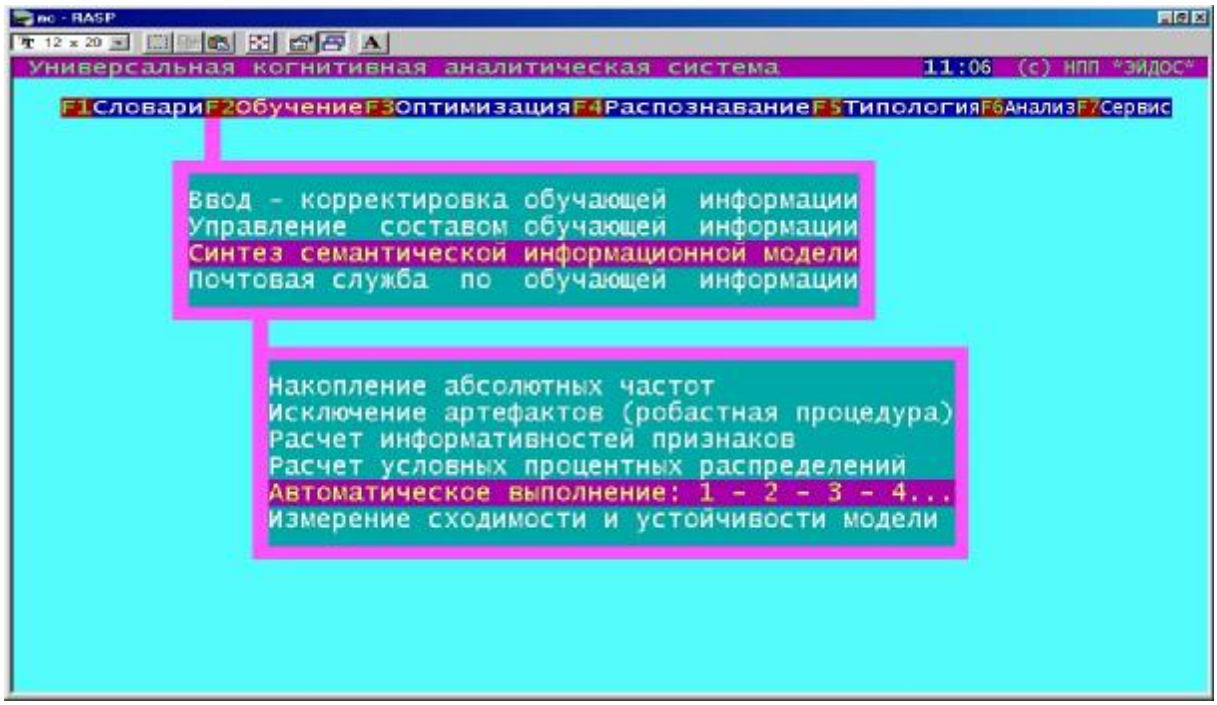


Рис. 5. Запуск режима:
"Синтез семантической информационной модели"

Стадия процесса синтеза отображается в ряде экранных форм, одна из которых приведена на рисунке 6.

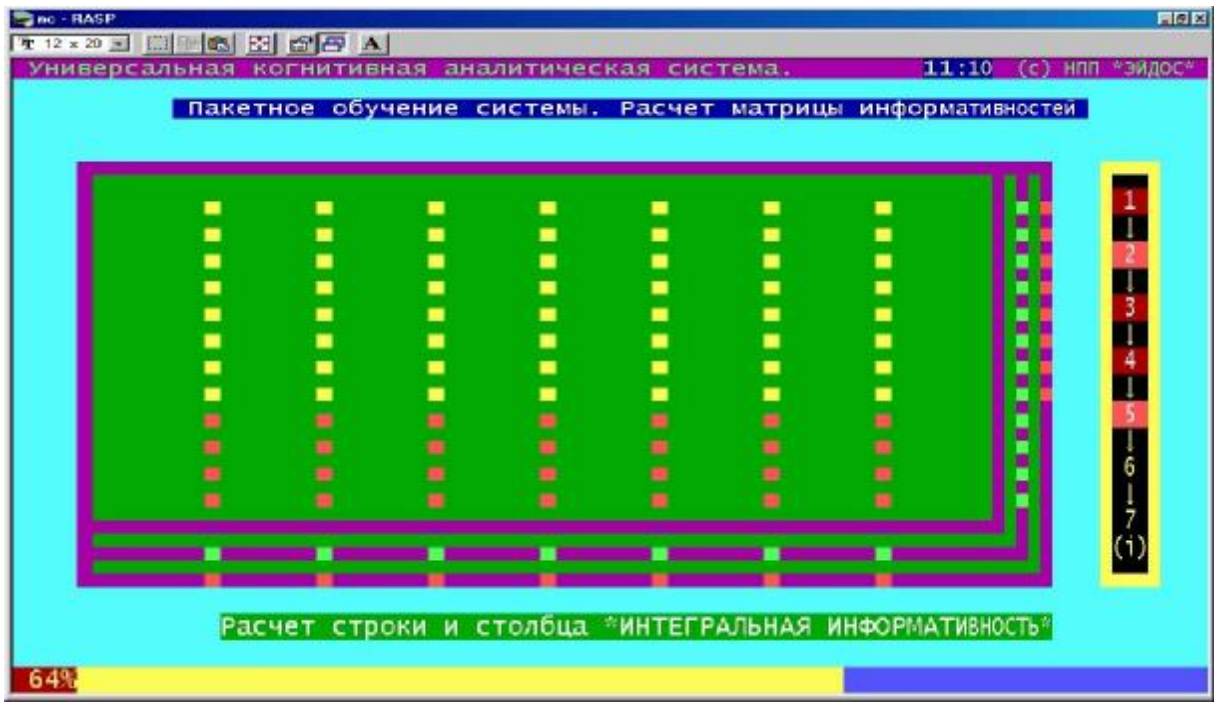


Рис. 6. Экранная форма, отображающая одну из стадий процесса синтеза семантической информационной модели

5. Оптимизировать семантическую информационную модель

Оптимизация модели представляет собой исключение из нее малозначащих признаков без потери адекватности модели. Эта операция осуществляется во 2-м режиме 3-й подсистемы (рис. 7).

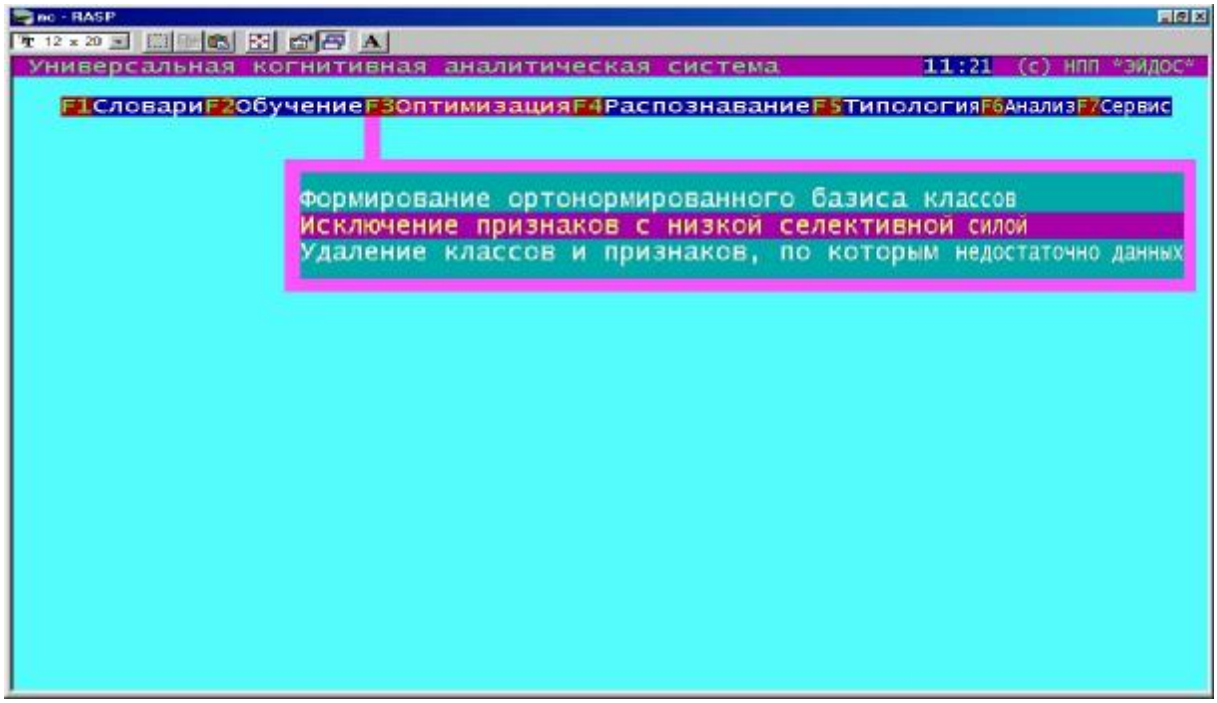


Рис. 7. Выход на режимы оптимизации модели

При том имеется возможность вывести график ценности признаков "нарастающим итогом", т. е. Паретто-диаграмму признаков (рис. 8).

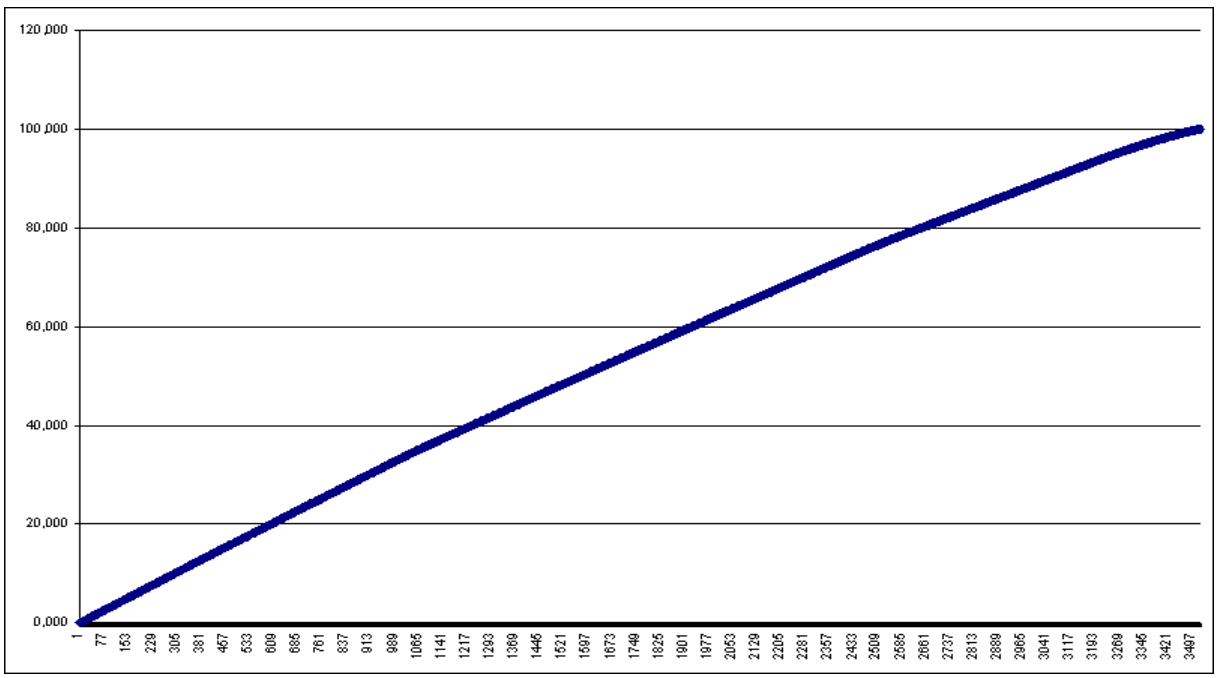


Рис. 8. Паретто-диаграмма признаков

Из рисунка 8 видно, что в системе нет признаков, имеющих очень малую или нулевую ценность. Это связано с тем, что все слова являются практически уникальными для фрагментов текстов, т. е. встречаются во всех текстах в основном от 1 до 5 раз (рис. 9).

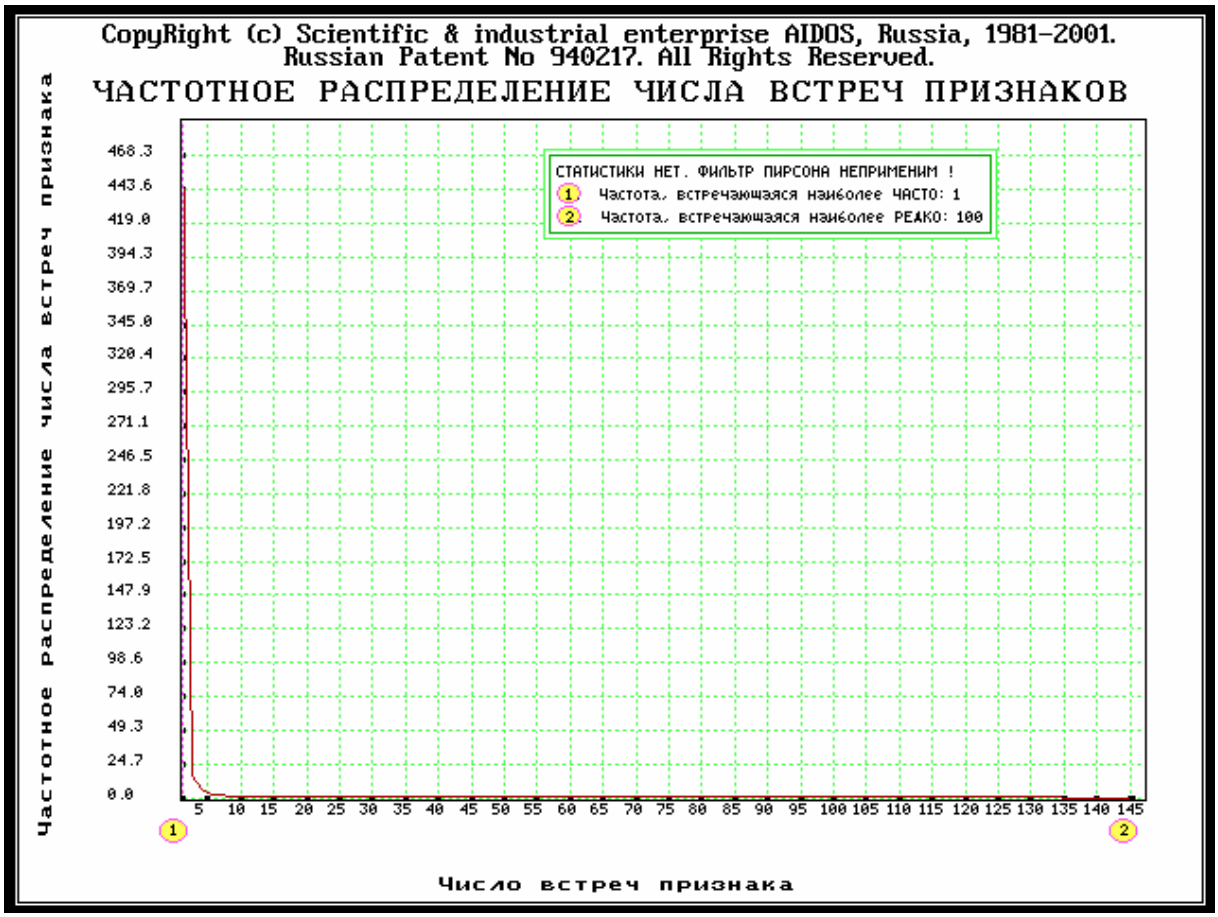


Рис. 9. Частотное распределение числа встреч признаков

6. Проверить семантическую информационную модель на адекватность, измерить внутреннюю и внешнюю, дифференциальную и интегральную валидность

6.1. Внутренняя дифференциальная и интегральная валидность

Под внутренней валидностью понимается способность модели верно идентифицировать объекты, входящие в обучающую выборку.

Для измерения адекватности модели необходимо выполнить следующие действия:

1. Скопировать обучающую выборку в распознаваемую (во 1-м режиме 2-й подсистемы, нажав клавишу F5).

2. Выполнить пакетное распознавание (во 2-м режиме 4-й подсистемы, задав 1-й критерий сходства) (рис. 10).

3. Измерить адекватность модели (во 2-м режиме 6-й подсистемы) (рис. 11 и 12).

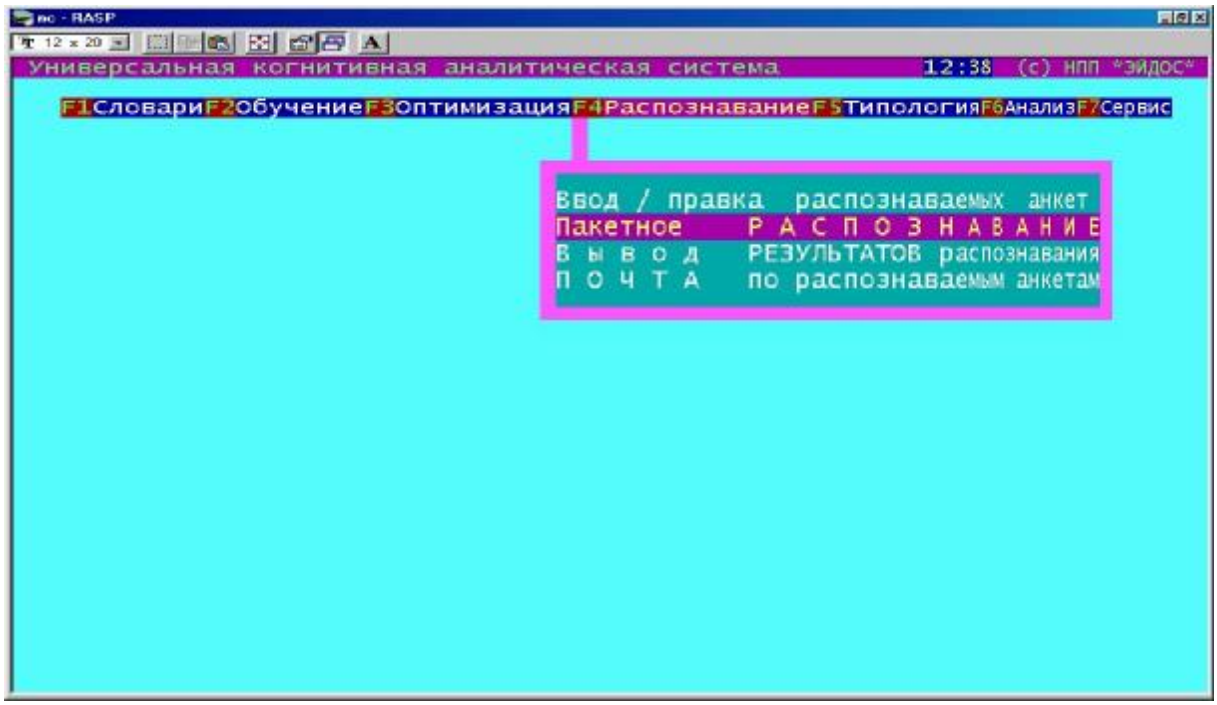


Рис. 10. Выход на режим пакетного распознавания

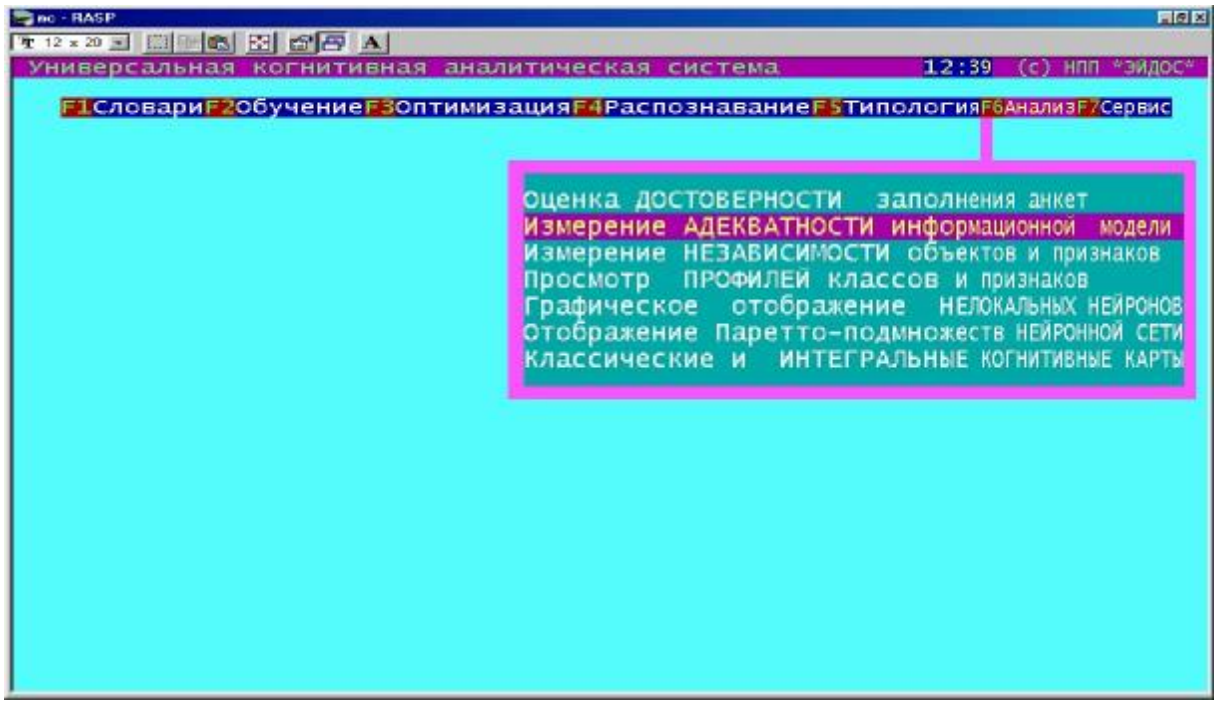


Рис. 11. Выход на режим измерения адекватности модели

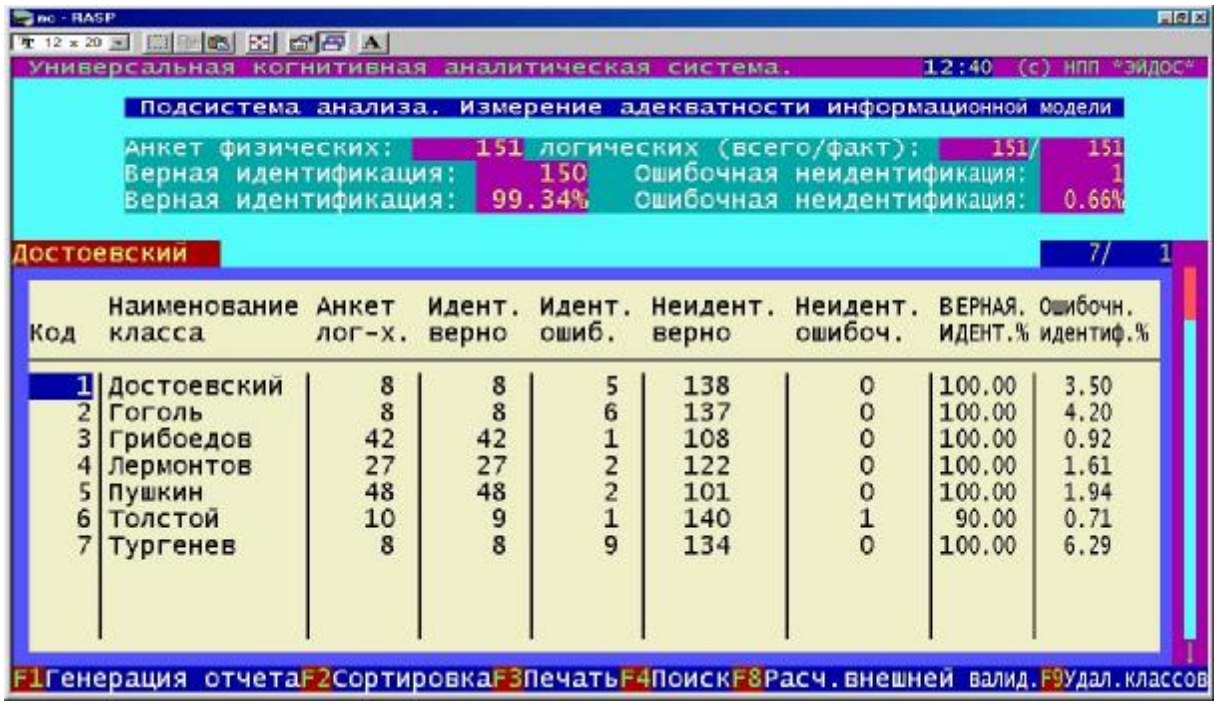


Рис. 12. Экранная форма управления измерением адекватности модели и отображения результатов

Эта форма может прокручиваться вправо-влево. В верхней части формы приведены показатели интегральной валидности (средневзвешенные по всей обучающей выборке), а в самой таблице – дифференциальной валидности, т. е. в разрезе по классам.

Кроме того, результаты измерения адекватности модели выводятся в форме файлов с именами ValidSys.txt (рис. 13) и ValAnkSt.txt (рис. 14) стандарта "TXT-текст DOS" в поддиректории TXT. Первый файл имеет следующий вид.

ИЗМЕРЕНИЕ АДЕКВАТНОСТИ СЕМАНТИЧЕСКОЙ ИНФОРМАЦИОННОЙ МОДЕЛИ
Анкет физических: 151 логических (всего/факт): 151/ 151
Верная идентификация: 150 Ошибочная неидентификация: 1
Верная идентификация: 99.34% Ошибочная неидентификация: 0.66%
Минимальный уровень сходства: 0.0 Максимальное кол-во классов: 99999
02-05-04 12:40:09 г.Краснодар

N п/п	Код класса	Наименование класса	Всего логич анкет	ИДЕНТИФИЦИР		Неидентифиц		ИДЕНТИФИЦИРОВ		Неидентифицир	
				ВЕРНО	Ошиб.	Верно	Ошиб.	ВЕРНО%	Ошиб.%	Верно%	Ошиб.%
1	1	Достоевский	8	8	5	138	0	100.00	3.50	96.50	0.00
2	2	Гоголь	8	8	6	137	0	100.00	4.20	95.80	0.00
3	3	Грибоедов	42	42	1	108	0	100.00	0.92	99.08	0.00
4	4	Лермонтов	27	27	2	122	0	100.00	1.61	98.39	0.00
5	5	Пушкин	48	48	2	101	0	100.00	1.94	98.06	0.00
6	6	Толстой	10	9	1	140	1	90.00	0.71	99.29	10.00
7	7	Тургенев	8	8	9	134	0	100.00	6.29	93.71	0.00

Универсальная когнитивная аналитическая система НПП *ЭЙДОС*

Рис. 13. Выходная форма ValidSys.txt с результатами измерения адекватности модели и отображения результатов

Рассмотрим, что означают графы этой выходной формы.

"Всего логических анкет" – это количество анкет (примеров текстов) в обучающей выборке, на основе которых формировался образ данного класса.

"Идентифицировано верно" – это количество анкет обучающей выборки, идентифицированных как классы, к которым они действительно относятся.

"Идентифицировано ошибочно" – это количество анкет обучающей выборки, идентифицированных как классы, к которым они в действительности не относятся (ошибка идентификации).

"Неидентифицировано верно" – это количество анкет обучающей выборки, неидентифицированных как классы, к которым они действительно не относятся.

"Неидентифицировано ошибочно" – это количество анкет обучающей выборки, неидентифицированных как классы, к которым они в действительности относятся (ошибка неидентфикации).

В правой части формы приведены те же показатели, но в процентом выражении:

– для анкет, идентифицированных верно и неидентифицированных ошибочно, за 100 % принимается количество логических анкет обучающей выборки по данному классу;

– для анкет, идентифицированных ошибочно и неидентифицированных верно, за 100 % принимается суммарное количество логических анкет обучающей выборки за вычетом логических анкет по данному классу.

```
А  Н  К  Е  Т  Ы   распознаваемой выборки
Класс распознавания          : 1 - ДОСТОЕВСКИЙ
Результат идентификации      : Верная идентификация
Минимальный уровень сходства:  0.0  Максимальное кол-во классов: 99999
02-05-04  12:40:09                                              г.Краснодар
```

Коды анкет распознаваемой выборки
2 3 4 5 6 7 8

```
Универсальная когнитивная аналитическая система                 НПП *ЭЙДОС*
```

Рис. 14. Фрагмент выходной формы ValAnkSt.txt с результатами измерения адекватности модели и отображения результатов

В данной форме приведены коды анкет обучающей выборки, которые были учтены в каждой графе предыдущей формы по каждому классу.

6.2. Внешняя дифференциальная и интегральная валидность

Под внешней валидностью понимается способность модели верно идентифицировать объекты, не входящие в обучающую выборку. Если объект идентифицирован верно, то это означает, что он входит в генеральную совокупность, по отношению к которой обучающая выборка репрезентативна.

Для измерения внешней валидности необходимо выполнить следующие действия:

1. В режиме измерения адекватности модели запустить режим измерения внешней валидности (нажав F8 Измерение внешней валидности) (см. рис. 12).

2. Выбрать один из режимов удаления объектов обучающей выборки, приведенный на экранной форме (рис. 15).

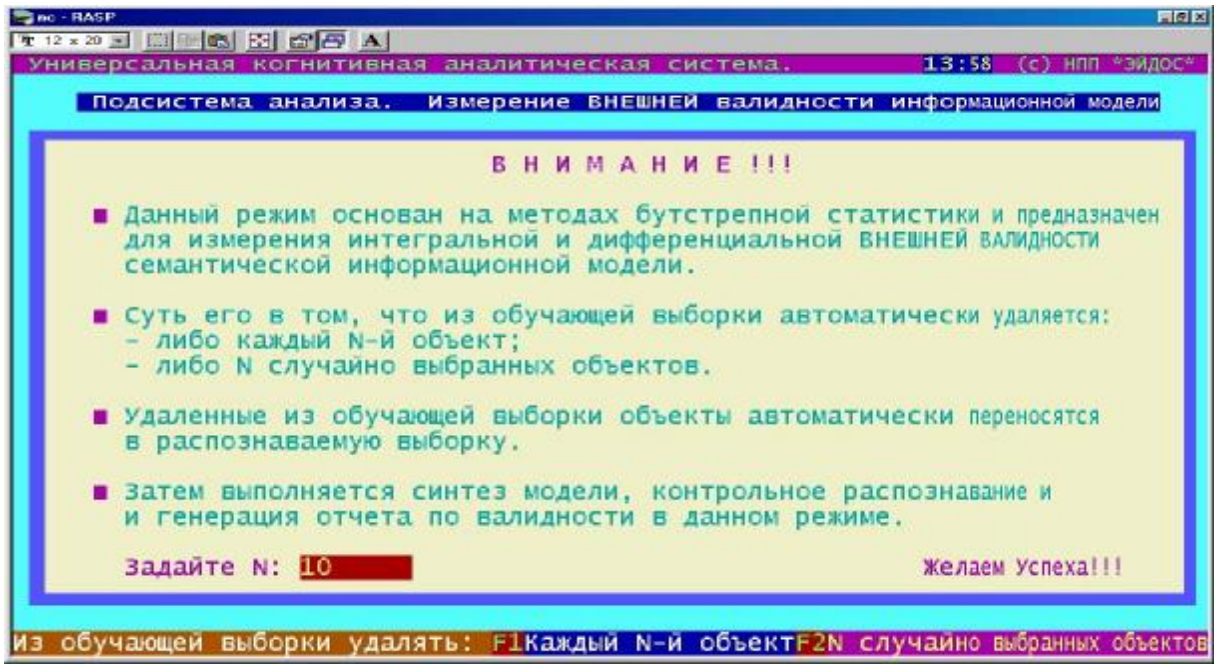


Рис. 15. Режим переноса анкет обучающей выборки в распознаваемую для измерения внешней валидности

Результат выполнения всех указанных на рисунке 15 действий приведен на рисунке 16.

ИЗМЕРЕНИЕ АДЕКВАТНОСТИ СЕМАНТИЧЕСКОЙ ИНФОРМАЦИОННОЙ МОДЕЛИ
Анкет физических: 75 логических (всего/факт): 75/ 75
Верная идентификация: 61 Ошибочная неидентификация: 14
Верная идентификация: 81.33% Ошибочная неидентификация: 18.67%
Минимальный уровень сходства: 0.0 Максимальное кол-во классов: 99999
09-05-04 08:20:09 г.Краснодар

N п/п	Код класса	Наименование класса	Всего логич анкет	ИДЕНТИФИЦИР		Неидентифиц		ИДЕНТИФИЦИРОВ		Неидентифицир	
				ВЕРНО	Ошиб.	Верно	Ошиб.	ВЕРНО%	Ошиб.%	Верно%	Ошиб.%
1	1	Достоевский	4	3	55	16	1	75.00	77.46	22.54	25.00
2	2	Гоголь	4	3	47	24	1	75.00	66.20	33.80	25.00
3	3	Грибоедов	21	19	43	11	2	90.48	79.63	20.37	9.52
4	4	Лермонтов	13	8	43	19	5	61.54	69.35	30.65	38.46
5	5	Пушкин	24	19	40	11	5	79.17	78.43	21.57	20.83
6	6	Толстой	5	5	45	25	0	100.00	64.29	35.71	0.00
7	7	Тургенев	4	4	53	18	0	100.00	74.65	25.35	0.00

Универсальная когнитивная аналитическая система НПП *ЭЙДОС*

Рис. 16. Выходная форма с результатами измерения внешней валидности методом бутстрепной статистики

При этом исходная выборка была разделена на две:

– в обучающей выборке остались только нечетные анкеты;

– в распознаваемую выборку были включены только четные анкеты;

– при распознавании был использован 2-й интегральный критерий: сумма количества информации.

Анализ отчета по внешней валидности, приведенного на рисунке 16, позволяет сделать вывод о высокой степени адекватности семантической информационной модели. Это значит, что взаимосвязи между словами, использованными в текстах, и принадлежностью этих текстов различным авторам, выявленные по примерам обучающей выборки, оказались имеющими силу и для других фрагментов текстов, не включенных в обучающую выборку, но входящих в распознаваемую выборку, по отношению к которой обучающая выборка репрезентативна.

7. Выполнить адаптацию модели и измерить, как изменилась ее адекватность

Под адаптацией модели понимается ее количественная модификация, осуществляемая путем включения в обучающую выборку дополнительных примеров реализации объектов, относящихся к тем же самым классам и описанным в той же системе признаков.

На первом этапе для изучения адаптивности модели осуществим ее синтез на основе обучающей выборки, состоящей из нечетных анкет, которая использовалась в примере для измерения внешней валидности. В отличие от этого примера эту же выборку будем применять и как распознаваемую.

На втором этапе осуществим синтез модели на основе полной обучающей выборки, включающей как четные, так и нечетные анкеты.

Адаптация модели повышает точность идентификации объектов той же самой генеральной совокупности.

8. Осуществить пересинтез модели и измерить, как изменилась ее адекватность

Под повторным синтезом (пересинтезом) модели понимается ее качественная модификация, осуществляемая путем включения в модель новых дополнительных классификационных и описательных шкал и градаций, представленных примерами в обучающей выборке.

Пересинтез модели обеспечивает возможность ее применения для идентификации объектов расширенной или новой генеральной совокупности.

Приведем пример синтеза новой модели, обобщающей предыдущую.

В модель добавлены новые классы распознавания (табл. 3).

Таблица 3. Классификационные шкалы

№	Наименования классов распознавания
1	Загадки о животных
2	А.П.Чехов "Вишневый сад"
3	Ф.М.Достоевский "Преступление и наказание"
4	Н.В.Гоголь "Ревизор"
5	А.С.Грибоедов "Горе от ума"
6	И.А.Крылов
7	М.Ю.Лермонтов "Мцыри"
8	Фольклорные загадки о природе
9	Некрасов "Кому на Руси жить хорошо"
10	Пословицы
11	А.С.Пушкин "Евгений Онегин"
12	Загадки о саде и огороде
13	В.Шекспир
14	М.А.Шолохов "Тихий Дон"
15	Скороговорки
16	Л.Н.Толстой "Война и мир"
17	И.С.Тургенев "Отцы и дети"

Описательные шкалы и градации не приводятся, т. к. их размерность составляет 6974 градации. Необходимо отметить, что текущая версия 11.7 системы "Эйдос" не имеет принципиальных ограничений на суммарное количество градаций классификационных и описательных шкал при синтезе модели и решении задач идентификации и прогнозирования, а также на количество объектов обучающей выборки. Реально решались задачи с объемом обучающей выборки до 25000 объектов с 1500 классами и 7000 признаками. При этом были осуществлены синтез и исследование моделей, содержащих до 25 миллионов фактов.

В программном интерфейсе импорта данных из 17 исходных текстовых файлов, посвященных различным темам (см. табл. 3), было сформировано 592 фрагмента, которые стали основой обучающей выборки.

После синтеза модели измеряется ее адекватность. Для этого обучающая выборка копируется в распознаваемую, после чего проводятся распознавание и измерение валидности (рис. 17). Продемонстрирована очень высокая внутренняя валидность новой модели.

ИЗМЕРЕНИЕ АДЕКВАТНОСТИ СЕМАНТИЧЕСКОЙ ИНФОРМАЦИОННОЙ МОДЕЛИ
Анкет физических: 592 логических (всего/факт): 592/ 592
Верная идентификация: 591 Ошибочная неидентификация: 1
Верная идентификация: 99.83% Ошибочная неидентификация: 0.17%
Минимальный уровень сходства: 0.0 Максимальное кол-во классов: 99999
09-05-04 11:25:19 г.Краснодар

N п/п	Код класса	Наименование класса	Всего логич анкет	ИДЕНТИФИЦИР		Неидентифиц		ИДЕНТИФИЦИРОВ		Неидентифицир	
				ВЕРНО	Ошиб.	Верно	Ошиб.	ВЕРНО%	Ошиб.%	Верно%	Ошиб.%
1	1	Загадки о животных	66	66	214	312	0	100.00	40.68	59.32	0.00
2	2	А.П.Чехов "Вишневый сад"	10	10	24	558	0	100.00	4.12	95.88	0.00
3	3	Ф.М.Достоевский "Преступление и наказание"	8	8	38	546	0	100.00	6.51	93.49	0.00
4	4	Н.В.Гоголь "Ревизор"	8	8	26	558	0	100.00	4.45	95.55	0.00
5	5	А.С.Грибоедов "Горе от ума"	42	42	33	517	0	100.00	6.00	94.00	0.00
6	6	И.А.Крылов	35	35	12	545	0	100.00	2.15	97.85	0.00
7	7	М.Ю.Лермонтов "Мцыри"	27	27	34	531	0	100.00	6.02	93.98	0.00
8	8	Фольклорные загадки о природе	31	31	263	298	0	100.00	46.88	53.12	0.00
9	9	Некрасов "Кому на Руси жить хорошо"	55	55	17	520	0	100.00	3.17	96.83	0.00
10	10	Пословицы	43	43	213	336	0	100.00	38.80	61.20	0.00
11	11	А.С.Пушкин "Евгений Онегин"	48	48	48	496	0	100.00	8.82	91.18	0.00
12	12	Загадки о саде и огороде	33	33	288	271	0	100.00	51.52	48.48	0.00
13	13	В.Шекспир	59	58	27	506	1	98.31	5.07	94.93	1.69
14	14	М.А.Шолохов "Тихий Дон"	7	7	34	551	0	100.00	5.81	94.19	0.00
15	15	Скороговорки	102	102	120	370	0	100.00	24.49	75.51	0.00
16	16	Л.Н.Толстой "Война и мир"	10	10	20	562	0	100.00	3.44	96.56	0.00
17	17	И.С.Тургенев "Отцы и дети"	8	8	61	523	0	100.00	10.45	89.55	0.00

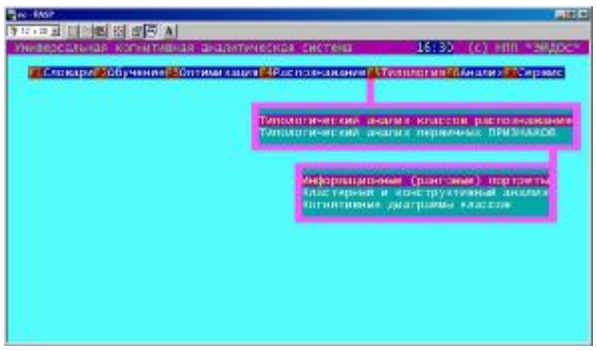
Универсальная когнитивная аналитическая система НПП «ЭЙДОС»

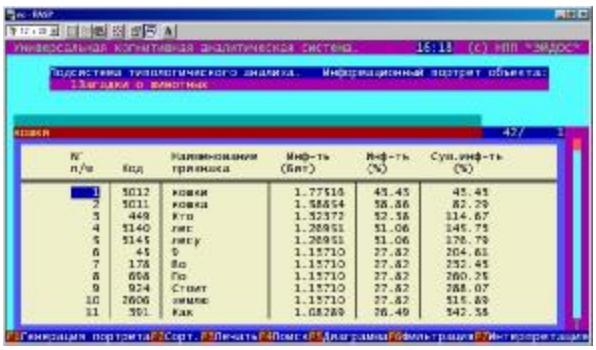
Рис. 17. Выходная форма с результатами измерения внутренней валидности после пересинтеза модели

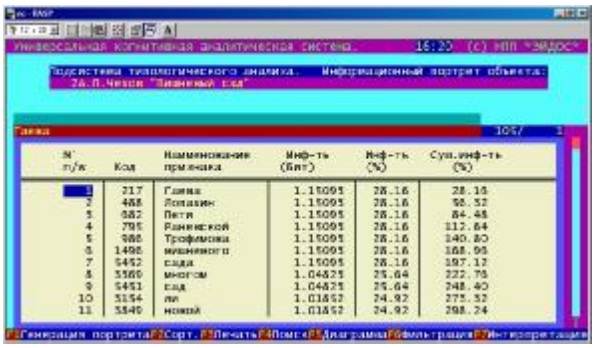
9. Вывести информационные портреты текстов и дать их интерпретацию

Информационный портрет класса представляет собой список признаков в порядке убывания количества информации, содержащегося в этих признаках, о принадлежности к данному классу.

Они генерируются в 1-м режиме 5-й подсистемы "Типология" (рис. 18). Информационные портреты классов отображаются системой "Эйдос" в виде экранных форм, круговых диаграмм и гистограмм, а также распечатываются в форме таблиц в поддиректории TXT. Графические формы записываются в поддиректории PCX.

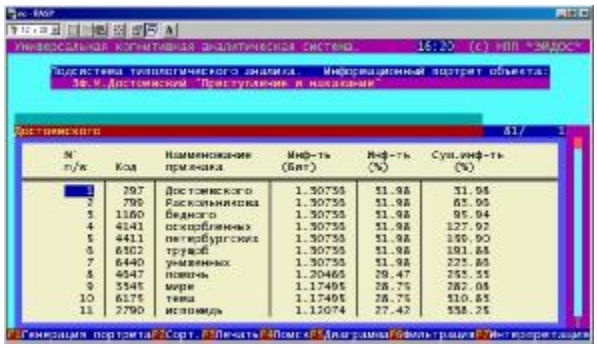






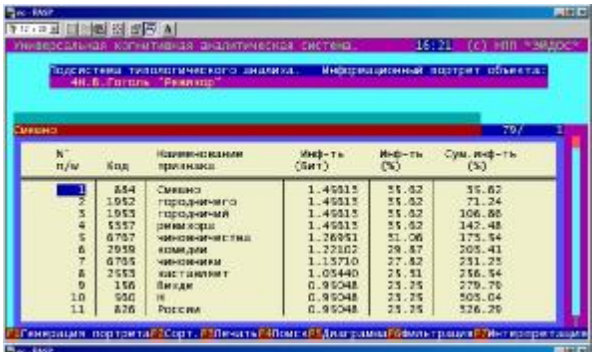
Подсистема типологического анализа. Информационный портрет объекта:
2А.П.Чехов "Вишневый сад"

N п/п	Код	Наименование признака	Инф-ть (Бит)	Инф-ть (%)	Сум.инф-ть (%)
1	217	Гаева	1.15095	28.16	28.16
2	468	Лопахин	1.15095	28.16	56.32
3	582	Пети	1.15095	28.16	84.48
4	795	Раневской	1.15095	28.16	112.64
5	986	Трофимова	1.15095	28.16	140.80
6	1498	именного	1.15095	28.16	168.96
7	5452	сада	1.15095	28.16	197.12
8	5569	многом	1.04825	25.64	222.76
9	5451	сад	1.04825	25.64	248.40
10	5154	ли	1.01652	24.92	273.32
11	5549	новой	1.01652	24.92	298.24



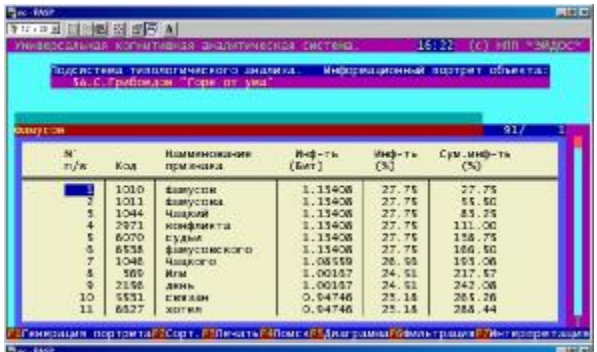
Подсистема типологического анализа. Информационный портрет объекта:
3Ф.М.Достоевский "Преступление и наказание"

N п/п	Код	Наименование признака	Инф-ть (Бит)	Инф-ть (%)	Сум.инф-ть (%)
1	297	Достоевского	1.30755	31.98	31.98
2	799	Раскольникова	1.30755	31.98	63.96
3	1160	Федора	1.30755	31.98	95.94
4	4141	оскорблённых	1.30755	31.98	127.92
5	4411	петербургских	1.30755	31.98	159.90
6	6502	трущоб	1.30755	31.98	191.88
7	6440	униженных	1.30755	31.98	223.86
8	4547	повесть	1.20465	29.47	253.33
9	5545	мире	1.17495	28.75	282.08
10	6175	тема	1.17495	28.75	310.83
11	2790	исповедь	1.12074	27.42	338.25



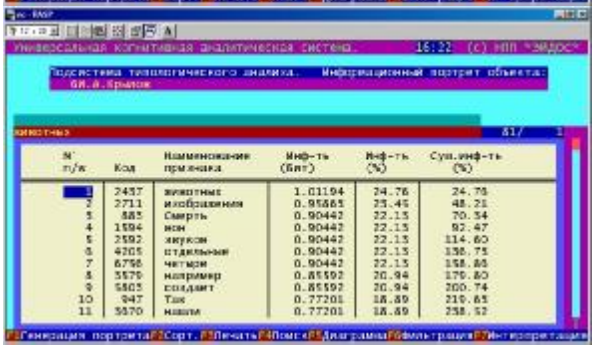
Подсистема типологического анализа. Информационный портрет объекта:
4Н.В.Гоголь "Ревизор"

N п/п	Код	Наименование признака	Инф-ть (Бит)	Инф-ть (%)	Сум.инф-ть (%)
1	884	Сквозник	1.49815	55.62	55.62
2	1952	городничего	1.49815	55.62	71.24
3	1953	городничий	1.49815	55.62	106.86
4	5557	ремизорш	1.49815	55.62	142.48
5	6787	чиновничества	1.26951	31.06	173.54
6	2939	комедии	1.22102	29.87	203.41
7	6795	чиновником	1.15710	27.82	231.23
8	2553	инспектит	1.05440	25.51	256.54
9	156	вице	0.95048	23.25	279.79
10	980	и	0.95048	23.25	303.04
11	826	России	0.95048	23.25	326.29



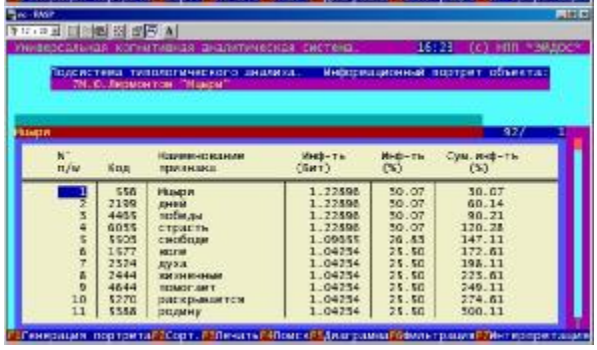
Подсистема типологического анализа. Информационный портрет объекта:
5А.С.Грибоедов "Горе от ума"

N п/п	Код	Наименование признака	Инф-ть (Бит)	Инф-ть (%)	Сум.инф-ть (%)
1	1010	фамусов	1.15406	27.75	27.75
2	1011	фамусова	1.15406	27.75	55.50
3	1044	Чацкой	1.15406	27.75	83.25
4	2971	конфликта	1.15406	27.75	111.00
5	6070	судьи	1.15406	27.75	138.75
6	6538	фамусовского	1.15406	27.75	166.50
7	1048	Чацкого	1.08559	26.10	193.06
8	588	Им	1.00157	24.51	217.57
9	2198	день	1.00157	24.51	242.08
10	5551	сказан	0.94746	25.18	265.26
11	6827	хотел	0.94746	25.18	288.44



Подсистема типологического анализа. Информационный портрет объекта:
6И.А.Крылов

N п/п	Код	Наименование признака	Инф-ть (Бит)	Инф-ть (%)	Сум.инф-ть (%)
1	2437	животных	1.01194	24.76	24.76
2	2711	изображения	0.95885	23.45	48.21
3	883	Смерть	0.90442	22.13	70.34
4	1594	мои	0.90442	22.13	92.47
5	2592	якуков	0.90442	22.13	114.60
6	4205	отдельные	0.90442	22.13	136.75
7	6798	четыре	0.90442	22.13	158.88
8	5579	например	0.85592	20.94	179.80
9	5805	создают	0.85592	20.94	200.74
10	947	Так	0.77201	18.89	219.63
11	5570	наши	0.77201	18.89	238.52

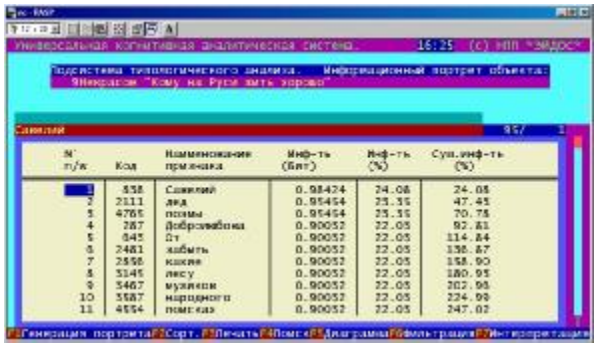


Подсистема типологического анализа. Информационный портрет объекта:
7М.Ю.Лермонтов "Мцыри"

N п/п	Код	Наименование признака	Инф-ть (Бит)	Инф-ть (%)	Сум.инф-ть (%)
1	558	Мцыри	1.22598	30.07	30.07
2	2139	дней	1.22598	30.07	60.14
3	4455	небыли	1.22598	30.07	90.21
4	6055	страсть	1.22598	30.07	120.28
5	5505	свободы	1.09555	26.85	147.11
6	1677	жили	1.04254	25.50	172.61
7	2524	духа	1.04254	25.50	198.11
8	2444	жизненные	1.04254	25.50	223.61
9	4644	помогает	1.04254	25.50	249.11
10	5270	раскрывается	1.04254	25.50	274.61
11	5588	родину	1.04254	25.50	300.11

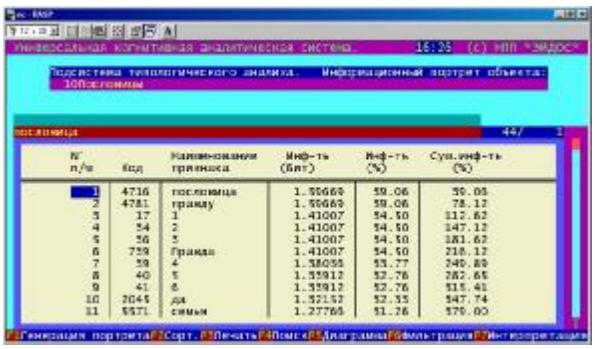
Подсистема типологического анализа. Информационный портрет объекта:
8Афоризмы загадки о природе

N п/п	Код	Наименование признака	Инф-ть (Бит)	Инф-ть (%)	Сум.инф-ть (%)
1	102	Без	1.72910	42.50	42.50
2	5709	нет	1.62659	39.79	82.09
3	5554	миру	1.54248	37.74	119.85
4	45	9	1.41007	34.50	154.35
5	564	Над	1.41007	34.50	188.85
6	658	По	1.41007	34.50	223.35
7	2606	земле	1.41007	34.50	257.85
8	4049	одним	1.41007	34.50	292.35
9	43	7	1.30755	31.98	324.31
10	44	8	1.30755	31.98	356.29
11	5460	сжат	1.30755	31.98	388.27



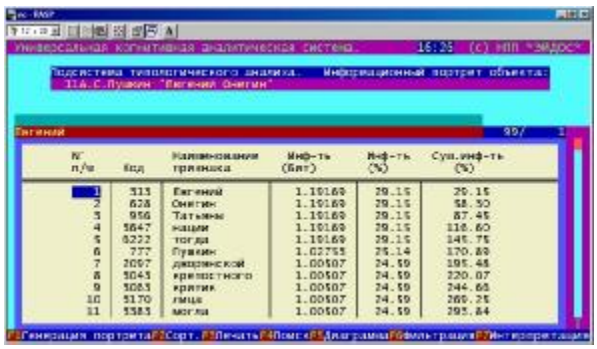
Подсистема типологического анализа. Информационный портрет объекта:
9Некрасов "Кому на Руси жить хорошо"

N п/п	Код	Наименование признака	Инф-ть (Бит)	Инф-ть (%)	Сум.инф-ть (%)
1	558	Савелий	0.98424	24.08	24.08
2	2111	дед	0.95454	23.35	47.45
3	4765	похмы	0.95454	23.35	70.78
4	787	Добросвйлова	0.90052	22.05	92.81
5	845	От	0.90052	22.05	114.84
6	2481	забыть	0.90052	22.05	136.87
7	2598	какие	0.90052	22.05	158.90
8	5145	лесу	0.90052	22.05	180.93
9	5467	мужики	0.90052	22.05	202.96
10	5587	народного	0.90052	22.05	224.99
11	4554	помехи	0.90052	22.05	247.02



Подсистема типологического анализа. Информационный портрет объекта:
10Пословицы

N п/п	Код	Наименование признака	Инф-ть (Бит)	Инф-ть (%)	Сум.инф-ть (%)
1	4716	пословица	1.59669	39.06	39.06
2	4781	праду	1.59669	39.06	78.12
3	17	1	1.41007	34.50	112.62
4	54	2	1.41007	34.50	147.12
5	56	3	1.41007	34.50	181.62
6	759	Правда	1.41007	34.50	216.12
7	59	4	1.38055	33.77	249.88
8	40	5	1.33912	32.76	282.65
9	41	6	1.33912	32.76	315.41
10	2045	да	1.32152	32.33	347.74
11	5571	самым	1.27765	31.26	379.00



Подсистема типологического анализа. Информационный портрет объекта:
11А.С.Пушкин "Евгений Онегин"

N п/п	Код	Наименование признака	Инф-ть (Бит)	Инф-ть (%)	Сум.инф-ть (%)
1	313	Евгений	1.19169	29.15	29.15
2	828	Онегин	1.19169	29.15	58.30
3	956	Татьяна	1.19169	29.15	87.45
4	5647	нации	1.19169	29.15	116.60
5	6222	тогда	1.19169	29.15	145.75
6	777	Пушкин	1.02755	25.14	170.88
7	2097	дворянской	1.00507	24.59	195.48
8	5043	крепостного	1.00507	24.59	220.07
9	5083	критик	1.00507	24.59	244.66
10	5170	лица	1.00507	24.59	269.25
11	5585	могла	1.00507	24.59	293.84

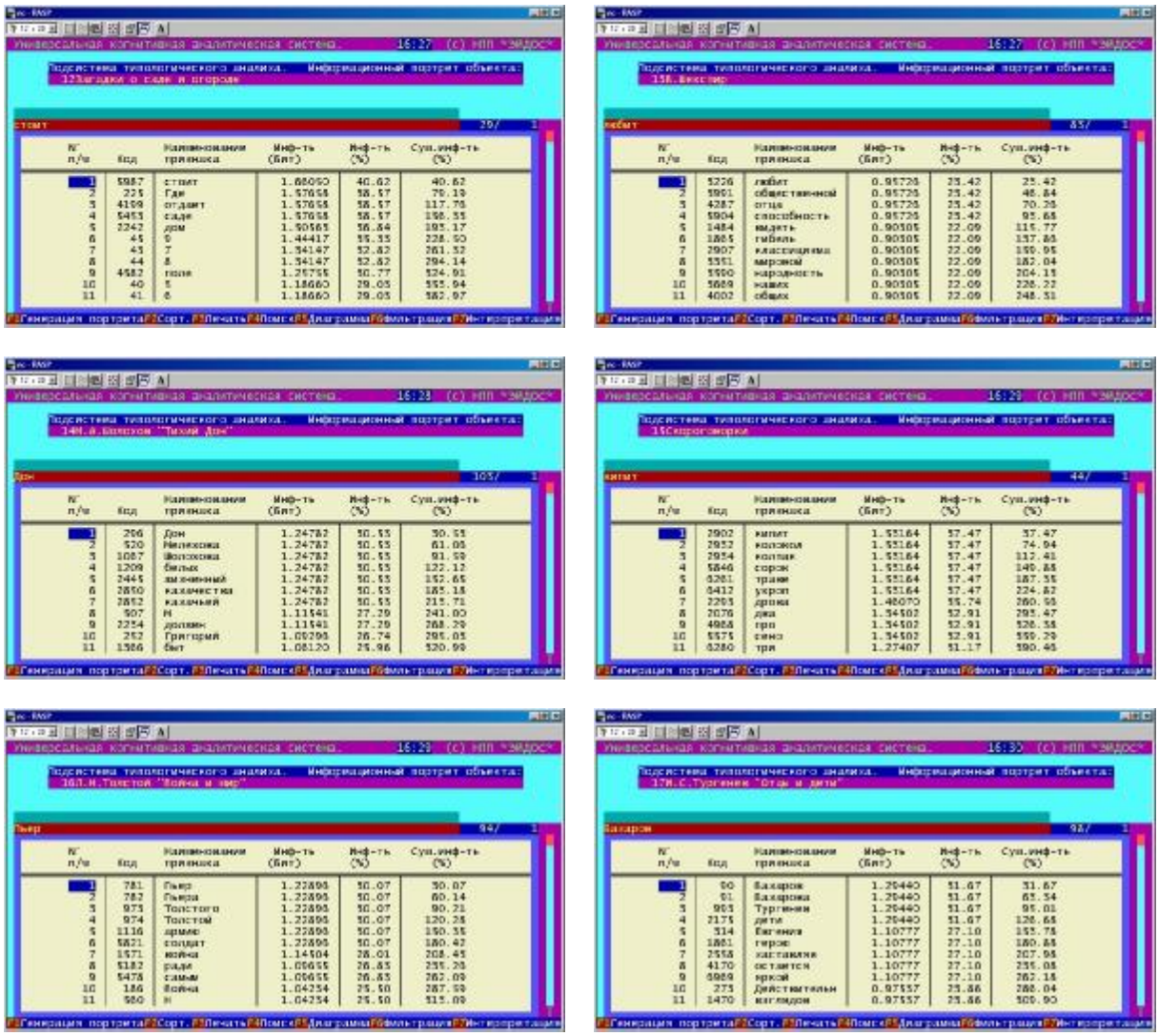


Рис. 18. Информационные портреты классов

10. Выполнить кластерно-конструктивный анализ модели

Кластерно-конструктивный анализ классов и признаков реализуется в 5-й подсистеме "Типология". В результате рассчитываются матрицы сходства классов и признаков, на основе которых генерируется и выводится ряд текстовых и графических форм. В данной статье мы приведем для примера лишь матрицу сходства классов (табл. 4) и отображающую ее в графической форме семантическую сеть классов (рис. 19).

Таблица 4. Матрица сходства классов

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	100,00	-9,55	-8,94	-8,16	-9,46	6,32	-10,08	38,11	-6,24	33,86	-4,34	41,96	-11,12	-8,17	11,70	-8,00	-11,05
2	-9,55	100,00	-5,11	-0,35	-2,47	-7,84	-2,95	-9,07	-5,50	-5,44	-6,47	1,20	-16,31	-5,43	-10,08	-8,26	2,04
3	-8,94	-5,11	100,00	-2,39	1,12	-3,97	-6,81	-5,61	-0,10	-6,56	-3,05	-10,13	-0,41	-7,36	-11,47	-6,04	-6,32
4	-8,16	-0,35	-2,39	100,00	2,52	9,73	1,33	-4,36	-6,23	-9,19	-4,94	-8,33	-9,55	-1,41	-9,14	-1,91	3,14
5	-9,46	-2,47	1,12	2,52	100,00	-8,31	-1,87	-5,24	-12,74	-1,25	-5,30	-4,55	-12,89	-8,85	-6,73	-9,59	-3,52
6	6,32	-7,84	-3,97	9,73	-8,31	100,00	-6,55	-5,05	-12,06	0,49	-7,34	-2,99	-15,19	-11,13	8,18	-3,10	-8,75
7	-10,08	-2,95	-6,81	1,33	-1,87	-6,55	100,00	-4,35	-1,04	-6,10	-10,14	-5,71	-7,09	-0,21	-9,40	-3,97	3,67

8	38,11	-9,07	-5,61	-4,36	-5,24	-5,05	-4,35	100,00	-2,38	34,04	-6,03	41,21	-6,48	-4,72	0,87	-8,50	-8,17
9	-6,24	-5,50	-0,10	-6,23	-12,74	-12,06	-1,04	-2,38	100,00	-1,85	-8,20	-6,28	-12,89	-1,18	-2,41	0,73	-3,53
10	33,86	-5,44	-6,56	-9,19	-1,25	0,49	-6,10	34,04	-1,85	100,00	-8,76	39,59	-9,83	-9,07	-1,63	-11,22	-7,73
11	-4,34	-6,47	-3,05	-4,94	-5,30	-7,34	-10,14	-6,03	-8,20	-8,76	100,00	-7,79	13,47	-3,96	-5,98	-11,77	-2,47
12	41,96	1,20	-10,13	-8,33	-4,55	-2,99	-5,71	41,21	-6,28	39,59	-7,79	100,00	-8,80	-8,13	5,09	-8,29	-5,24
13	-11,12	-16,31	-0,41	-9,55	-12,89	-15,19	-7,09	-6,48	-12,89	-9,83	13,47	-8,80	100,00	-3,67	-3,20	-1,92	1,77
14	-8,17	-5,43	-7,36	-1,41	-8,85	-11,13	-0,21	-4,72	-1,18	-9,07	-3,96	-8,13	-3,67	100,00	-11,07	-0,69	-3,25
15	11,70	-10,08	-11,47	-9,14	-6,73	8,18	-9,40	0,87	-2,41	-1,63	-5,98	5,09	-3,20	-11,07	100,00	-8,44	-12,23
16	-8,00	-8,26	-6,04	-1,91	-9,59	-3,10	-3,97	-8,50	0,73	-11,22	-11,77	-8,29	-1,92	-0,69	-8,44	100,00	-5,50
17	-11,05	2,04	-6,32	3,14	-3,52	-8,75	3,67	-8,17	-3,53	-7,73	-2,47	-5,24	1,77	-3,25	-12,23	-5,50	100,00

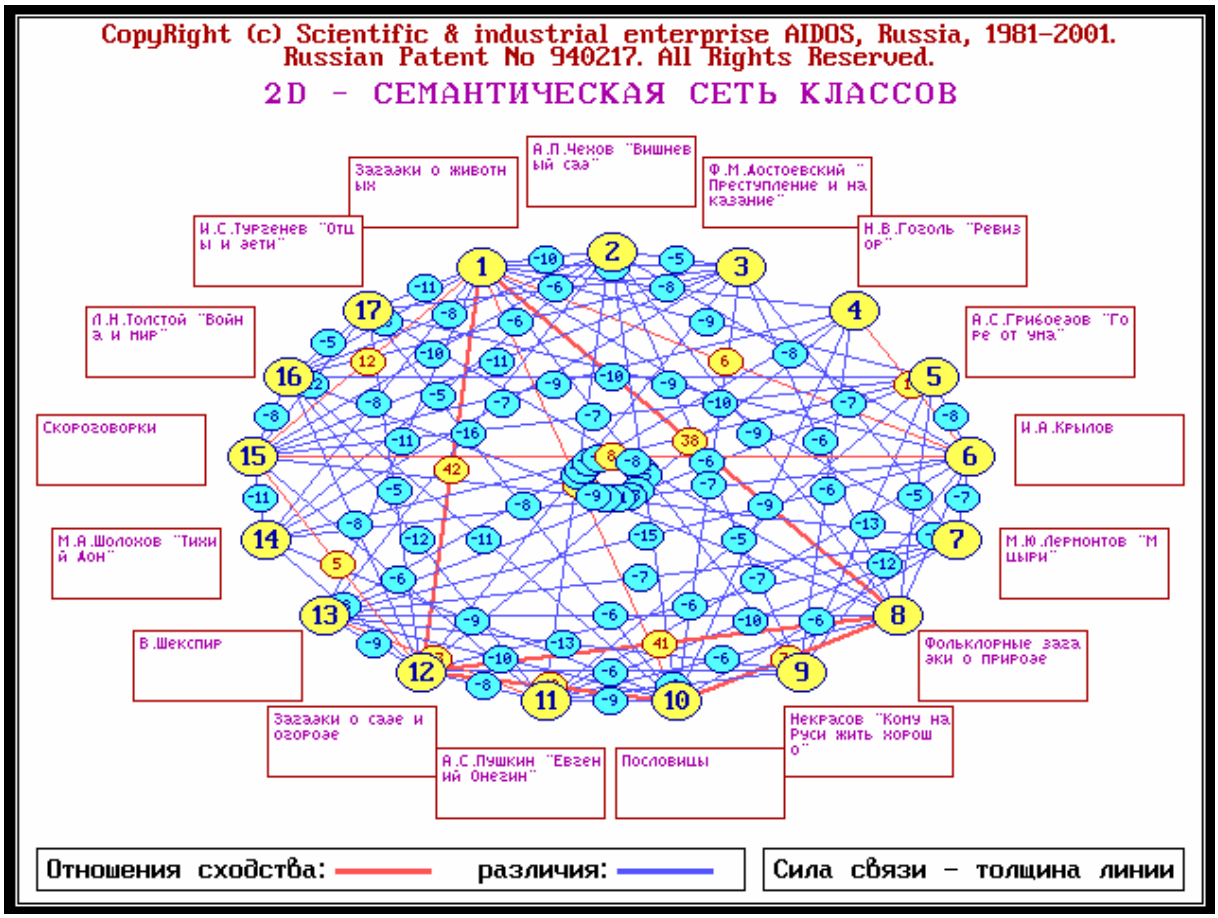


Рис. 19. Отображение матрицы сходства классов в графической форме семантической сети классов (отображены связи значимостью более 5 %)

Выводы

Продемонстрирована возможность и эффективность применения технологии и инструментария системно-когнитивного анализа для решения ряда задач атрибуции текстов.

Приведен подробный численный пример (с большим количеством конкретных иллюстративных материалов) реализации всех этапов СК-анализа при атрибуции текстов:

– когнитивной структуризации и формализации предметной области;

– формирования обучающей выборки;
– синтеза семантической информационной модели;
– оптимизации и измерения адекватности модели;
– адаптации и пересинтеза модели;
– типологического и кластерно-конструктивного анализа модели.

Статья может представлять интерес для специалистов по атрибуции и контент-анализу текстов на естественном языке. Материал может быть также использован в качестве руководства к лабораторной работе по дисциплине: "Интеллектуальные информационные системы".

Список литературы

1. Марусенко М.А. Атрибуция анонимных и псевдонимных литературных произведений методами теории распознавания образов. – Л.: ЛГУ, 1990. – 164 с.

2. Луценко Е.В. Теоретические основы и технология адаптивного семантического анализа в поддержке принятия решений (на примере универсальной автоматизированной системы распознавания образов "ЭЙДОС-5.1"). – Краснодар: КЮИ МВД РФ, 1996. – 280 с.

3. Луценко Е.В. Автоматизированный системно-когнитивный анализ в управлении активными объектами (системная теория информации и ее применение в исследовании экономических, социально-психологических, технологических и организационно-технических систем): Монография (научное издание). – Краснодар: КубГАУ, 2002. – 605 с.

4. Луценко Е.В. Атрибуция текстов как обобщение задач идентификации и прогнозирования // Научный журнал КубГАУ. – 2003.– № 2 (2). –19 с. http://ej.kubagro.ru.

5. Пат. № 2003610986 РФ. Универсальная когнитивная аналитическая система "ЭЙДОС" / Е.В. Луценко (Россия); Заяв. № 2003610510 РФ. Опубл. от 22.04.2003. – 50 с.