

УДК 001.18

Методика исследования динамики научно-технических циклов

06, июнь 2012

Месропян В.Р.

Студент,

Кафедра «Компьютерные системы автоматизации производства»

Научный руководитель: Овсянников М.В.

к.т.н. , доцент кафедры «Компьютерные системы автоматизации производства»

МГТУ им. Н.Э. Баумана

mesr0pyan@yahoo.com

В свете социально-политических тенденций последних пяти лет в нашей стране проблемы эффективности НИОКР и определения приоритетных направлений развития становятся наиболее актуальными. Для того чтобы стимулирование научно-технического потенциала страны представляло собой сбалансированный и последовательный комплекс мероприятий и государственных программ, необходимо построить или принять на вооружение существующие методики создания дорожных карт и сценариев развития науки и техники. Западный опыт реализации подобных проектов построения научно-технических прогнозов, привел к появлению обширного спектра методов прогнозирования, объединенных методологией форсайта. Исследование, освещенное в статье, имеет своей целью разработку методики прогнозирования научно-технических циклов на основе анализа библиометрической информации. Следует отметить, что данное направление анализа технической информации относится к количественным методам форсайта и развивается в последнее время наиболее интенсивно, о чем свидетельствует активность мирового академического сообщества.

Для удобства дальнейшего повествования введем следующие термины:

- Интеллектуальная среда – это самоорганизующаяся и управляемая распределено совокупность интеллектуального информационного пространства (далее – ИИП), организационных ресурсов (субъектов: исследователей, вендоров, и т. д.) и рынка высоких технологий;

- ИИП - совокупность современных электронных ресурсов, аккумулирующих информацию о результатах научно-исследовательской деятельности всего мирового академического сообщества.

В открытом доступе в сети интернет существует огромное количество сервисов, предоставляющих любому желающему доступ к базам данных научных публикаций специализированных журналов, сборников статей конференций и патентных агентств. Средства эти отличаются разнообразием форм и простой доступностью, а так же, что наиболее важно, предоставляют возможность аналитикам и разработчикам по всему миру обращаться через стандартные интерфейсы к содержимому своих баз данных и использовать результаты запросов для автоматизированного анализа. За последние несколько лет на базе ведущих научно-технических издательств и электронных ресурсов были разработаны эффективные информационно-аналитические инструменты. Отличительной особенностью данных ресурсов является их доступность в режиме онлайн в сети интернет, а так же встроенные гибкие инструменты для анализа. Наибольшей популярностью сегодня обладают следующие системы с платной подпиской и разнообразными модификациями, удовлетворяющими потребностям разных групп пользователей: система SciVal фирмы Elsevier и система ESSENTIAL SCIENCE INDICATORS фирмы Tomson Reuters. Единственный в своем роде проект, построенный на основе веб-ориентированной клиент-серверной технологии, – система STING [1], создана в Институте компьютерных технологий Греции. Данная разработка позволяет проводить специализированный поиск по базам данных патентных агентств, строить выборки в различных разрезах, проводить их кластеризацию, а так же вводит оригинальный набор индикаторов (индикатор развития анализируемых технологий в разрезе стран, ученых и изобретателей; индикатор использования технологии в разрезе отраслей промышленного применения и т д). Однако ее применение ограничивается, по ряду объективных причин, только экспертным сообществом инвестиционных аналитиков в достаточно редких случаях. В свою очередь, в данной работе предложены подходы к обработке данных по принципиально иной методике в рамках концепции жизненного цикла технологий, комбинированное использование всех источников интеллектуальной среды (патенты, статьи, публикации и другие) и реализацию результатов построения индикаторов инновационных трендов для выстраивания приоритетных направлений деятельности.

Прибегнув к использованию распространенных сервисов доступа к интеллектуальной среде [2], был проделан первичный анализ данных с целью выяснения

77-51038/467498

характера развития технических тенденций в данной среде. Используя патентную базу данных Global Patent Index (далее – GPI), доступ к которой предоставляет, как в пользовательском режиме на официальном сайте [3], так и в режиме открытого сервиса для разработчиков (SOAP/WSDL), европейское патентное агентство, была получена выборка патентов, относящихся к технологии ЭЛТ (в англоязычной среде – Cathod Ray Tube). Затем отсортировали выборку по времени публикации патентов и построили гистограмму (Рисунок 1).

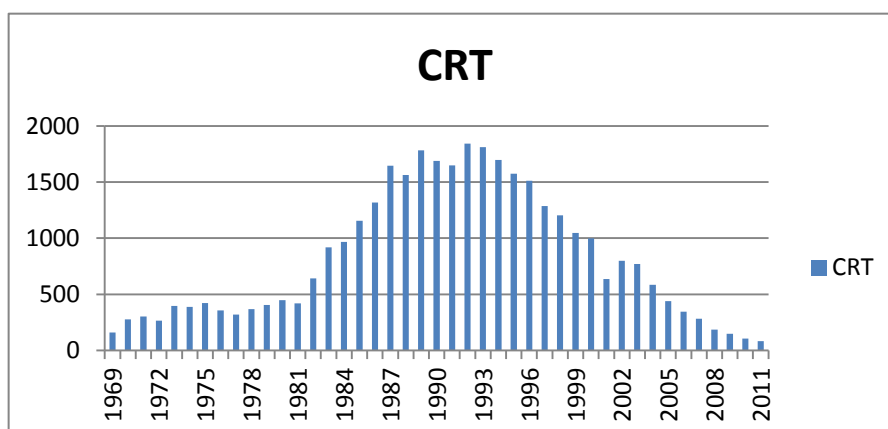


Рисунок 1. Распределение патентов, относящихся к технологии ЭЛТ

На данном примере отчетливо видно, что развитие технической идеи электронной лучевой трубки имело ясно выраженный восходящий тренд до 1989 года и устойчивый нисходящий тренд с 1992 года. Далее посредством аналогичных операций была построена гистограмма выборки по технологии жидкокристаллического дисплея (LCD), которая имеет тот же характер что и приведенная выше, но находится в начале нисходящего тренда. В общем и целом на сегодняшний день уже проанализировано порядка 100 полученных таким способом эквидистантных временных рядов, представляющих собой ежегодный прирост количества патентов в базе GPI, и все они имеют характерный тренд. Здесь немало важным будет сразу оговорить факторы задания критериев поиска: они должны быть достаточно ясными и прозрачными для выделения «чистой» выборки, дабы не происходило аддитивное наложение двух и более семантико-лексически близких совокупностей. Так же подчеркнем, что во всех вышеперечисленных экспериментах производился полнотекстовый поиск в обязательном разделе патентов «Abstract», т е в кратком содержании.

Интересна сама форма гистограммы временного ряда ЭЛТ, очевидно, что она отвечает вполне логичным соображениям о развитии рынка товаров и услуг. И даже вписывается в эконометрическую модель [4] удовлетворения спроса и жизненного цикла

технологий, введенного великим русским экономистом Н.Д. Кондратьевым [5] (Рисунок 2).

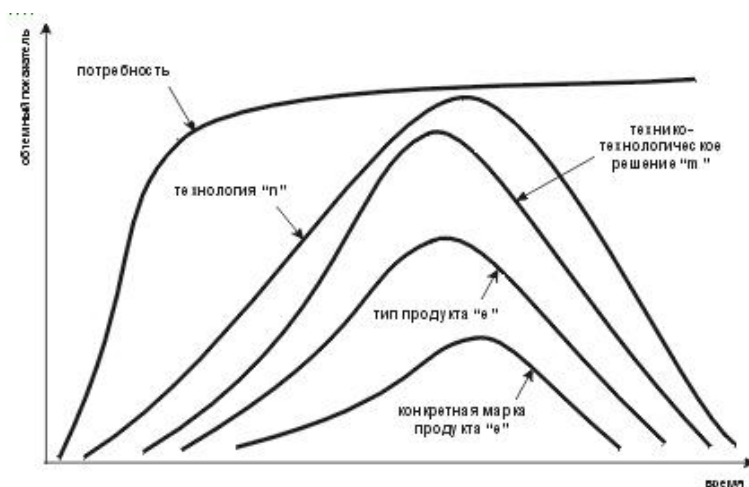


Рисунок 2. Модель жизненного цикла спроса и технологий

Перейдем непосредственно к вопросам анализа эквидистантных временных рядов, полученных обработкой выборок. Поскольку они могут так же трактоваться как сгруппированные в интервалы выборки наблюдений случайной величины, то в первом приближении нами были выполнены проверки статистической гипотезы (далее – H_0) об их принадлежности нормальному закону распределения случайных величин. Данные тесты проводились по критерию согласия хи-квадрат с помощью статистических функций автоматизированной проверки гипотез в программной среде MS Excel, и во всех случаях на уровне значимости $\alpha=0.95$ была доказана невозможность отвергнуть гипотезу H_0 . По результатам данного эксперимента следующим шагом стало решение применить по отношению к известным временным рядам элементы одномерного регрессионного анализа, а именно найти аппроксимативную гладкую параметрическую функцию. Сразу заметим, что искать её предполагалось среди гауссового семейства моделей (1) по аналогии с известной функцией плотности вероятности распределения нормального закона.

$$f = a \times \exp\left(-\frac{(x - b)^2}{c}\right); (1)$$

Однако в ходе экспериментов были проверены и другие типы гладких параметрических функций (например, линейные комбинации синусоид, экспонент, прямых), но результаты оказывались значительно хуже как с точки зрения оптимизации

функционала суммы квадратов невязок (метод наименьших квадратов), так и статистического распределения величин невязок относительно подобранной функции.

Далее приведен графический результат аппроксимации выше озвученного эквидистантного временного ряда, относящегося к технологии ЭЛТ, функцией Гаусса в модуле Curve Fitting Tool системы MATLAB с помощью алгоритма Левенберга-Марквардта [6] (Рисунок 3):

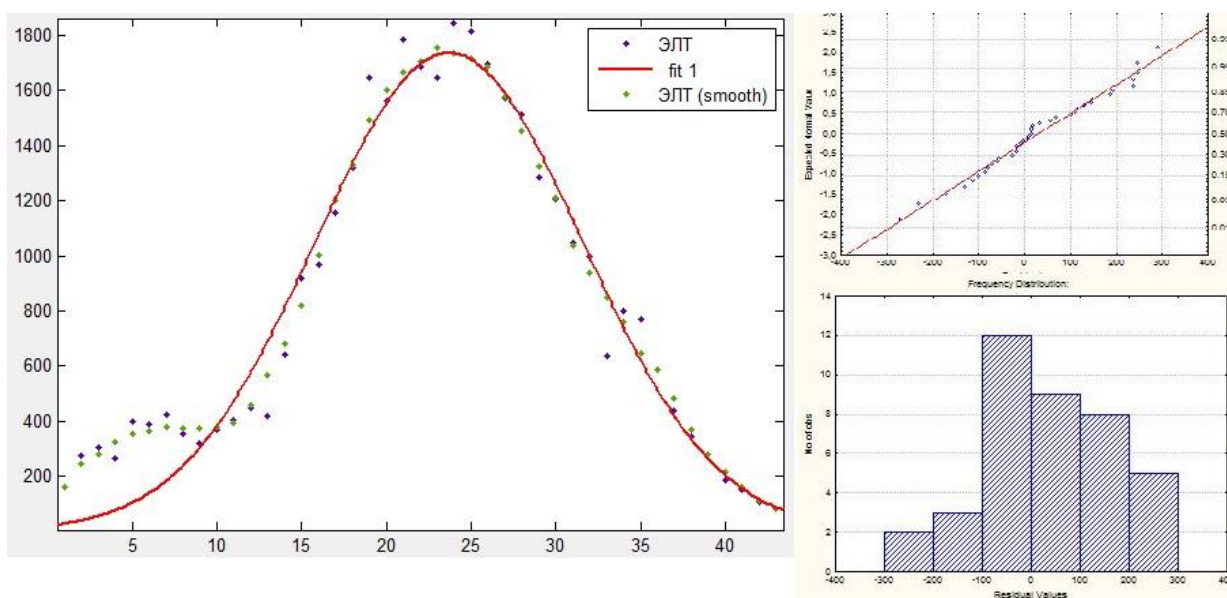


Рисунок 3. Аппроксимация ряда ЭЛТ функцией Гаусса. Распределение остатков модели

Так же получены оценки искомых параметров функции на уровне достоверности 95% и границы их доверительных интервалов: $a=1735(1646, 1825)$; $b=23.7(23.24, 24.17)$; $c=11.15(10.48, 11.82)$. В данном эксперименте для удобства вычисления была проведена замена внутри временного интервала (1969;2011) на интервал (1;43), таким образом, пользуясь тем, что унимодальная функция Гаусса имеет один экстремум в точке, заданной значением $x=b_1=23.7$, можно определить, что смена тренда произошла в 1991 году. А значение параметра $a_1=y_{\max}=1735$ равно величине экстремума данной математической модели. В правой части рисунка 5 показано распределение отклонений от модели аппроксимации относительно нормального распределения и их гистограмма. Модель показывает хорошие результаты, так как остатки с большой вероятностью распределены по нормальному закону с математическим ожиданием равным нулю. Аналогично были построены 25 аппроксимативных моделей различных выборок и все они выявили преимущество выбранной модели над другими распространенными гладкими параметрическими функциями. По результатам эксперимента нами была выдвинута гипотеза о том, что чистые эквидистантные временные ряды, извлеченные из базы GPI, с

77-51038/467498

достаточной точностью аппроксимируются функцией Гаусса и они инвариантны к линейным\аффинным преобразованиям.

Для подтверждения гипотезы для всех 25 выше озвученных аппроксимативных моделей были проведены эксперименты по их оптимизации с помощью открытого программного обеспечения MVR Composer реализованного и исполняемого в среде MATLAB [7]. Данный продукт изобретен и развивается специалистами Вычислительного центра РАН. Суть работы алгоритма заключается в том, что для экспериментального временного ряда задается начальное множество моделей аппроксимации, требующих уточнения, и множество гладких параметрических функций, из которых строятся аппроксимации нового поколения посредством механизмов мутации и естественного отбора. Алгоритм генерирует новые поколения аппроксимативных моделей, в которые входят лучшие модели из предыдущих поколений и мутировавшие потомки в определенном соотношении, до выполнения критериев останова. Данный алгоритм фактически является развитием хорошо известного метода группового учета аргументов, однако он много более эффективен, так как использует развитые технологии генетических алгоритмов при составлении композиций функций потомков. В результате эксперимента было получено 25 пять оптимизированных моделей, в которых функция Гаусса осталась наиболее чувствительным элементом суперпозиции функций. По результатам численного моделирования была принята гипотеза об аппроксимации эквидистантных рядов из базы GPI функциями Гаусса и о том, что данная аппроксимация инвариантна к аффинным преобразованиям.

Для прогнозирования развития тенденций разработан алгоритм (Рисунок 4), который обеспечивает приемлемые оценки параметров функции аппроксимации (особенно «a» и «b») на начальных этапах формирования тенденции, а, следовательно, позволяющий строить валидные прогнозы. Алгоритм будет рассмотрен на примере LCD технологии и ее предшественницы – ЭЛТ:

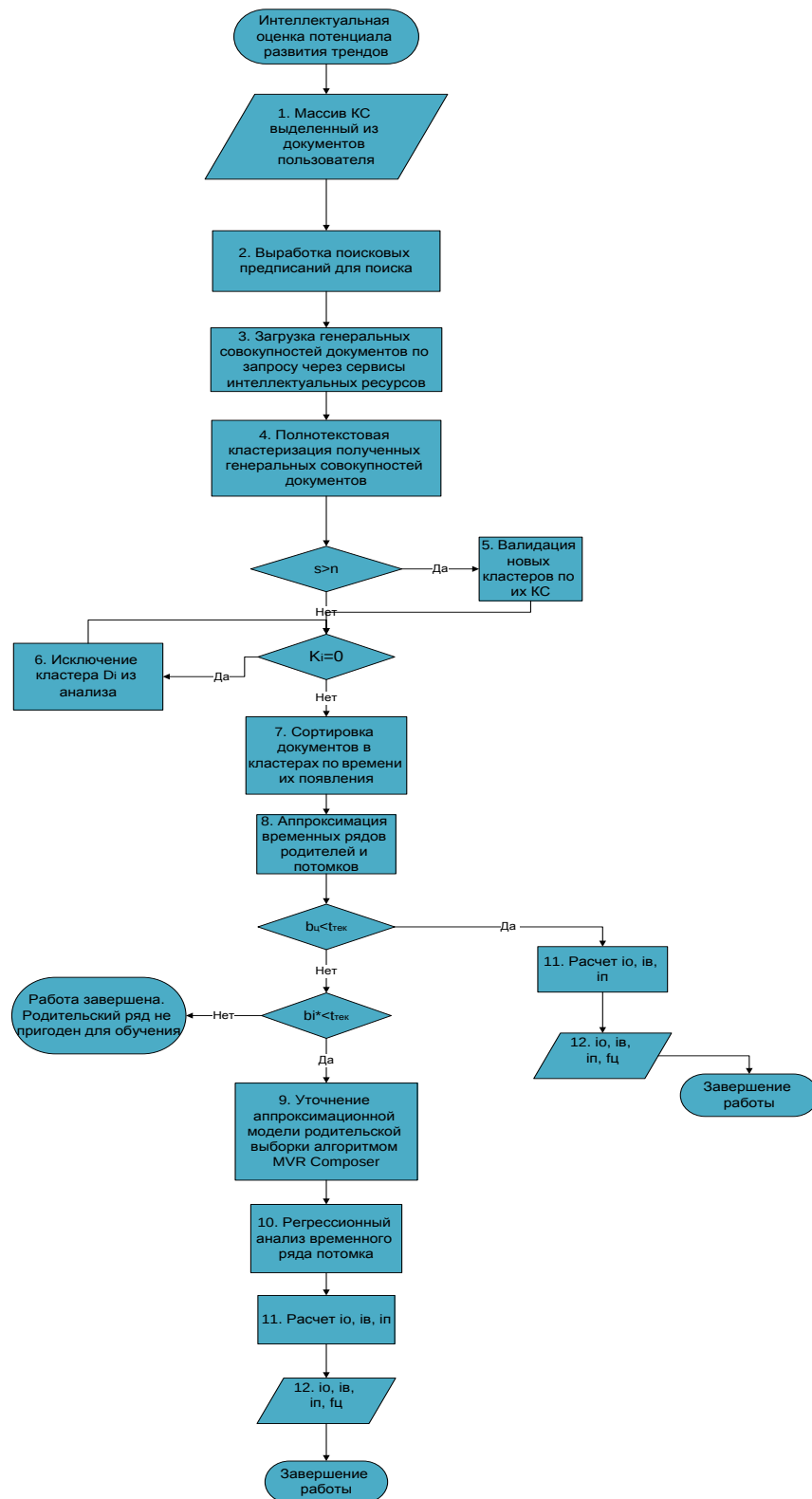


Рисунок 4. Алгоритм прогнозирования развития технических тенденций и расчета индексов инноваций

Приведем результаты работы алгоритма на временном ряде технологии LCD, состоящем из 27 проанализированных годовых значений, а оценку параметра «с» возьмем из расчетов по ЭЛТ, приведенных выше. В ходе эксперимента были построены модели по 77-51038/467498

первым 17, 18, 20, 22 и 27 значениям ряда и отслежено изменение их параметров в таблице 1 (Рисунок 5).

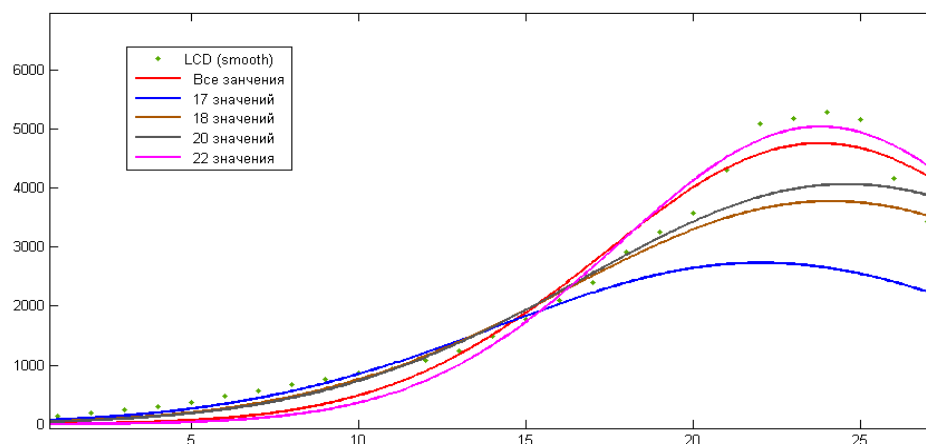


Рисунок 5. Комплексный графический результат работы алгоритма

Таблица 1. Оценки параметров модели по экспериментальным данным

Количество значений	Оценка «а»	Оценка «b»	Оценка «с»
17	2733 (2284,3183)	22.04 (20.78, 23.3)	11.15 (Зафиксировано на границе)
18	3776 (2020, 5532)	24.11 (17.58, 30.64)	11.15 (7.363, 14.94)
20	4065 (3518, 4612)	24.59 (23.53, 25.66)	11.15 (Зафиксировано на границе)
22	5038 (4507, 5570)	23.82 (22.25, 25.39)	8.525 (6.487, 10.56)
27 (полная выборка)	4758 (4459,5056)	23.79 (22.74, 24.84)	9.137 (7.805, 10.47)

Необходимо отметить, что при прохождении временного порога b^* (то есть реального, эталонного значения параметра b), оценки «а» и «b» становятся «стабильными» и меняются слабо. Данный алгоритм обладает достаточной точностью для определения потенциала развития тренда: оценки параметра «b» при приближении к точке разворота тренда (можно считать этим временем значение параметра b^* в эксперименте с полной выборкой) входят в 10% коридор. Так уже на 17 шаге алгоритм показал достаточно реалистичный прогноз времени разворота тренда, приняв который к сведению, мы бы имели 7 лет в «буфере безопасности» до реального события. Оценка параметра «а» так же сходится к эталонному значению по мере приближения к точке b^* .

Введем следующие важные индексы в рамках параметрической регрессионной модели (1):

$$\text{➤ Индекс запаса по объему: } i_O = h(b_u) * \frac{a_u - a}{a_u}; \quad (2)$$

$$\text{➤ Индекс запаса по времени: } i_B = h(b_u) * \frac{b_u - b}{b_u}; \quad (3)$$

$$\text{➤ Индекс потенциала развития: } i_{\Pi} = h(b_u) * \frac{b_u - b}{b_u} * \frac{a_u - a}{a_u}; \quad (4)$$

$$\text{➤ } h(b_u) = \begin{cases} 1, b \leq b_u \\ -1, b \geq b_u \end{cases} \quad (5)$$

Где a_u и b_u - прогнозные параметры регрессионной модели (1);

a – объем выборки за последний анализируемый год;

b – анализируемый момент времени.

Нетрудно видеть, что все значения введенных индексов лежат в пределах $(-1;1)$ и в развитии реагируют на изменение произведенных оценок регрессионных моделей. Описанные выше подходы отнюдь не являются эталонными, а в рамках развития данного исследования выполняется работа по уточнению моделей и усовершенствованию математического обеспечения. В частности планируется введение в модели дополнительных данных, а именно – других эквидистантных временных рядов. В эконометрике такие данные называются экзогенными переменными, тогда как основная анализируемая выборка эндогенной.

Основываясь на положениях теории цикличной динамики развития экономических сред и опыте использования технологий Patent mining в методологии форсайта, в данной работе были объединены два этих направления и доказана эффективность данного синтеза по отношению к развивающимся в интеллектуальном информационном пространстве научно-техническим тенденциям. Полученный по результатам исследования алгоритм и регрессионные модели позволяют с достаточной точностью спрогнозировать будущее развитие трендов и построить систему индикаторов этого развития, которая позволит предрекать кризисные фазы развития технологий и выявлять наиболее перспективные направления роста.

Список литературы

1. S. Sirmakessis, K. Markellos, P. Markellou, G. Mayritsakis, STING: Evaluation of Scientific & Technological Innovation and Progress in Europe Through Patents: Statistical data mining and knowledge discovery / edited by Hamparsum Bozdogan. CRC Press Company 2004. – pp. 566-584
2. Онлайн сервис поиска патентов и публикаций «Sciverse». Официальный сайт [Электронный ресурс]. – Режим доступа: <http://www.sciverse.com> (20.04.12)
3. Сервис доступа к базе данных GPI на официальном сайте европейского патентного агентства [Электронный ресурс]. – Режим доступа: <http://www.epo.org/searching/subscription/expert.html> (20.04.12).
4. Жизненные циклы спроса и технологий [Электронный ресурс]. – Режим доступа: <http://www.emcon.ru/bk-ans-2-2-2.html> (20.04.12).
5. Кондратьев Н.Д. Большие циклы конъюнктуры и теория предвидения. М.: Экономика, 2002, с. 511.
6. Алгоритм Левенберга-Макгвардта. Электронный учебник MATLAB [Электронный ресурс]. – Режим доступа: http://matlab.exponenta.ru/optimiz/book_1/13.php (20.04.12).
7. MVR Composer [Электронный ресурс]. – Режим доступа: http://www.machinelearning.ru/wiki/index.php?title=MVR_Composer (20.04.12).