УДК 51-77

ВЫДЕЛЕНИЕ ИНФОРМАЦИИ И АНАЛИЗ ДАННЫХ ДЛЯ СЛАБОФОРМАЛИЗОВАННЫХ ПРЕДМЕТНЫХ ОБЛАСТЕЙ

М.А. Гузев[1)], Е.Ю. Никитина[2)]

[1)] Институт прикладной математики ДВО РАН

guzev@iam.dvo.ru

[2)] Дальневосточный федеральный университет

nikitina@imcs.dvgu.ru

Описан метод семиотического представления предметной области с помощью словаря знаков. Такое представление делает возможным проведение построения ранговых распределений для различных характеристик, используя подход В.П.Маслова. Решены задачи выявления влияющих факторов и анализа структуры распределения для криминологии и библиометрии.

Ключевые слова: ранговое распределение, закон Ципфа, частотный словарь.

Во многих слабо-формализованных предметных областях выделение информации из накопленных данных часто проводится с целью определения влияющих факторов, а также изучения поведения моделируемых объектов. Важным инструментарием в обработке данных являются статистические методы, но сегодня количество накопленных данных в различных областях человеческой деятельности настолько огромно, что экспертам трудно анализировать весь объём предоставляемой им информации за приемлемое для принятия решения время. Например, в криминологии широко используются методы статистического анализа, применяемые в социологии, психологии, биологии и психиатрии. Однако, выявление факторов, влияющих на преступность и личность преступника, имеет часто субъективный аспект, что приводит к неопределенности в анализе ситуаций. Следует также заметить, что актуальна проблема латентности преступности, тогда как для объективного анализа важно иметь чёткое представление об объёме и полноте анализируемых данных.

Поэтому актуальной является задача обработки информации для выявления объективных закономерностей при исследовании объектов различной природы с целью моделирования их поведения и поддержки процесса принятия решений. В данной работе предлагается подход к выделению информации, основанный на представлении массива данных как семиотической системы.

Рассмотрим некоторую предметную область и выделим отдельные классы объектов $\{O_1, O_2, ...O_k\}$, каждый из которых обладает некоторым набором характеристик, имеющих конкретные значения и зависящих от нескольких параметров. Значения таких характеристик представим набором $\{f_j(x_1, x_2, ...x_m)\}$, где f_j - имя характеристики, а $x_1, x_2, ...x_m$ – значения ее параметров, которые могут быть количественными или качественными. В предметной области выделяем множество факторов, которые могут влиять на значения характеристик объектов предметной области $\{\varphi_i(y_1, y_2, ...y_l)\}$, где φ_i - название фактора, а $y_1, y_2, ...y_l$ – значения его параметров. При этом значение параметра характеристики также может быть количественным или качественным. Каждый из выше введённых параметров x_i, y_i имеет некоторый содержательный смысл в предметной области.

Таким образом, предметная область характеризуется набором $S = \{f_i\} \cup \{\varphi_j\} \cup \{x_k\} \cup \{y_n\}$. Такое представление позволяет интерпретировать данные с позиций семиотики как знаковую систему, в которой информация порождается некоторым источником, жизненный период которого определяется объективными законами. Соответствующий набор S принято называть словарем, в котором первым элементом набора является термин словаря, остальные элементы задают значение характеристик некоторого объекта, обозначенного термином, а также информацию о том, когда и как было получено данное значение.

Соответствующие закономерности в семиотических системах исследовались многими авторами. Широко известен в научной литературе закон Ципфа, впоследствии уточнённый Мандельбротом [1]. Как известно, закон Ципфа,

описывает соотношение между частотой встречаемости слова - знаком - и его порядковым номером в словаре. В современной зарубежной научной литературе на это соотношение ссылаются как «power law». Анализ поведения различных систем показал, что для них также существуют закономерности ципфовского типа [2].

Проблема понимания внутренней природы ранговых распределений и возможностей их использования для анализа явлений в различных предметных областях представлена в работах академика В.П.Маслова [3]. Доказанная им теорема позволяет параметризовать набор моделей рангового распределения.

Используя идею подхода В.П. Маслова, мы исследовали задачи для криминологии, основным предметом изучения которой являются преступления и различные факторы, которые могут влиять на состояние преступности.

Преступление в описываемой модели — класс объектов предметной области. Характеристиками преступления можно считать состав преступления, место и время совершения преступления, мотив преступления, личность преступника, и т.д. Тогда информация о численности преступлений разного состава в разные годы представляется таблицей следующего вида:

Год	Состав преступления 1 типа	…	Состав преступления *s* типа
1926	Кол-во преступлений типа 1	…	Кол-во преступлений типа *s*
…	…	…	…

В терминах ранговых распределений знаком является преступление определенного состава, количество зарегистрированных преступлений одного состава является частотой, рангом является номер состава преступления в словаре составов преступлений. Аналогично описываются факторы, предположительно оказывающие влияние на состояние преступности.

Задача анализа представленной информации состоит в определении, какие из факторов влияют на преступность. Естественная гипотеза состоит в том, что фактор влияет на характеристику, если параметры их распределений совпадают. Была построена модель для криминологической ситуации в Японии [4]. В ней используется составленный словарь факторов, определенных экспертами, предположительно оказывающих существенное влияние на

преступность. Ранговый анализ позволил выявить наиболее значимые для описания преступности в этой стране факторы — уровень безработицы и количество потребляемых наркотических веществ (Рис. 1). При превышении количества зарегистрированных безработных уровня в 2,5 млн. человек (Рис. 2), наблюдается резкий рост количества зарегистрированных преступлений.

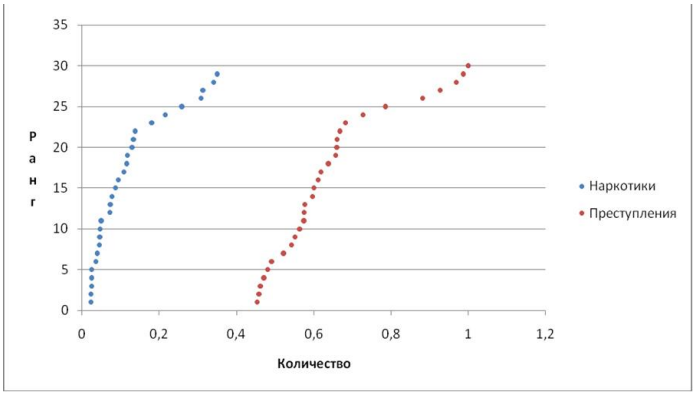

Рис. 1 Ранговое распределение преступлений и наркотических веществ

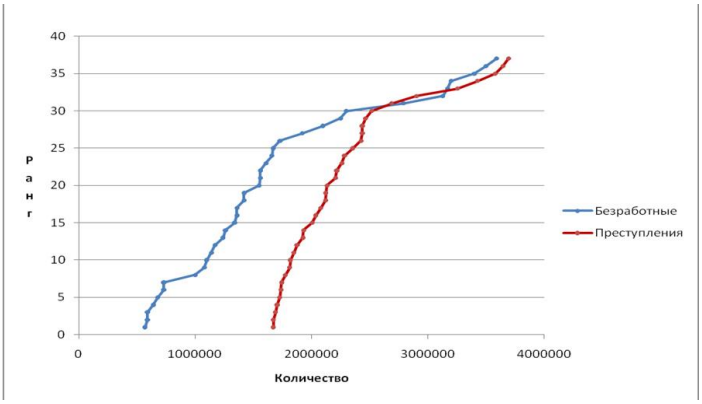

Рис. 2 Ранговое распределение количества безработных и преступлений

Модель ранговых распределений была использована для анализа полноты и системности структуры Уголовного Кодекса Российской Федерации (УК РФ) в различных редакциях [5]. Входными данными задачи является словарь, задающий названия мер наказаний и частоту упоминаний этих наказаний в статьях УК РФ. Было установлено, что словарь наказаний 1996 года содержит всего 23 различных меры наказаний, причем частот среднего и большого ранга соответствующих тяжёлым наказаниям, связанным с лишением свободы на длительные сроки — большинство (рис.3). Словарь наказаний 2009 года лишен указанного недостатка — различных мер наказаний 37 и распределены они более равномерно по количеству наказаний разной тяжести (рис.4).

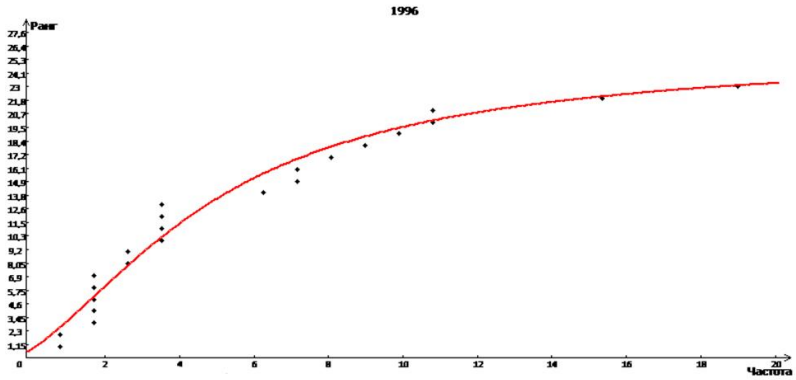

Рис. 3. Зависимость частоты наказания от ранга для УК РФ 1996 г.

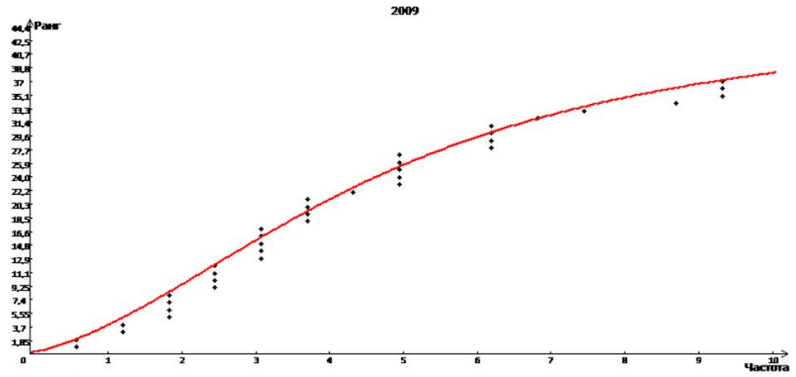

Рис. 4. Зависимость частоты наказания от ранга для УК РФ 2009 г.

Следует указать, что модель описывает объективную ситуацию, полностью адекватную действиям, направленным на изменение УК, проводимым законотворческими органами РФ.

Следующим примером применения метода извлечения информации является изучение публикационной активности научных работников институтов, входящих в состав ДВО РАН. Было проведено построение ранговых распределений данных о количестве научных публикаций в институтах ДВО РАН (всего 33 единицы) за период 2005-2009 гг. Был составлен словарь институтов, который ранжировался по количеству научных работников и по количеству публикаций. Установлено, что общее количество научных публикаций находится в прямой связи с количеством научных сотрудников, занимающих финансируемые РАН штатные ставки. Из Рис.5-6 видно, что увеличение объема финансирования научно-исследовательских институтов в 2007 году привело к увеличению общего количества публикаций.

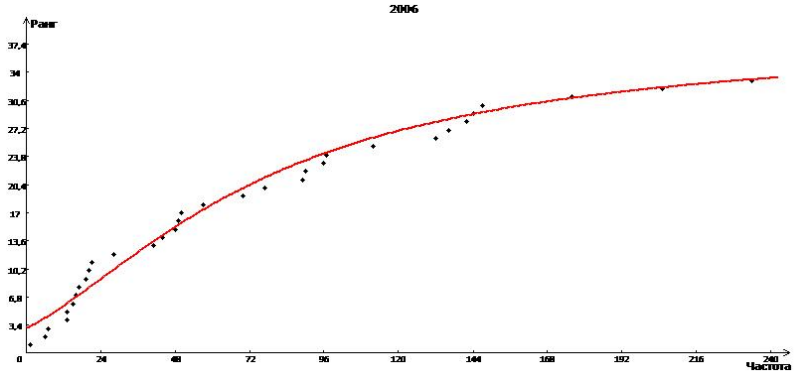

Рис. 5 Общее количество публикаций в 2006 году

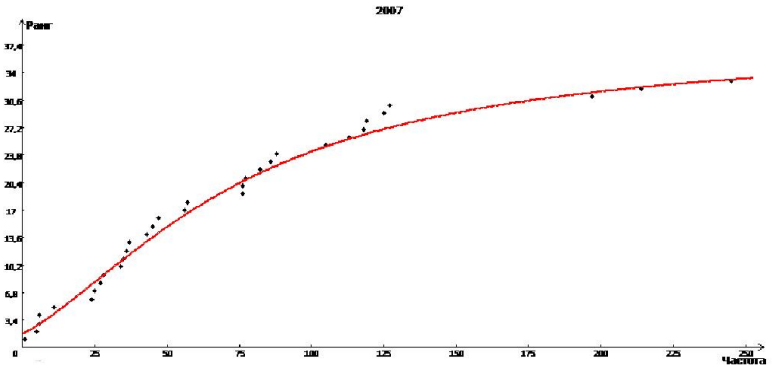

Рис. 6 Общее количество публикаций в 2007 году

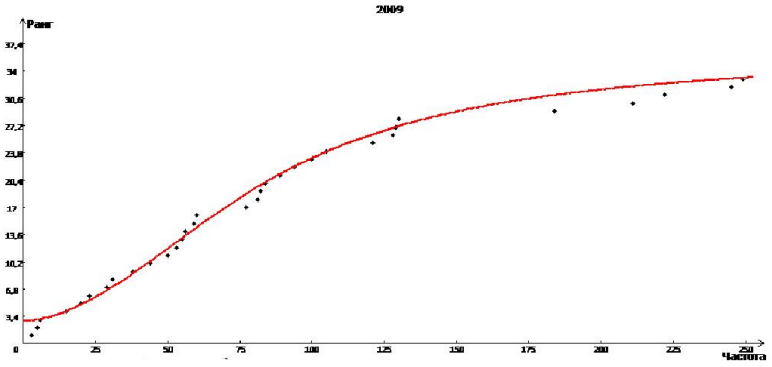

Рис. 7 Общее количество публикаций в 2009 году

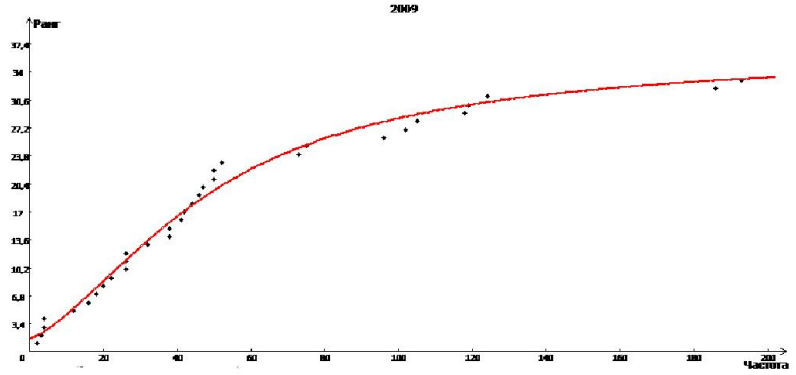

Рис. 8 Количество статей в рецензируемых журналах в 2009 году

Анализ распределений общего количества публикаций и публикаций в рецензируемых журналах показывает неравномерное распределение вторых, на графике видны лакунарные области, где должны были располагаться данные, соответствующие общей публикационной активности для этих рангов. Это означает, что в таких институтах необходимо предпринимать шаги, стимулирующие научных сотрудников к таким публикациям (Рис. 7-8).

Количество публикаций в Web of Science подчиняется другой зависимости (Рис. 9), что указывает на то, что количество цитирований публикаций в зарубежных источниках не зависит от численности научных сотрудников.

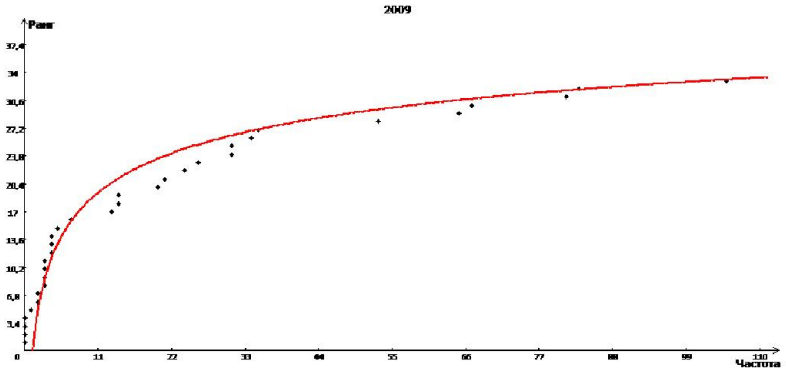

Рис 9 Количество публикаций в Web of Science в 2009

Подводя итог, можно утверждать, что построение ранговых словарей для описания предметной области позволяет выявить закономерности в поведении семиотической системы. На практике описанный подход можно применять для проведения анализа структуры распределений наборов данных, подгонки и исследования поведения параметров распределения для влияющих факторов, выдачи рекомендаций по корректировке и получению новых данных.

Список литературы

1. *Мандельброт Б.* Фрактальная геометрия природы // Институт компьютерных исследований, М., 2002 г.

2. *Clauset A., Shalizi C.R. and Newman M.E.J.* Power-law distributions in empirical data. E-print (2007). http://cse.nd.edu›courses/cse40768/www/Handout4.pdf (15.09.2011)

3. Персоналии. Маслов Виктор Павлович // База данных Math-Net.Ru URL: http://www.mathnet.ru/ (25.05.2011).
4. *М.А.Гузев, Е.Ю.Никитина* Применение математических методов при исследовании криминологических данных (на примере Японии) // Россия и АТР, № 2 (2009). С. 77-85.
5. *М.А.Гузев, Е.Ю.Никитина* Ранговый анализ Уголовного Кодекса РФ (на примере экономических преступлений) // Дальневост. матем. журн., 10:2 (2010). С. 117–129

M.A.Guzev

E.Y.Nikitina

ALLOCATION OF THE INFORMATION AND THE ANALYSIS OF THE DATA FOR POORLY FORMALIZED SUBJECT DOMAINS

The method of semiotics representation of subject domain by means of the dictionary of signs is described. In these terms it is possible to perform rank distributions for various characteristics using V.P.Maslov's approach. Problems of revealing of influencing factors and the analysis of structure of distribution in criminology and bibliometrics are solved.

Keywords: rank distribution, Zipf's law, frequency word book.