УДК 004.02

ПРИМЕНЕНИЕ ЛАТЕНТНО-СЕМАНТИЧЕСКОГО АНАЛИЗА ДЛЯ АВТОМАТИЗАЦИИ ПРОЦЕССА ПРОФОРИЕНТАЦИИ

Макарова Мария Юрьевна

Аспирант кафедры «Физика и прикладная математика»;
Муромский институт (филиал) ФГБОУ ВПО «Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых»;
602264 г. Муром, Владимирской области, ул. Орловская, д.23;
e-mail: marisabel.science@gmail.com.

Данная статья посвящена рассмотрению вопроса применения латентно-семантического анализа для автоматизации процесса профориентации. В статье приводится математическая модель профориентации и обосновывается применение данного метода на основе обзора существующих алгоритмов классификации текстовой информации. Рассмотрен пример использования латентно-семантического анализа для установления сходства компетенций направления обучения описанию профессии, и на основе представленных результатов сделан вывод о целесообразности применения данного метода.

Ключевые слова: информационная система по профориентации, латентно-семантический анализ, компетенции.

THE USING OF THE LATENT SEMANTIC ANALYSIS FOR AUTOMATION OF VOCATIONAL GUIDANCE PROCESS

Makarova Maria

Graduate student of the department of «Physics and Applied Mathematics»;
The Murom Institute (Branch) of the federal government budget educational institutions of higher education «Vladimir State University named after Alexander G. and Nicholas G. Stoletovs»
602264, Murom, Vladimir reg., Orlovskaya str., 23;
e-mail: marisabel.science@gmail.com.

This article is devoted to consideration of a question of using of the latent semantic analysis for automation of vocational guidance process. In article the mathematical model of vocational guidance is given and on the basis of the review of existing algorithms of classification of text information using of this method is substantiated. The example of using of the latent semantic analysis for similarity definition of competences of the direction of training to the profession description is considered, and on the basis of the presented results the conclusion is drawn about expediency of application of this method.

Keywords: information system for career guidance, latent semantic analysis, competence.

Введение

В последнее время вопросы профессиональной ориентации приобрели особую актуальность: на рынке труда появляется все больше молодых специалистов с «модными» профессиями, что обостряет конкуренцию, а множество вакансий продолжают оставаться невостребованными. Из-за отсутствия общепринятых критериев в построении системы профессиональных ориентаций у большинства соискателей возникают психологические трудности в создании образа профессии, оценке соответствия этому образу и осуществлении профессионального выбора. Профориентация как научное направление, призванное решать прикладные задачи оптимального регулирования процесса профессионального самоопределения личности в ее интересах и в интересах общества [1], должна решать эти трудности.

В связи с повсеместным развитием инфокоммуникационных технологий стало возможным и является наиболее предпочтительным автоматизация процессов профориентации. Существующие информационные системы по профориентации [1, 4-8] предоставляют возможность определить индивидуальные профессиональные качества с помощью профориентационных методик и получить в соответствии с ними рекомендации по выбору будущей профессии. К сожалению, эти системы не универсальны, включают определенный набор тестов по профориентации и, соответственно, не отвечают требованиям современного общества, являются коммерческими продуктами и, зачастую, не позволяют автономно получить профориентационную помощь. Все это обусловило необходимость разработки информационной системы, автоматизирующей процессы профориентации на любых уровнях профессионального становления личности. Создание такой системы предполагает построение математической модели профориентации и разработку на ее основе алгоритмов и методов для автоматизации процессов профориентации.

1. Математическая модель процесса профориентации

Под профориентацией понимается согласование трех факторов: «Хочу» (интересы, склонности, пожелания личности), «Могу» (знания, умения, квалификация личности), «Надо» (требования рынка труда). С точки зрения математики каждый из описанных факторов представляет собой множество параметров, а процесс профориентации подразумевает установление пересечений этих множеств:

$$G = P \bigcap S \bigcap R ,$$

где P – множество «Хочу»; S – множество «Могу»; R – множество «Надо»; G – множество данных, которое содержит те элементы, которые присутствуют и в множестве P, и в множестве S, и в множестве R одновременно (результат профориентации).

Для установления пересечения среди описанных множеств необходимо представить их элементы в едином формате. Иными словами, каждое множество должно быть подмножеством множества данных, с помощью которого можно будет описать и склонности человека, и имеющиеся у него знания и умения, и требования рынка труда. Также следует помнить, что процесс профориентации должен быть реализован в соответствии с ГОС третьего поколения, что позволит учитывать современные требования работодателей к квалификации соискателей. Именно поэтому в качестве единого формата для описания данных были выбраны компетенции. Тогда каждое из описанных множеств P, S, R будет представлять собой соответствующие множества компетенций, а результатом профориентационного процесса будет множество C_G, содержащее только те компетенции, которые присутствуют во всех трех множествах одновременно:

$$C_G = C_P \bigcap C_S \bigcap C_R ,$$

где C_P – множество компетенций, соответствующих множеству «Хочу»; C_S – множество компетенций, соответствующих множеству «Могу»; C_R – множество компетенций, соответствующих множеству «Надо».

Тогда алгоритм реализации процесса профориентации в информационной системе можно описать следующим образом:

– формируются множества P, S, R;

– полученные множества преобразуются в соответствующие множества компетенций C_P , C_S и C_R;

– проводится операция пересечения преобразованных множеств;

– полученное множество компетенций C_G преобразуется в ряд профессий, рекомендуемых пользователю.

Данная математическая модель и предложенный алгоритм позволят использовать в системе любые профориентационные методики для определения индивидуальных особенностей пользователей и ориентировать на соответствующие им актуальные профессии. Иными словами, данная информационная система по профориентации будет отличаться от существующих разработок универсальностью и позволит учитывать современные требования рынка труда.

2. Обзор методов для автоматизации процесса профориентации

Для автоматизации профориентационного процесса необходим метод, позволяющий преобразовывать исходные данные трех множеств «Хочу», «Могу» и «Надо» в единый формат (множество компетенций). В таком случае процесс профориентации, подразумевающий логическое увязывание индивидуальных особенностей и способностей личности с требованиями рынка труда, будет сводиться к тривиальной задаче установления пересечения трех множеств, содержащих данные одного типа.

Поскольку каждое из указанных множеств («Хочу», «Могу», «Надо») представляет собой текстовые данные, задачу установления соответствия между данными или, иными словами, задачу установления сходства документов можно решить с использованием методов классификации текстовой информации.

Наибольшее распространение получили следующие методы классификации текстовых документов:

– иерархические методы (метод ближайшего соседа, метод наиболее удаленных соседей, попарное среднее, суффиксные деревья) заключаются в получении древовидной структуры кластеров, каждый узел такого дерева содержит все объекты кластеров-потомков;

– метод квадратичной ошибки (метод *k*-средних) выделяет группы, расположенные на возможно больших расстояниях друг от друга, каждый документ присваивается тому кластеру, чей центр является наиболее близким документу;

– методы теории графов (алгоритм, основанный на построении минимального остовного дерева) заключаются в построении остовного дерева с последующим удалением ребер с наибольшими длинами для генерации кластеров;

– методы, основанные на концепции плотности (алгоритм DBSCAN) заключаются в обнаружении кластеров на основе предположения о том, что внутри каждого кластера наблюдается типичная плотность объектов, которая значительно выше плотности объектов за его пределами;

– нейросетевые методы (самоорганизующиеся карты Кохонена, алгоритмы теории адаптивного резонанса) представляют собой алгоритмы с обучением, основанные на свойствах человеческого мозга изучать новые понятия, сравнивая с уже существующими;

– эволюционные методы (генетические алгоритмы) используют эволюционные операторы и популяцию решений для получения глобально оптимального разделения данных;

– методы понижения размерности пространства (латентно-семантический анализ, метод главных компонент) основываются на том, что документы, содержащие семантически близкие термины, сгущаются в определенных местах пространства терминов.

В настоящей работе необходим алгоритм, который позволит выявить компетенции, описанные в государственном образовательном стандарте (ГОС), наиболее подходящие для каждой профессии из единого тарифно-квалификационного справочника. Иными словами, необходимо установить степень сходства документа, содержащего описание профессии, с документами, содержащими компетенции. Детальный анализ описанных методов позволил сделать вывод, что для решения данной задачи лучше всего подходит метод латентно-семантического анализа.

Латентно-семантический анализ основан на идее, что совокупность всех контекстов, в которых встречается и не встречается данное слово, задает множество обоюдных ограничений, которые в значительной степени позволяют определить похожесть лексических значений слов между собой. В качестве исходной информации используется матрица, содержащая частоты использования каждого терма в документах. Используя разложение матрицы по сингулярным выражениям, каждый терм и документ представляются в виде векторов в общем пространстве размерности. Для определения меры сходства текстовых документов удобнее всего использовать косинусную меру сходства, основанную на вычислении значения косинуса между двумя векторами документов.

3. Пример использования латентно-семантического анализа для автоматизации процесса профориентации

Рассмотрим пример использования латентно-семантического анализа, чтобы установить компетенции, подходящие для профессии «Врач-специалист». Описание данной профессии взято из единого квалификационного справочника профессий [3]. В требованиях к квалификации данной профессии указано наличие обязательного профессионального образования по направлениям «Лечебное дело», «Педиатрия», «Стоматология», «Медицинская биофизика», «Медицинская биохимия» или «Медицинская кибернетика». Соответственно, компетенции из данных ГОС должны наибольшим образом соответствовать этой профессии, то есть общая оценка соответствия компетенций из указанных ГОС описанию профессии «Врач-специалист» будет достаточно высока. Возьмем компетенции ГОС по направлению «Лечебное дело» (32 компетенции).

Исходную текстовую информацию необходимо подвергнуть предварительной обработке: исключить стоп-слова и провести операцию стемминга. К стоп-словам относятся слова, не несущие смысловой нагрузки (предлоги, союзы, частиц). Для стемминга (выделения основы слова) использовался алгоритм Портера [9]. Затем из полученного массива термов были исключены термы, встречающиеся один раз. Таким образом, было получено 135 различных термов.

Следующим шагом является составление частотной матрицы термов. Строки матрицы – полученные термы (индексируемые слова), столбцы матрицы – исходные документы (описание профессии и 32 компетенции). В каждой ячейке матрицы указывается, сколько раз этот терм встречается в каждом документе. В нашем примере получилась матрица размерности 135×33.

После составления частотной матрицы ее следует нормализовать. Стандартный способ нормализации матрицы – TF-IDF [2]. Мера TF-IDF является произведением двух сомножителей: TF и IDF.

TF – отношение числа вхождения некоторого слова к общему количеству слов документа (важность слова в пределах отдельного документа):

$$TF = \frac{n_i}{\sum_k n_k},$$

где n_i – число вхождений слова в документ, $\sum_k n_k$ – общее число слов в данном документе.

IDF – инверсия частоты, с которой некоторое слово встречается в документах коллекции.

$$IDF = \log \frac{|D|}{\left|\left(d_i \supset t_i\right)\right|},$$

где $|D|$ – количество документов в корпусе, $\left|\left(d_i \supset t_i\right)\right|$ – количество документов, в которых встречается слово.

Выбор основания логарифма в формуле не имеет значения, поскольку изменение основания приводит к изменению веса каждого слова на постоянный множитель, что не влияет на соотношение весов.

Таким образом, все документы будут представлены в виде векторов, отражающих важность использования каждого слова в документе. Совокупность таких векторов – получившаяся матрица, часть которой представлена на рисунке 1.

	Описание профессии	1 компе-тенция	2 компе-тенция	3 компе-тенция	4 компе-тенция	5 компе-тенция	...
насел	0,0089	0,0366	0,0000	0,0000	0,0000	0,0000	...
медицин	0,0266	0,0366	0,0000	0,0549	0,0000	0,0366	...
помощ	0,0210	0,0000	0,0000	0,0000	0,0000	0,0000	...
специальн	0,0307	0,0000	0,0000	0,0000	0,0000	0,0000	...
использов	0,0044	0,0000	0,0732	0,0275	0,0000	0,0000	...
соврем	0,0136	0,0000	0,0000	0,0000	0,0000	0,0561	...
метод	0,0136	0,0000	0,0000	0,0000	0,0000	0,0000	...
профилакт	0,0248	0,0000	0,0000	0,0000	0,0000	0,0000	...
диагност	0,0185	0,0000	0,0000	0,0000	0,0000	0,0000	...
...	...	...	...	...	...	...	...

Рис. 1. Часть нормализованной частотной матрицы

Следующим шагом является сингулярное разложение матрицы. Сингулярное разложение позволяет выделять ключевые составляющие матрицы, игнорируя шумы, и заключается в разложении исходной матрицы (M) на три составляющие:

$$M = U \cdot W \cdot V^T,$$

где U и V^T – ортогональные матрицы, W – диагональная матрица, элементы которой упорядочены по убыванию (сингулярные числа).

В нашем примере далее используются первые три строки матрицы V^T, содержащей значения векторов-документов. В ходе испытаний применение именно трехмерной декомпозиции показало наилучшие результаты. Размерность может быть и больше, но чем больше размерность, тем большее влияние шумов. На рис. 2 представлено расположение точек, соответствующих векторам (исходных документов):

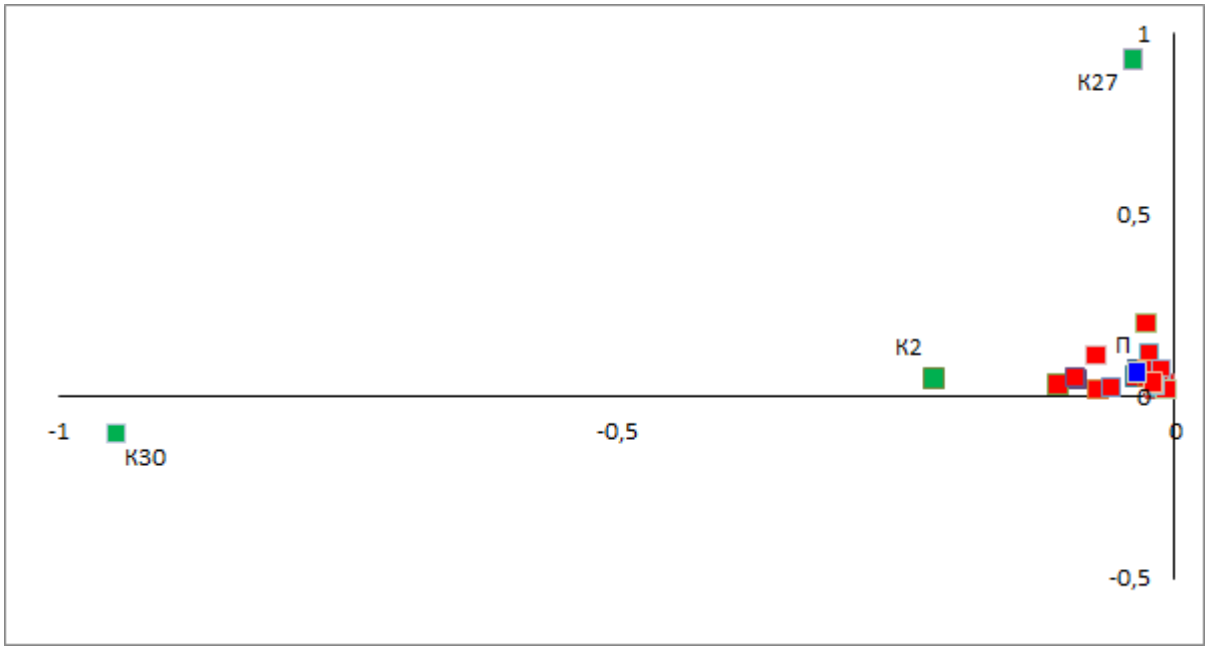

Рис. 2. График расположения документов

Как видно из рис. 2 описание профессии (точка П) и большинство точек сосредоточены в одном месте, за исключением точек К2, К27, К30. Эти точки обозначают соответствующие компетенции. Для подтверждения наглядного результата проведем следующие расчеты.

Для оценки близости исходного документа (вектора-профессии) и документов, содержащих описание компетенций (векторов-компетенций), воспользуемся косинусной мерой, определяемой по следующей формуле:

$$c_i = \frac{\sum_{k=0}^{2}(V_{k,0} \cdot V_{k,i})}{\sqrt{\sum_{k=0}^{2}(V_{k,0})^2} \cdot \sqrt{\sum_{k=0}^{2}(V_{k,i})^2}},$$

где $V_{k,0}$ – значения вектора–профессии, $V_{k,i}$ – значения векторов-компетенций.

На рис. 3 представлены получившиеся результаты:

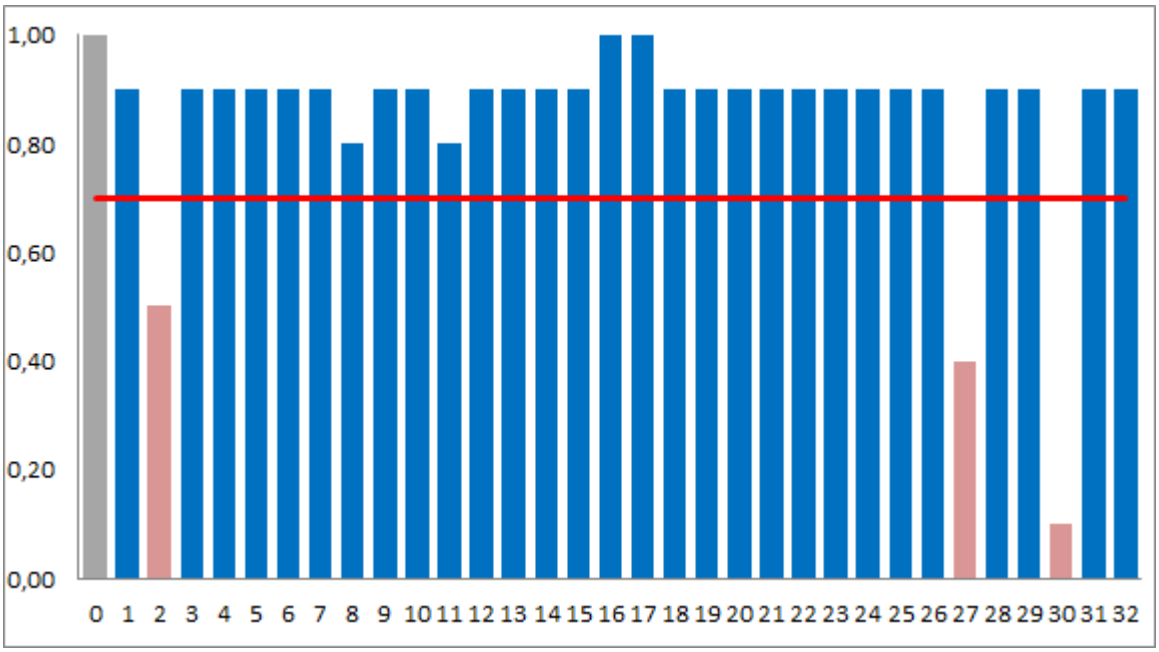

Рис. 3. Результаты расчета степени сравнения описания профессии «Врач-специалист» и компетенций направления «Лечебное дело»

Как видно из рис. 3, только 3 компетенции из 32 (2, 27 и 30) получили степень сравнения менее 0,7. Нулевой столбик – описание профессии. Соответственно, можно говорить, что ГОС по направлению «Лечебное дело» соответствует профессии «Врач-специалист». Рассчитаем общую оценку соответствия (*OCR* – overall conformity rating) по следующей формуле:

$$OCR = \frac{|c_i > L|}{|C|} \cdot 100\%,$$

где $|c_i > L|$ – количество компетенций, для которых значение косинусной меры больше некоторого порога L (в нашем случае это значение равно 0,7, $|C|$ – общее количество компетенций.

Таким образом, с использованием латентно-семантического анализа было установлено, что степень соответствия компетенций направления «Лечебное дело» профессии «Врач-специалист» равна 91%. Следовательно, можно говорить, что для этой профессии действительно необходимо освоить данное направление обучения.

Для сравнения приведем пример оценки соответствия компетенций направления «Прикладная математика и информатика» (17 компетенций) данной профессии «Врач-специалист». На первый взгляд, эти области совершенно различны, докажем это с помощью латентно-семантического анализа.

В данном случае была получена частотная матрица размерности 81×18. После нормализации матрицы и ее сингулярного разложения была также применена косинусная мера. Результаты расчетов приведены на рис. 4:

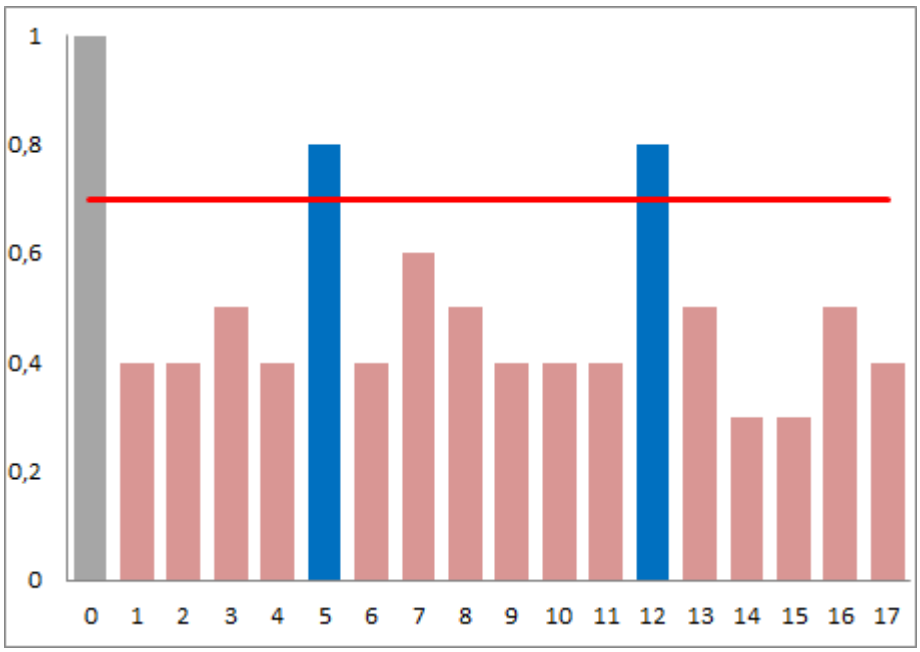

Рис. 4. Результаты расчета степени сравнения описания профессии «Врач-специалист» и компетенций направления «Прикладная математика и информатика»

Как видно из рис. 4, только 2 компетенции из 32 (5 и 12) получили оценку соответствия более 0,7. Общая оценка соответствия (OCR) равна 12%. Это очень низкое значение, следовательно, можно сделать вывод, что направление «Прикладная математика и информатика» не подходит для освоения профессии «Врач-специалист».

Заключение

В данной статье был рассмотрен вопрос применения латентно-семантического анализа для автоматизации процесса профориентации и установлено, что данный метод подходит для решения этой задачи. Автоматизация процесса профориентации подразумевает логическое увязывание трех основных множеств («Хочу», «Могу», «Надо»), необходим алгоритм, который позволит привести элементы этих множеств в единый формат для установления пересечения. В качестве единого формата были выбраны компетенции, так как они наилучшим образом способны отразить и профессиональные интересы и умения личности, и требования работодателей. Поскольку данные множества представляют собой текстовую информацию, а задача преобразования такой информации из одного вида в другой может быть сведена к задаче распределения компетенций по соответствующим профессиям, были рассмотрены существующие методы классификации текстовой информации по категориям. Для решения данной задачи был выбран метод латентно-семантического анализа.

В качестве примера использования метода латентно-семантического анализа была рассмотрена задача установления сходства компетенций направления «Лечебное дело» описанию профессии «Врач-специалист». Полученная общая оценка соответствия равна 91%, что говорит о высокой степени сходства данных документов. Кроме того, в статье приведен другой пример расчета оценки соответствия компетенций направления «Прикладная математика и информатика» той же профессии. В результате была получена оценка, равная 12%, что свидетельствует о том, что данное направление обучения не подходит для освоения профессии «Врач-специалист». Данные результаты являются вполне логичными, что говорит о целесообразности применения метода латентно-семантического анализа для автоматизации процесса профориентации.

Список литературы

1. Живага А.Ю. Особенности профессиональных ориентаций студенческой молодежи (на материалах Сахалинской области): Автореферат дис. … канд. соц. наук. – М., 2008. – С. 31.
2. TF-IDF. – [Электронный ресурс]. URL: http://ru.wikipedia.org/wiki/TF-IDF.
3. Единый квалификационный справочник должностей руководителей, специалистов и служащих. – [Электронный ресурс]. URL: http://base.consultant.ru/cons/cgi/online.cgi?req=doc;base=LAW;n=97378.
4. Диагностический компьютерный тестовый комплекс Профориентатор. – [Электронный ресурс]. URL: http://teletesting.ru/modules/tests.
5. Комплекс Effecton Studio. Пакет Профориентация. – [Электронный ресурс]. URL: http://www.effecton.ru/730.html.
6. Компьютерная система психодиагностики СМОЛ-Эксперт. – [Электронный ресурс]. URL: http://www.nmc-radix.com/r_se.html.
7. ПДК «Эксперт+». – [Электронный ресурс]. URL: http://www.expert-plus.ru.
8. Программный комплекс КАТАРСИС. – [Электронный ресурс]. URL: http://katarsisprog.ru.
9. Стеммер Портера. – [Электронный ресурс]. URL: http://ru.wikipedia.org/wiki/Стеммер_Портера.