Несколько лучшими свойствами обладают секторные интервалы. Хотя перенос на них теоремы Бека-Никеля в прямом виде тоже неверен, существует возможность построения её аналога для оценки значений функций, пусть и осложнённого свойствами операций с самими секторными интервалами. Утверждается, что алгоритм SIVIA допускает перенос на случай этого базового объекта с модификацией порядка дробления из-за того, что комплексные числа не представляют собой упорядоченного поля.

Наибольшим удобством с точки зрения оценок функций на них из описанных объектов обладают круговые интервалы, сохраняющие все важные для этой части алгоритма полезные свойства действительных. К сожалению, они же обладают наименьшим удобством построения разбиений. Выпуклость круговых интервалов обеспечивает возможность оценки близости различных разбиений и покрытий множества, но построение разбиений множества на основе круговых интервалов утрачивает свою вычислительную лёгкость.

Библиографический список

1. Candau Y., Raissi T., Ramdani N., Ibos L. Complex interval arithmetic using polar form – Reliable Computing. 2006. №1

2. Жолен Л., Кифер М., Дидри О., Вальтер Э. Прикладной интервальный анализ – М.-Ижевск: Институт компьютерных исследований, 2007.

3. Дронов В.С. Об аналогах теоремы Бека-Никеля в комплексном случае // МАК 2012 : материалы пятнадцатой конференции по математике. – Барнаул: Изд-во. Алт. ун-та, 2012.

УДК 519.237.7

Анализ данных Российского индекса научного цитирования интервальным методом главных компонент

С.И. Жилин, П.А. Ледомский
АлтГУ, г. Барнаул

Метод главных компонент (МГК) широко используется при обработке многомерных экспериментальных данных для первичного разведочного анализа, а также в задачах распознавания образов для понижения размерности признакового пространства, устранения мультиколлинеарности и шумов в данных [1]. В практике моделирования не-

редко приходится сталкиваться с необходимостью обработки интервальных экспериментальных данных. При этом интервальная форма данных может находить существенно отличающиеся интерпретации и использоваться в различных целях.

Одной из проблем, в преодолении которых может помочь интервальная форма данных, является визуальное представление результатов МГК-моделирования при обработке выборок значительного объема. Традиционный подход состоит в отображении каждого из объектов выборки в виде точки в той или иной системе координат. Именно в этой манере строится один из наиболее полезных при МГК-анализе графиков – график счетов. График счетов отражает взаимное расположение объектов выборки при их проецировании на плоскость, образованную парой тех или иных главных компонент, и именно он позволяет исследователю составить представление о структуре данных, наличии выбросов и некоторых закономерностей в размещении объектов в признаковом пространстве. Однако подобный анализ существенно затрудняется или становится совершенно невозможным при работе с выборками из тысячи и более объектов, которые загромождают график многочисленными наложениями.

Естественный выход из этой ситуации может состоять в некотором агрегировании исходных данных и последующем анализе этих агрегатов. Конечно, такой анализ может терять в детальности выводов, но, тем не менее, позволяет выявить некоторые крупномасштабные тенденции и структурные свойства совокупности данных, на поиск которых зачастую и направлены усилия аналитика.

Идея анализа данных, предварительно агрегированных в той или иной форме, положена в основу подхода, именуемого «анализ символьных данных» (Symbolic Data Analysis, SDA) [2]. Брус (интервальный вектор) является одним из наиболее простых и наглядных способов обобщения данных и широко употребляется в SDA. Брус вбирает в себя как информацию о центральной тенденции, так и о степени разброса вокруг нее агрегируемых наблюдений. При этом способы построения бруса по совокупности объектов в признаковом пространстве не сводятся исключительно к охвату точек интервальной оболочкой, но могут быть и иными, например, намеренно огрубляющими получаемое представление.

Для анализа интервальных данных, построенных агрегированием точечных, известно несколько вариантов метода главных компонент: центровой МГК (Centers PCA, CPCA), вершинный МГК (Verteces PCA, VPCA) и «полноинформационный» (Complete-Information PCA, CIPCA) [3]. Названия этих методов отражают используемые в них

приемы сведения анализа интервальных данных к классическому точечному МГК. Так CPCA и VPCA предписывают для этого представлять брусы точками, являющимися их центрами и вершинами соответственно. Метод CIPCA обладает более высокой описательной способностью в сравнении с CPCA и VPCA, поскольку не отождествляет интервальные наблюдения с отдельными точками, а задействует интегральную информацию об их внутренности, полагая брусы совокупностью равномерно распределенных бесконечно плотных точек. Следствием этой гипотезы являются отличия в способе вычисления скалярного произведения интервальных векторов и анализируемой далее ковариационной матрицы.

Все три указанных варианта МГК для анализа интервальных данных реализованы авторами в виде высокоуровневого решателя с использованием Java-библиотеки для интервальных вычислений JInterval [4]. С использованием решателя анализу методом CIPCA были подвергнуты данные, предоставляемые информационно-аналитической системой Российский индекс научного цитирования (РИНЦ) [5], аккумулирующей публикации российских авторов и информацию об их цитировании.

В исходную выборку была включена информация о 929 журналах, входящих в список рекомендуемых ВАК и представленных в РИНЦ сведениями за 2010 год (последний из тех, для которых рассчитан двухлетний импакт-фактор). Каждый из журналов характеризуется четырьмя показателями: общим количеством публикаций за год (Пуб), средним количеством ссылок в публикации (Слк), двухлетним импакт-фактором РИНЦ (Имп), общим количеством цитирований публикаций журнала (Цит). Журналы были объединены в 8 категорий (таблица 1), представляемых брусами, порожденными декартовым произведением интерквартильных интервалов основных показателей (таблица 2).

Таблица 1. Категории журналов, индексируемых РИНЦ

Номер категории	Категория журналов	Количество журналов	Коды ГРНТИ
1	Физико-математические науки	107	2, 28, 33, 57
2	Химические науки	28	61
3	Науки о жизни	355	3, 15, 30, 39, 51, 58
4	Науки о Земле	59	9–12
5	Технические науки	230	1, 4, 7, 13, 16, 22, 24, 25, 29, 31, 32, 40–43, 45, 49, 50, 53, 55, 56, 60, 63, 64–66
6	Науки об информации	21	17, 20
7	Управление и экономика	117	38, 46, 54, 62
8	Комплексные исследования	12	21, 34, 37

Таблица 2. Интерквартильные интервалы показателей РИНЦ

Номер категории	Количество публикаций	Количество ссылок	Импакт-фактор РИНЦ	Количество цитирований
1	[36.00, 111.50]	[9.925, 17.700]	[0.1065, 0.32675]	[18.25, 421.75]
2	[85.50, 170.00]	[13.000, 22.200]	[0.2380, 0.50150]	[133.50, 1696.00]
3	[39.00, 125.75]	[8.125, 21.175]	[0.0590, 0.27175]	[17.00, 216.25]
4	[29.25, 95.75]	[7.875, 21.725]	[0.1095, 0.44150]	[27.50, 422.50]
5	[40.00, 128.00]	[4.300, 8.100]	[0.0650, 0.18400]	[18.00, 140.00]
6	[33.25, 121.25]	[6.400, 11.950]	[0.0590, 0.27775]	[10.75, 111.00]
7	[35.75, 147.25]	[3.475, 8.500]	[0.0260, 0.21850]	[7.75, 72.50]
8	[43.50, 114.00]	[8.150, 19.650]	[0.0415, 0.17300]	[4.50, 128.50]

Две первые главные компоненты для стандартизованных интервальных данных (рис. 1) объясняют более 77% общей вариации данных и могут интерпретироваться как «показатель цитируемости» и «объем журнала». На графике счетов в этих координатах (рис. 2)

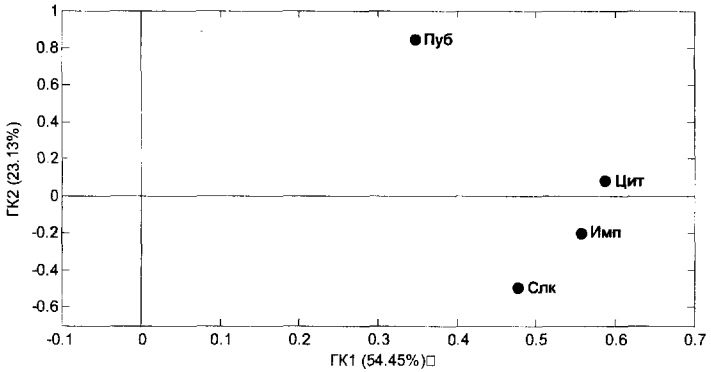

Рис. 1. График нагрузок для данных о журналах, индексируемых в РИНЦ. Пуб – количество публикаций; Слк – количество ссылок в статье; Имп – импакт-фактор РИНЦ; Цит – общее количество цитирований

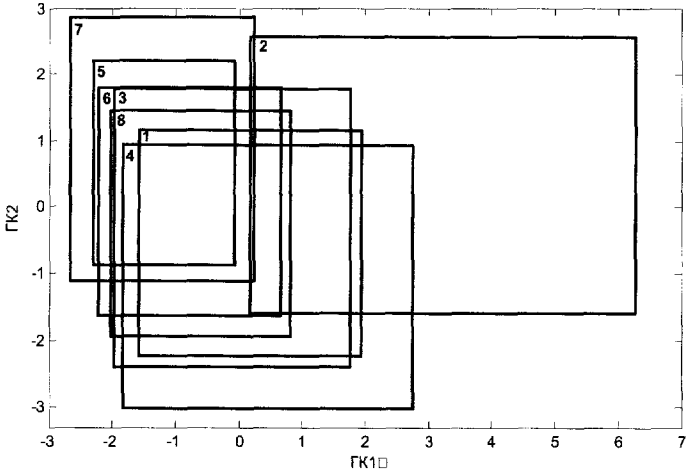

Рис. 2. График счетов для обобщенных категорий журналов, индексируемых в РИНЦ:
1 – физико-математические науки; 2 – химические науки; 3 – науки о жизни; 4 – науки о Земле; 5 – технические науки; 6 – науки об информации; 7 – управление и экономика; 8 – комплексные исследования

каждая из категорий представлена прямоугольником, центр которого соответствует усредненному значению фактора, а размер – разбросу журналов в данной категории. Среди множества содержательных выводов, вытекающих из графика счетов, здесь отметим лишь заметное превышение и разброс цитируемости химических журналов над прочими, а также контрастирующие с ними низкие значения цитируемости журналов из категории «Управление и экономика» при наибольших объемах этих журналов.

Библиографический список

1. Эсбенсен К. Анализ многомерных данных. Избранные главы. – Барнаул: Изд-во Алт. ун-та, 2003.

2. Billard L., Diday E. Symbolic Data Analysis: Conceptual Statistics and Data Mining. Chichester, John Wiley & Sons, 2006.

3. Wang H., Guang R., Wu J. CIPCA: Complete-Information-based Principal Component Analysis for interval-valued data // Neurocomputing. – 2012. – V. 86. – P. 158–169.

4. Nadezhin D.Ju, Zhilin S.I. JInterval Library: Principles, Development, and Perspectives // 15th GAMM-IMACS International Symposium on Scientific Computing, Computer Arithmetics and Verified Numerics SCAN-

2012, September 23–29, 2012. Book of Abstracts, Novosibirsk, Institute of Computational Technologies, 2012. – P. 117–118.

5. Научная электронная библиотека. – http://elibrary.ru.

УДК 004.89

Преимущества использования Акторного Пролога для реализации семантического поиска

О.Н. Половикова

АлтГУ, г. Барнаул

Одной из основных задач концепции Semantic Web является решение проблем, связанных с индексацией и поиском информации по смысловому содержанию. Практическая реализация семантических поисковых систем напрямую связана с разработкой и использованием специализированных языков для встраивания знаний непосредственно в сам документ, либо для создания отдельных от ресурса описаний-заменителей. Данные языки призваны обеспечивать реализацию всех компонентов модели поиска: способа представления информационных документов (или их заменителей), способа формирования запросов, критерия релевантности web-документов запросу.

При поступлении в систему пользовательского запроса для него также строится соответствующее представление, а метод его построения аналогичен методу построения представлений документов. Логический вывод позволит достроить необходимые цепочки метаданных хранимых документов, процесс поиска будет заключаться в построении соответствий между сравнимыми контентами.

Разметка документов с помощью метаформатов или онтологических терминов позволит производить автоматическую обработку их семантического содержания. Среди специальных язык запросов, которые умеют работать с семантическим содержанием, следует выделить SPARQL и RDF Query, которые базируются на обработке направленных графов (RDF-графов).

Построение базы знаний термов, описывающих знания нескольких информационных ресурсов, может быть реализовано логическими языками программирования, например Акторным Прологом. Программы-агенты объектно-ориентированного Акторного Пролога позволяют извлекать данные из документов, опубликованных в сети Интернет, посредствам предопределённого класса Receptor, преобразовывать их в различного рода термы (множества, списки, миры, структуры