МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное бюджетное образовательное учреждение
высшего образования
«УЛЬЯНОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»
Российская ассоциация нечетких систем и мягких вычислений

Прикладные информационные системы (ПИС-2019)

Сборник научных трудов
VI Всероссийской научно-практической конференции
с международным участием

(Россия, г. Ульяновск 27 мая – 09 июня 2019 г.)

Ульяновск
УлГТУ
2019

УДК 004.62 (082)
ББК 32.973.202я43
П 75

УДК 004.62 (082)

Прикладные информационные системы (ПИС-2019) : сборник научных трудов VI Всероссийской научно-практической конференции с международным участием (Россия, г. Ульяновск, 27 мая – 09 июня, 2019 г.) – Ульяновск, УлГТУ, 2019. – 116 с.

В сборнике опубликованы доклады участников VI Всероссийской научно-практической конференции «Прикладные информационные системы (ПИС-2019)».

Материалы сборника предназначены для специалистов по информационных систем и технологий, магистрантам и аспирантам информационно-технических специальностей вузов.

Конференция проведена в соответствии с научным проектом 2.1182.2017/4.6 «Разработка методов и средств автоматизации производственно-технологической подготовки агрегатно-сборочного самолетостроительного производства в условиях мультипродуктовой производственной программы», выполняемый научным коллективом Ульяновского государственного технического университета в рамках государственного задания Минобрнауки РФ.

Статьи представлены в авторской редакции.

ISBN 978-5-9795-1986-9

Оргкомитет ПИС–2019:

Председатель:

Ярушкина Н.Г., д.т.н., проф., УлГТУ, г. Ульяновск

Члены организационного комитета:

Романов А.А., к.т.н., доцент, УлГТУ, г. Ульяновск
Филиппов А.А., к.т.н., УлГТУ, г. Ульяновск
Гуськов Г.Ю., УлГТУ, г. Ульяновск
Эгов Е.Н., УлГТУ, г. Ульяновск
Григоричева М.С., УлГТУ, г. Ульяновск
Долгановская А.Ю., УлГТУ, г. Ульяновск

Программный комитет ПИС–2019:

Председатель:

Ярушкина Н.Г., д.т.н., проф., УлГТУ, г. Ульяновск

Заместитель Председателя:

Афанасьева Т.В., д.т.н., проф., УлГТУ, г. Ульяновск

Члены программного комитета:

Наместников А.М., д.т.н., доцент, УлГТУ, г. Ульяновск
Воронина В.В., к.т.н., доцент, УлГТУ, г. Ульяновск
Евсеева О.Н., к.т.н., доцент, УлГТУ, г. Ульяновск
Романов А.А., к.т.н., доцент, УлГТУ, г. Ульяновск
Филиппов А.А., к.т.н., УлГТУ, г. Ульяновск
Мошкин В.С., к.т.н. УлГТУ, г. Ульяновск
Мошкина И.А., к.т.н. УлГТУ, г. Ульяновск

VI Всероссийская научно-практическая конференция с международным участием Прикладные информационные системы-2019

VI Всероссийская научно-практическая конференция с международным участием «Прикладные информационные системы-2019» проведена на базе Ульяновского государственного технического университета при поддержке Российской ассоциации нечетких систем и мягких вычислений.

Конференция «Прикладные информационные системы-2019» была проведена в соответствии с планом научного проекта 2.1182.2017/4.6 «Разработка методов и средств автоматизации производственно-технологической подготовки агрегатно-сборочного самолетостроительного производства в условиях мультипродуктовой производственной программы», выполняемый научным коллективом Ульяновского государственного технического университета в рамках государственного задания Минобрнауки РФ.

В рамках конференции было представлено более двадцати научных докладов по следующим тематикам: «Интеллектуальный анализ данных», «Инженерия знаний, онтологий, управления знаниями», «Нечеткие системы и мягкие вычисления», «Проектирование информационных систем». Также в рамках конференции была проведена Молодежная школа-семинар, в которой участвовали бакалавры и магистранты технических специальностей.

Организационный комитет благодарит авторов докладов, приславших свои работы на конференцию, а также ректорат Ульяновского государственного технического университета, обеспечивший проведение конференции и издание ее материалов.

Председатель
организационного комитета конференции доктор
технических наук, профессор Н.Г. Ярушкина

СОДЕРЖАНИЕ

CONTENTS

УДК 004.94

АВТОМАТИЗИРОВАННАЯ ДИАГНОСТИКА МЕДИЦИНСКИХ ОБСЛЕДОВАНИЙ

Иванова Н.П.(np.iva9ova@ulstu.ru)
Ульяновский государственный технический университет, Ульяновск

Рассмотрено исследование результатов медицинских наблюдений на основе кластерного анализа качественных шкал данных пациентов. Кодирование количественных показателей в лингвистические параметры позволило получить частотную характеристику по кластерам и провести качественное резюмирование.

Ключевые слова: кластерный анализ, качественные шкалы, автоматизация, диагностика, медицинские обследования, лингвистические оценки, резюмирование.

Введение

Широкое использование кластерных методов анализа обусловлено тем, что создание классификаций имеет объективный характер и легко воспроизводится. Кластерный анализ позволяет сгруппировать многомерные объекты. В статье [Кузнецов, 2006] рассматривается как часть многомерного статистического анализа. В книге [Кобринский, 2013] включает этот анализ в один из этапов при разработке алгоритма диагностического анализа заболеваний (состояний, симптомов). В настоящее время для получения диагностического алгоритма в большинстве случаев используют известные статистические пакеты: SPSS, Stastica и т. д.

А в работе [Семененко, 2014] применяет методы кластерного анализа для совершенствования диагностики заболевания и работы врача. В статье [Чопоров и др., 2015] по исследуемой предметной области рассказывает о многоуровневом мониторинге и предлагает несколько методов построения прогностических и классификационных моделей, в том числе и кластеризацию. Причем важным выводом из этого исследования

является необходимость использования процедур предварительной обработки информации.

Также хочется отметить коммерческую разработку АС «КоМеД», в которой оценивается состояние каждого конкретного медицинского сотрудника до и после осмотра, используя данные медицинского измерения.

1 Средства моделирования кластерного анализа

В соответствии с целью проекта, где необходимо найти различия и сходства качественных данных медицинского наблюдения по кластерам, будут использованы в качестве средств моделирования простейшие методы и инструменты для наглядности исследования, такие как графические и табличные представления данных.

Анализ будет проводиться при помощи операций из MS Excel и программы на языке python с использованием библиотеки pyplot.

2 Проведение кластерного анализа

2.1 Отбор выборки для кластеризации

В качестве основных исследуемых показателей взяты данные:

1. деменция
2. хроническая болезнь почек
3. систолическое артериальное давление
4. диастолическое артериальное давление
5. образование
6. семейное положение
7. социальный статус
8. индекс массы тела
9. объем талии
10. индекс массы миокарда левого желудка
11. фракция выброса левого желудочка
12. анемия
13. триглицериды
14. холестерин
15. артериальная гипертензия
16. фибрилляция предсердий
17. инфаркт миокарда
18. язвенная болезнь желудка
19. острое нарушение мозгового кровообращения
20. сахарный диабет 2 типа
21. креатинин

22. скорость клубочковой фильтрации
23. курение
24. ограничение поваренной соли
25. физическая активность
26. приверженность к изменению образа жизни и к медикаментозной терапии
27. уровень физического здоровья, баллы (шкала по SF36)
28. уровень психического здоровья, баллы (шкала по SF36)
29. уровень депрессии, баллы (по опроснику mmpi)
30. акцентуация характера (тревожность).

Имеется база данных пациентов по выше описанным медицинским показателям. Для проведения качественной оценки состояния по этим данным кодируются количественные показатели в лингвистические параметры, благодаря чему будет упрощена дальнейшая кластеризация. Разделение показателей на:

- два параметра: N(норма) и A(не норма)
- четыре и пять параметров: N(нормальный), H (высокий), F (первый), S (второй), T (третий).

Причем уровни здоровья по шкале SF36 (п.27,28) не рассматриваются в дальнейшем и диапазоны не выделяются.

2.2 Применение метода кластерного анализа для создания групп сходных объектов и их оценки

Предварительно для кластеризации решено вручную определить количество кластеров – шесть, и уже по ходу исследования проверить и выяснить их оптимальное количество.

Полученная мощность (количество объектов) каждого кластера:

1. 14 пациентов,
2. 17 пациентов,
3. 10 пациентов,
4. 15 пациентов,
5. 26 пациентов,
6. 18 пациентов.

Для этого на основе качественных данных и открытой базы данных будут проведены анализы:

- по всем пациентам (таблица 1),
- по пациентам в каждом кластере (таблица 2).

Таблица 1 – Фрагмент анализа по всем пациентам

Деменция		САД					Образование				ИММ ЛЖ	
N	A	N	H	F	S	T	H	T	M	Z	N	A
74	26	29	0	40	25	6	53	39	6	2	1	99

Таблица 2 – Фрагмент анализа по пациентам в каждом кластере

Показатели и их код	№ кластера	1	2	3	4	5	6
Деменция	N	9	8	9	7	24	17
	A	5	9	1	8	2	1
САД	N	12	1	2	5	4	5
	H	0	0	0	0	0	0
	F	2	4	2	5	17	10
	S	0	11	6	1	4	3
	T	0	1	0	4	1	0
Образова-ние	H	10	14	5	1	13	10
	T	4	3	5	6	13	8
	M	0	0	0	6	0	0
	Z	0	0	0	2	0	0
ИММ ЛЖ	N	0	0	1	0	0	0
	A	14	17	9	15	26	18

Следовательно, проведенные проверки дали достаточно разнородные символьные значения симптомов в каждом кластере, потому решено не менять их количество. Однако выделяются почти однородные записи следующих показателей:

- индекс массы миокарда левого желудка,
- артериальная гипертензия,
- уровень депрессии, баллы (по опроснику mmpi),
- акцентуация характера (тревожность).

Их решено исключить из дальнейшего исследования, как и показатели, не отображающие медицинских данных:

- образование,
- семейное положение,
- социальный статус,
- индекс массы тела.

2.3 Лингвистическое резюмирование

Этот этап включает в себя частотный анализ, проводимый по каждому лингвистическому значению показателя. Для равномерности анализа каждого показателя решено было объединить параметры (H и F, S и T), сократив их до трех – N, HF, ST. Зная мощность каждого кластера, легко рассчитать по каждому кластеру частоты, приведенные в Таблице 4.

Таблица 4 – Фрагмент частотный анализа в каждом кластере

Показатели и их код	№ кластера	1	2	3	4	5	6
Деменция	N	64,29	47,06	90	46,67	92,31	94,44
	A	35,71	52,94	10	53,33	7,69	5,56
САД	N	85,71	5,88	20	33,33	15,38	27,78
	HF	14,29	28,57	14,29	35,71	121,3	71,43
	ST	0	85,71	42,86	35,71	35,71	21,43

На основе этих данных будет составляться лингвистическое резюмирование пациентов по отдельному показателю симптома, для которого будет использован базовый подход Ягера [Jager 1982].

В рамках базового подхода рассматривается множество из n объектов в виде записей *F {S1,…, gn}* в базе данных *D {y (g1),…,y (gn)}* с одним атрибутом (свойством) *y*, при этом *y(gi)* – это значение показателя симптома для пациента в кластере ***gi***.

Ядром лингвистического резюмирования является набор предложений, которые могут быть записаны в абстрактном виде:

Q объектов исследования g`s имеют значение V по показателю y

Результаты характеристики доли значений лингвистических переменных по кластерам будет осуществляться с помощью лингвистической шкалы, строящейся из трапецеидальных функций принадлежности для нечетких множеств (рис. 1).

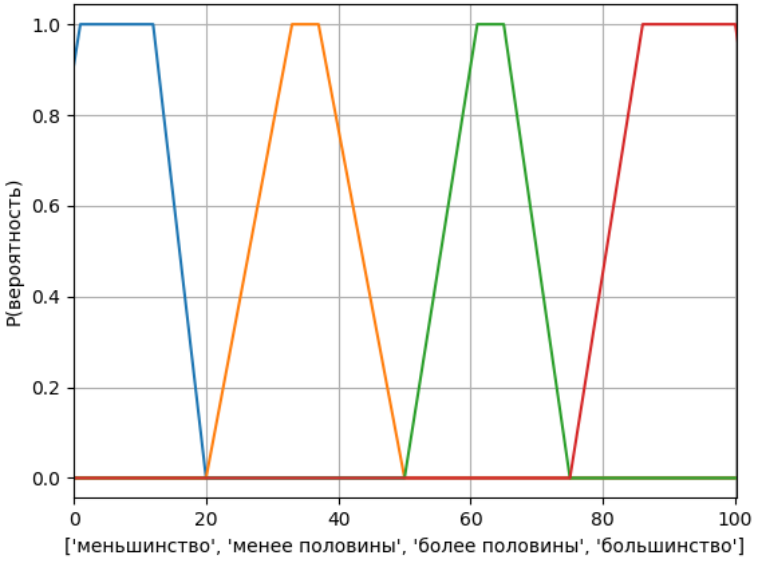


Рисунок 1 – Лингвистические шкалы для резюмирования

По графику (рис. 1) в результате получены степени истинности с соответствующими термами и сформировано резюмирование, представленное на рис. 2.

В результате по данному массиву данных пациентов выявлено здоровое состояние большинства объектов в каждом кластере по показателям. Однако выделяются такие показатели по выделяющимся аномальным уровням, как холестерин, объем талии, индекс массы миокарда левого желудка, диастолическое артериальное давление. Отдельно стоит выделить объекты четвертого кластера, обладающие в большей половине показателей аномальными уровнями по медицинским наблюдениям.

```
   1 -- нормальный уровень по показателю деменция
более половины объектов кластера 1 - с вероятностью 1.00
менее половины объектов кластера 2 - с вероятностью 0.23
большинство объектов кластера 3 - с вероятностью 1.00
менее половины объектов кластера 4 - с вероятностью 0.26
большинство объектов кластера 5 - с вероятностью 1.00
большинство объектов кластера 6 - с вероятностью 1.00
   2 -- аномальный уровень по показателю деменция
менее половины объектов кластера 1 - с вероятностью 1.00
более половины объектов кластера 2 - с вероятностью 0.27
меньшинство объектов кластера 3 - с вероятностью 1.00
более половины объектов кластера 4 - с вероятностью 0.30
меньшинство объектов кластера 5 - с вероятностью 1.00
меньшинство объектов кластера 6 - с вероятностью 1.00
   5 -- нормальный уровень по показателю САД
большинство объектов кластера 1 - с вероятностью 0.97
меньшинство объектов кластера 2 - с вероятностью 1.00
большинство объектов кластера 3 - с вероятностью 1.00
менее половины объектов кластера 4 - с вероятностью 0.58
меньшинство объектов кластера 5 - с вероятностью 0.60
менее половины объектов кластера 6 - с вероятностью 0.71
   6 -- аномально средний уровень по показателю САД
меньшинство объектов кластера 1 - с вероятностью 0.66
менее половины объектов кластера 2 - с вероятностью 0.71
меньшинство объектов кластера 3 - с вероятностью 1.00
менее половины объектов кластера 4 - с вероятностью 0.36
большинство объектов кластера 5 - с вероятностью 0.91
более половины объектов кластера 6 - с вероятностью 0.97
   7 -- аномально высокий уровень по показателю САД
меньшинство объектов кластера 1 - с вероятностью 0.55
большинство объектов кластера 2 - с вероятностью 1.00
менее половины объектов кластера 3 - с вероятностью 1.00
менее половины объектов кластера 4 - с вероятностью 0.11
менее половины объектов кластера 5 - с вероятностью 1.00
менее половины объектов кластера 6 - с вероятностью 0.11
```

Рисунок 2 – Фрагмент лингвистического резюмирования

Заключение

В ходе автоматизированной диагностики медицинских обследований выбраны средства моделирования и уточнены значимые показатели из массива данных пациентов, которые преобразованы в качественные данные симптомов в соответствии с введенными лингвистическими параметрами.

Эта обработка данных позволила провести более упрощенный кластерный анализ и получить частотную характеристику по кластерам. Данная характеристика симптомов необходима при получении резюмирования. А именно каждая лингвистическая переменная связывает числовые характеристики кластеров с лингвистическими термами, определенными с помощью трапециевидной функции.

Данное исследование является новым, благодаря точечному применению кластерного анализа, без других методов, обычно применяемых при диагностике. В итоге были получены качественные результаты, которые расширяют возможности для итогового заключения.

В дальнейшем предполагается расширить это исследование и провести более углубленную диагностику путем изменения применяемых методов и средств при исследовании в сторону сокращения объема результатов.

Список литературы

[Кобринский, 2013] Кобринский Б.А., Зарубина Т.В.. Медицинская информатика// учеб. для студ. учреждений высш. проф. образования. – 4-е изд., перераб и доп. – М. : Издательский центр «Академия», 2013.

[Кузнецов, 2006] Кузнецов Д.Ю., Трошина Т.Л. Кластерный анализ и его применение // Ярославский педагогический вестник. – 2006.

[Семененко, 2014] Семененко М. Г. Разработка функций пользователя в Excel 2013: приложения нечеткой логики // Современные информационные технологии и ИТ-образование. 2014. №10.

[Чопоров и т.д., 2015] Чопоров О.Н., Болгов С.В., Манакин И.И. Особенности применения методов интеллектуального анализа данных и многоуровневого мониторинга при решении задачи рационализации медицинской помощи // Моделирование, оптимизация и информационные технологии. – Воронеж: ВИВТ, 2015. – № 1 (8).

[Joaquim, 2015] Joaquim L. Viegas, Susana M. Vieira, Joao M.C. Sousa. Fuzzy clustering and prediction of electricity demand based on household characteristics // International Joint Conferece IFSA-EUSFLAT (16th World Congress of the International Fuzzy Systems Association (IFSA), 9th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT)), Gijon (Asturias) Spain, 2015.

[Yager, 1982] Yager, R. R. A new approach to the summarization of data // Information Sciences, vol. 28, 1982.

AUTOMATED DIAGNOSIS OF MEDICAL EXAMINATION

Ivanova N.P.(np.iva9ova@ulstu.ru)
Ulyanovsk State Technical University, Ulyanovsk

The paper considers the study of the results of medical observations based on cluster analysis of qualitative scales of patients ' data. The coding of quantitative indicators into linguistic parameters allowed us to obtain a frequency response for clusters and conduct a qualitative summary.

Keywords: cluster analysis, quality scales, automation, diagnostics, medical examinations, linguistic assessments, summarizing

УДК 004.94

РАЗРАБОТКА СЕРВИСА ДЛЯ ФОРМИРОВАНИЯ СОЦИАЛЬНОГО ПОРТРЕТА ПОЛЬЗОВАТЕЛЯ НА ОСНОВЕ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА СОДЕРЖИМОГО ДАННЫХ В ОТКРЫТЫХ ИСТОЧНИКАХ*

Григоричева М.С.(gms4295@mail.ru)
Ульяновский государственный технический университет, Ульяновск

Описывается анализ существующих, выбор и адаптация алгоритма кластеризации для решения задачи формирования социального портрета пользователя на основе интеллектуального анализа содержимого данных в открытых источниках на примере социальной сети ВКонтакте. Такой анализ позволит провести сегментацию аудитории социальной сети и получить «усредненный» социальный портрет типового пользователя для каждого сегмента.

Ключевые слова: социальный портрет, кластеризация, социальные сети

Введение

На территории Ульяновской области существуют организации, занимающиеся разработкой и продвижением программных систем на рынок: ITECH.group, MST digital agency, СимбирСофт, и др. Для автоматизации процесса анализа целевой аудитории необходимо использовать интеллектуальные системы анализа данных. Большие объемы данных, многообразие форм их представления и их неструктурированное представление не позволяют оперативно сформировать социальный портрет потенциального пользователя продукта компании. Данная проблема увеличивает сроки процессов анализа требований, идеи, конкурентов и целевой аудитории.

1) *Работа выполнена при финансовой поддержке ФСИ (номер договора 13655ГУ/2018).

Основной целью проекта является автоматизация процессов построения социального портрета пользователя социальной сети для снижения времени и повышения качества анализа его социального окружения и предпочтений. Помимо анализа данных конкретного пользователя, интересным может оказаться анализ данных сообщества (группы) пользователей.

Каждый пользователь оставляет в сети интернет множество цифровых «следов», сопоставив которые можно составить социальный портрет человека. Этот портрет будет содержать информацию о возрасте, поле, образовании, социальном статусе, родственных связях, друзьях, интересах, политических и религиозных предпочтениях, местах жительства, учебы и работы, а иногда и об уровне доходов. Большинство людей добровольно указывают всю эту информацию о себе в социальных сетях [safe-surf, 2013].

Даже если человек не ведет страницу в соцсети или заведомо не выкладывает информацию о себе, о таком человеке все равно можно многое узнать, например, путем анализа данных о его друзьях. На основе полученного портрета появляется возможность спрогнозировать поступки человека, а в определенных случаях и повлиять на его поведение и принятие решений. В современную эпоху всеобщей цифровизации и «больших данных» все это стало реальностью [safe-surf, 2013].

Одной из задач, которые нужно решить в процессе разработки сервиса формирования социального портрета пользователя социальной сети ВКонтакте, является разработка метода интеллектуального анализа данных, содержащихся на странице пользователя (профиль) социальной сети.

В работе будет описан анализ существующих, выбор и адаптация алгоритма кластеризации для решения задачи анализа данных о пользователе социальной сети. Такой анализ позволит провести сегментацию аудитории сообщества пользователей и сегментацию аудитории социальной сети и получить «усредненный» социальный портрет типового пользователя для каждого сегмента.

1 Обзор алгоритмов и методов кластеризации

Формальной задачей кластеризации является разбиение заданного множества объектов на непересекающиеся подмножества, называемые кластерами, так, чтобы кластеры состояли из похожих объектов, а объекты разных кластеров существенно отличались. Кластеризация является одной из фундаментальных задач в Data Mining и часто выступает первым шагом при анализе данных: выделение групп похожих

объектов помогает понять структуру данных и использовать свой подход к обработке каждой группы [Варламов, 2012].

Если обобщить различные классификации методов кластеризации, то можно выделить ряд групп [Филиппов, 2013]:

1. Вероятностный подход. Предполагается, что каждый рассматриваемый объект относится к одному из k классов.
   - k-средних (K-means);
   - k-medians;
   - em-алгоритм;
   - алгоритмы семейства FOREL;
   - дискриминантный анализ.
2. Подходы на основе систем искусственного интеллекта.

• метод нечеткой кластеризации C-средних (C-means);

   - нейронная сеть Кохонена;
   - генетический алгоритм.

3. Логический подход. Построение дендрограммы осуществляется с помощью дерева решений.
4. Теоретико-графовый подход.
5. Иерархический подход.
6. Другие методы. Не вошедшие в предыдущие группы.

Все методы разбиения множества на кластеры можно разделить на:

- неиерархические;
- иерархические.

Неиерархические методы кластерного анализа сложны в настройке, так как требуют определения различных параметров, например, количество кластеров, критерий остановки, и т. д. Однако данные методы более устойчивы к шумам, выбору метрики и включению незначимых переменных в набор исходных данных кластерного анализа [Филиппов, 2013].

Иерархические методы кластерного анализа, в отличие от неиерархических, строят полное дерево вложенных кластеров (дендрограмм) и не требуют указания числа кластеров, но при этом данные методы накладывают ограничение на объем набора данных, зависимы от выбора меры близости и имеют меньшую гибкость полученных решений [Филиппов, 2013].

Иерархические методы кластерного анализа подразделяются на [Филиппов, 2013]:

- агломеративные, характеризуемые последовательным объединением исходных элементов и соответствующим уменьшением числа кластеров (построение кластеров снизу-вверх);

- дивизимные (делимые), в которых число кластеров возрастает, начиная с одного, в результате чего образуется последовательность расщепляющих групп (построение кластеров сверху вниз).

Для целей исследования был выбран метод нечеткой кластеризации c-means. Этот метод позволяет учесть тот факт, что пользователь относится к более чем одному кластеру с разной степенью принадлежности. Алгоритм Fuzzy C-Means является обобщением алгоритма k-means с учетом представления кластеров нечеткими множествами, каждая точка которых принадлежит различным кластерам с различной степенью принадлежности. Точка относится к тому или иному кластеру по критерию максимума принадлежности данному кластеру [Филиппов, 2013].

Основная идея алгоритма: кластеры представляют собой нечеткие множества, т.е. каждый объект принадлежит всем кластерам с разной степенью принадлежности [А. Кулажский].

Достоинства: нечеткость при определении объекта в кластер позволяет найти объекты, которые находятся на границе, в кластеры [А. Кулажский].

Недостатки: [А. Кулажский]

- вычислительная сложность,
- задание количества кластеров,
- возникает неопределенность с объектами, которые удалены от центров всех кластеров.

Алгоритм:

1. Задать случайным образом k центров кластеров c_j , $j = 1 \dots k$;

2. Рассчитать матрицу принадлежности элементов к кластерам r. В случае нормального распределения: $r_{ij} = \frac{N(d(x_i,c_j)|\mu=0,\sigma)}{\sum_j^k N(d(x_i,c_j)|\mu=0,\sigma)}$,

где x_i $-i$-й элемент множества, c_j $-$цr_{ij}ентр кластераj, $d(x_i, c_j)$– расстояние между точкамиx_i и c_j, $N-$ это плотность вероятности нормального распределения в точке$d(x_i, c_j)$;

3. Переместить центры кластеров $\mathrm{c}_j \leftarrow \frac{\sum_i r_{ij} x_i}{\sum_i r_{ij}}$;

4. Рассчитать функцию потерь. В случае нормального распределения функция потерь будет равна:$J = \sum_j^k \sum_i^N d(x_i, c_j)^2 \, r_{ij}$;

5. Если значение функции потерь уменьшается, то повторить цикл с п.2.

2 Адаптация алгоритма кластеризации для решения задачи анализа данных группы пользователей социальной сети

Кластерный анализ позволит провести сегментацию аудитории группы и получить социальный портрет пользователя из каждого сегмента, опираясь на данные, представленные центром каждого кластера.

Как было сказано ранее, в п. 2 настоящей статьи, первым этапом алгоритма нечеткой кластеризации Fuzzy c-means является инициализация. Этот этап включает в себя выбор метрики расстояния от объекта до центра кластера.

Метрика расстояния – метод определения расстояния между входными векторами (определение их сходства или различия). Выбор способа вычисления расстояния зависит от природы исследуемых объектов и непосредственно влияет на результат [Филиппов, 2013].

В нашем случае входной вектор будет представлен набором характеристик, содержащихся в профиле пользователя социальной сети Вконтакте.

Для целей исследования было принято решение работать со следующими данными, содержащимися в профиле пользователя:

- пол,
- день рождения,
- образование,
- город.

Эти данные можно трансформировать в следующие характеристики:

- пол -> пол,
- день рождения -> возраст,
- образование -> образование,
- город -> географическое местоположение.

Все эти характеристики имеют различный тип данных (таблица 1).

Таблица 1 – Типы данных характеристик объекта кластеризации

Характеристика	Тип данных
Пол	Булево
Возраст	Дата, Число
Образование	Составной тип данных
Город	Гео-координаты

Таким образом, целесообразно рассмотреть особенности выбора метрики расстояния для каждой характеристики в отдельности.

Учитывая, что характеристика Пол может принимать одно из двух значений «Мужской» или «Женский», ее тип данных – булево. Тогда

расстояние между значениями координат можно найти как совпадение или не совпадение значений. В том случае если они совпадают, то расстояние будет равным 0. В другом случае расстояние будет равно 1.

Данные о возрасте в профиле могут быть представлены разными способами. В первом случае пользователь может указать дату своего рождения. Во втором случае пользователь может ничего не указать в поле Дата рождения. В таком случае можно будет с некоторой степенью вероятности получить характеристику Возраст с использованием другой информации со страницы пользователя, например, Дата окончания школы. Тогда расстояние между значениями координат можно найти как разницу между датами. Значения возраста для каждого объекта будут получены в виде чисел в результате предварительной предобработки данных из профиля путем перевода даты рождения в числовое значение возраста.

Характеристика Образование включает в себя большой объем информации со сложной структурой: Уровень образования, Город, Название вуза, Специальность и т. д. Возможность вычисления меры расстояния этой характеристики будет достигнута с использованием графовой структуры. В рамках данной статьи будет рассмотрен пример вычисления расстояния по элементу этой структуры – Уровень образования. Значениями характеристики по этому элементу могут быть: «Среднее», «Средне-специальное» или «Высшее». Для удобства расчетов присвоим каждому качественному значению характеристики числовое следующим образом: «Среднее» - 0, «Средне-специальное» - 1, «Высшее» - 2.

Характеристика Город получается путем преобразования наименования города из профиля пользователя в географические координаты. Это даст возможность вычислить расстояние между объектами. Кратчайшее расстояние между двумя точками на земной поверхности (если принять ее за сферу) определяется зависимостью:

$$\cos d = \sin \varphi_A * \sin \varphi_B + \cos \varphi_A * \cos \varphi_B * \cos(\lambda_A - \lambda_B), \qquad (1)$$

где φ_A и φ_B– широты, λ_A и λ_B–долготы данных пунктов, d– расстояние между пунктами, измеряемое в радианах длиной дуги большого круга земного шара.

Расстояние между пунктами, измеряемое в километрах, определяется по формуле

$$L = d * R, \qquad (2)$$

где $R = 6371$ км – средний радиус земного шара.

Таким образом, функция определения расстояния между объектами будет иметь вид

$$D=\sum_{i=1}^{4}(n_i - r_i), \quad (3)$$

где n_i – значение *i-й* характеристики первого объекта, *а* r_i – значение *i-й* характеристики второго объекта.

В качестве примера рассмотрим сообщество пользователей социальной сети ВКонтакте, участники которого увлекаются музыкой. В этом сообществе состоят пользователи со следующими значениями характеристик, указанных выше:

1. Профиль_1{Мужской, 60, Высшее, Ульяновск};
2. Профиль_2{Мужской, 23, Среднее, Москва};
3. Профиль_3{Женский, 53, Высшее, Москва};
4. Профиль_4{Женский, 57, Высшее, Димитровград};
5. Профиль_5{Мужской, 21, Высшее, Ногинск};
6. Профиль_6{Женский, 22, Высшее, Димитровград};
7. Профиль_7{Мужской, 56, Среднее специальное, Барыш};
8. Профиль_8{Женский, 25, Среднее, Красногорск}.

Рассмотрим пример нахождения расстояния между первым и остальными объектами. Все объекты представлены характеристиками, рассмотренными ранее. Тогда расстояние между первым (Профиль_1) и вторым объектом (Профиль_2) можно рассчитать следующим образом:

Профиль_1(Пол)= «Мужской», Профиль_2(Пол)= «Мужской». Значения характеристики первого и второго объекта совпадают, следовательно, расстояние будет равно 0.

Профиль_1(Возраст)= 60, Профиль_2(Возраст)= 23. Расстояние между первым и вторым объектом по характеристике Возраст можно найти путем разности значений. В данном случае получим 37.

Профиль_1(Образование)= Высшее, Профиль_2(Образование)= Среднее. С учетом алгоритма поиска расстояния по характеристике Образование, описанному выше, в данном случае оно будет равно 1.

Профиль_1(Город)= Ульяновск, Профиль_2(Город)= Москва. Географические координаты города Ульяновск: 54°19,00' северной широты и 48°23,00' восточной долготы. Координаты города Москва: 55° 45' 07" северной широты 37° 36' 59" восточной долготы. После подстановки этих значений в формулу 1 получим расстояние между объектами по формуле 2, равное 703,648 км.

Полученные значения расстояния между объектами по характеристикам «Возраст», «Образование» и «Город», перед расчетом итогового расстояния, необходимо нормировать с целью получения значений в диапазоне от 0 до 1. Нормирование значений производится с помощью наиболее общеупотребляемого способа приведения критериев к безразмерному виду – линейной трансформации:

$$y(x) = \frac{x - x_{min}}{x_{max} - x_{min}}, \quad (4)$$

где x – значение i-й характеристики объекта, x_{min} – минимальное значение i-й характеристики объекта из всей выборки, x_{max} – максимальное значение i-й характеристики объекта из всей выборки.

Подставив полученные после нормировки по формуле (4) значения расстояний по каждой характеристике в формулу (3), получим, что расстояние между объектом Профиль_1 и объектом Профиль_2: $D = 0 + 0{,}81 + 0{,}5 + 0{,}97 = 2{,}28$.

Подобным образом были произведены расчеты расстояний между объектом Профиль_1 и остальными объектами, результаты которых представлены в таблице 2.

Таблица 2 – Расстояния между объектом Профиль_1 и остальными объектами

Профиль	Расстояние Пол	Расстояние Возраст	Расстояние Образование	Расстояние Город	Итоговое расстояние
Профиль_2	0	0,81	0,5	0,97	2,28
Профиль_3	1	0,1	0	0,97	2,07
Профиль_4	1	0	0	0	1
Профиль_5	0	0,86	0	0,89	1,75
Профиль_6	1	0,83	0	0	1,83
Профиль_7	0	0,23	1	0,05	1,28
Профиль_8	1	1	0,5	1	3,5

Заключение

В статье был проведен анализ существующих, выбор и адаптация алгоритма кластеризации для решения задачи анализа данных группы пользователей социальной сети при разработке сервиса для формирования социального портрета пользователя на основе интеллектуального анализа содержимого социальных сетей. Такой анализ позволит провести сегментацию аудитории группы и получить социальный портрет пользователя из каждого сегмента.

Был проведен обзор методов кластеризации и обоснован выбор метода нечеткой кластеризации Fuzzy c-means. Для этого метода были описаны особенности выбора метрики расстояния между объектами.

Благодарности. Авторы выражают благодарность Фонду содействия инновациям за помощь в финансировании исследования.

Список литературы

[safe-surf, 2013] safe-surf, Социальный портрет пользователя. Кто и зачем его составляет. [Электронный ресурс] //Безопасность пользователей в сети интернет: [сайт].URL: https://safe-surf.ru/users-of/article/611617/ (дата обращения: 11.07.2019).

[Варламов, 2012] Сравнение алгоритма кластеризации на основе отношения α-квазиэквивалентности с классическими иерархическими алгоритмами на синтетических наборах данных [Электронный ресурс] //[сайт].URL: http://seminar.at.ispras.ru/wp-content/uploads/2012/07/Varlamov-thesis.pdf (дата обращения: 11.07.2019).

[Филиппов, 2013] Формирование навигационной структуры электронного архива технических документов на основе онтологических моделей / Филиппов А.А.:
диссертация на соискание ученой степени кандидата технических наук / Ульяновский государственный технический университет. – Ульяновск, – 2013.

[А. Кулажский] А. Кулажский, Персональный блог Артема Кулажского [Электронный ресурс] //Алгоритмы: [сайт].URL: http://blogs.it-claim.ru/akulazhski/algoritmy/ (дата обращения: 11.07.2019).

DEVELOPMENT OF A SERVICE FOR THE FORMATION OF A USER'S SOCIAL PORTRAIT ON THE BASIS OF INTELLECTUAL ANALYSIS OF THE CONTENTS OF DATA IN OPEN SOURCES

Grigoricheva M.S. (gms4295@mail.ru)
Ulyanovsk State Technical University, Ulyanovsk

The paper describes the analysis of existing, selection and adaptation of the clustering algorithm for solving the problem of creating a social portrait of a user based on the intellectual analysis of data content in open sources using the example of the social network VKontakte. Such an analysis will allow a segmentation of the social network audience and a “average” social portrait of a typical user for each segment.

Keywords: social portrait, clustering, social networks

УДК 004.93'14

РАСПОЗНАВАНИЕ СЛИВОВ ТОПЛИВА В СЕЛЬСКОХОЗЯЙСТВЕННОЙ ТЕХНИКЕ

Анашкина Ю.В.(julyaanashkin@yandex.ru)

БГТУ им. В.Г.Шухова, Белгород

Хищение топлива – распространенная проблема в сельском хозяйстве. В работе описывается алгоритм распознавания, основанный на методе экспоненциального сглаживания. Разработанный метод с достаточной точностью определяет слив или заправку топлива, учитывает колебание уровня топлива при движении и стоянке. Применение алгоритма распознавания позволит повысить достоверность определения слива и сократить время обработки данных.

Ключевые слова: хищение топлива, алгоритмы распознавания, оценка Розенблатта-Парзена, метод экспоненциального сглаживания

Введение

Распространенным видом хищения в сельском хозяйстве является слив топлива [Барышников и др., 2018]. Одним из способов предотвращения хищения является анализ данных, получаемых с датчиков уровня топлива (ДУТ), которые фиксируют текущий уровень топлива в баке. Показания ДУТ образуют временной ряд, представляющий собой последовательность $(t_i, dut_i)_{i=\overline{1,n}}$, где dut_i – уровень топлива в баке в момент времени t_i. Аномалии в поведении этого ряда интерпретируются как подозрения на слив или заправку. Задача состоит в разработке алгоритма автоматического распознавания фактов заправки и сливов топлива и их значений.

Разработка алгоритма распознавания заправок и сливов топлива по показаниям ДУТ является актуальной задачей, так как позволяет компаниям экономить значительные средства путем выявления фактов хищения топлива. Основная проблема при разработке такого алгоритма состоит в значительных колебаниях показаний ДУТ в процессе движения

и стоянок транспортного средства, которые могут достигать 20% объемов бака, что составляет 100 – 1200 литров для сельскохозяйственной техники. Поэтому разрабатываемый алгоритм должен распознавать заправки и сливы топлива с приемлемой на практике достоверностью определения самого факта заправки или слива и их значений. Основным подходом в решении этой задачи является сглаживание показателей колебания уровня топлива и оценка их фактических значений.

Одним из распространенных алгоритмов сглаживания является алгоритм скользящее среднее [Сарычев, 2009]. Этот алгоритм относится к семейству функций, значения которых в каждой точке определения равны среднему значению исходной функции за предыдущий период:

$$ma_j = \frac{\sum_{i=j-\frac{w}{2}}^{j+\frac{w}{2}-1} y_i}{w},$$

где w – окно скользящего среднего: чем оно больше, тем больше данных участвует в расчете среднего, тем более гладкой получается кривая [Сарычев, 2009].

Такой подход применяется с данными временных рядов для сглаживания краткосрочных колебаний и позволяет выделить основные тенденции.

Для решения задачи распознавания слива топлива данный метод не подошел. Так как алгоритм скользящего среднего сглаживает мелкие неровности, тем самым практически повторяя исходную кривую данных. Алгоритм не способен учесть резкого скачка уровня топлива при торможении или начале движения, из-за чего невозможно обеспечить достаточную точность распознавания.

Подход, предлагаемый в настоящей статье, состоит в следующем. Фактический расход топлива аппроксимируется суммой кусочно-линейной функции и случайной величины с неизвестным распределением, моделирующей всевозможные колебания уровня топлива в баке. Плотность распределения восстанавливается методом Розенблатта-Парзена, а затем данные ДУТ сглаживаются при помощи экспоненциального фильтра.

1 Восстановление плотности распределения

Метод Розенблатта-Парзена был предложен в 1956 году Мюрреем Розенблаттом и позже обобщен Эммануэлем Парзеном [Добронец и др., 2017]. В основе метода лежит ядерная оценка плотности – непараметрический способ оценки плотности случайной величины. Это метод аппроксимации данных, в котором делается заключение о

распределении генеральной совокупности, основываясь на конечных выборках данных.

Метод Розенблатта-Парзена основан на предположении, что плотность распределения вероятности связана с функцией распределения через оператора дифференцирования [Бардасов, 2017], который заменяется конечной разностью:

$$f'(x) = \frac{dF(x)}{dx} \approx \frac{F(x+h) - F(x-h)}{2h}. \quad (1)$$

Формула (1) берется за основу при построении оценки плотности распределения вероятности $f_n(x)$.

Функции $F(x+h)$, $F(x-h)$ заменяются простейшими кусочно-постоянными оценками, построенными на основе выборки $x_1, \dots, x_n$ для одномерной случайной величины X. Тогда оценка плотности вероятности приобретает вид:

$$f_n(x) = \frac{\frac{1}{n}\sum_{i=1}^{n} 1(x+h-x_i) - \frac{1}{n}\sum_{i=1}^{n} 1(x-h-x_i)}{2h} = \\ = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} \frac{1(x+h-x_i) - 1(x-h-x_i)}{2}. \quad (2)$$

В формуле (2) выражение под знаком суммы называют селектирующим прямоугольным ядром:

$$\frac{1(x+h-x_i) - 1(x-h-x_i)}{2} = I\left(\frac{x-x_i}{h}\right).$$

Ядро $I(z)$ симметричное, и площадь под кривой $I(z)$ равна единице. Эта функция ничем не отличается от плотности распределения вероятности случайной величины z, равномерно распределенной в интервале [-1;1]. В результате оценка для плотности распределения вероятности имеет вид:

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} I\left(\frac{x-x_i}{h}\right).$$

Заменим в оценке $f_n(x)$ прямоугольное ядро $I(z)$ на произвольное $K(z)$ и получим формулу оценки Розенблатта-Парзена:

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x-x_i}{h}\right),$$

где $K(z)$ – произвольная четная функция, называемая функцией ядра или окна, h – коэффициент размытости ядра или ширина окна. Степень гладкости оценки плотности зависит от степени гладкости ядра, поэтому на практике обычно используются более гладкие функции. Ширина окна

сильно влияет на качество восстановления плотности [Мещеряков и др., 2017] и, как следствие, классификации. При слишком малом окне плотность концентрируется вблизи экстремумов. При слишком большом окне плотность вырождается в константу.

2 Распознавание сливов и заправок топлива

2.1 Алгоритм распознавания. Заметим, что в задаче распознавания сливов топлива требуется не восстанавливать плотность распределения случайных колебаний, а сглаживать их. Поэтому метод Розенблатта-Парзена был применен в сочетании с экспоненциальным фильтром. В результате был получен следующий алгоритм:

1. Для каждой пары $(t_i, dut_i)_{i=\overline{1,n}}$ вычисляем дискретную производную по формуле:
$$f'(t_i) = \frac{dut_{i+p} - dut_i}{t_{i+p} - t_i}, i = \overline{1,n}.$$
2. Производим нормировку величины $f'(t_i)$, чтобы определить относительную выраженность некоторых показателей уровня топлива относительно других. Ищем максимальную по модулю производную и делим на нее все элементы $f'(t_i)$:
$$\mathrm{f'(t_i)} = \frac{\mathrm{f'(t_i)}}{\max\left(|\mathrm{f'(t)}|\right)}.$$
3. Произведем сглаживание распределения с помощью экспоненциального фильтра $f_n\big(f'(t_i)\big)$:
$$f_n\big(f'(t_i)\big) = \frac{2 \cdot e^{k f'(t_i)} - 2}{e^k + e^{-k}},$$
где $k > 0$ – параметр сглаживания.
4. Производим фильтрацию результатов оценки плотности распределения вероятности. Сравниваем модуль оценки $f_n\big(f'(t_i)\big)$ с максимальным по модулю значением оценки на всем временном отрезке, умноженным на границу максимальной разности – border. То есть из исходного временного ряда выделяем события, удовлетворяющие условию:
$$\big|f_n\big(f'(t_i)\big)\big| \geq border * max\big(\big|f_n\big(f'(t)\big)\big|\big).$$
5. Затем классифицируем подозрительные события на заправку и слив в зависимости от знака $f_n\big(f'(t_i)\big)$. Если оценка положительная, то была произведена заправка. Если оценка отрицательная, то был произведен слив.

2.2 Результаты распознавания

Алгоритм фильтра с экспоненциальным сглаживанием был программно реализован и протестирован на реальных и имитационных данных.

На следующих рисунках представлены результаты работы предложенного алгоритма на двух наборах реальных данных при $k = 2{,}5$ и $border = 0{,}6$ (значения параметров выбраны эмпирически). Данные различаются объемом выборки и характером передвижения транспортного средства (ТС).

На графиках показан результат оценки изменения уровня топлива для двух состояний ТС: «ТС в движении» и «ТС в покое». Кривая черного цвета отображает оценку методом экспоненциального сглаживания. Серая кривая – первая производная по текущему и предыдущему нормализованному значению ДУТ, это первый шаг метода экспоненциального сглаживания.

Тестовые данные №1:

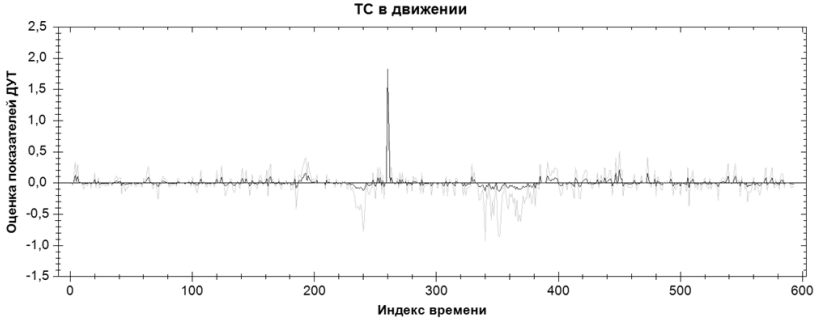


Рисунок 1 – Тестовые данные №1, ТС в движении

На рисунке 1 показан результат оценки изменения уровня топлива, когда ТС находится в движении. Хорошо выраженный максимум черной кривой в районе 259 отметки времени соответствует значительному изменению уровня топлива в баке. Серая кривая в этой области возрастает – алгоритм обнаружил заправку.

Заметим, что резкие колебания топлива также имеются в районе 340 отметки времени. При использовании скользящего среднего алгоритм интерпретирует их как слив топлива, а при экспоненциальной фильтрации эти колебания игнорируются. Фактически в это время техника работала в поле, т. е. колебания показаний ДУТ вызваны естественными причинами.

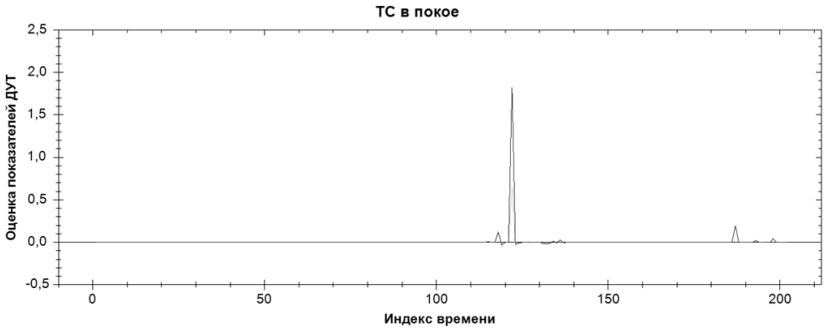


Рисунок 2 – Тестовые данные №1, ТС в покое

На рисунке 2 показан результат оценки изменения уровня топлива, в период, когда транспортное средство не использовалось. Отчетливо видно, что в районе 121 отметки времени алгоритм фиксирует заправку.

Тестовые данные №2:

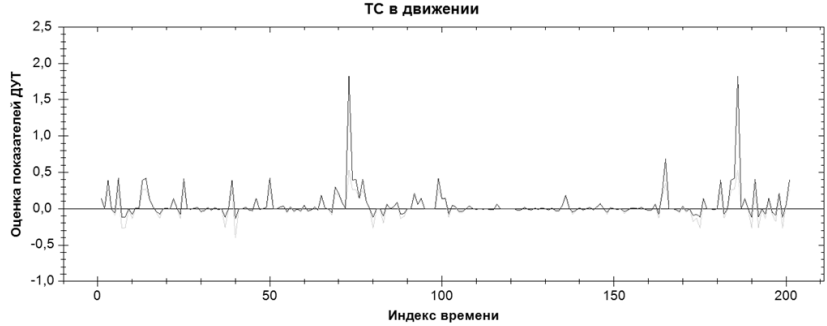


Рисунок 3 – Тестовые данные №2, ТС в движении

На рисунке 3 показан результат оценки изменения уровня топлива в период работы транспортного средства. Значительное изменение уровня топлива наблюдается в районе 72 и 185 отметок времени – алгоритм зафиксировал 2 заправки.

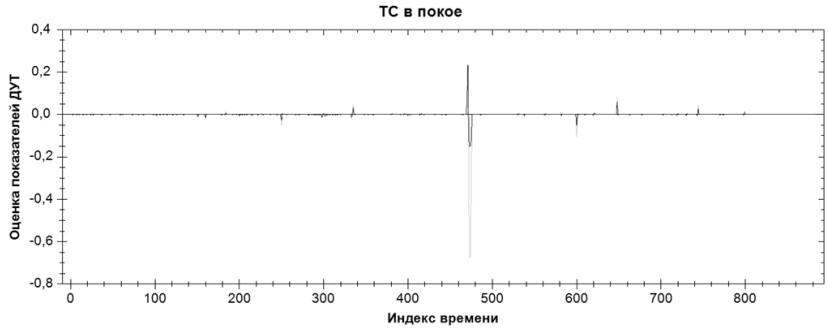


Рисунок 4 – Тестовые данные №2, ТС в покое

На рисунке 4 показан результат оценки изменения уровня топлива в период, когда транспортное средство не использовалось. Отчетливо видно, что в районе 470 отметки времени алгоритм фиксирует незначительную заправку, а в районе 473 отметки времени был зафиксирован слив. В действительности зафиксированный слив был необходим производству для дальнейшего ремонта бака транспортного средства.

Алгоритм фильтра с экспоненциальным сглаживанием так же был протестирован и на имитационных данных, получаемых с генератора показаний ДУТ. Было проведено 100 испытаний, из которых в 86 случаях были правильно распознаны сливы или заправки топлива. На основании проведенных испытаний можно утверждать, что точность распознавания предложенного метода распознавания сливов или заправок топлива составляет 86,3%. Точность распознавания можно улучшить путем автоподбора параметров k и $border$ методом кросс-валидации.

Предлагаемый алгоритм позволяет не только достаточно точно обнаруживать факты заправок или сливов топлива, но и защищен от ложных срабатываний, что позволяет уберечь компании от потери денежных средств и сэкономить время, затраченное аналитиком на поиск сливов топлива.

Заключение

Контактное лицо: Анашкина Юлия Владимировна, студентка 2 курса магистратуры.

Список литературы

[Бардасов, 2017] Бардасов С.А. Модифицированный метод перекрестной оценки ширины окна ядерной функции Розенблатта-Парзена // Инновации в науке. – 2017. – № 4 (65).

[Барышников и др., 2018] Барышников Н.Г., Мурзин Д.А. Контроль горюче-смазочных материалов в сельхозорганизациях // Модели, системы, сети в экономике, технике, природе и обществе. – 2018. – № 3 (27).

[Добронец и др., 2017] Добронец Б.С., Попова О.А. Численные операции над ядерными оценками в задаче восстановления функции плотности вероятности // Марчуковские научные чтения – 2017: Труды Международной научной конференции, 2017.

[Мещеряков и др., 2017] Мещеряков М.В., Кузнецова Е.А. Исследование влияния ядерной функции на качество аппроксимации в методе Парзена-Розенблатта // Телекоммуникационные устройства и системы. – 2017. – Т. 7. – № 2.

[Сарычев, 2009] Сарычев В.В. Оценка эффективности сглаживания сигнала по регулярным и нерегулярным отсчетам // Известия ЮФУ. Технические науки. – 2009. – № 1 (90).

RECOGNITION OF FUEL PLUMES IN AGRICULTURAL MACHINERY

Anashkina Yu.V.(julyaanashkin@yandex.ru)
BSTU them. V.G.Shukhova, Belgorod

Fuel theft is a common problem in agriculture. The recognition algorithm described in the paper is based on a method of exponential data smoothing. The developed method determines fuel draining or fueling up with sufficient accuracy, while taking into account deviations of fuel level while in motion or standstill. The use of the recognition method will enable increase of reliability in determining fuel drain and reduce duration of data processing.

Keywords: fuel theft, recognition algorithms, Rosenblatt-Parzen estimation, method of exponential data smoothing

УДК 004.94

РАЗРАБОТКА СИСТЕМЫ ИНТЕГРАЦИИ РАБОТЫ С ГРАНТАМИ С КАФЕДРАЛЬНЫМ ПРИЛОЖЕНИЕМ

Белоусова Т.С. (zirael36@gmail.com)
Ульяновский государственный технический университет, Ульяновск

Описываются объект автоматизации, цель, задачи и актуальность разработки системы интеграции работы с грантами с кафедральным приложением «NG-Tracker».

Ключевые слова: интеграция, система, NG-Tracker, грант

Введение

Разработка системы интеграции работы с грантами с кафедральным приложением «NG-Tracker» производилась в рамках работы над выпускной квалификационной работой по специальности «Программная инженерия» кафедры «Информационные системы» Ульяновского государственного технического университета. «NG-Tracker» – информационная система для автоматизации деятельности научной группы в различных областях, таких как гранты, статьи, конференции и др. Цели создания данной информационной системы:

- получить удобный инструмент для сокращения времени управления текущими активностями научной группы;
- создать новую функцию автоматизированной системы управления задачами – интеллектуальную постановку задач исполнителям;
- создать систему хранения и трансляции опыта между участниками научной группы.

На рисунке 1 представлена домашняя страница кафедрального приложения. Она содержит модули, представляющие различные области деятельности сотрудников кафедры. В данной статье описывается разработка системы, отвечающей за работу с грантами.

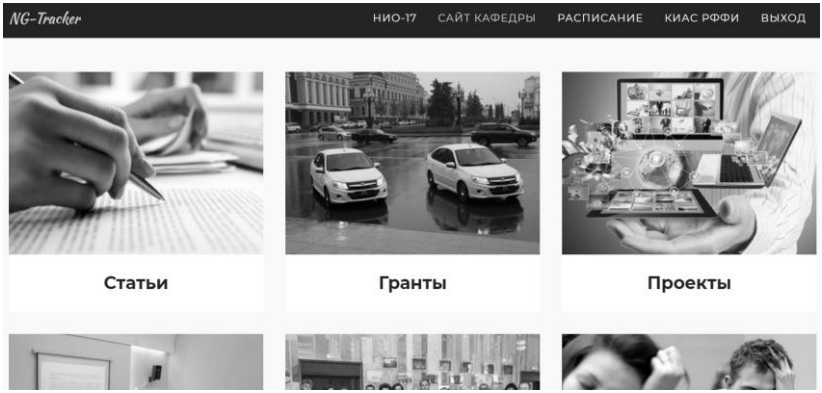


Рисунок 1 – Главная страница

1 Общие положения

Система интеграции работы с грантами с кафедральным приложением предназначена для повышения производительности труда работников кафедры и автоматизации их деятельности, в части исполнения следующих процессов, относящихся к грантам:

- создание и управление грантами;
- планирование календарного плана работ по гранту;
- подбор членов рабочей группы по заданным критериям, учитывая загруженность другими проектами;
- оперативное отслеживание всех изменений в гранте путем отправки уведомлений об изменениях на почту членам рабочей группы гранта;
- надежное хранение информации и файлов в общем доступе;
- контролирование выполнения научных и публикационных показателей гранта, заявленных при создании.

Объектом автоматизации являлась рабочая группа по участию в конкурсах на получение грантов, а предметом автоматизации – учет информации по заявкам на гранты, их статусам и выполняемым работам.

Необходимость создания такой системы обусловлена тем, что кафедра «Информационные системы» нуждалась в реализации собственного модуля по работе с грантами как части полноценной системы по автоматизации многих деятельностей кафедры. Ранее работа с грантами осуществлялась на основе использования бумажных документов и устных договоренностей относительно плана работ и координации действий сотрудников. В сравнении с предыдущим новое решение позволило

исключить всю бумажную работу, обеспечить надежное хранение данных, а также скоординировать деятельность сотрудников посредством календарного плана работ, сформированного автоматически в системе.

Основной целью программного продукта является автоматизация работы сотрудников кафедры, повышение оперативности информирования о статусе работ по гранту и снижение трудоемкости работы благодаря осуществлению автоматического планирования задач по грантам. Такое решение приведет к повышению производительности труда.

2 Основные технические решения

Вопрос выбора технологий не был актуальным, так как работа велась над одним из модулей кафедрального приложения, в котором другие модули уже частично были реализованы, соответственно стек технологий был заранее определен. Было реализовано веб-приложение, языком реализации был выбран Java. Стек технологий:

1. WEB – основной компонент для разработки web-приложения;
2. JPA – технология, требуемая для работы с базами данных и обеспечивающая объектно-реляционное отображение объектов;
3. Hibernate – популярная реализация ORM-модели;
4. Spring Boot – помогает облегчить разработку приложений за счет настройки зависимостей между компонентами приложения;
5. Thymeleaf – современный серверный механизм Java-шаблонов.

На рисунке 2 представлена форма редактирования гранта. Функции создания и редактирования гранта были принято разместить на одной странице – при открытии уже существующего гранта все поля автоматически заполняются сохраненной ранее информацией, при создании нового – все поля пустые.

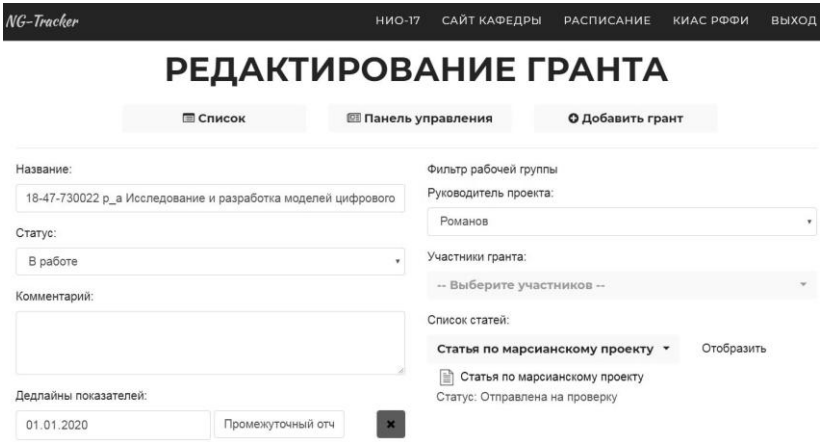


Рисунок 2 – Страница просмотра существующего гранта

На данной странице осуществляется практически все управление грантом и его показателями. Отображается название, статус и комментарий к гранту. В дедлайнах фиксируется этапность выполнения научных и публикационных показателей гранта, они рассматриваются как средство уведомления. Фильтр рабочей группы применяется при выборе сотрудников для участия в гранте. Фильтр распространяется и на поле руководителя гранта, и на поле участников и является свернутым, если фильтры не были применены. Так же к гранту можно прикрепить статьи, которые еще не являются завершенными и опубликованными. Далее по их статусу производится контроль выполнения показателей – если статус статьи «завершена», значит, показатель выполнен. Возможен так же и быстрый переход к прикрепленной статье – здесь уже задействуется модуль «Статьи» кафедрального приложения. Для хранения файла заявки на грант, а также других необходимых в процессе работы над грантом файлов было реализовано прикрепление и отображение файлов. Интеграция выполнялась с сайтом https://kias.rfbr.ru, оттуда в автоматическом режиме в систему загружались гранты за текущий год, заявки на которые принимались в текущий момент.

Алгоритм создания гранта:

1. Авторизоваться в системе;
2. Перейти на страницу грантов;
3. Выбрать функцию «Добавить грант», переход на страницу создания и редактирования гранта;
4. Заполнить все поля, поля «Название», «Руководитель гранта» и «Дедлайн» являются обязательными;

5. Нажать кнопку «Сохранить»;
6. Проверка уникальности гранта по названию. Если такого гранта еще не существует в системе, то в списке грантов отобразится только что созданный грант, иначе отображение пользователю ошибки.

Алгоритм интеграции с системой КИАС:

1. Автоматический вызов планировщиком функции загрузки грантов по заданному расписанию;
2. Запуск браузера в headless режиме;
3. Переход на страницу с конкурсами, установка фильтров – текущий год и возможность подать заявку на грант, автоматизированный сбор информации о грантах;
4. Формирование списка грантов, проверка их уникальности в системе;
5. Уникальные гранты сохраняются.

Заключение

Была разработана система интеграции работы с грантами с кафедральным приложением, работа сотрудников кафедры в отношении грантов автоматизирована. Программный продукт реализует все бизнес-задачи, которые требовалось реализовать по плану.

DEVELOPMENT OF THE SYSTEM OF INTEGRATION OF WORK WITH GRANTS WITH THE CATHEDRAL APPLICATION

Belousova T.S. (ziraeI36@gmail.com)
Ulyanovsk State Technical University, Ulyanovsk

The paper describes the automation object, the purpose, objectives and the relevance of developing a system of integration of work with grants with the NG-Tracker cathedral application.

Keywords: integration, system, NG-Tracker, grant

УДК 665.6

РАЗРАБОТКА АВТОМАТИЗИРОВАННОЙ СИСТЕМЫ ПРОМЫШЛЕННОЙ БЕЗОПАСНОСТИ ДЛЯ ПРЕДПРИЯТИЙ НЕФТЕГАЗОВОГО СЕКТОРА

Илюшин П.Ю. (ilushin-pavel@yandex.ru)
Лекомцев А.В. (alex.lekomtsev@mail.ru)
Галкин С.В. (gnfd@pstu.ru)
Пермский национальный исследовательский политехнический университет, Пермь

Приведено описание разработанной автоматизированной системы промышленной безопасности (АСПБ), успешно опробованной на одной из производственных площадок ООО «ЛУКОЙЛ-ПЕРМЬ». В результате проведенного тестирования прототипа АСПБ обеспечена автоматизация процесса в области допуска на проведение работ повышенной опасности, подготовка и тестирование специалистов в области проведения работ повышенной опасности.

Ключевые слова: система управления, промышленная безопасность, охрана труда.

Введение

Особенности современного производства обусловливают необходимость создания адекватной корпоративной модели управления промышленной безопасностью и безопасностью труда [Чуйко, 2011]. Для этого целесообразно разрабатывать системы управления, направленные на обеспечение оперативного и эффективного процесса производственной деятельности предприятия в сфере охраны труда и промышленной безопасности [Сюч, 2008]. Создание автоматизированной системы управления промышленной безопасностью и охраной труда должно базироваться на современных требованиях к подобным системам управления, сформулированных в стандартах ISO 9001, ISO 14001, OHSAS 18001 и других документах [Марков, 2011].

1 Система промышленной безопасности

Группой авторов разработана Автоматизированная система промышленной безопасности (АСПБ), предназначенная для руководителей и специалистов предприятия, которая позволяет оперативно получать информацию о техническом состоянии оборудования, устройств, зданий и сооружений на опасном объекте; о характеристиках и соответствии персонала и лиц, допущенных на опасный объект; о ситуации с хранением и транспортировкой опасных веществ; о результатах проверок производственного контроля и о статусе выполнения корректирующих мероприятий и предписаний; о внутренних нормативах и документах по промышленной безопасности.

Основными функциональными задачами при эксплуатации АСПБ являются:

- создание единой базы данных по промышленной безопасности на предприятии;
- функционирование действенной системы контроля над соблюдением правил промышленной безопасности;
- оперативный контроль текущего состояния опасного объекта и соответствия требованиям промышленной безопасности;
- оперативный «план-факт» контроль выполнения мероприятий и предписаний производственного контроля;
- контроль соблюдения персоналом внутренних регламентов и правил выполнения процессов;
- снижение рисков аварий и аварийных ситуаций по причинам несоблюдения требований промышленной безопасности;
- выполнение требований законодательства по функционированию системы управления промышленной безопасностью.

Система АСПБ направлена на автоматизацию процесса согласования выдачи наряд-допуска [Белослудцев, 2019]. Она является многопользовательской с различными уровнями доступа согласно занимаемой должности и состоит из двух укрупненных функциональных блоков: нормативно-справочная информация и журнал наряд-допусков. Рабочее пространство системы показано на рисунке 1.

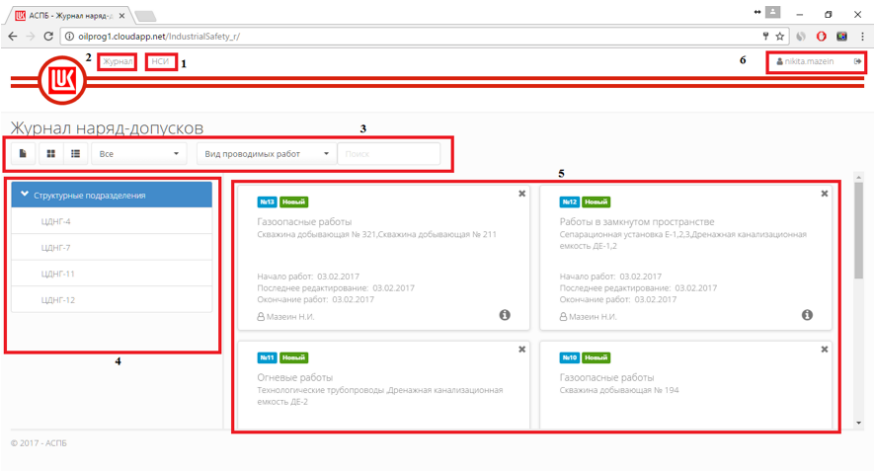


Рисунок 1 – Элементы рабочей поверхности системы

На рисунке 1 представлены элементы рабочей поверхности, разбитые на блоки от 1 до 6.

Блок 1. Ведение нормативно-справочной информации;

Блок 2. Журнал наряд-допусков;

Блок 3. Панель для создания и фильтрации наряд-допусков;

Блок 4. Иерархическое дерево для фильтрации наряд-допусков;

Блок 5. Отображение наряд-допусков;

Блок 6. Данные учетной-записи, выход из системы.

Интерфейс системы разработан с учетом особенностей стандартов и документооборота предприятия ООО «ЛУКОЙЛ-ПЕРМЬ», и при незначительных изменениях может быть адаптирован под другие предприятия отрасли.

Блок «Нормативно-справочная информация» (НСИ) позволяет получить оперативный доступ к документам предприятия и информации о исполнителях работ (рисунок 2).

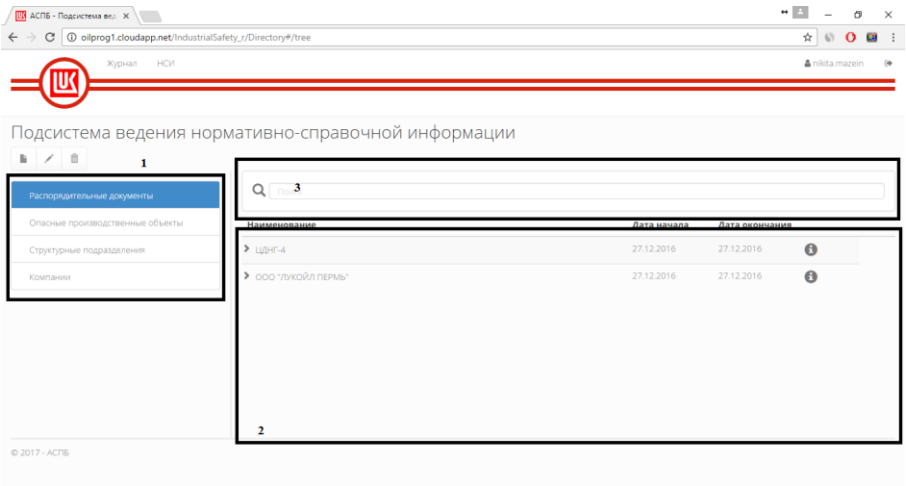


Рисунок 2 – Функциональные блоки НСИ

Основные блоки НСИ.

Блок 1. Группы нормативно-справочной информации;

Блок 2. Документы, относящиеся к выбранной группе.

Блок 3. Поиск нормативно-справочной информации.

Группы нормативно-справочной информации содержат более подробную информацию по следующим разделам:

Распорядительные документы: Положения, Инструкции, Распоряжения о назначении, Соглашения;

Опасные производственные объекты: Список опасных производственных объектов;

Структурные подразделения компании;

Информация по подрядным (сервисным) и субподрядным организациям.

Основной целью автоматизированной системы управления промышленной безопасностью является создание, просмотр, хранение наряд-допусков на проводимые работы.

Форма заполнения наряд-допуска имеет встроенные справочники по типам работ, объектам, специалистам, подрядным организациям, что значительно упрощает работу с наряд-допусками. Наряд-допуски имеют привязку ко времени и утверждающему лицу.

Сформированный наряд-допуск отображается по форме предприятия, сверстывается для печати (рисунок 3).

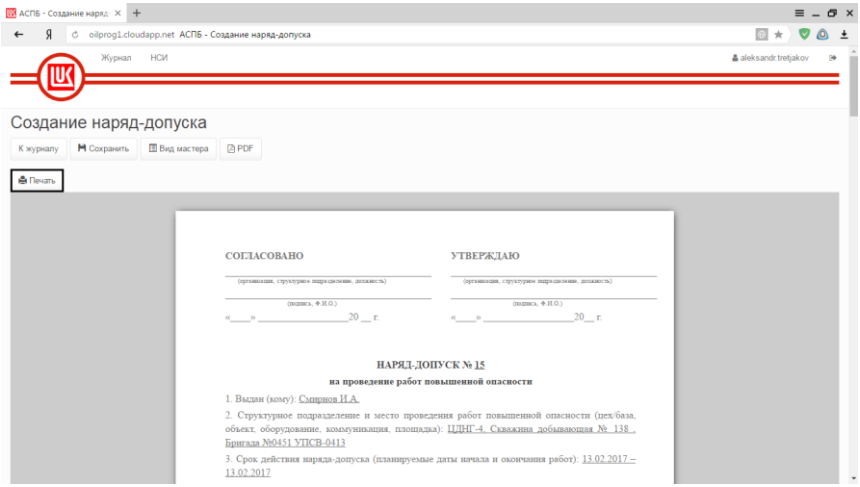


Рисунок 3 – Сформированный вид наряд-допуска

Наряд-допуск в процессе выполнения работ возможно присваивать различный статус, по выполнению работы выполняется закрытие наряд-допуска.

2 Описание общего алгоритма информационной системы

Общий алгоритм (сценарий действий пользователей – рисунок 4) в рамках процесса следующий.

Исполнитель в подразделении/подрядной организации вносит информацию о наряд-допуске на работу повышенной опасности в Модуль и определяет согласующие стороны. После этого исполнитель отправляет наряд-допуск на согласование. Согласовывающие лица могут согласовать документ или отправить документ на доработку исполнителю, написав в примечании о причинах отклонения. При этом если документ редактируется в процессе согласования, то все флаги «Согласованно» снимаются. Если наряд-допуск успешно согласован всеми участниками процесса, то исполнителю необходимо закрыть документ для редактирования и распечатать документ. После этого идет подписание и утверждение бумажного документа. В случае обнаружения ошибки в закрытом для редактирования документе наряд-допуск аннулируется в Модуле. После того как наряд-допуск подписан, в модуле участник роли «Исполнитель от подразделения/подрядной организации» присваивает ему статус «Оформлен». Для начала работ необходимо открыть наряд-допуск, эта процедура проводится в те дни, когда производятся работы по

документу в пределах временного периода, определенного в наряд-допуске. По окончании работ или при истечении срока действия наряд-допуска участник роли «Исполнитель от подразделения/подрядной организации» должен провести операцию закрытия наряд-допуска (рисунок 5). Ниже приведена структурная схема реализации информационной системы.

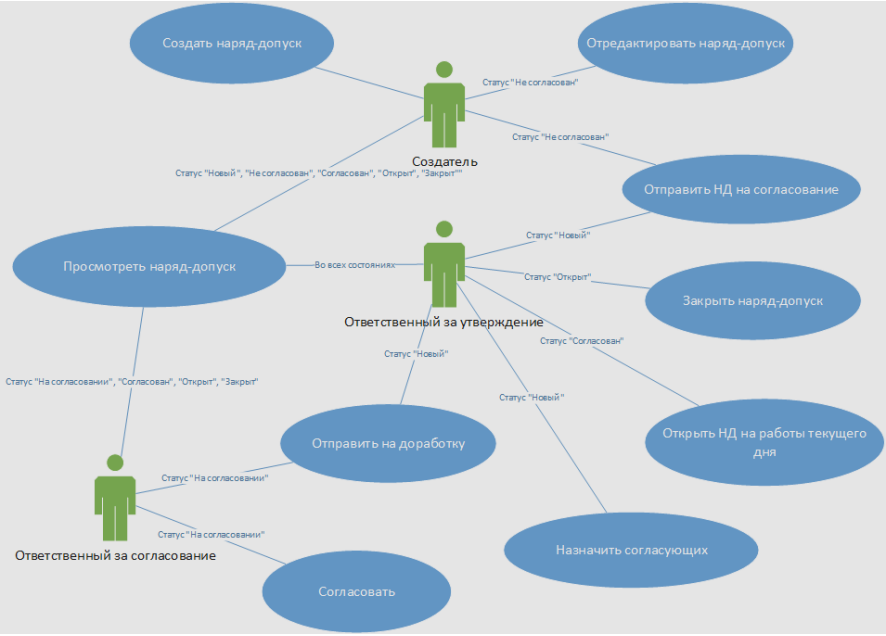


Рисунок 4 – Сценарии действий пользователей

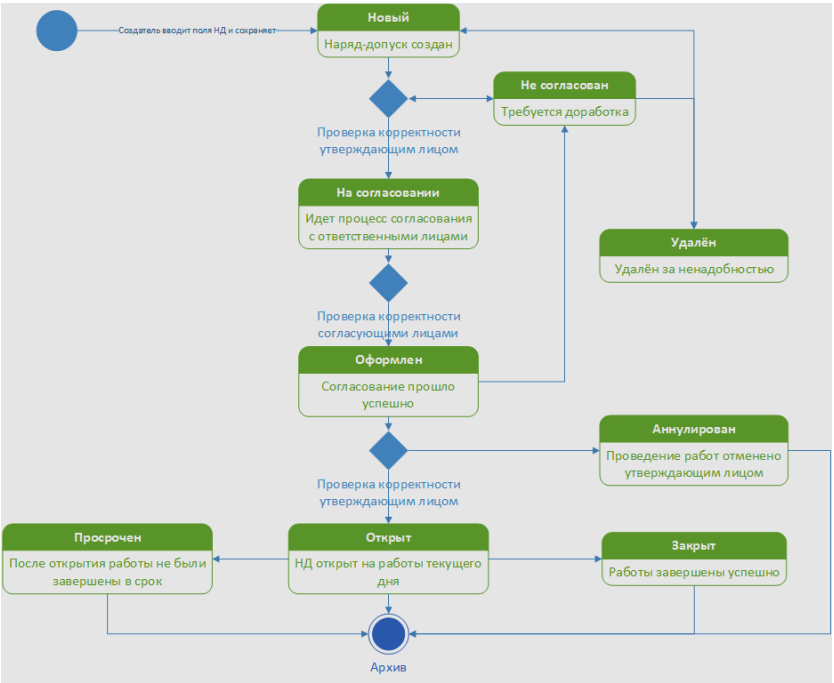


Рисунок 5 – Структурная схема реализации информационной системы

3 Средства реализации приложения и требования к серверу приложения

Серверная часть: Среда разработки Microsoft Visual Studio 2017, язык программирования C#. Тип проекта ASP.NET MVC. База данных реляционная MS SQL. Инструмент для взаимодействия с базой данных – EntityFramework.

Клиентская часть: Язык разметки HTML, язык программирования JavaScript. Взаимодействие с сервером осуществляется посредством AJAX запросов с использованием библиотеки JQuery. Отделение бизнес логики клиентской части приложения от интерфейса пользователя осуществляется посредством библиотеки Knockout.

Рекомендуемые требования к аппаратному обеспечению:

4-ядерный 64-разрядный процессор с тактовой частотой 3,0 ГГц

Оперативная память 16 ГБ

Свободный объем жесткого диска не менее 50 ГБ

Сетевой интерфейс 100MBps Ethernet

Рекомендуемые требования к программному обеспечению:

ОС Microsoft® Windows Server 2012 R2 / 2014 R2;

.NET Framework 4.5.2

СУБД Microsoft SQL Server 2014 Standard/Business Intelligence/Enterprise

Требования к персональному компьютеру пользователя

Рекомендуемые требования к аппаратному обеспечению:

Разрешение экрана 1280×720 и выше.

Рекомендуемые требования к программному обеспечению: Один из перечисленных браузеров: Microsoft Edge, Google Chrome, Mozilla Firefox, Opera.

Для проверки в промышленных условиях работоспособности, надежности и удобства эксплуатации средств для автоматизации процесса допуска на проведение работ повышенной опасности, подготовки и тестирования специалистов в области промышленной безопасности проведены опытно-промышленные испытания системы на объекте УППН «Каменный лог» ЦДНГ №4 ООО «ЛУКОЙЛ-ПЕРМЬ». В результате проведенного тестирования прототипа АСПБ обеспечена автоматизация процесса в области допуска на проведение работ повышенной опасности, подготовка и тестирование специалистов в области проведения работ повышенной опасности.

Заключение

Разработанная система может быть применена в качестве инструмента для обучения студентов нефтегазовых специальностей в области безопасного проведения работ повышенной опасности.

Список литературы

[Белослудцев, 2019] Белослудцев А.. «Наряд-допуск сменил бумагу на «цифру» // Нефть России. – 2019. – № 3-4. – С. 24-26.

[Марков, 2011] Марков В.К. Инновации как вектор стратегического развития нефтегазового комплекса России // Вестник Саратовского государственного социально-экономического университета. – 2011. №1. – С. 29-32.

[Сюч, 2008] Сюч Э.О. Эффективное управление производством в нефтегазовой промышленности // Экспозиция Нефть Газ. – 2008. – № 5/Н. – С. 27-30.

[Чуйко, 2011] Чуйко С.А. Основные направления совершенствования механизма формирования инвестиционной стратегии предприятия нефтегазового комплекса (процесс управления развитием

нефтегазового комплекса) // Известия Волгоградского государственного технического университета. – 2011. – №4 (77). – С.142-149.

DEVELOPMENT OF THE AUTOMATED INDUSTRIAL SAFETY SYSTEM FOR THE ENTERPRISES OF PETROLEUM INDUSTRY

Ilyushin P.Y. (ilushin-pavel@yandex.ru)
Lekomtsev A.V. (alex.lekomtsev@mail.ru)
Galkin S.V. (gnfd@pstu.ru)
Perm National Research Polytechnic University, Perm

The article describes the developed automated industrial safety system (AISS), successfully tested on the production sites of LLC LUKOIL-PERM. As a result of the carried out testing of the AISS prototype, the automation of the process in the area of access to high-risk work, training and testing of specialists in the field of high-risk work was ensured.

Keywords: management system, industrial safety, labor protection.

УДК 004.94

ГЕНЕТИЧЕСКАЯ ОПТИМИЗАЦИЯ ПОДБОРА ИСПОЛНИТЕЛЕЙ ИЗ СПИСКА ПОТЕНЦИАЛЬНЫХ УЧАСТНИКОВ ДЛЯ ПРОЕКТОВ

Синдюкова М.О.(sindyukova.m@gmail.com),
Горлова Е.А. (gorlova.k@mail.ru)
Ульяновский государственный технический университет, Ульяновск

Рассматривается способ подбора сотрудников государственных учреждений на региональные проекты, разрабатываемые в Ульяновской области для исполнения отдельных контрольных точек проекта, на основе анализа проектной деятельности региона. Целью работы является оптимизация процесса назначения исполнителя на конкретную задачу на основании данных выполненных ими, с применением генетического алгоритма.

Ключевые слова: проектное управление, генетический алгоритм, онтология, подбор персонала

Введение

На сегодняшний день на уровне региональных и федеральных органов власти все больше внимания уделяется управлению проектами. Одним из главных приоритетов сегодня является масштабное внедрение и профессиональное использование проектного подхода при реализации федеральных, ведомственных и региональных приоритетных проектов. В условиях повышенных требований к ценности создаваемых результатов, роль профессиональных методов проектного управления возрастает и при реализации крупных инфраструктурных и индустриальных проектов.

Одной из задач при формировании нового проекта является объективный выбор куратора, администратора и ответственных исполнителей контрольных точек проекта. Как правило, руководитель принимает решение относительно выбора сотрудников для работы над

проектом (критериями отбора обычно выступают опыт и компетенции сотрудника в определенной предметной области).

В статье рассмотрен способ определения опыта и компетенций сотрудников в определенной предметной области, основанный на онтологическом анализе. Полученные данные используются для автоматизированного подбора сотрудников на проекты с помощью генетического алгоритма.

Анализ и сбор информации о компетенциях исполнителей в различных предметных областях формируются путем анализа базы «Контрольные поручения». «Контрольное поручение» – документ, с помощью которого исполнители проектов отчитываются по проектам, далее – контрольное поручение. Эти данные выступают критериями отбора для генетического алгоритма. Задачей данного алгоритма при формировании команды является подбор членов команды, которые обеспечивали бы:

1. Соответствие количественного и качественного состава команды целям проекта;
2. Эффективную работу выбранной группы по управлению проектом.

Целью данной работы является создание алгоритмов и средств, позволяющих автоматизировать процесс подбора исполнителей для проектов, исходя из задач в рамках данного проекта, опыта работы сотрудников и их компетенций в различных предметных областях.

1 Сбор и анализ информации, необходимой для формирования кандидатов для работы над проектом

Для контроля выполнения поручений по проектам, разрабатываемым в Ульяновской области, используется механизм электронного документооборота. Руководитель формирует документ «контрольное поручение», исполнитель формирует отчет о выполнении поручения. Контрольное поручение – структурированный документ, содержащий доклад об исполнении поручений исполнителем (ответственным исполнителем), который несет персональную ответственность за своевременное и качественное исполнение поручений на проектах, реализуемых на территории Ульяновской области [Распоряжение, 2010]. Данный документ содержит следующие сведения о поручении: наименование, дата и номер, статус, заголовок, поручение, ответственный исполнитель, контрольный срок, фактический срок, информация о выполнении поручения (доклад). Анализ базы контрольных поручений позволит сформировать список сотрудников, которые выполняли какие-либо работы по проектам, оценить их компетенции в различных

предметных областях. Контрольные поручения загружаются пользователем в систему в формате *.docx.

Информацию о сотрудниках, должностях, компетенциях и предметных областях проектов можно представить в виде онтологии. Формально модель онтологии сотрудника можно определить следующим образом:

$$O_i^S = \left\langle S_i, E^S, P^S, C^S, L^S, R_E^S, R_O^S \right\rangle,$$

где S_i – конкретный сотрудник;

$E^S = \{e_1, e_2, ..., e_n\}$ – множество ведомств или структур гос.службы, в которых работал сотрудник;

$P^S \subseteq P$ – множество должностей сотрудника;

$C^S \subseteq C$ – множество компетенций сотрудника;

$L^S \subseteq L$ – множество предметных областей, в которых работает сотрудник;

R_E^S – отношение типа «объект-объект», связывающее специалиста с должностью и компетенциями;

R_O^S = {степень, направление} – конечное множество свойств, характеризующих образование.

На данном этапе работы онтология построена вручную. В дальнейшем развитие проекта предполагает доработку модуля составления онтологии в автоматизированном режиме.

Анализ каждого документа разбивается на три этапа. Первый – извлечение данных об исполнителе проекта – его ФИО, место работы и должность, данных о самом поручении – номер, наименование, сроки исполнения и самого доклада поручения. Данные извлекаются путем перебора ключевых слов документа.

На следующем этапе производится предобработка исходного текста документа и его морфологический анализ с помощью программы Mystem. Mystem работает на основе словаря, приводит слова к начальной форме и может формировать морфологические гипотезы о неизвестных ей словах. Данная особенность создает трудности при анализе и определении предметной области документа. В связи с этим, перед тем как анализировать текст программой Mystem, система производит поиск и исключение аббревиатур, городов, фамилий, инициалов, слов из других языков, чисел.

Заключительным этапом анализа текста является определение компетенций сотрудника и предметных областей, в которых он работал.

Компетенции сотрудника извлекаются из документа по заранее известным ключевым словам. Для определения предметной области, из текста извлекаются термины, относящиеся к заранее определенному в онтологии набору понятий предметной области, и подсчитывается их частота и степень выраженности с применением следующего выражения:

$$k = f(x) / N,$$

где k – степень выраженности предметной области в обрабатываемом тексте;
$f(x)$ – количество ключевых слов в тексте, относящихся к конкретной предметной области (частота встречаемости);
N – общее число слов обработанного текста.

Список исполнителей записывается в таблицу, и в дальнейшем передается в модуль подбора исполнителей для проектов.

Обработанные данные записываются в базу данных с целью дальнейшего использования в качестве исходных данных в системе назначения ответственных исполнителей на проекты (рисунок 1).

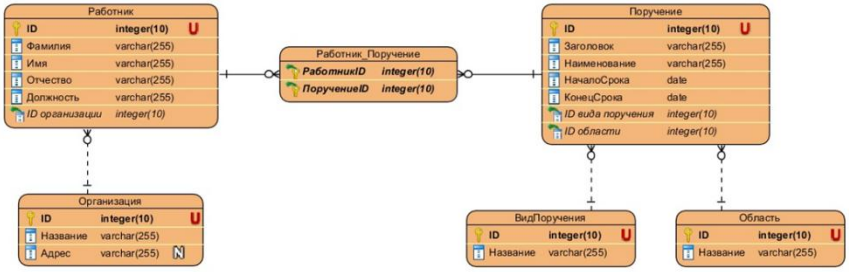


Рисунок 1 – ER- диаграмма базы данных системы

2 Корпоративная социальная сеть

Для решения описанной выше проблемы было выбрано взаимодействие в формате корпоративной социальной сети для руководителей проектов и потенциальных исполнителей проектов. Ее модель можно формально представить с помощью следующего выражения:

$$ESN = \langle Pr, T, Ex, R \rangle,$$

где Pr – множество проектов;

T – множество команд проектов;

Ex – множество специалистов;

R – множество отношений типа «объект-объект», определяемых следующим образом:

$$R = \{R^{PT}, R^{TE}\},$$

где R^{PT} – отношение, связывающее проект и проектную команду («выполнен»);

R^{TE} – отношение, связывающее исполнителя проекта и проектную команду («исполнял»).

Более подробно главные объекты корпоративной социальной сети рассмотрим ниже:

Объект «Проект» – это текущий проект, для которого нужно подобрать исполнителей, а также уже реализованные проекты. Данный объект содержит в себе следующие свойства: наименование проекта; область проекта (для поиска исполнителей, обладающих компетенциями в заданной области); вид проекта (например, мероприятие, дорожная карта, отчет и т. д.)

Объект «Работник» описывает сотрудника, который может работать или уже работает над проектами. «Работник» содержит в себе следующие свойства: Фамилия; Имя; Отчество; Должность; Дата рождения; Пол; Стаж; Контактные данные (адрес электронной почты, телефон); Компетенции работника.

3 Генетический алгоритм для подбора исполнителей на проекты

В настоящий момент создается корпоративная социальная сеть, в задачи которой включены подбор и составление оптимального состава проектной команды на основе компетенций сотрудников. Для оптимизации работы команды в описываемой подсистеме подразумевается использование методов генетического алгоритма.

В базе данных социальной сети хранится информация о работниках, их компетенциях и проектах. При поступлении нового проекта определяется количество позиций, на которые нужно определить сотрудников. Процедура сравнения компетенции и стажа сотрудников с требуемыми и последующий выбор конкретного человека на конкретный проект происходит с помощью генетического алгоритма. Настраиваемыми параметрами алгоритма являются:

- размер популяции *p_size*;
- размер формируемой проектной команды υ*_count*;
- массив весов $\{w_{st};\ w_{cp}\}$;
- количество элитных хромосом *elit*;

- вероятность кроссинговера *cros*
- вероятность мутации *mut*;
- пороговое изменение функции приспособленности *delta*
- максимальное количество итераций алгоритма *max_step*.

При распределении кандидатов существует ограничение, о котором нельзя забывать – отпуск, а также один сотрудник может принимать участие не в более чем пяти проектах.

Функция приспособленности содержит в себе показатели стажа, компетенций человека и имеет следующий вид:

$$F = \frac{1}{s}\sum_p \left(1 - \mathrm{C}_p\right) \to min, p = \overline{1, s}, \quad (1)$$

где S – количество требуемых людей в проекте;

C_p – уровень квалификации кандидата на соответствующую должность, вычисляемый с помощью следующего выражения:

$$\mathrm{C}_p = w_{st} * \frac{st - st_v}{st_v} + w_{cp} * \sum cp + h, \quad (2)$$

где w_{st}, w_{cp} – веса коэффициентов стажа и уровня компетенций сотрудника;

st, st_v – стаж кандидата и требуемый стаж (стаж вакансии);

cp – степень обладания требуемым уровнем компетенции работника соответственно,

h – вероятность взятия отпуска на время проекта (0 – большая вероятность; 1 – наименьшая вероятность)

Потенциальное решение (хромосома) генетического алгоритма формирования проектной команды имеет следующий вид:

$$a_i^t = \left(\langle v, c \rangle\right), v = \overline{1, v\_count},$$

где *v* – номер вакансии в формируемой проектной команде;

c – номер сотрудника, претендующего на вакансию;

v_count – количество вакансий;

i – номер хромосомы;

t – номер поколения.

Первый этап осуществляется с помощью генерации начального поколения: для каждого вакантного места случайным образом выбирается сотрудник. Если он не подходит по требованиям проекта, то продолжается выполнение случайного распределения до того момента, пока на каждую вакансию не будет назначен сотрудник или не будет принято решение об оставлении этой вакансии не заполненной.

После рассчитывается общая функция приспособленности для созданного распределения (выражение 1).

В новое поколение переходят только *elit* первых хромосом из полученной популяции. С помощью оператора скрещивания формируется остальное количество, при этом в новое поколение выбирается только одна хромосома из двух, полученных в результате скрещивания. Точка скрещивания назначается случайным образом, потом осуществляется взаимообмен частями хромосомы (рисунок 2).

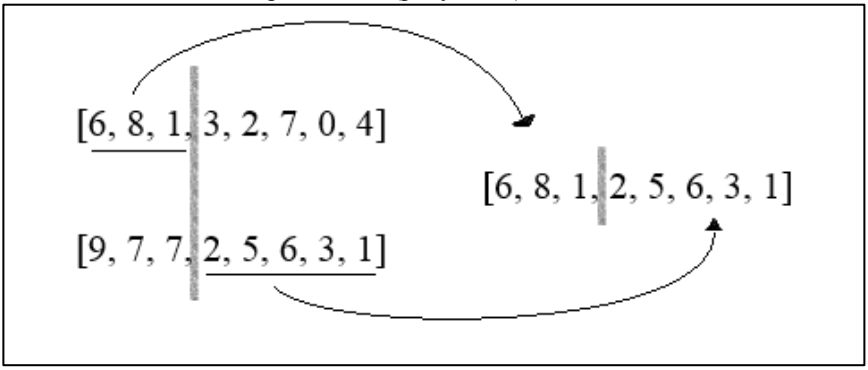


Рисунок 2 – Наглядный пример оператора скрещивания

Следом за завершением процесса генерации популяции, применяется оператор мутации (рис. 2). Он предполагает случайный выбор номера вакансии, затем принимается решение об уменьшении или увеличении номера сотрудника на данную должность. Номер кандидата изменяется на единицу.

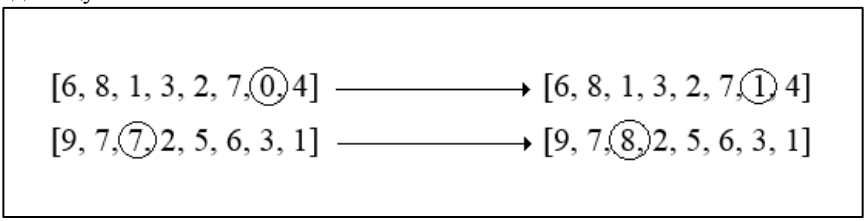


Рисунок 3 – Наглядный пример оператора мутации

Для решения проблемы, при которой один кандидат может участвовать в не более чем пяти разных проектах, после выполнения операций скрещивания и мутации активируется функция случайного выбора сотрудника. Этот процесс будет повторяться до тех пор, пока на данную должность не найдется работник, который занят на менее чем

пяти других проектах, или не будет принято решение об оставлении этой вакансии не заполненной.

После выполнения мутации для каждого распределения кандидатов рассчитывается значение функции приспособленности.

Если количество итераций алгоритма достигло значения *max_step* или разница между значениями функции приспособленности исходного и сформированного поколения меньше *delta*, то распределение последнего поколения с минимальным значением функции приспособленности будет наилучшим.

В результате работы генетического алгоритма кандидаты наилучшим образом распределяются на соответствующие должности в рамках поступающих новых проектов.

Заключение

В данной работе рассмотрена проблема неэффективного расходования времени на назначение исполнителей на новый проект. Был предложен способ определения опыта и компетенций сотрудников в определенной предметной области, основанный на онтологическом анализе, чтобы далее использовать полученные данные в генетическом алгоритме для подбора исполнителей на проекты.

В статье были описаны модели онтологии сотрудника и корпоративной социальной сети. Для генетического алгоритма были определены функция приспособленности и потенциальное решение (хромосома), а также выбраны операции скрещивания и мутации для решения рассмотренной проблемы.

В дальнейшем развитие проекта предполагает доработку модуля составления онтологии в автоматизированном режиме и реализацию генетического алгоритма.

Список литературы

[Распоряжение, 2010] Об утверждении Порядка исполнения поручений Губернатора - Председателя Правительства Ульяновской области МЭРИЯ ГОРОДА УЛЬЯНОВСКА РАСПОРЯЖЕНИЕ от 4 июня 2010 г. N 85-Р.

[Воронина и др., 2015] Воронина В.В., Мошкин В.С. Разработка приложений для анализа слабоструктурированных информационных ресурсов: учебное пособие. – Ульяновск : УлГТУ, 2015. – 162 с.

[Соловьев и др.] Соловьев В.Д., Добров Б.В., Иванов В.В., Лукашевич Н.В. Онтологии и тезаурусы: учебное пособие. – М., 2006. – 157 с.

[Скурихин 1995] Скурихин А.Н. Генетические алгоритмы // Новости искусственного интеллекта №4. – М.: АСИ, 1995. – С. 6-17.

[Панченко, 2007] Панченко Т.В. Генетические алгоритмы: учебно-методическое пособие / под ред. Ю.Ю. Тарасевича. – Астрахань: ИД «Астраханский университет», 2007. – 87 с.

[Стецюра, 1998] Стецюра Г.Г. Эволюционные методы в задачах управления, выбора, оптимизации // Приборы и системы управления. – Б.м.: Б.и. – 1998. – С. 54-62.

[Иванов и др., 2014] Иванов В.К., Мескин П.И. Реализация генетического алгоритма для эффективного документального тематического поиска. // Программные продукты и системы №4 (108) – Тверь : Научно-исследовательский институт «Центрпрограммсистем», 2014. – С.118-126.

[Pathak et al, 2000] Pathak, P. Gordon, M. Fan, W. Effective information retrieval using genetic algorithms based matching functions adaption // Proc. 33rd Hawaii International Conference on Science (HICS), Hawaii, USA. pp. 8-15.

[Шипилов и др., 2014] Шипилов В.В., Сахаров О.В. Моделирование подбора и расстановки кадров с учетом их профессиональных навыков для выполнения проектов // Вопросы теории безопасности и устойчивости систем №16. – М.: ВЦ РАН. 2014. – С. 153-164.

GENETIC OPTIMIZATION OF SELECTION OF EXECUTORS FROM THE LIST OF POTENTIAL PARTICIPANTS FOR PROJECTS

Sindyukova M.O. (sindyukova.m@gmail.com),
Gorlova E.A. (gorlova.k@mail.ru)
Ulyanovsk State Technical University, Ulyanovsk

The paper discusses the method of selection of government employees for regional projects developed in the Ulyanovsk region for the execution of individual control points of the project, based on the analysis of project activities in the region. The aim of the work is to optimize the process of appointing the contractor to a specific task based on the data performed by them, using a genetic algorithm

Keywords: project management, genetic algorithm, ontology, recruitment

УДК 622.691

ПРИМЕНЕНИЕ НЕЧЕТКИХ МНОЖЕСТВ ДЛЯ ОРГАНИЗАЦИИ АВТОМАТИЗАЦИИ ПРОЦЕССА ОДОРИЗАЦИИ ПРИРОДНОГО ГАЗА

Даев Ж.А.(zhand@yandex.ru)
Баишев университет, Актобе, Казахстан
Султанов Н.З. (sultanov@mail.osu.ru)
Оренбургский государственный университет, Оренбург

Описывается метод и система автоматического регулирования одоранта природного газа. Отличительной особенностью предлагаемой системы является применение нечетких множеств для решения задач одоризации газа. Для решения задачи авторами вводятся лингвистические переменные для концентрации одоранта, потребляемого газа и времени года с соответствующими нечеткими множествами. В статье даются рекомендации, связанные с реализацией системы и метода на программируемых логических контроллерах.

Ключевые слова: нечеткие множества, одоризация, газ, система.

Введение

Одним из важнейших способов доставки энергоносителей является трубопроводный транспорт. Практически весь добываемый природный газ транспортируется по магистральным газопроводам [Алиев и др., 1988]. Преимуществами такого способа доставки является возможность непрерывной подачи, простота организации технологических процессов, эффективность контроля над параметрами транспортировки. Во время организации трубопроводного транспорта природного газа по магистральным газопроводам доставка энергоносителя до конечного потребителя состоит из серии технологических процессов таких, как осушка газа, удаление механических примесей, нагнетание, редуцирование, измерение количества газа, одоризация и т. д.

В данной статье рассматривается автоматизация процесса одоризации природного газа. В соответствии с работой [Алиев и др., 1988] природный газ, очищенный от серосодержащих соединений, не имеет никакого цвета

и запаха, поэтому обнаружить его утечку довольно трудно. Для того чтобы придать природному газу резкий и неприятный запах, его одорируют. Во время одоризации в природный газ вводят специальные вещества, называемые одорантами. Одоранты и продукты их сжигания должны быть физиологически безвредными, достаточно летучими, не вызывать коррозию металла трубы, химически взаимодействовать с газом, поглощаться водой, сорбироваться почвой или предметами, находящимися в помещении [Алиев и др., 1988]. Этим требованиям удовлетворяют меркаптаны. В частности, одорант изготавливают из смеси, состоящей из метилмеркаптана, этилмеркаптана, пропилмеркаптана [Данилов, 2014]. Например, в работе [Данилов, 2014] указывается, что одорант марки СПМ представляет собой смесь, которая состоит из 30% этилмеркаптана, 50-60% изо- и н-пропилмеркаптанов, 10-20% изобутилемркаптанов. Контроль количества одоранта в составе природного газа является одним из важных процессов во время доставки природного газа. Чрезмерное увеличение количества одоранта в газе влияет на здоровье человека, особенно в условиях, когда сероводородные соединения могут содержаться в транспортируемом газе, с одной стороны. С другой стороны, недостаток одоранта не обеспечивает своевременное обнаружение утечек газа, что тоже влияет на безопасную эксплуатацию газовых приборов.

Устройства, которые выполняют одоризацию, называются одоризационными установками. В зависимости от расхода установки добавляют определенное количество одоранта. Некоторые виды таких установок описаны в работах [Алиев и др., 1988; Темников, 2016; Кузь и др., 2014; Negaresh et al. 1988; Данилов, 2014].

Степень одоризации природного газа оценивают в соответствии с нормативным документом [ГОСТ, 1986]. В данном документе рассматривается два способа оценки степени одоризации: камерный и приборный. Сущность обоих методов заключается в оценке интенсивности запаха газовоздушной смеси, создаваемой в камере или приборе, по следующей пятибалльной шкале [ГОСТ, 1986]:

0 – запаха нет;

1 – запах очень слабый, неопределенный;

2 – запах слабый, но определенный;

3 – запах умеренный;

4 – запах сильный;

5 – запах очень сильный.

Последняя пятибалльная шкала представляет собой неявно представленное нечеткое множество, связанное с интенсивностью запаха. Приведенные баллы являются термами, которые отражают интенсивность

запаха одоранта в газе. Поэтому в рамках данной статьи ставится задача моделирования автоматической системы одоризации природного газа, работа которой основана на методах теории нечетких множеств. Выбор систем регулирования одоранта, основанный на нечетких моделях, обусловлен необходимостью порционно-дискретного контроля интенсивности запаха одоранта в газе. Последнее позволяет системе своевременно и просто реагировать на изменение количества газа, а также обеспечивает удобный контроль для восприятия оператора газораспределительной станции. С другой стороны, решение данной задачи с помощью нечетких множеств позволяет учесть другие факторы, влияющие на дозирование одоранта, например, сезонные погодные изменения.

1 Введение нечетких множеств для процесса одоризации газа

Представленную в документе [ГОСТ, 1986] шкалу можно описать в виде лингвистической переменной x «Интенсивность запаха газа», которая представлена на рисунке 1. В качестве термов данной переменной введем треугольные нечеткие множества, которые часто применяются в аналогичных задачах, как это описано в работе [Пегат, 2013].

На рисунке 1 нечеткие множества введенной лингвистической переменной представлены через дискретные баллы. Для решения нашей задачи заменим данные баллы массовыми концентрациями одоранта.

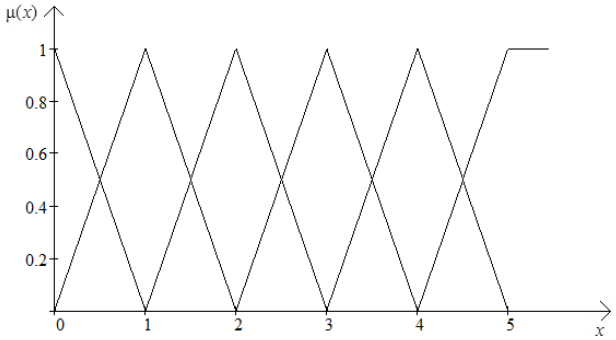


Рисунок 1 – Лингвистическая переменная «Интенсивность запаха газа»

В соответствии с работой [Данилов, 2014], рекомендуемая норма среднегодовой концентрации одоранта, которая представляет собой смесь меркаптанов для 1000 м^3, должна быть следующая:

- минимальная норма 5 г;

- максимальная норма 16 г.

Попробуем связать данные требования по концентрации с лингвистической переменной «Интенсивность запаха газа» на рисунке 1. Для этого определим границы нечетких множеств, которые определяют термы лингвистической переменной следующим образом: A_1={x, $\mu_{A_1}(x)$} – запаха нет, A_2={x, $\mu_{A_2}(x)$} – запах очень слабый, неопределенный, A_3={x, $\mu_{A_3}(x)$} – запах слабый, но определенный, A_4={x, $\mu_{A_4}(x)$} – запах умеренный, A_5={x, $\mu_{A_5}(x)$} – запах сильный, A_6={x, $\mu_{A_6}(x)$} – запах очень сильный, где x – массовая концентрация одоранта на 1000 м^3, $\mu(x)$ – функция принадлежности нечетких множеств. Преобразованная лингвистическая переменная представлена на рисунке 2.

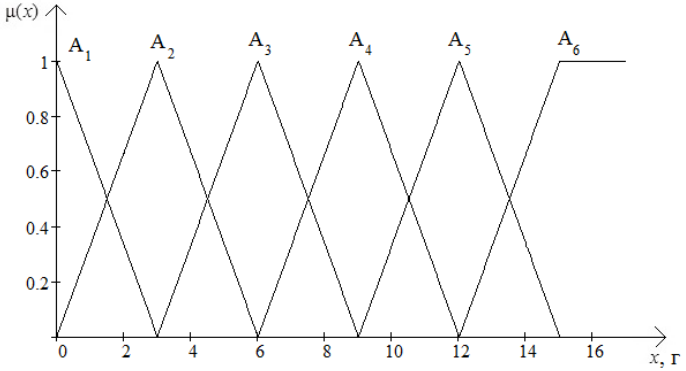


Рисунок 2 – Преобразованная лингвистическая переменная «Интенсивность запаха газа»

С учетом введенной лингвистической переменной нечеткий регулятор может контролировать величину дозирования, чтобы интенсивность запах газа оставалась необходимой. С другой стороны, в работах [Алиев и др., 1988; Данилов, 2014] говорится о том, что величина концентрации, указанная на рисунке 2, является среднегодовой, а в летнее время года норма одоранта должна снижаться в два раза. Также величина расхода одоранта также должна зависеть от количества транспортируемого газа. Это обусловлено тем, что на каждом населенном пункте потребление газа сильно отличается. В поселках количество газа может составлять несколько сотен кубических метров зимой, а летом расход вовсе уменьшаться до десятков кубических метров в час. В городах количество

потребления также сильно меняется в зависимости от количества жителей города. Поэтому для расхода газа, отпускаемого в населенный пункт, можно было бы ввести лингвистическую переменную y «Потребляемый расход газа». Нечеткие множества для такой лингвистической переменной можно выполнить в соответствии с рекомендациями нормативного документа [МИ, 2007]. В данном документе замерные узлы расхода и количества газа делятся в зависимости от величины измеряемого объемного расхода газа на следующие категории:

- свыше 100 тыс. м^3/ч – большой производительности;
- от 20 тыс. до 100 тыс. м^3/ч – средней производительности;
- от 1 тыс. до 20 тыс. м^3/ч – малой производительности;
- до 1 тыс. м^3/ч – минимальной производительности.

Если использовать данные категории в качестве термов введенной лингвистической переменной, то ее можно представить следующим образом:

$B_1=\{y,\ \mu_{B_1}(y)\}$ – минимальный расход газа; $B_2=\{y,\ \mu_{B_2}(y)\}$ – малый расход газа; $B_3=\{y,\ \mu_{B_3}(y)\}$ – средний расход газа; $B_4=\{y,\ \mu_{B_4}(y)\}$ – большой расход газа. Лингвистическую переменную с ее термами можно изобразить через трапецеидальные нечеткие множества, как на рисунке 3.

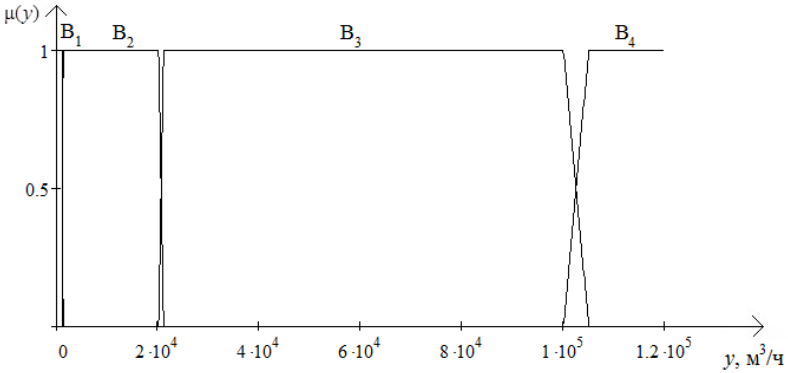


Рисунок 3 – Лингвистическая переменная «Потребляемый расход газа»

После введения необходимых лингвистических переменных с соответствующими нечеткими множествами разработаем нечеткую базу правил, которая будет положена в основу работы нечеткого регулятора, выполняющего дозирование одоранта в зависимости от потребляемого расхода и времени года. Для учета времени года введем в рассмотрение лингвистическую переменную t «Время года». Введем термы, связанные с

временем года с соответствующими нечеткими множествами: C_1={t, $\mu_{C_1}(t)$ } – зимнее время года; C_2={t, $\mu_{C_2}(t)$ } – весенне-летнее время года; C_3={t, $\mu_{C_3}(t)$ } – осеннее время года.

Данная переменная будет отражать времена года. Отсчет по шкале абсцисс начнем с декабря, который будет характеризоваться интервалом от нуля до единицы, и так далее до одиннадцатого месяца. Графическое представление данной переменной отображается трапецеидальными нечеткими множествами, которое представлено на рисунке 4.

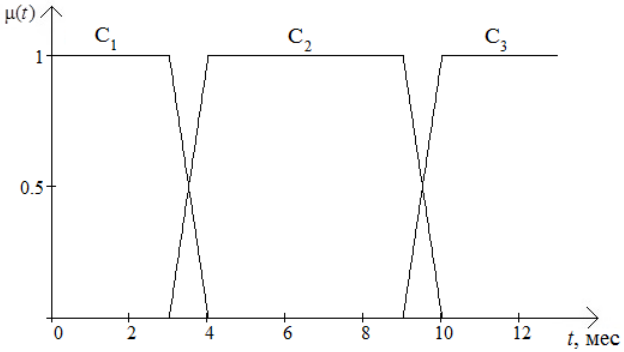


Рисунок 4 – Лингвистическая переменная «Время года»

2 Нечеткая база правил и модель системы

В соответствии с работами [Пегат, 2013; Mamdani, 1974] концепция лингвистической нечеткой модели управления динамическими системами была предложена Мамдани. Целью методики Мамдани является разработка модели, выполняющей такое отображение своих входов в выход, которое обеспечивало бы как можно более точную аппроксимацию реальной системы [Пегат, 2013]. В нашем случае система регулирования количества одоранта должна выполнять дозирование в зависимости от потребляемого расхода и времени года. Введение нечетких множеств позволит также варьировать величину одоризации с учетом времени года так, чтобы интенсивность запаха газа была довольно ощутимой и менее вредной для потребителя.

На вход системы должен поступать сигнал о потребляемом количестве газа и значение переменной, связанной с временем года, а на выходе системы должно выполняться дозирование одоранта. Этот процесс можно смоделировать, а затем реализовать в нечетком регуляторе путем введения нечеткой базы правил, которую составим ниже:

R1: ЕСЛИ $(y = B_1)$ И $(t = C_1)$ ТО $(x = A_5)$,
R2: ЕСЛИ $(y = B_1)$ И $(t = C_2)$ ТО $(x = A_4)$,
R3: ЕСЛИ $(y = B_1)$ И $(t = C_3)$ ТО $(x = A_5)$,
R4: ЕСЛИ $(y = B_2)$ И $(t = C_1)$ ТО $(x = A_6)$,
R5: ЕСЛИ $(y = B_2)$ И $(t = C_2)$ ТО $(x = A_4)$, (1)
R6: ЕСЛИ $(y = B_2)$ И $(t = C_3)$ ТО $(x = A_5)$,
R7: ЕСЛИ $(y = B_3)$ И $(t = C_1)$ ТО $(x = A_6)$,
R8: ЕСЛИ $(y = B_3)$ И $(t = C_2)$ ТО $(x = A_4)$,
R9: ЕСЛИ $(y = B_3)$ И $(t = C_3)$ ТО $(x = A_5)$,
R10: ЕСЛИ $(y = B_4)$ И $(t = C_1)$ ТО $(x = A_6)$,
R11: ЕСЛИ $(y = B_4)$ И $(t = C_2)$ ТО $(x = A_4)$,
R12: ЕСЛИ $(y = B_4)$ И $(t = C_3)$ ТО $(x = A_5)$.

Идентичность выходных воздействий позволяет модифицировать базу правил (1) путем агрегации условий с помощью оператора ИЛИ с целью уменьшения нагрузки на логический решатель и сокращения базы правил. Это можно сделать следующим образом, как описано в работах [Пегат, 2013; Ярушкина, 2009]:

R1: (ЕСЛИ $(y = B_1)$ И $(t = C_1)$) ИЛИ
(ЕСЛИ $(y = B_1)$ И $(t = C_3)$) ИЛИ
(ЕСЛИ $(y = B_2)$ И $(t = C_3)$) ИЛИ
(ЕСЛИ $(y = B_3)$ И $(t = C_3)$) ИЛИ
(ЕСЛИ $(y = B_4)$ И $(t = C_3)$) ТО $(x = A_5)$,
R2: (ЕСЛИ $(y = B_1)$ И $(t = C_2)$) ИЛИ (2)
(ЕСЛИ $(y = B_2)$ И $(t = C_2)$) ИЛИ
(ЕСЛИ $(y = B_3)$ И $(t = C_2)$) ИЛИ
(ЕСЛИ $(y = B_4)$ И $(t = C_2)$) ТО $(x = A_4)$,
R3: (ЕСЛИ $(y = B_2)$ И $(t = C_1)$) ИЛИ
(ЕСЛИ $(y = B_3)$ И $(t = C_1)$) ИЛИ
(ЕСЛИ $(y = B_4)$ И $(t = C_1)$) ТО $(x = A_6)$.

Применяемые операции И, ИЛИ выполняют с помощью s-норм и t-норм, которые приводятся в работах [Пегат, 2013; Ярушкина, 2009]. Необходимо учитывать рекомендации при использовании данных операторов при построении моделей типа Мамдани по правилам (1) и (2),

т. е. не использовать операторы SUM и MEAN. Операторы s-норм и t-норм могут быть организованы в контроллерах, которые поддерживают стандарт МЭК 61131-7, либо такие операторы могут быть реализованы путем написания программ на языках стандарта МЭК 61131-3.

Архитектура системы может быть выполнена в соответствии с существующими решениями и на базе существующих одоризационных установок [Данилов, 2014; Кузь и др., 2014]. Новизной и отличительной особенностью является предлагаемый метод решения задачи, который будет реализован в логическом решателе (ПЛК), который полностью основан путем введения нечетких множеств и формирования нечеткой базы правил.

На примере данной модели показана формализация решения задачи автоматизации процесса одоризации природного газа. Путем введения группы лингвистических переменных и соответствующих нечетких множеств легко без привлечения сложных математических моделей можно реализовать вполне адекватную систему, которая позволяет выполнять одоризацию газа, перед тем как отправлять его конечному потребителю.

Заключение

Таким образом, в рамках настоящей статьи рассмотрено решение задачи об автоматизации процесса одоризации природного газа с помощью нечетких множеств. Показана модель автоматической системы одоризации, которая выполняет дозирование одоранта в зависимости от потребляемого количества газа, и учитывает время года, благодаря чему снижается расход одоранта. Модель автоматической системы одоризации полностью построена на аппарате нечеткого моделирования. В статье также даются рекомендации по построению и реализации нечеткой системы с помощью программируемых контроллеров.

Список литературы

[Алиев и др., 1988] Алиев Р.А., Белоусов В.Д., Немудров А.Г. Трубопроводный транспорт нефти и газа. – М.: Недра, 1988. – 368 с.

[Темников, 2016] Темников А.А. Одоризация углеводородного топлива. Одоризационные установки // Science Time. – 2016. – Т. 28, – №4. – С. 834 – 837.

[Кузь и др., 2014] Кузь Н.В., Шевчук М.О. Измерительная система объема природного газа и степени его одоризации //Мир измерений. – 2014. – №5. – С. 11 – 12.

[Negaresh et al., 1988] Negaresh, M., Farrokhnia, M., Mehranbod, N., Modeling and control of natural gas bypass odorizer, Journal of Natural Gas Science and Engineering (2018), doi: 10.1016/j.jngse.2017.12.010.

[Данилов, 2014] Данилов А.А. Автоматизированные газораспределительные станции: Справочник. – СПб.: Химиздат, 2004. – 544 с.

[ГОСТ, 1986] ГОСТ 22387.5–77. Газ для коммунально-бытового потребления. Методы определения интенсивности запаха. – М.: Издательство стандартов, 1986. – 6 с.

[Пегат, 2013] Пегат А. Нечеткое моделирование и управление. – М.: Бином. Лаборатория знаний, 2013. – 798 с.

[МИ, 2007] МИ 3082–2007. Выбор методов и средств измерений расхода и количества потребляемого природного газа в зависимости от условий эксплуатации на узлах учета. Рекомендации по выбору рабочих эталонов для их поверки. – Казань: ФГУП «ВНИИР», 2007. – 42 с.

[Mamdani, 1974] Mamdani E.H. Application of fuzzy algorithms for control of simple dynamic plant // Proc. IEEE. – 1974. – V. 121, №12. – P. 1585–1588.

[Ярушкина, 2009] Ярушкина Н.Г. Основы нечетких и гибридных систем. – М.: Финансы и статистика, 2009. – 320 с.

THE USE OF FUZZY SETS FOR AUTOMATION OF THE PROCESS ODORIZATION OF NATURAL GAS

Dayev Zh.A. (zhand@yandex.ru)
Baishev University, Aktobe, Kazakhstan
Sultanov N.Z. (sultanov@mail.osu.ru)
Orenburg State University, Orenburg

The paper proposes a method and system of automatic control of odorant. The proposed system is based on fuzzy set methods. To solve the problem, we introduce linguistic variables for odorant concentration, gas consumption and season with their fuzzy sets. The article provides recommendations related to the implementation of the system and method on programmable logic controllers.

Keywords: fuzzy sets, odorization, gas, system.

УДК 519.248:681.518.5

АНАЛИЗ ВЗАИМОСВЯЗЕЙ ПОКАЗАТЕЛЕЙ КАЧЕСТВА ДИАГНОСТИКИ ОБЪЕКТА ПРИ БИНАРНОЙ КЛАССИФИКАЦИИ*

Жуков Д.А. (zh.dimka17@mail.ru)
Клячкин В.Н. (v_kl@mail.ru)
Ульяновский государственный технический университет, Ульяновск

Для распознавания состояния технического объекта могут быть использованы методы машинного обучения (нейронные сети, бэггинг деревьев решений, логистическая регрессия, методы бустинга и другие). В статье рассматриваются меры качества распознавания исправного состояния объекта: доля ошибок при кросс-валидации, F-мера, площадь под кривой ошибок и другие. На примерах реальных объектов анализируются взаимосвязи между этими показателями.

Ключевые слова: техническая диагностика, машинное обучение, бинарная классификация, показатели качества, корреляции

Введение

Одна из задач технической диагностики – определение вида состояния объекта: является объект исправным или неисправным [Биргер, 1978]. Для распознавания состояния объекта может использоваться бинарная классификация методами машинного обучения. Практическая реализация такого подхода проводилась на базе библиотеки инструментов Statistics and Machine Learning Toolbox в пакете Matlab [Жуков и др., 2018].

При распознавании используются результаты измерений показателей функционирования в процессе предшествующей эксплуатации. В качестве исходных данных рассматривается множество прецедентов: объекты с заданными показателями функционирования и соответствующими состояниями. На основе этих данных надо

* Работа выполнена при финансовой поддержке РФФИ и Правительства Ульяновской области, проект 18-48-730001

восстановить зависимость между показателями функционирования и состоянием объекта [Воронцов, 2016]. Это частный случай одной из задач машинного обучения – бинарная классификация с обучением по прецедентам.

Для оценки качества классификации могут использоваться различные критерии: доля ошибок в контрольной выборке, площадь под кривой ошибок, *F*-мера и другие. Представляет интерес проанализировать степень взаимосвязи различных мер качества машинного обучения на реальных технических объектах с целью выбора критериев, обеспечивающих объективное распознавание состояния конкретного объекта.

1 Критерии качества классификации

Для оценки качества классификатора могут быть использованы различные критерии [Соколов, 2016]. Самой простой и понятной метрикой является доля ошибок [Воронина и др., 2017]. При этом может быть использована кросс-валидация, когда выборка разбивается на N частей. $(N - 1)$ часть используется для обучения, а одна – для контроля. Последовательно перебираются все возможные варианты. Среднее значение ошибки по всем вариантам разбиения при кросс-валидации и характеризует обобщающую способность алгоритма и, по существу, является оценкой качества диагностики технического объекта по рассматриваемому алгоритму.

Однако использование этой метрики в задачах с неравными классами (а в задачах технической диагностики количество прецедентов с исправными состояниями, как правило, значительно больше, чем с неисправными) может привести к неправильной интерпретации полученных результатов. Более информативными критериями являются точность и полнота. Точность – доля объектов, классифицированных как положительные, в действительности являющихся положительными. Полнота – доля положительных объектов, выделенных классификатором. Один из методов объединения точности и полноты в одну метрику качества – *F*-мера (среднее гармоническое этих двух характеристик) [Davis et al., 2006].

Принадлежность объекта к конкретному классу определяется по вероятности с заданным порогом. Логичным кажется порог, равный 0.5, но часто он не является оптимальным [Witten et al., 2005]. Чтобы оценить модель, не привязываясь к выбору порога, можно использовать площадь под кривой ошибок, которая характеризует качество алгоритма: чем ближе ее значение к единице, тем качество классификации лучше.

2 Диагностика системы водоочистки

Для проведения испытаний использовались данные по системе водоочистки, работа которой определяется восьмью показателями функционирования – содержанием различных веществ в водоисточнике; получено 348 наблюдений (из них в 246 система оказалась исправной). В качестве классификаторов применялись 11 методов машинного обучения: логистическая регрессия (ЛР), дискриминантный анализ (ДА), наивный байесовский классификатор (БК), нейронные сети (НС), метод опорных векторов (МОВ), бэггинг деревьев решений (БДР), пять методов бустинга [Neykov et al., 2016]: градиентный бустинг (GrB), AdaBoost (AB) и другие.

На рисунке 1 представлены полученные результаты: процент ошибок при однократной контрольной выборке (*KB*), процент ошибок при кросс-валидации (*CV*), и другие. Анализ полученных значений показывает, что наилучшим методом обучения оказался БДР.

	Метод	Ошибка по к/в	Ошибка кросс-вал	Дисперсия по крос-вал	Точнос...	Полнота	F-мера	AUC
1	ЛР	14.3939	18.2122	10.1503	0.9171	0.8532	0.8839	0.7835
2	ДА	16.7000	21.0618	2.7687	0.8626	0.8616	0.8615	0.7707
3	БК	16.7000	20.2970	7.1120	0.8625	0.8689	0.8652	0.7125
4	НС	17.4242	17.8348	24.9689	0.9182	0.8593	0.8866	0.6660
5	МОВ	14.4000	20.3071	4.1672	0.8553	0.8749	0.8648	0.8534
6	БДР	13.6000	14.9824	14.8319	0.9277	0.8836	0.9039	0.9149
7	GrB	21.2121	18.5982	5.2848	0.9522	0.8281	0.8858	0.8637
8	AB	15.9000	17.0802	1.8317	0.9126	0.8694	0.8903	0.8548
9	LB	15.9000	18.4088	4.8734	0.9026	0.8624	0.8812	0.8351
10	GB	15.2000	19.7360	12.7112	0.8879	0.8573	0.8716	0.8049
11	RB	19.7000	19.5495	3.8693	0.8388	0.8993	0.8668	0.8257

Рисунок 1 – Показатели качества классификации при диагностике системы водоочистки

В таблице 1 показаны соответствующие коэффициенты корреляции. Наиболее сильная отрицательная связь имеет место между долей ошибок кросс-валидации (*CV*) и *F*-мерой (коэффициент корреляции r = - 0,93).

Таблица 1 – Корреляции показателей качества при различных методах классификации состояния системы водоочистки

	KB	*CV*	*D*	*P*	*R*	*F*	*AUC*
KB	1						
CV	0,26	1					
D	-0,18	-0,41	1				
P	0,01	-0,72	0,42	1			
R	-0,18	-0,06	-0,08	-0,63	1		
F	0,03	-0,93	0,45	0,85	-0,18	1	
AUC	-0,12	-0,41	-0,46	0,21	0,13	0,25	1

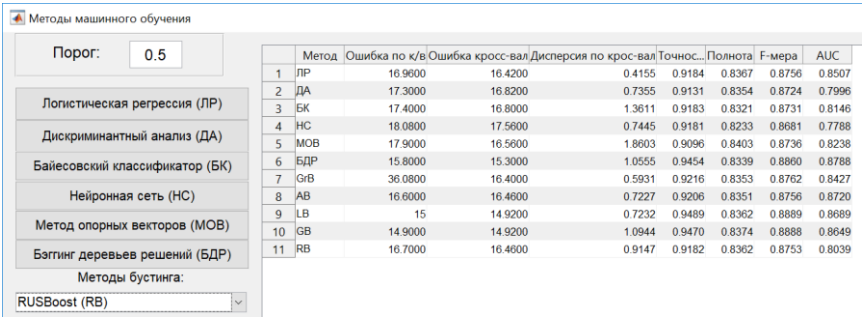
Учитывая, что контрольная выборка формируется случайным образом, и это обстоятельство может повлиять на результаты классификации, испытания были повторены 30 раз с усреднением результатов. Из таблицы 2 видно, что результаты по рассматриваемым мерам практически не изменились.

Таблица 2 – Корреляции показателей качества при различных вариантах случайного отбора контрольной выборки

	KB	*CV*	*D*	*P*	*R*	*F*	*AUC*
KB	1						
CV	-0,18	1					
D	0,02	-0,09	1				
P	0,38	-0,88	-0,03	1			
R	-0,32	-0,57	0,13	0,13	1		
F	0,26	-0,95	-0,04	0,95	0,38	1	
AUC	0,28	-0,61	0,02	0,52	0,29	0,57	1

3 Вибродиагностика гидроагрегата

Система управления гидроагрегатом фиксировала показания датчиков биения вала и вибраций гидроагрегата. Исходные данные содержат 10 показателей: вибрации нижнего генераторного подшипника верхнего бьефа и другие, а также бинарную оценку исправности гидроагрегата. Всего имелись данные по 1557 прецедентам (исправное состояние в 1204 случаях). Результаты расчета представлены на рисунке 2.



	Метод	Ошибка по к/в	Ошибка кросс-вал	Дисперсия по крос-вал	Точнос...	Полнота	F-мера	AUC
1	ЛР	16.9600	16.4200	0.4155	0.9184	0.8367	0.8756	0.8507
2	ДА	17.3000	16.8200	0.7355	0.9131	0.8354	0.8724	0.7996
3	БК	17.4000	16.8000	1.3611	0.9183	0.8321	0.8731	0.8146
4	НС	18.0800	17.5600	0.7445	0.9181	0.8233	0.8681	0.7788
5	МОВ	17.9000	16.5600	1.8603	0.9096	0.8403	0.8736	0.8238
6	БДР	15.8000	15.3000	1.0555	0.9454	0.8339	0.8860	0.8788
7	GrB	36.0800	16.4000	0.5931	0.9216	0.8353	0.8762	0.8427
8	AB	16.6000	16.4600	0.7227	0.9206	0.8351	0.8756	0.8720
9	LB	15	14.9200	0.7232	0.9489	0.8362	0.8889	0.8689
10	GB	14.9000	14.9200	1.0944	0.9470	0.8374	0.8888	0.8649
11	RB	16.7000	16.4600	0.9147	0.9182	0.8362	0.8753	0.8039

Рисунок 2 – Показатели качества классификации для различных методов обучения при вибродиагностике гидроагрегата

Из рисунка видно, что лучшими методами по большинству мер качества оказались варианты бустинга – методы LogitBoost (LB) и GentleBoost (GB). В таблице 3 показана соответствующая корреляционная матрица, из которой, как и в предыдущем опыте, видна сильная отрицательная корреляция между *CV* и *F*-мерой.

Таблица 3 – Корреляции показателей качества при различных методах классификации состояния гидроагрегата

	KB	*CV*	*D*	*P*	*R*	*F*	*AUC*
KB	1						
CV	0,23	1					
D	-0,22	-0,03	1				
P	-0,29	-0,88	-0,06	1			
R	0,11	-0,4	-0,07	-0,04	1		
F	-0,13	-0,91	-0,25	0,89	0,29	1	
AUC	-0,08	-0,8	-0,13	0,69	0,38	0,83	1

Заключение

Для оценки качества диагностики технического объекта могут использоваться различные метрики качества бинарной классификации. Проведенное исследование показывает неоднозначность взаимосвязей показателей по различным критериям. Конечно, два рассмотренных примера не могут дать объективное заключение о характере взаимосвязей, однако показывают возможные направления для выбора метрики. При

неравных объемах прецедентов для исправного и неисправного состояний, характерных для технических объектов, по-видимому целесообразно в качестве основной метрики использовать F-критерий, а при совпадающих или близких значениях этого показателя как дополнительный критерий – площадь под кривой ошибок.

Список литературы

[Биргер 1978] Биргер И.А. Техническая диагностика. – М. : Машиностроение, 1978. – 240 с. (2-е изд.: М. : URSS, 2019)

[Воронина и др., 2017] Воронина В.В. Теория и практика машинного обучения : учебное пособие / В. В. Воронина, А. В. Михеев, Н. Г. Ярушкина, К. В. Святов. –Ульяновск : УлГТУ, 2017. – 290 с.

[Воронцов, 2016] Воронцов К.В. Машинное обучение. Композиция классификаторов https://yadi.sk/i/FItIu6V0beBmF

[Жуков и др., 2018] Жуков Д.А., Клячкин В.Н. Диагностика исправности технического объекта с использованием пакета Matlab // Перспективные информационные технологии: труды Международной научно-технической конференции. – Самара: Изд. Самарского научного центра РАН, 2018. – С. 55-57.

[Клячкин и др., 2018] Клячкин В.Н., Кувайскова Ю.Е., Жуков Д.А. Диагностика технического состояния аппаратуры с использованием агрегированных классификаторов // Радиотехника. – 2018. – №6.

[Соколов, 2016] Соколов Е.А. Машинное обучение [Электронный ресурс] http://wiki.cs.hse.ru/ Машинное_обучение _1/2016_ 2017

[Davis et al,, 2006] Davis J., Goadrich M. The relationship between Precision-Recall and ROC curves / Proceedings of the 23rd international conference on Machine learning. – Pittsburgh, 2006.

[Neykov et al., 2016] Neykov, M. On the Characterization of a Class of Fisher-Consistent Loss Functions and its Application to Boosting / M. Neykov, Jun S. Liu, Tianxi Cai // Journal of Machine Learning Research. 2016. №17(70).

[Witten et al., 2005] Witten I.H., Frank E. Data mining: practical machine learning tools and techniques. 2nd ed. – San Francisco : Morgan Kaufmann Publishers, 2005. – 525 p.

ASSOCIATION ANALYSIS OF THE QUALITY INDICATORS OF THE OBJECT'S DIAGNOSIS DURING THE BINARY CLASSIFICATION

Zhukov D.A.(zh.dimka17@mail.ru)
Klyachkin V.N. (v_kl@mail.ru)
Ulyanovsk state technical university, Ulyanovsk

To recognize the technical condition of an object can be used methods of machine learning (neural networks, begging of decision trees, logistic regression, boosting methods, etc.). The article deals with measures of the quality of its object recognition: percentage of errors during cross-validation, F-measure, area under the curve of bugs and others. Based on examples of real objects examines the relationship between these indicators.

Keywords: technical diagnostics, machine learning, binary classification, quality indicators, correlations

УДК 004.4

РАЗРАБОТКА МОДУЛЯ АВТОМАТИЗАЦИИ РАБОТЫ С КОНФЕРЕНЦИЯМИ В КАФЕДРАЛЬНОМ ПРИЛОЖЕНИИ

Зарайский В.И. (v-zar@list.ru)
Ульяновский государственный технический университет, Ульяновск

Описывается разработка модуля автоматизации работы с конференциями в кафедральном приложении. Статья включает описание системы, ее функциональность, архитектуру, а также технологии и инструменты, использованные при разработке.

Ключевые слова: автоматизация, кафедра, веб-приложение, клиент-сервер, конференция, spring boot, java.

Введение

Актуальность данной разработки подтверждается тем, что кафедра «Информационные системы» постоянно принимает участие в конференциях, а также проводит конференции внутри университета. Сотрудники кафедры не имеют централизованной системы по учету всей информации о конференциях. Каждому члену рабочей группы необходимо самому вести учет информации в произвольной форме. Информирование участников конференции осуществляется с помощью звонков, электронных сообщений и при личной встрече. Система оповещения об изменениях, а также оповещение о приближающихся сроках отсутствует, что влечет за собой просрочку некоторых задач по датам.

Необходимо разработать модуль для кафедрального приложения, который позволил бы централизованно хранить информацию о конференциях, в которых кафедра когда-либо принимала участие. Также пользователи нуждаются в системе управления информацией о конференциях для поддержания их в актуальном состоянии. Модуль должен в автоматическом режиме оповещать сотрудников кафедры об изменениях в конференциях и о приближении крайних сроков конференции.

Целью создания модуля является упрощение учета информации по конференциям и срокам участия в них за счет минимизации количества выполняемых операций, а также централизации всех данных о конференции в одной базе данных.

1 Описание модуля

Модуль для работы с конференциями является частью кафедрального веб-приложения. Приложение представляет собой систему контроля работы кафедры «Информационной системы», позволяющие проще, быстрее и эффективнее ставить, отслеживать и выполнять задачи.

Кафедральное приложение предназначено для:

- сокращения времени управления текущими активностями научной группы;
- автоматизированного управления задачи – интеллектуальной постановки задач исполнителям;
- обучения бакалавров современными технологиями разработки;
- хранения и трансляции опыта между участниками научной группы.

Модуль предназначен для автоматизации деятельности кафедры по принятию участия в конференциях. Упрощение достигается за счет централизованного хранения списка конференций и поддержания его в актуальном состоянии силами научной группы. Также модуль обладает функциональностью оповещать пользователей системы об изменении информации или приближении дат крайних сроков.

Данные нововведения позволяют повысить эффективность работы кафедры с помощью упрощения и сокращения времени отслеживания информации по конференциям.

1.1 Функции системы

Модуль имеет следующие функции:

- создание конференции;
- редактирование конференции:
    - изменение текстовых полей наименование, описания, URL адрес и дат проведения конференции;
    - добавление крайнего срока;
    - редактирование крайнего срока;
    - удаление крайнего срока;
    - прикрепление существующих статей;
    - добавление новых статей;
    - откреплление статей;
    - принятие участие в конференции текущему пользователю;
    - выбор формата и вида участия;

- удаление конференции;
- просмотр всего списка конференции;
- просмотр актуального списка конференции;
- фильтрация списка конференций по годам и участникам;
- «привлечение внимания» участников конференции;
- отображение крайних сроков конференций на графике событий пользователя;
- автоматическое оповещение пользователей веб-приложения на электронную почту о создании новой конференции;
- автоматическое оповещение участников конференции на электронную почту:
  - о приближающихся крайних сроках;
  - об изменении информации о крайних сроках;
  - об изменении сроков проведения конференции;
- автоматическая отправка статистики по «привлечению внимания» участникам конференции на электронную почту.

«Привлечение внимания» конференции служит для повышения приоритета конференции. Также, за счет отправки статистики по «привлечению внимания» всем участникам конференции, может являться дополнительной системой оповещения в случае, если были внесены какие-либо информационные изменения.

1.2 Архитектура модуля

Модуль обладает трехуровневой архитектурой, представленной с помощью нотации UML диаграммой развертывания (рисунок 1). На верхнем уровне находится слой клиента, представленный в виде браузера. Далее идет слой бизнес-логики и слой данных.

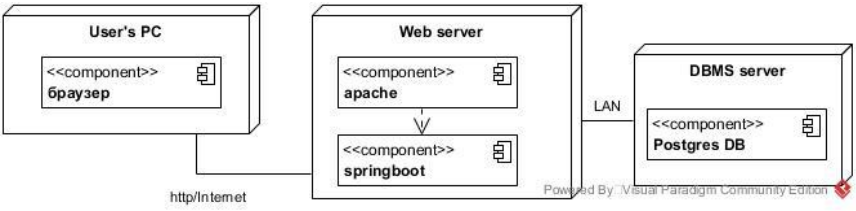


Рисунок 1 – Диаграмма развертывания

Взаимодействие клиента и веб-сервера осуществляется посредством REST (Representational state transfer) запросов и специального шаблона проектирования DTO (Data Transfer Object) для передачи данных между слоями.

Серверная часть обрабатывает данные, пришедшие со стороны клиента, взаимодействует с другими модуля системы и осуществляет запросы к базе данных. На стороне сервера используется архитектурный паттерн MVC (Model View Controller), позволяющий отделить логику обработки запросов от представления и сущностей базы данных.

Диаграмма классов представлена на рисунке 2. Класс ConferenceController отвечает за взаимодействие между клиентской и серверной частями модуля. Каждый метод имеет набор аннотаций, с помощью которых фреймворк определяет из URL адреса запроса, какой метод необходимо вызвать.

Контроллер передает управление сервису модуля ConferenceService, содержащий основную бизнес-логику. Здесь происходит обработка всей информации, а также взаимодействие с другими модулями приложения.

После завершения обработки информации, сервис с помощью интерфейса ConferenceRepository сохраняет изменения в базу данных.

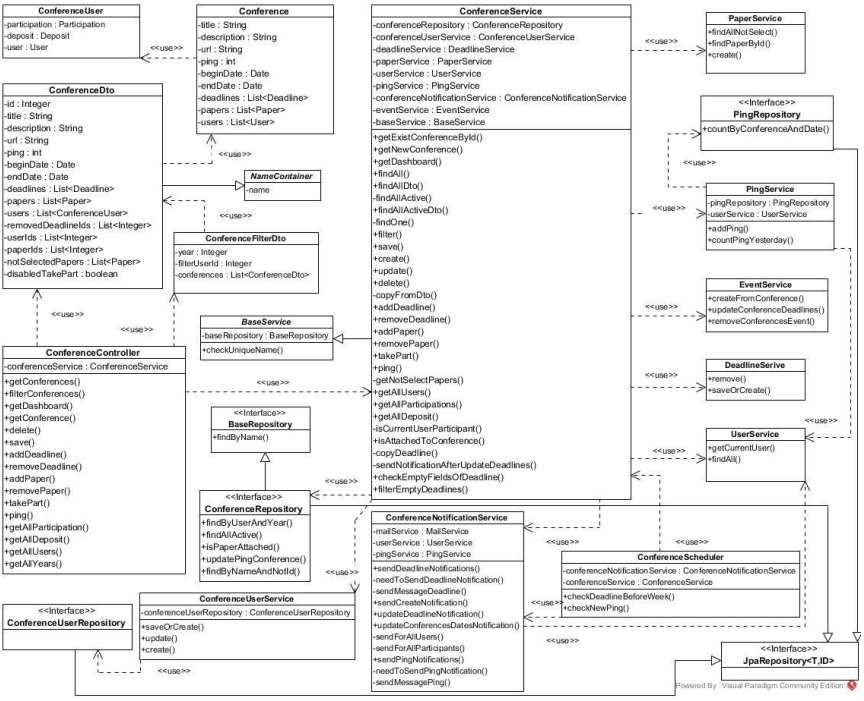


Рисунок 2 – Диаграмма классов

2 Описание реализации

Для создания использовалась интегрированная среда разработки программного обеспечения для многих языков программирования IntelliJ IDEA, так как обладает бесплатной версией, включающая в себя инструменты для проведения тестирования приложения, работы с распределенной системой управления версиями Git, поддержку системы автоматической сборки Gradle.

Сам проект имеет собственное закрытое хранилище на сайте Gitlab. Данный сайт позволяет разграничить разработку модулей для кафедрального приложения на этапы (milestones), к которым привязываются проблемы (issues) по модулям. Вся разработка происходит в основной ветке dev. Gitlab, благодаря удобному UI, дает возможность создавать новые ветки от dev, которые привязаны к определенной проблеме для ее решения, а также отслеживать все изменения, которые будут внесены после слияния с основной веткой. Данный подход к организации разработки модулей позволил разработчикам абстрагироваться от других частей создания системы, контролировать ход выполнение собственных задач, тем самым повысив эффективность разработки. Модуль работы с конференциями не стал исключением.

Для реализации серверной части использовался универсальный фреймворк с открытым исходным кодом Spring Framework, написанным на Java. Фреймворк обладает широкой функциональностью, поэтому тяжело определить наиболее значимые структурные элементы.

Для взаимодействия сервера с базой данных использовался интерфейс JpaRepository. Данный интерфейс уже обладает базовой функциональностью, так как наследуется от двух других интерфейсов PagingAndSortingRepository и CrudRepository. Особенностью интерфейса является то, что JpaRepository способен анализировать широкий спектр названий методов и по этим результатам создавать запросы к базе данных самостоятельно, упрощая разработку. Если же есть потребность написать сложный SQL запрос, то можно воспользоваться аннотацией @Query. Также интерфейс позволяет абстрагироваться разработчику от того, с какой базой данных приходится взаимодействовать.

Для реализации клиентской части использовался шаблонизатор Thymeleaf. Шаблонизатор интегрирован в Spring и хорошо подходит для обслуживания HTML5 на уровне представлений веб-приложения на основе MVC. Thymeleaf позволяет реализовать компонентный подход для разработки приложения, который упрощает виденье проекта, а также уменьшает порог вхождения в проект новых разработчиков.

Возможностями шаблонизатора также являются инкапсуляция и повторное использование элементов DOM.

Особенностью модуля является «привлечение внимания» пользователя. Преподавателю необходимо нажать на кнопку «Ping участникам» на странице просмотра информации о конференции. Сделать это он может несколько раз, тем самым повысив приоритет конференции. Результаты того, что преподаватель нажал на кнопку, заносятся в специальную таблицу Ping базы данных. Ежедневно по расписанию сервис ConferenceScheduler производит проверку, необходимо ли «привлечь внимание» по какой-либо конференции. Если так и есть, то всем участникам конференции будет отправлено сообщение на электронную почту с соответствующей информацией. Пример оповещения представлен на рисунке 3.

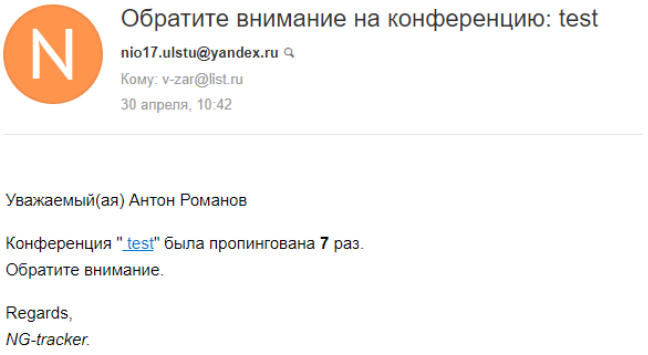


Рисунок 3 – Оповещение о «привлечении внимания»

Также после повышения приоритета конференции на странице с актуальными конференциями она будет визуально отличаться. Конференция считается актуальной, если ее дата окончания проведения больше текущей даты. Чем больше приоритет, тем ярче будет выглядеть карточка конференции. Пример представлен на рисунке 4.

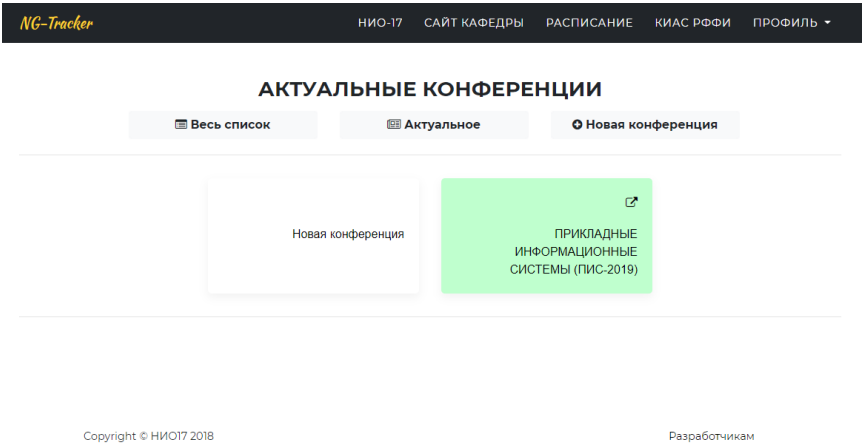


Рисунок 4 – Список актуальных конференций

На рисунке 5 показана начальная страница модуля. Здесь можно видеть меню навигации сверху, список конференций, а также параметры фильтров.

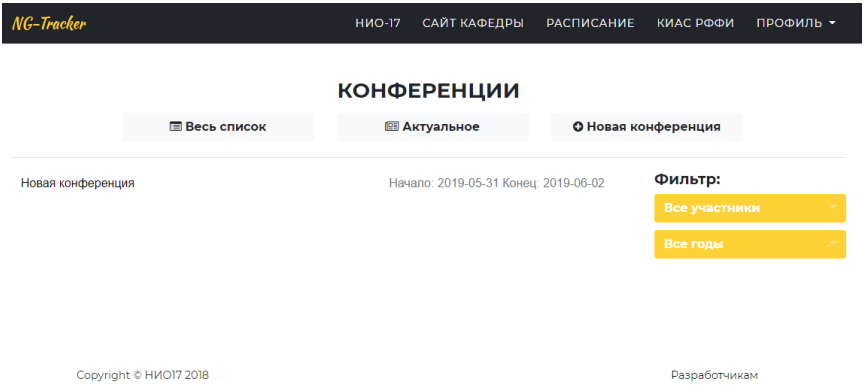


Рисунок 5 – Начальная страница модуля

Страница с просмотром и редактированием информации о конференции представлена на рисунке 6.

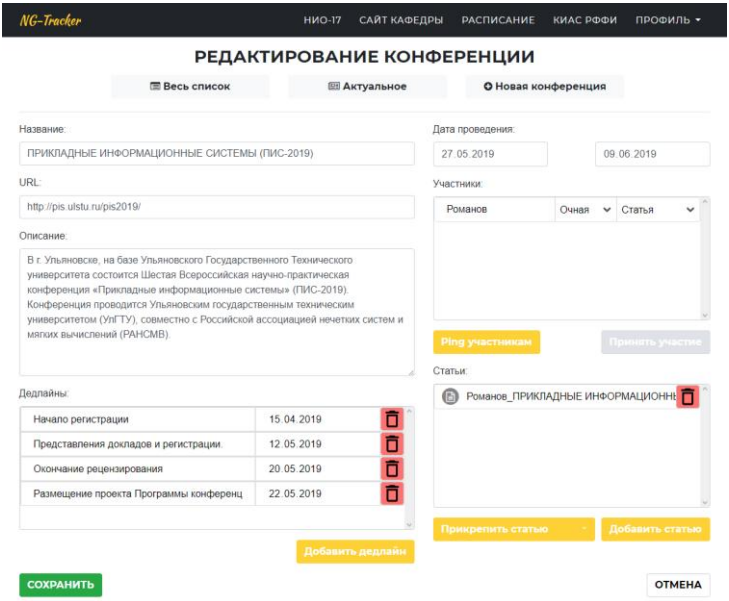


Рисунок 6 – Страница просмотра и редактирования информации о конференции

Заключение

По результатам работы разработан модуль автоматизации работы с конференциями для кафедрального приложения. Функциональность приложения в полном объеме покрывает потребности сотрудников кафедры при участии в конференциях. Теперь членам рабочей группы не придется самим вести учет информации по конференциям, а система оповещения позволяет сократить время, уделяемое конференциям, на организацию участия и поддержания актуальности информации

Модуль обладает большим потенциалом, так как не только кафедра «Информационные системы» принимает участие в конференциях, но и все кафедры университета. Также систему можно расширять дополнительными возможностями, не затрагивая основную функциональность. Модуль может являться хорошим примером для начинающих бакалавров для освоения современных средств создания приложений.

Благодарности. Автор считает своим приятным долгом поблагодарить тех, кто внимательно прочитал данную статью.

Список литературы

[Романов 2016] Романов А.А, Конструирование программного обеспечения: учебное пособие. – Ульяновск: УлГТУ. 2016. – 127 с.

[Spring boot docs 2019] Документация Spring boot [Электронный ресурс]. Spring Boot Docs 2.1.5.RELEASE API: [сайт].URL: https://docs.spring.io/spring-boot/docs/current/api/ (дата обращения: май 2019 год)

[Gradle 2019] Gradle [Электронный ресурс]. Wikipedia: [сайт].URL: https://ru.wikipedia.org/wiki/Gradle (дата обращения: май 2019 год)

[HTTP 2019] HTTP [Электронный ресурс]. Wikipedia: [сайт].URL: https://ru.wikipedia.org/wiki/HTTP (дата обращения: май 2019 год)

[Java Persistence API 2019] Java Persistence API [Электронный ресурс]. Wikipedia: [сайт].URL: https://ru.wikipedia.org/wiki/Java_Persistence_API (дата обращения: май 2019 год)

[DTO 2019] DTO [Электронный ресурс]. Wikipedia: [сайт].URL: https://ru.wikipedia.org/wiki/DTO (дата обращения: май 2019 год)

[Thymeleaf 2019] Thymeleaf [Электронный ресурс]. Wikipedia: [сайт].URL: https://en.wikipedia.org/wiki/Thymeleaf (дата обращения: май 2019 год)

[REST 2019] REST [Электронный ресурс]. Wikipedia: [сайт].URL: https://ru.wikipedia.org/wiki/REST (дата обращения: май 2019 год)

[Spring Framework 2019] Spring Framework [Электронный ресурс]. Wikipedia: [сайт].URL: https://ru.wikipedia.org/wiki/Spring_Framework (дата обращения: май 2019 год)

SYSTEM OF CUSTOMIZATION OF CUISINE DESIGN

Zaraysky V.I.(v-zar@list.ru)
Ulyanovsk State Technical University, Ulyanovsk

In work describes the development of the automation module working with conferences in the cathedral application. The article includes a description of the system, its functionality, architecture, and technologies and tools used in the development.

Keywords: automation, chair, web application, client-server, conference, spring boot, java.

УДК 004.89

РАЗРАБОТКА АВТОМАТИЗИРОВАННОЙ СИСТЕМЫ КЛАСТЕРИЗАЦИИ ПРОГРАММНЫХ РЕПОЗИТОРИЕВ КРУПНЫХ ПРОЕКТНЫХ ОРГАНИЗАЦИЙ

Савельев Я.К.(Kpuccmajl@gmail.com)
Ульяновский государственный технический университет, Ульяновск

Описывается алгоритм работы и общее описание программного средства для кластеризации программных репозиториев крупных проектных организаций

Ключевые слова: разработка, кластеризация, валидация, программное средство, репозиторий, ПИС–2019

Введение

Программным репозиторием является программный код, расположенный на удаленном сервере, поддерживаемый системой контроля версий [2].

В ходе работы инженером-программистом в качестве разработчика части репозиториев большого проекта (более двух миллионов строк кода) для автоматизированных систем управления надводных кораблей, зачастую возникала необходимость использования репозиториев сторонних разработчиков. Из-за большого количества зависимостей, тяжело контролировать качество финального кода.

Для менеджера проекта эта задача является особенно трудной по той причине, что проект состоит из множества репозиториев, со своими зависимостями и особенностями. Программное средство контроля над проектами Jenkins позволяет контролировать процесс общей сборки проекта и анализировать код на ошибки, но не позволяет группировать модули проекта по критериям качества программного кода и его количества. Для реализации данного функционала было разработано программное средство, описанное в данной статье. В его основе лежит алгоритм кластеризации k-means. После кластеризации происходит валидация полученных кластеров с использованием минимаксного критерия до тех пор, пока значение индекса валидации не станет

максимальным. Вывод результатов происходит в виде процентного отношения расстояния от репозитория до центроида ближайшего кластера, с расстоянием до центроидов остальных кластеров.

Целью работы является разработка и анализ применимости метода валидации кластеров с использованием минимаксного критерия для кластеризации программных репозиториев крупных проектных организаций.

1 Алгоритм работы кластеризации и валидации

В качестве основных алгоритмов кластеризации были выбраны методы «k-means» и «maximum capturing» [6].

Таблица 1 – Сравнение алгоритмов кластеризации

Алгоритм	k-means	Maximum capturing
Количество кластеров	Заранее известно	Вычисляется в ходе алгоритма
Необходимость валидации	Требуется валидация	Валидация не требуется
Работа с вещественными метриками	Оптимально	Нежелательно (требуется применение коэффициента схожести)

Метод «maximum capturing» не требует валидации полученных кластеров, но имеет недостаток в том, что для работы с вещественными индексами приходится использовать коэффициент схожести, который задается вручную, что негативно влияет на гибкость работы алгоритма.

Поэтому в качестве основного алгоритма кластеризации был выбран метод «k-means». Единственным управляющим параметром является число классов, на которые проводится разбиение $S = (S_1, \dots, S_k)$ выборки X. В результате получается несмещенное разбиение $S^* = (S_1^*, \dots, S_k^*)$.

Математическое описание алгоритма:

1. Выберем начальное разбиение $S^0 = (S_1^0, \dots, S_k^0)$, где

$$S_i^0 = \{x_{i1}^0, \dots, x_{jn}^0\}, U_1^k S_i^0 = X, S_i^0 \cap S_{i`}^0 = \emptyset, i \neq i`.$$

2. Пусть построено m-е разбиение $S^m = (S_1^m, \dots, S_k^m)$. Вычислим набор средних $e^m = (e_1^m, \dots, e_k^m)$, где $e^m = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}^m$.
3. Построим минимальное дистанционное разбиение порождаемое набором e^m возьмем его в качестве$S^{m+1} = (S_1^{m+1}, \dots, S_k^{m+1})$, т. е.:
$$S_1^{m+1} = \{X \in X: d(X, e_i^m) = \min_{1 \le i \le k} d(X, e_i^m)\},$$
где $d(X, e) = ||X - e||$
4. Если $S^{m+1} \neq S^m$,то переходим к п.2, заменив m на m+1, если $S^{m+1} = S^m$, то полагаем $S^m = S^*$ и заканчиваем работу алгоритма [4].

Преимуществом алгоритма являются быстрота и простота реализации. К его недостаткам можно отнести неопределенность выбора начальных центров кластеров, а также то, что число кластеров должно быть задано изначально, что может потребовать некоторой априорной информации об исходных данных [1].

Алгоритм был реализован на языке C++ с использованием библиотеки Qt. Для валидации кластеров был использован минимаксный критерий, состоящий из следующих этапов:

1. Для каждого кластера вычисляется разница между максимальным и минимальным расстоянием от точек, входящих в кластер до центроида.
2. Вычисляется среднее максимальное расстояние между центроидами кластера.
3. Результат вычисления на этапе 2 делится на результат вычисления на этапе 1.
4. Алгоритм повторяется заданное число N итераций, в ходе которых кластеризация по методу k-средних начинается заново.
5. Результат кластеризации с наименьшим значением минимаксного критерия является лучшим.

Данный алгоритм валидации разработан на основе метода валидации Дэвиса-Болдина, вычисляющем индекс валидации по формуле

$$DB = \frac{1}{k}\sum_{i=1}^{k} \max_{i \neq j} \left\{\frac{S_k(Q_i)+S_k(Q_j)}{S(Q_i,Q_j)}\right\},$$

где K – количество кластеров, S_K – среднее расстояние от объектов до центра кластера, $S(Q_i,Q_j)$ – расстояние между центрами кластеров. Чем меньше значение этого индекса, тем кластеры компактнее и удаленнее друг от друга [4].

Таблица 2 – Сравнение методов валидации

Метод валидации	Метод Дэвиса-Болдина	Минимаксный критерий
Влияние количества кластеров на валидацию	+	-
Методика вычисления суммы расстояний	Расчет отдельно для каждого кластера	Сумма расстояний во всех кластерах
Индекс на выходе	Средний разброс кластеров	Минимальный разброс разницы расстояний между кластерами

В качестве основного метода валидации был выбран метод по минимаксному критерию по той причине, что кластеры, прошедшие валидацию с использованием индекса Дэвиса-Болдина показывали большее отклонение от среднего значения индекса, в сравнении с минимаксным критерием. Также эмпирическим путем было выявлено, что кластеры, прошедшие валидацию по методу минимаксного критерия наиболее применимы к данной предметной области.

2 Описание разработанного программного средства

Для сбора метрик репозиториев использовались утилиты git, cppcheck и sloccount, включенные в состав операционной системы специального назначения Astra Linux 1.6.

Утилита git представляет собой распределенную систему управления версиями. Данная утилита используется для контроля версий репозиториев. С помощью данной утилиты происходит сбор метрики **Support**, представляющую собой индекс поддерживаемости репозитория от 0.0 до 1.0. Наименьший индекс присваивается репозиториям, не обновлявшимся более 90 дней. Наивысший присваивается репозиториям, которые обновлялись в течение суток. В случае, если обновление репозитория происходило в периоде от 1 до 90 дней, то индекс поддерживаемости рассчитывается как промежуточное значение.

Утилита sloccount представляет собой средство сбора метрики программного обеспечения, основанного на подсчете количества строк исходного кода [3]. С помощью данной утилиты происходит сбор метрик **Size** и **C%**, представляющих собой индекс размера репозитория от 0.0 до 1.0, и индекс количества строк кода, написанных на языке C. Чем выше количество строк кода в репозитории, тем выше индекс метрики **Size**, достигающий максимума при количестве строк кода свыше 50000. Индекс

метрики **C%** вычисляется как отношение количества строк на языке C к общему объему кода.

Утилита cppcheck представляет собой статический анализатор кода программы. После анализа кода репозитория утилита выводит отчет в виде количества стилистических ошибок, предупреждений и ошибок кода (например, вероятные утечки памяти). Метрика **Quality** высчитывается, опираясь на количество данных ошибок. Индекс качества представляется в виде вещественного числа от 0.0 до 1.0. Наименьший индекс присваивается репозиториям, в которых количество ошибок превышает 200. Наивысший индекс присваивается репозиториям без ошибок. В случае, если количество ошибок в интервале от 0 о 200, то индекс качества рассчитывается как промежуточное значение.

После сбора метрик с помощью данных утилит, для репозитория выделяют метрику **mQuality**, представляющую собой сумму метрик **Quality** и **Support.** Также выделяют метрику **mQuantity,** представляющую собой сумму метрик **Size** и **C%.**

После разработки алгоритма и проверки его работы на тестовых данных было разработано графическое программное средство для удобства работы оператора. Разработка осуществлялась на языке C++ с использованием библиотеки Qt**.**

На вход программе подается конфигурационный файл с адресами репозиториев в формате xml в следующем виде:

<repository link="@gitlab.lan.ru:project/repository.git "/>

По нажатию на кнопку «Analyze» происходит загрузка репозиториев с последующим сбором метрик.

Репозитории заносятся в таблицу вместе со значениями метрик и отображаются на графике.

Запуск кластеризации происходит по нажатию на кнопку «Clusterize»**.** Алгоритм k-средних требует на вход количество конечных кластеров, для этого на главном окне расположено соответствующее поле.

Количество конечных кластеров устанавливается эмпирически, но рекомендуемое значение количества кластеров отображается в отчете кластеризации и равно корню квадратному от половины общего количества точек.

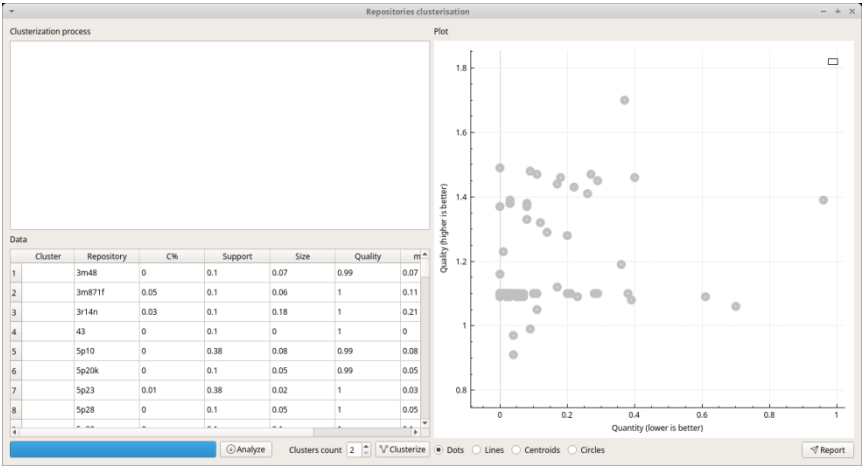


Рисунок 1 – Проанализированный список репозиториев в табличном и графическом виде

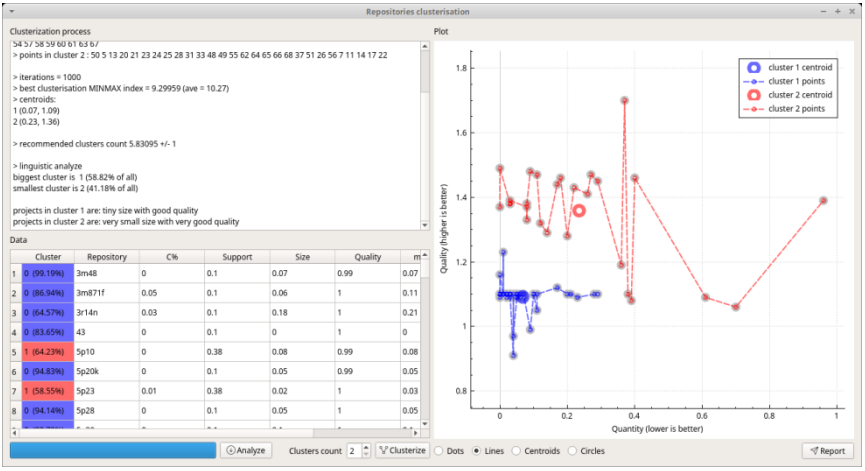


Рисунок 2 – Отображение центроидов кластеров и процентного отношения репозиториев к кластерам (для двух кластеров)

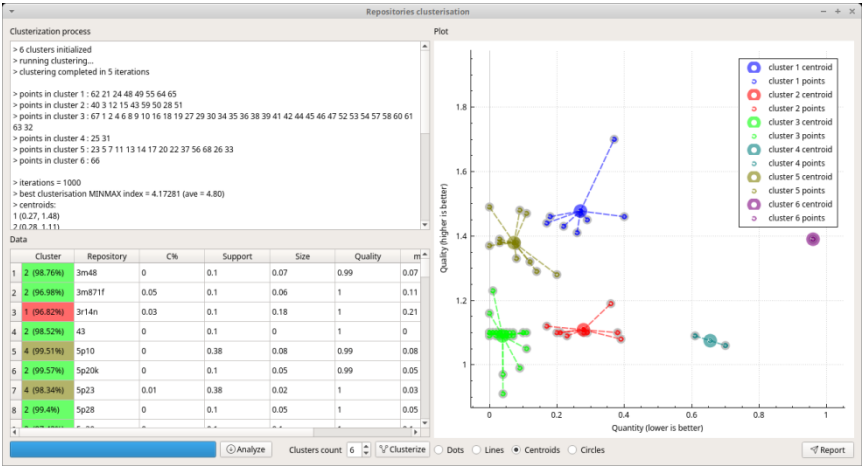


Рисунок 3 – Рекомендуемое количество кластеров – 6. На рисунке изображена кластеризация для данного количества кластеров

Было выявлено, что для данного набора репозиториев шесть кластеров являются оптимальным количеством.

На основе лингвистического анализа по метрикам центроидов кластеров установлено:

1. Репозитории, входящие в третий и второй кластеры малы, по объему кода, качество кода самое плохое. Примечательно, что в данные кластеры отнеслось самое большое количество репозиториев (более 60%).
2. В первый и пятый кластеры вошли репозитории с относительно малым количеством кода и лучшим качеством.
3. В четвертый кластер попали репозитории с большим объемом кода и с наиболее низким индексом качества.
4. В шестой кластер попал только один репозиторий, в нем объем кода самый большой с довольно высоким качеством.

Для вывода текстового отчета кластеризации и лингвистического анализа в файл необходимо нажать на кнопку «Report».

Заключение

В ходе работы было достигнуто следующее:

1. Разработан алгоритм сбора количественных и качественных метрик программного репозитория для дальнейшей кластеризации.

2. Разработан метод валидации полученных кластеров на основе минимаксного критерия.
3. Разработана объектно-ориентированная архитектура программного средства для автоматизированной кластеризации программных репозиториев.
4. Разработан алгоритм лингвистического анализа кластера в контексте программных репозиториев крупных проектных организаций.

Удобный лингвистический анализатор кластеров позволяет выделить репозитории, требующие контроля над качеством программного кода из общего числа, а также репозитории с лучшим качеством кода и минимальным объемом.

Благодарности. Считаю своим долгом поблагодарить тех, кто прочитал данную статью, а также Наместникова А.М., Эгова Е.Н. и Филиппова А.А. за помощь в ее написании.

Список литературы

[Loginom, 2016] Метод k-средних [Электронный ресурс] [сайт].URL: https://wiki.loginom.ru/articles/k-means.html (дата обращения: 24.04.2019).

[Потапов Михаил Сергеевич, 2016] Репозиторий как средство предоставления программного обеспечения [Электронный ресурс][сайт].URL: https://sibac.info/journal/student/16/83704 (дата обращения: 23.04.2019).

[David Wheeler, 2004] Sloccount user's guide [Электронный ресурс] [сайт].URL: https://dwheeler.com/sloccount/sloccount.html (дата обращения: 24.04.2019)

[Андросова Т.Е., Курочкин В.М., Болдырев А.С. [и др.]] Применение алгоритма K-MEANS для эскизного проектирования местоположения транспортных объектов // Молодежный научный форум: Технические и математические науки: электр. сб. ст. по мат. XLI Междунар. студ. науч.-практ. URL: https://nauchforum.ru/archive/MNF_tech/1(41).pdf (дата обращения: 20.04.2019)

[Афанасьева Т.В., 2018] Применение методов интеллектуального анализа данных и процессов [Текст] / сост. Т. В. Афанасьева. – Ульяновск : УлГТУ, 2018. – 51 с.

[Chintakindi Srinivasa, Vangipuram Radhakrishnab, Dr.C.V.Guru Rao, 2014] Clustering and Classification of Software Component for Efficient Component Retrieval and Building Component Reuse Libraries [Электронный ресурс].[сайт].URL:https://www.semanticscholar.org/paper/Clustering-and-Classification-of-Software-Component-Srinivas-Radhakrishna/67ac2fc8979e84b4e57bac4ccab3af8e71f821ec (дата обращения: 19.04.2019)

DEVELOPMENT OF AUTOMATED CLUSTERING SYSTEM OF SOFTWARE REPOSITORIES IN MAJOR PROJECT ORGANIZATIONS

Savelyev Y.K. (Kpuccmajl@gmail.com)
Ulyanovsk state technical university, Ulyanovsk

This article describes the algorithm of work and a general description of the software for clustering software repositories of major project organizations.

Keywords: development, clustering, software, repository, AIS'2019

УДК 004.8

КЛАССИФИКАЦИЯ СНИМКОВ КОМПЬЮТЕРНОЙ ТОМОГРАФИИ С ЦЕЛЬЮ ВЫЯВЛЕНИЯ РАКА ЛЕГКИХ

Полежаев П.П.(polezhaev.ds@gmail.com)
Усанова А.А. (usanova.anastasiya.99@mail.ru)
Оренбургский государственный университет, Оренбург

Данная работа посвящена одному из подходов к выявлению рака легкого на основе снимков компьютерной томографии. В статье приведена архитектура сверточной нейронной сети, осуществляющей классификацию легочных образований. Данная сеть была обучена на тренировочном наборе данных и продемонстрировала достаточно высокую точность на валидационном и тестовом наборах.

Ключевые слова: прогнозирование рака легкого, злокачественные узелковые образования, машинное обучение, сверточные нейронные сети

Введение

Рак легкого является серьезной медицинской и социальной проблемой. Методы машинного обучения также активно применяются для обнаружения и классификации раковых поражений в медицинских изображениях, что значительно помогает радиологам улучшить точность принимаемых решений при идентификации рака [Yang et. al., 2016]. Наиболее подходящим данными для выявления рака легких являются снимки компьютерной томографии грудной клетки [Gomathi et. al., 2010]. Нейронные сети, в частности, сверточные нейронные сети, как вид машинного обучения, особенно хорошо проявляют себя в решении задач распознавания и обработки изображений. Поэтому актуально их применение в работе со снимками компьютерной томографии.

На данный момент по рассматриваемой теме существует несколько исследований, например, в работе [Yang et. al., 2016] сжато освещена архитектура нейронной сети, классифицирующей легочные узелки. В основном статья посвящена аугментации данных – подготовке набора

данных для обучения сети. В [Cesar et. al., 2016] приведено описание сверточной нейронной сети, имеющей архитектуру, отличающуюся от предложенной в настоящей работе, также описаны методы и результаты ее обучения.

Создание нейронной сети, классифицирующей легочные образования

В качестве первичной задачи необходимо обучить нейронную сеть классифицировать отдельные части компьютерной томографии как изображение со злокачественным узелковыми образованиями или изображения без злокачественных образований. Следующей задачей будет являться применение данной нейронной сети к снимкам компьютерной томографии с целью составления карт вероятностей наличия узелковых образований с последующей их бинарной классификацией (есть рак – нет рака).

Для обучения такой сети использовался набор данных Lung Nodule Malignancy [Lung Nodule Malignancy, 2017], состоящий из 6691 изображений разрешением 64×64 пикселя, на каждом из которых различные части легочной томографии. 2526 изображений данного набора содержат злокачественные образования (помечены классом 1), 4165 изображений содержат доброкачественные образования или другие неузелковые части томографии (класс 0).

Была построена архитектура сверточной нейронной сети, получающей на вход изображение разрешением 64×64 в оттенках серого, и выдающей на выходе вероятности принадлежности классам 1 и 0 (см. рисунок 1).

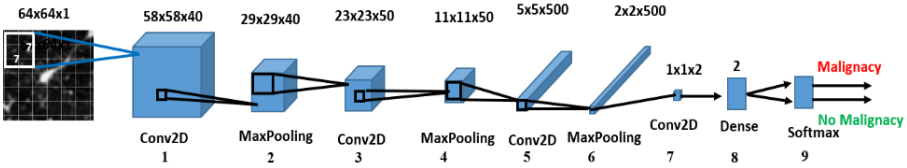


Рисунок 1 – Архитектура модели сверточной нейронной сети

Описание слоев сети приведено в таблице 1.

Таблица 1 – Описание основных слоев модели

№	Название слоя	Описание слоя
1	Conv2D	Сверточный слой с ядром свертки 7×7. На выходе 40 карт признаков размером 58×58.
2	MaxPooling	Слой имеет фильтр 2×2, шаг обработки 2x2. На выходе 40 карт признаков размером 29×29.
3	Conv2D	Сверточный слой с ядром свертки 7×7. На выходе 50 карт признаков размером 23×23.
4	MaxPooling	Слой имеет фильтр 2×2, шаг обработки 2x2. На выходе 50 карт признаков размером 11×11.
5	Conv2D	Сверточный слой с ядром свертки 7×7. На выходе 500 карт признаков размером 5×5.
6	MaxPooling	Слой имеет фильтр 2×2, шаг обработки 2×2. На выходе 500 карт признаков размером 2×2.
7	Conv2D	Сверточный слой с ядром свертки 2×2. На выходе выдает 2 карты признаков размером 1×1.
8	Dense	Полносвязный слой, имеющий 2 выхода.
9	Softmax	Слой, преобразующий выходы сети в вещественные числа в диапазоне [0,1], соответствующие вероятностям двух классов.

Перед полносвязным слоем идет слой Flatten, который выравнивает тензор, поступающий ему на вход, преобразуя все его измерения в одно. После каждого сверточного слоя следует слой BatchNormalization, за которым идет слой нелинейности, представляющий собой функцию активации RELU (REctified Linear Unit).

Реализация модели осуществлялось на языке Python при помощи библиотеки Keras (фрагмент кода приведен на рисунке 2). Для обучения модели использовалось 70% данных всего набора, из которых 20% отводилось на валидацию на каждой эпохе обучения. Оставшиеся 30% образцов всего набора использовались для конечного тестирования модели. Обновление весов происходило с помощью оптимизатора Adam.

```
model = Sequential()
model.add(Conv2D(40, kernel_size = 7, padding = 'valid'))
model.add(BatchNormalization())
model.add(Activation('relu'))
model.add(MaxPooling2D(pool_size = (2,2), strides = (2,2)))
model.add(Conv2D(50, kernel_size = 7, padding = 'valid'))
model.add(BatchNormalization())
model.add(Activation('relu'))
model.add(MaxPooling2D(pool_size = (2,2), strides = (2,2)))
model.add(Conv2D(500, kernel_size = 7, padding = 'valid'))
model.add(BatchNormalization())
model.add(Activation('relu'))
model.add(MaxPooling2D(pool_size = (2,2), strides = (2,2)))
model.add(Conv2D(2, kernel_size = 2, padding = 'valid'))
model.add(BatchNormalization())
model.add(Activation('relu'))
model.add(Flatten())
model.add(Dense(2))
model.add(Activation('softmax'))
```

Рисунок 2 – Исходный код модели нейронной сети

В рамках эксперимента модель обучалась на 150 эпохах. Уже на 70-й эпохе значение функции ошибки (logloss) начало медленно расти на валидационном наборе (см. рисунок 2), что свидетельствует о переобучении модели. Оптимальное состояние модели, полученное на 61-й эпохе, было сохранено в качестве результата обучения. Максимальная верность (accuracy), которую удалось достичь на валидационных данных, составила 93.92%, минимальное значение функции ошибки – 0.2122. На тестовых данных, не участвующих в обучении, модель показала верность 93.37%, значение функции ошибки – 0.2097.

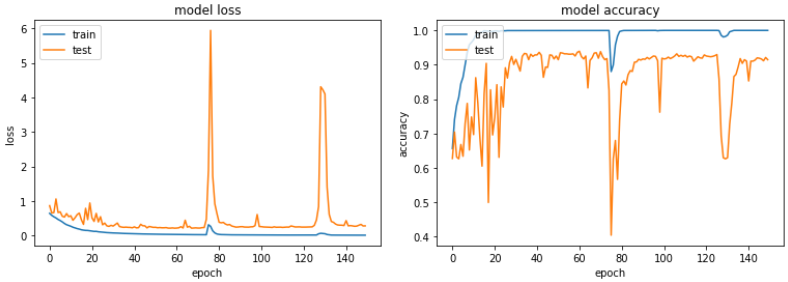


Рисунок 3 – Зависимости ошибки обучения и верности обучаемой модели от номера эпохи

На рисунке 4 представлены зависимости значения классификационных метрик Precision и Recall от номера эпохи для тренировочного и валидационного наборов данных. Несмотря на некоторые колебания значений данных метрик, в начале заметен их рост с дальнейшей

стабилизацией средних значений. Также для медицинский задач, таких как выявление рака легких, более важно снижение доли ложно-негативных срабатываний модели, что соответствует увеличению значений Recall. На лучшей 61-й эпохе значение Recall равно 96.42%. На тестовых данных значение Recall для результирующей модели – 95.31%, а Precision – 94.02%.

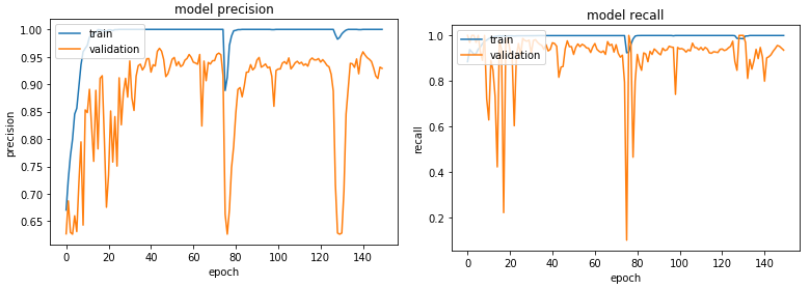


Рисунок 4 – Зависимости метрик Precision и Recall от номера эпохи

Заключение

В итоге была обучена модель нейронной сети, получающая на вход изображение в оттенках серого разрешением 64×64 и выдающая вероятность того, что на этом изображении есть злокачественное легочное образование.

С помощью данной нейронной сети возможно получение двумерных карт вероятностей злокачественных узелковых образований в определенных местах томографии для отдельных срезов сканирования. В дальнейшем планируется получить трехмерные карты вероятности для каждого сканирования из набора Data Science Bowl 2017 [Data Science Bowl 2017, 2017] в формате DICOM [PS3.1: DICOM PS3.1 2019a, 2019] и с помощью алгоритма градиентного бустинга классифицировать их в целом.

Исследование выполнено при финансовой поддержке РФФИ в рамках проекта 18-07-01446.

Список литературы

[Cesar et. al., 2016] Cesar J., Bobadilla M., Pedrini H. Lung Nodule Classification. Based on Deep Convolutional Neural Networks. // Institute of Computing - University of Campinas Campinas – 2016.

[Data Science Bowl 2017, 2017] Data Science Bowl 2017 [Электронный ресурс] // Официальный сайт kaggle: [сайт].URL:

https://www.kaggle.com/c/data-science-bowl-2017/data (дата обращения: 26.02.2019).
[Gomathi et. al., 2010] Gomathi M., Thangaraj P. Lung Nodule Detection using a Neural Classifier // IACSIT International Journal of Engineering and Technology. – 2010. – № 3. – С. 1.
[Lung Nodule Malignancy, 2017] Lung Nodule Malignancy [Электронный ресурс]. // Официальный сайт kaggle: [сайт].URL: https://www.kaggle.com/kmader/lungnodemalignancy (дата обращения: 04.03.2019).
[PS3.1: DICOM PS3.1 2019a, 2019] PS3.1: DICOM PS3.1 2019a [Электронный ресурс] // Официальный сайт DICOM: [сайт].URL: http://dicom.nema.org/medical/dicom/current/output/pdf/part01.pdf (дата обращения: 26.04.2019).
[Yang et. al., 2016] Yang H., Yu H., Wang G. Deep Learning for the Classification of Lung Nodules. // Department of Mathematics, The Ohio State University. – 2016. – С. 1.

CLASSIFICATION OF COMPUTED TOMOGRAPHY IMAGES FOR LUNG CANCER DETECTION

Polezhaev P.N.(polezhaev.ds@gmail.com)
Usanova A.A.(usanova.anastasiya.99@mail.ru)
Orenburg State University

This research is devoted to one of the approaches to lung cancer detection using computed tomography images. The paper describes the architecture of convolutional neural network for classification of lung nodules. This network was trained on the training set and showed high accuracy on the validation and testing sets.

Keywords: lung cancer prediction, malignant nodules, machine learning, convolutional neural networks

УДК 004.89

ЭКСПЕРТНАЯ СИСТЕМА ПОДДЕРЖКИ РАННЕГО СЕМЕЙНОГО ВОСПИТАНИЯ

Филиппова Л.И. (lilivesna@yandex.ru)
Ульяновский государственный педагогический университет, Ульяновск

Представлены результаты формирования базы знаний экспертной системы, предназначенной для поддержки процесса раннего семейного воспитания. Сформированы требования и параметры методики диагностики для использования в качестве основы разрабатываемой экспертной системы, представлен подход к формированию базы знаний экспертной системы. В качестве основы базы знаний используется прикладная онтология в формате OWL и набор правил на языке SWRL.

Ключевые слова: раннее семейное воспитание, степень развития, дети раннего возраста, экспертная система, база знаний, онтология, логический вывод

Введение

Целью данного исследованию является создание экспертной системы, позволяющей получить вероятное заключение и рекомендации на основе данных, собранных в процессе диагностики степени развития детей раннего возраста, и не требующей от родителей специальных знаний в области психологии, педагогики и информационных технологий.

Ранний возраст является уникальным и наиважнейшим периодом в развитии ребенка. Определение текущего состояния развития ребенка позволяет своевременно выбрать оптимальные варианты поддерживающих мероприятий развития детей раннего возраста.

Основная сложность разработки подобной экспертной системы заключается в необходимости совмещения подходов из следующих научных направлений: информатика, математика, педагогика и психология. Так как поведение и состояние ребенка невозможно формализовать без существенной потери точности полученной модели, разрабатываемая экспертная система будет формировать возможные, приближенные заключения и рекомендации.

1 Используемая методика диагностики степени развития детей раннего возраста

На современном этапе наблюдается разнообразие методов и научных подходов к проблеме диагностики детей раннего возраста.

В ходе анализа были рассмотрены 10 распространенных диагностических методик, направленных на изучение развития детей раннего возраста, – пять отечественных и пять зарубежных методик диагностики:

- диагностика уровня нервно-психического развития ребенка раннего возраста (К.Л. Печора, Г.В. Пантюхина, Л.Г. Голубева) [Печора, 2016];
- диагностика психического развития детей от рождения до 3 лет (Е.О. Смирнова, Л. Н. Галигузова, Т.В. Ермолова, С.Ю. Мещерякова) [Смирнова, 2005];
- ранняя диагностика умственного развития (Е.А. Стребелева) [Стребелева, 1994];
- карты развития детей от 0 до 3 лет (В.К. Загвоздкин, И.Е. Федосова, Е.Ю. Мишняева) [Карты развития, 2016];
- дневник развития ребенка от рождения до трех лет (А.М. Казьмин, Л.В. Казьмина) [Казьмин, 2008];
- мюнхенская функциональная диагностика развития детей [1];
- шкалы психомоторного развития детей от рождения до 42 месяцев (Н. Бэйли) [Бэйли];
- таблицы сенсомоторного и социального развития от 0 до 4 лет (Эрнст Й. Кипхард) [Кипхард, 2016];
- наблюдение за развитием детей от 3 до 48 месяцев (У. Петерман, Ф. Петерман, У. Коглин) [Петерман, 2016];
- система Играй-Обучайся K`s Kids [K`s Kids].

Однако данные методики не предполагают возможности автоматизированной обработки результатов наблюдения. В частности, «Диагностика уровня нервно-психического развития ребенка раннего возраста» не учитывает диапазон нормативов развития, верхнюю границу «нормы», содержит специальную терминологию, а перечень показателей подвергается критике, так как представляет собой излишне формализованные требования к умениям и навыкам [Смирнова, 2005; Хохрякова]. «Карты развития детей от 0 до 3 лет» не имеют четких инструкций по интерпретации результатов наблюдения, а направления развития ребенка привязаны к ФГОС ДО, что делает их менее гибкими.

«Таблицы сенсомоторного и социального развития от 0 до 4 лет» охватывают не все направления развития ребенка и не учитывают диапазон нормативов развития детей раннего возраста так же, как и методика «Наблюдение за развитием детей от 3 до 48 месяцев».

Таким образом, можно выдвинуть следующие требования к методике диагностики степени развития детей раннего возраста для последующего использования в экспертной системе [Yarushkina, 2018; Майданкина, 2013]:

1. Возможность использования для проведения целенаправленного наблюдения за развитием детей в возрасте от 12 до 36 месяцев.
2. Возможность организации наблюдения как в дошкольной образовательной организации, так и в домашних условиях за счет низкого уровня временных и финансовых затрат, использование доступных материалов.
3. Охват ключевых направлений развития ребенка: крупная моторика, мелкая моторика, речевое развитие, когнитивное развитие, социальное развитие и эмоциональное развитие.
4. Понятные задания с точно определенными критериями их выполнения. В каждом из перечисленных направлений предлагается от 3 до 4 заданий для детей второго года жизни и по 5 заданий для детей третьего года жизни. При составлении заданий обращалось внимание на то, смогут ли педагоги в условиях, ограниченных рамками ДОУ, и родители, не имеющие специальной подготовки, пронаблюдать, справляется ли ребенок с предложенным заданием.
5. Четкие инструкции по интерпретации выполнения заданий.
6. Диапазон нормативов развития. В предлагаемой методике указанный возраст является наиболее поздним вариантом нормы овладения навыками, проверяемыми предложенными заданиями.
7. Возможность регулярного наблюдения за развитием ребенка. Промежуток между наблюдениями зависит от возраста ребенка: для детей от 1 до 2 лет он составляет три месяца, после двух лет – полгода.
8. Возможность автоматизированной обработки результатов наблюдения: быстрый подсчет и анализ результатов наблюдений по каждому направлению развития.

Предлагаемая в рамках данного исследования методика диагностики позволяет проводить целенаправленное наблюдение за развитием детей в возрасте от 12 до 36 месяцев как в дошкольной образовательной организации, так и в домашних условиях.

Методика диагностики охватывает ключевые направления развития ребенка: крупная моторика, мелкая моторика, речевое развитие,

когнитивное развитие, социальное развитие и эмоциональное развитие. В каждом из перечисленных направлений предлагается от 3 до 4 заданий для детей второго года жизни и по 5 заданий для детей третьего года жизни. При составлении заданий обращалось внимание на то, смогут ли педагоги, в условиях ограниченных рамками ДОУ, и родители, не имеющие специальной подготовки, пронаблюдать, справляется ли ребенок с предложенным заданием.

Формально текущий продиагностированный уровень развития (УР) ребенка можно представить в виде следующего выражения:

$$D = \langle Speech, Sensor, Game, Move, Independence, Invention, Social \rangle,$$

где $Speech$ – значение продиагностированного показателя речевого развития;

$Sensor$ – значение продиагностированного показателя сенсорного развития;

$Game$ – значение продиагностированного показателя поведения ребенка в игре и действиях с предметами;

$Move$ – значение продиагностированного показателя двигательных умений;

$Independence$ – значение продиагностированного показателя навыков самостоятельности;

$Invention$ – значение продиагностированного показателя навыков в конструктивной и изобретательной деятельности;

$Social$ – значение продиагностированного показателя социального развития.

Для определения перечня развивающих занятий, направленных на корректировку показателей УР, используется следующая функция:

$$F: \{Age, \widetilde{D}\} \to \hat{L}^{Age},$$

где Age – возраст ребенка, диагностика которого проводится;

$\widetilde{D} \in D$ – множество показателей текущего продиагностированного УР, значение которых меньше нормы $\widetilde{D}_i \in \widetilde{D}, \widetilde{D}_i < N_i$;

$\hat{L}^{Age} \in L^{Age}$ – множество развивающих занятий, подобранных на основании степени отклонения УР от нормы для возраста Age.

2 База знаний экспертной системы

База знаний экспертной системы формировалась в редакторе онтологий Protégé 5.5 и представляет собой онтологию в формате OWL с набором правил логического вывода на языке SWRL.

Рассмотри сущности онтологии:

1. Классы онтологии описывают основные понятия предметной области:

   Диагностика

   МоторикаОбратитьВнимание

   МоторикаПлохо

   РечьОбратитьВнимание

   РечьПлохо и т.д.

   Ребенок

   Возраст01

   Возраст12

   Возраст23

   Возраст3

   Занятия

   Моторика

   Моторика01

   Моторика01ОбратитьВнимание

   Моторика01Плохо

   Моторика12

   Моторика12ОбратитьВнимание

   Моторика12Плохо и т.д.

   Речь

   Речь01

   Речь01ОбратитьВнимание

   Речь01Плохо и т.д.

   Класс *Диагностика* описывает различные характеристики показателей УР после проведения диагностики УР, например, обратить внимание на развитие речи (*РечьОбратитьВнимание*) или плохой уровень развития моторики (*МоторикаПлохо*).

   Класс *Ребенок* описывает детей разных возрастов, например, дети до одного года (*Возраст01*).

   Класс *Занятия* описывает множество занятий для некоторого возраста ребенка с определенным значением показателя УР, например, занятия для детей до одного года с плохим уровнем развития моторики (*Моторика01Плохо*).

   Классы онтологии *Диагностика* и *Ребенок* определены как непересекающиеся, также непересекающимися классами являются классы, описывающие перечень занятий для развития показателей УР: *Речь*, *Моторика* и т. д.

2. Объектные свойства онтологии позволяют описывать связи между объектами онтологии, например, диагностикой и ребенком (*Диагностика1* ***диагностикаРебенка*** *ИвановИИ*).

3. Свойства данных онтологии позволяют описывать связи между объектами онтологии и конкретными значениями, например, определять дату диагностики (*Диагностика1* ***датаПроведения*** *2019-04-01*), значение показателя двигательных умений (*Диагностика1* ***значениеПоказателяМоторика*** *5*), возраст ребенка (*ИвановИИ* ***имеетВозраст*** *2*).
4. Объекты онтологии представляют конкретные объекты предметной области, например, диагностика (*Диагностика1*) или ребенок (*ИвановИИ*).

Рассмотрим правила логического вывода на языке SWRL:

1. Правила для отнесения ребенка к определенному классу возраста, например,
   имеетВозраст(?child, ?age) ^ swrlb:greaterThanOrEqual(?age, 1)
   ^ swrlb:lessThan(?age, 2) -> Возраст12(?child)
2. Правила для отнесения диагностики к определенному классу показателей УР ребенка, например,
   Диагностика(?diagnostic)
   ^ значениеПоказателяМоторика(?diagnostic, ?value)
   ^ swrlb:equal(?value, 4) ->
   МоторикаОбратитьВнимание(?diagnostic)
3. Правила для определения перечня развивающих занятий для ребенка в зависимости от класса возраста и класса показателей УР, например,
   МоторикаОбратитьВнимание(?diagnostic)
   ^ диагностикаРебенка(?diagnostic, ?child)
   ^ Возраст01(?child) ->
   Моторика01ОбратитьВнимание(?child)

Таким образом, представленная онтология и набор SWRL правил позволяют определить множество необходимых занятий, учитывая возраст ребенка и его показатели УР.

Заключение

Применение онтологического подхода для построения базы знаний экспертной системы позволяет обеспечить педагогов и родителей универсальным инструментарием для диагностики уровня развития ребенка. Использование разрабатываемой экспертной системы способствует ускорению процесса диагностики уровня развития ребенка, повышению компетентности родителей в вопросах воспитания и развития детей раннего возраста. При этом разрабатываемая экспертная система не требует от пользователей навыков в области инженерии знаний и онтологического анализа.

Список литературы

[Казьмин, 2008] Казьмин, А. Дневник развития ребенка от рождения до трех лет [Текст]/ А. Казьмин, Л. Казьмина. – М.: Когито-Центр, 2008. – 80 с.

[Карты развития, 2016] Карты развития детей от 0 до 3 лет. – М.: Национальное образование, 2016. – 128 с.

[Кипхард, 2016] Кипхард, Эрнст Й. Как развивается ваш ребенок? Таблицы сенсомоторного и социального развития: От рождения до 4-х лет [Текст] / Эрнст Й. Кипхард. – Изд. 4-е. – М.: Теревинф, 2016. – 112 с.

[Петерман, 2016] Петерман, У. Наблюдение за развитием детей от 3 до 48 месяцев и протоколирование результатов: учебно-практическое пособие для педагогов дошкольного образования [Текст] / У. Петерман, Ф. Петерман, У. Коглин; под ред. С.Н. Бондаревой. – М.: Национальное образование, 2016. – 132 с.

[Печора, 2016] Печора, К.Л. Диагностика развития детей раннего возраста. Развивающие игры и занятия [Текст] / К.Л. Печора, Г.В. Пантюхина. – М.: ТЦ Сфера, 2016. – 80 с.

[K's Kids] Система Играй-Обучайся K`s Kids. [Электронный ресурс]. URL: http://www.kskids.com/ru/advice (дата обращения: 10.04.2019).

[Смирнова, 2005] Смирнова, Е.О. Диагностика психического развития детей от рождения до 3 лет: Методическое пособие для практических психологов [Текст] / Е.О. Смирнова, Л. Н. Галигузова, Т.В. Ермолова, С.Ю. Мещерякова. – 2-е изд. испр. и доп. – СПб.: ДЕТСТВО-ПРЕСС, 2005. – 144 с.

[Стребелева, 1994] Стребелева, Е.А. Методические рекомендации к психолого-педагогическому изучению детей (2-3 лет): Ранняя диагностика умственного развития [Текст] / Е.А. Стребелева. – М.: Компания «Петит», 1994. – 32 с.

[Хохрякова] Хохрякова, Ю. М. Проблема диагностики психического развития в педагогике раннего возраста // Сибирский педагогический журнал. – 2010. – №9. [Электронный ресурс]. URL: http://cyberleninka.ru/article/n/problema-diagnostiki-psihicheskogo-razvitiya-v-pedagogike-rannego-vozrasta (дата обращения: 25.04.2019).

[Бэйли] Шкалы психомоторного развития детей от рождения до 42 месяцев Нэнси Бэйли. [Электронный ресурс]. URL: https://www.mniip.org/science/scales/scales_children.php (дата обращения: 03.05.2019).

[Yarushkina, 2018] N.G.Yarushkina, A.A.Filippov, V.S.Moshkin, L.I.Filippova Application of the fuzzy knowledge base in the construction

of expert systems // Information Technology in Industry, Vol. 6, #2, 2018, pp. 32-37.

[Майданкина, 2013] Н.Ю.Майданкина, Л.И.Зиязетдинова, Л.Л. Лебедь. Особенности обеспечения педагогической поддержки раннего семейного воспитания в условиях ДОУ общеразвивающего вида // Современный детский сад. – 2013. – №2. – С. 64-66.

THE EXPERT SYSTEM FOR EARLY FAMILY EDUCATION SUPPORT

Filippova L.I.(lilivesna@yandex.ru)
Ulyanovsk State Pedagogical University, Ulyanovsk

The process of development of the knowledge base of the expert system for early family education support are presented in this article. Requirements and parameters of the technique of diagnostics for use as a basis of the developed expert system are created. The review of analogs of the developed expert system is submitted. The knowledge base contains OWL-ontology and SWRL-rules.

Keywords: early childhood family education, child development degree, early childhood, expert system, knowledge base, ontology, inference

УДК 81.322.2

РАЗРАБОТКА СИСТЕМЫ РЕФЕРИРОВАНИЯ СООБЩЕНИЙ ЭЛЕКТРОННЫХ СМИ

Шигабутдинов И. М. (isl23@mail.ru)
Ульяновский государственный технический университет, Ульяновск

Описываются алгоритмы экстракционного реферирования, а также проводится оценка их эффективности по группе метрик ROUGE.

Ключевые слова: автоматическое реферирование, статистический алгоритм, TextRank, KMeans, LSA, ROUGE.

Введение

Человеку в современном мире необходимо работать с огромными объемами информации, которые постоянно растут вследствие информатизации общества. В условиях информационной перегрузки на поиски и обработку необходимой информации у человека уходит все больше времени. Сегодня автоматическая обработка текстов на естественных языках представляет собой крайне важную задачу.

Разработка системы автоматического реферирования позволит сократить время на поиск и обработку информации за счет уменьшения объема обрабатываемой информации.

Целью данной работы является исследование и сравнение эффективности алгоритмов автоматического реферирования.

1 Предметная область

Реферирование – это процесс извлечения из текста основного содержания с целью формирования общего представления об исходном тексте в кратком изложении. Важно отметить, что реферирование не ставит перед собой цель заменить первоисточник, а формирует общее представление о нем, что позволяет пользователю оценить целесообразность обращения к нему.

Алгоритмы автоматического реферирования классифицируются по типу получаемого реферата на экстракцию и абстракцию.

Экстракция представляет собой процесс извлечения из текста ключевых блоков информации. К характерным чертам алгоритмов данного класса можно отнести наличие оценочной функции, которая определяет степень значимости информационных блоков.

В процессе абстракции производится построение нового текста, содержательно обобщающего текст первоисточника. В данном подходе можно выделить три этапа: анализ исходного текста с генерацией внутреннего представления, семантическое сжатие внутреннего представления и синтез реферата. Системы данного класса более сложны в реализации в силу необходимости учета семантики в процессе реферирования, однако генерируемые тексты более естественны и предложения более связанны.

В силу относительной простоты реализации, а также имеющихся временных ресурсов, в разработанной системе автоматического реферирования были реализованы экстракционные алгоритмы.

2 Реализация

Для исследования были выбраны следующие экстракционные алгоритмы автоматического реферирования:

- Статистический алгоритм;
- TextRank;
- KMeans;
- LSA.

В статистическом алгоритме предполагается, что слова, наиболее часто встречающиеся в тексте, имеют наибольшее значение. Производится частотный анализ слов, на основе которого определяется значимость предложений. В реферат попадают предложения с наибольшей значимостью.

TextRank реализует графовый алгоритм. Строится полный, взвешенный, неориентированный граф, вершинами которого являются предложения. Вес вершины отражает значимость предложения, а вес ребер графа отражает степень схожести двух предложений. Вес вершин определяется по формуле

$$W(V_i) = (1-d) + d * \sum_{v_j \in Inc(V_i)} \frac{w_{ij}}{\sum_{V_k \in Inc(V_j)} w_{jk}} * W(V_j),$$

где $W(V_i)$ – вес *i-й* вершины *V*, w_{ij} – вес ребра между *i-й* и *j-й* вершинами, а *d* – коэффициент затухания.

Вычисление веса вершин производится итеративно, алгоритм останавливается при достижении необходимого уровня точности, т. е.

когда новое значение веса вершины незначительно отличается от своего предыдущего значения.

KMeans – использует алгоритм кластеризации k-средних. Предложения представляются в виде вектора длины n, где n – количество уникальных слов в тексте, *i-й* элемент вектора равен значению TF-IDF *i-го* уникального слова текста. Показатель TF-IDF зависит от частоты слова в предложении и в тексте. Наибольшее значение показателя TF-IDF получают слова, наиболее часто встречающиеся в предложении и при этом наиболее редко встречающиеся в тексте.

В результате кластеризации предложения разбиваются на центроиды. Предложения, находящиеся наиболее близко к центрам, попадают в реферат.

LSA – данный алгоритм можно сравнить с простым видом трехслойной нейросети. Заполняется матрица A размерности $n \times m$, где n – количество уникальных слов в тексте, а m – количество предложений. Элемент матрицы a_{ij} равен частоте *i-го* уникального слова текста в *j-м* предложении.

К матрице A применяется сингулярное разложение: $A = USV^T$, где U – ортонормированная матрица размера $n \times m$, элементы которой u_{ij}, S – диагональная матрица элементы которой σ_{jj}, V^T – ортонормированная транспонированная матрица, элементы которой v_{ij}.

Вес предложений s_k определяется по формуле

$$s_k = \sqrt{\sum_{i=1}^{m} {v_{ik}}^2 \cdot {\sigma_i}^2}.$$

В реферат попадают предложения с наибольшим весом.

3 Исследование

Для оценки эффективности алгоритмов реферирования была выбрана группа метрик ROUGE, а именно метрики ROUGE-N1, ROUGE-N2, ROUGE-L, ROUGE-S.

ROUGE используется для оценки качества автоматического реферата. Оценки, полученные с использованием данной метрики, тесно коррелируют с ручными оценками людей. Однако для применения данной метрики необходимо иметь ручные рефераты текстов.

Фактически данная группа метрик определяет степень схожести двух текстов в данном случае автоматически сгенерированного и эталонного реферата. Данная группа метрик принимает вещественные значения в диапазоне [0;1].

Метрика ROUGE-N заключается в расчете общих *n*-грамм эталонного и сгенерированного рефератов и нормировании данной величины на суммарное количество *n*-грамм двух текстов. В данной работе были использованы метрики ROUGE-1 и ROUGE-2, которые в своих расчетах используют униграммы и биграммы соответственно.

ROUGE-L – вычисление данной метрики опирается на длину наибольшей общей подпоследовательности слов и рассчитывается по формулам:

$$R_{lcs} = \frac{LCS(X,Y)}{m},$$

$$P_{lcs} = \frac{LCS(X,Y)}{n},$$

$$ROUGE - L = \frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}},$$

где *X* – последовательность слов эталонного реферата длины *n*, *Y* – последовательность слов автореферата длины *m*, *LCS(X,Y)* длина наибольшей общей подпоследовательности *X* и *Y*. Коэффициент β в расчетах принимался за единицу.

ROUGE-S рассчитывается по формулам:

$$R_{skip2} = \frac{SKIP2(X,Y)}{C_m^2},$$

$$P_{skip2} = \frac{SKIP2(X,Y)}{C_n^2},$$

$$ROUGE - S = \frac{(1+\beta^2)R_{skip2}P_{skip2}}{R_{skip2} + \beta^2 P_{skip2}},$$

где *SKIP2(X,Y)* – количество биграмм с пропусками, являющихся общими в текстах эталонного (*X* длины *n*) и автоматического (*Y* длины *m*) рефератов.

В качестве тестовых данных выступили политические новости с портала «газета.ru». Данный ресурс имеет вступительную часть, в которой раскрывается содержание новости. Всего было собрано 988 новостей с их «рефератами», из них было отобрано 811 новостей для исследования.

Средний объем образцового реферата составляет ~2,9 предложений, а средний объем новости равен ~30,2. Таким образом, в среднем производилось сжатие текста примерно в ~10,3 раза

На рисунках 1-4, а также в таблицах 1-4 представлены результаты оценки алгоритмов по группе метрик ROUGE.

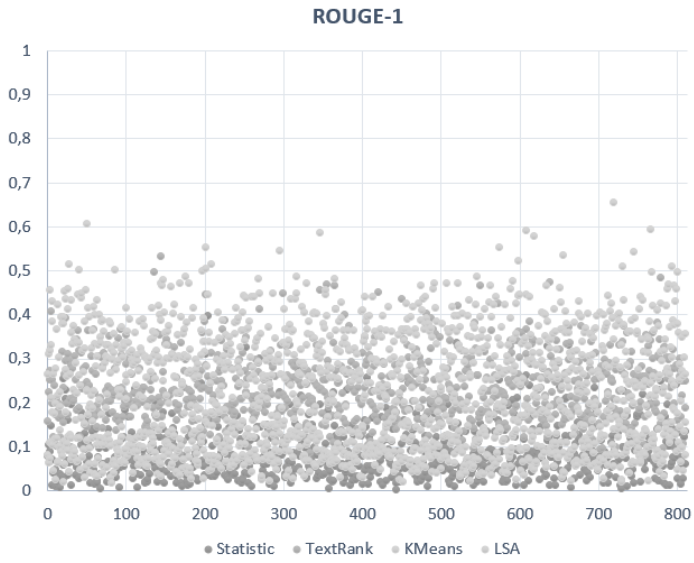


Рисунок 1 – метрика ROUGE-1

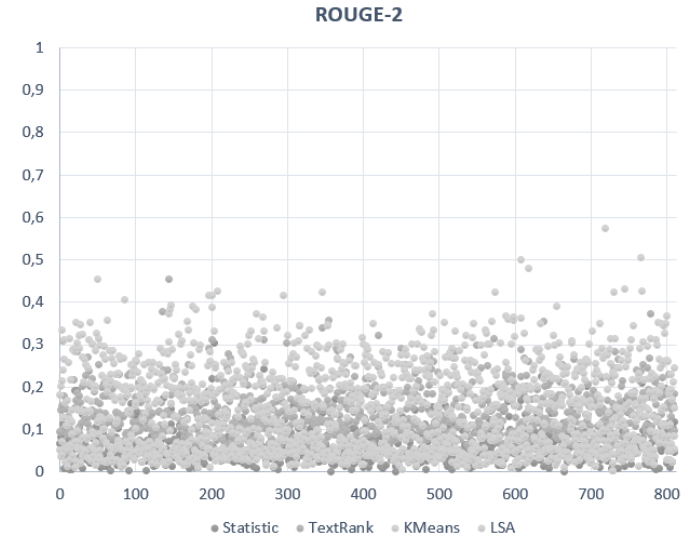


Рисунок 2 – метрика ROUGE-2

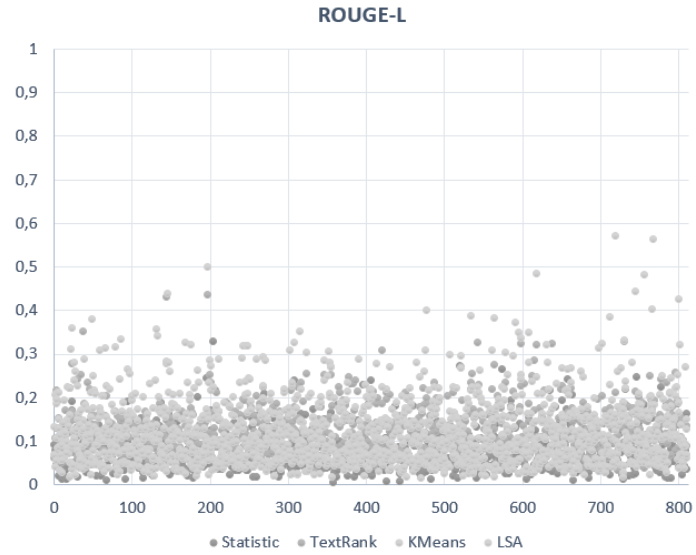


Рисунок 3 – метрика ROUGE-L

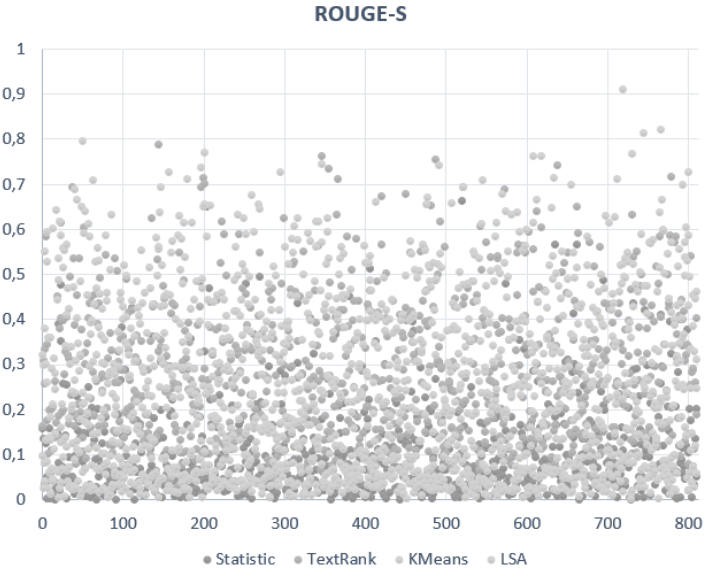


Рисунок 4 – метрика ROUGE-S

Таблица 1 – ROUGE-1

	Statistic	TextRank	Kmeans	LSA
Min	0,004	0,075	0,018	0,08
Max	0,4	0,533	0,316	0,655
Avg	0,087	0,233	0,107	0,324
Median	0,074	0,224	0,098	0,32

Таблица 2 – ROUGE-2

	Statistic	TextRank	Kmeans	LSA
Min	0	0,028	0,005	0,047
Max	0,305	0,454	0,207	0,575
Avg	0,068	0,142	0,059	0,22
Median	0,059	0,131	0,052	0,212

Таблица 3 – ROUGE-L

	Statistic	TextRank	Kmeans	LSA
Min	0,005	0,028	0,017	0,037
Max	0,33	0,436	0,32	0,571
Avg	0,078	0,124	0,079	0,156
Median	0,068	0,115	0,068	0,142

Таблица 4 – ROUGE-S

	Statistic	TextRank	Kmeans	LSA
Min	0	0,032	0,003	0,034
Max	0,664	0,787	0,651	0,91
Avg	0,107	0,288	0,099	0,372
Median	0,074	0,266	0,067	0,361

На рисунке 5 и в таблице 5 представлен итог оценки алгоритмов реферирования по группе метрик ROUGE. Исходя из полученных данных, в результате исследовательской работы лучшим алгоритмом был признан LSA со средним значением, равным ~0,268. Следующим по эффективности оказался алгоритм TextRank со средним значением метрик, равным ~0,196. Алгоритм KMeans имеет среднее значение метрик, равное ~0,086. Статистический алгоритм показал худший результат, он равен ~0,085.

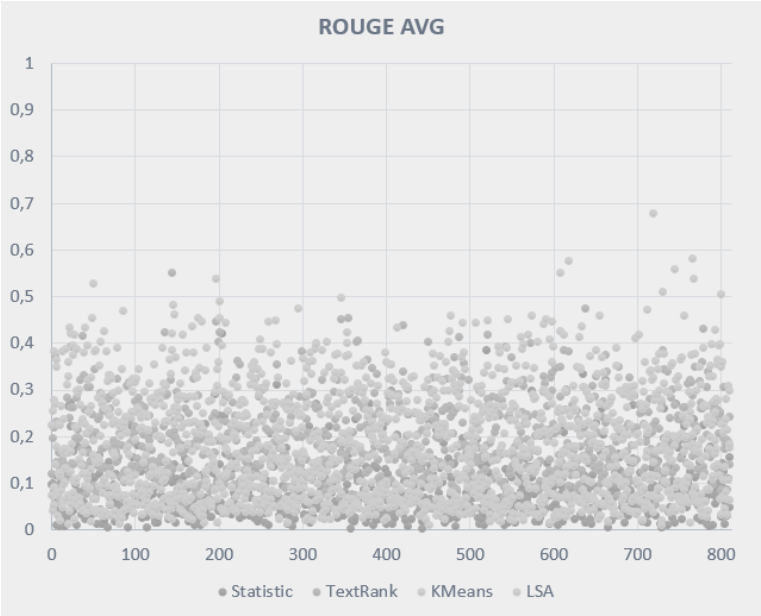


Рисунок 5 – ROUGE среднее

Таблица 5 – ROUGE среднее

	Statistic	TextRank	Kmeans	LSA
Min	0,002	0,043	0,011	0,054
Max	0,421	0,552	0,36	0,678
Avg	0,085	0,196	0,086	0,268
Median	0,069	0,185	0,071	0,26

Чтобы действительно убедиться в эффективности алгоритмов реферирования, а также в корректности группы метрик ROUGE, был проведен эксперимент с алгоритмом *n* случайных предложений из текста, где *n* – желаемое число предложений реферата.

Результат алгоритма N случайных предложений представлен в таблице 6. Среднее значение группы метрик ROUGE для него составляет ~0,127 данное значение более чем в 2 раза меньше значения алгоритма LSA, однако превосходит значение статистического алгоритма и KMeans.

Таблица 6 – N случайных предложений

	Rouge-1	Rouge-2	Rouge-L	Rouge-S	Rouge AVG
Min	0,009	0,005	0,005	0,002	0,007
Max	0,542	0,391	0,333	0,791	0,486
Avg	0,157	0,092	0,091	0,17	0,127
Median	0,148	0,083	0,08	0,14	0,115

Исходя из результата данного эксперимента, можно сделать вывод о том, что алгоритмы реферирования LSA и TextRank действительно имеют эффективность по сравнению с алгоритмом случайных предложений. Низкие значения алгоритма KMeans обусловлены сильно ограниченным количеством кластеров. При составлении рефератов бóльших объемов алгоритм KMeans имеет тенденцию улучшения качества получаемых рефератов.

Заключение

В результате проделанной работы были реализованы экстракционные алгоритмы автоматического реферирования. Была проведена исследовательская работа по оценке эффективности выбранных алгоритмов по группе метрик ROUGE.

Список литературы

[Inderjeet Mani, Udo Han, 2000] The Challenges of Automatic Summarization IEEE Computer, Ноябрь 2000, – С. 29-36.

[Тарасов С. Д., 2010] Современные методы автоматического реферирования // Научно-технические ведомости СПбГПУ. Информатика. Телекоммуникации. Управление. – 2010. – №6 (113). URL: https://cyberleninka.ru/article/n/sovremennye-metody-avtomaticheskogo-referirovaniya (дата обращения: 21.05.2018).

[Попова С. В., Ходырев И. А., 2013] Извлечение и ранжирование ключевых фраз в задаче аннотирования // Научно-технический вестник информационных технологий, механики и оптики. – 2013. – №1 (83). URL: https://cyberleninka.ru/article/n/izvlechenie-i-ranzhirovanie-klyuchevyh-fraz-v-zadache-annotirovaniya (дата обращения: 11.09.2018).

[Хомоненко А.Д., Краснов С.А., 2012] Применение метода латентно-семантического анализа для автоматической рубрикации документов // Известия Петербургского университета путей сообщения. 2012. №2 (31). URL: https://cyberleninka.ru/article/n/primenenie-metoda-latentno-

semanticheskogo-analiza-dlya-avtomaticheskoy-rubrikatsii-dokumentov (дата обращения: 22.10.2018).
[Lin, Chin-Yew 2004] ROUGE: a Package for Automatic Evaluation of Summaries.

DEVELOPMENT OF SYSTEM AUTOMATIC SUMMARIZATION MEDIA MESSAGES

Shigabutdinov I. M. (isl23@mail.ru)
Ulyanovsk State Technical University, Ulyanovsk

The paper describes the algorithms of extraction summarization and evaluating their effectiveness by the group of ROUGE metrics.

Keywords: automatic summarization, Statistic algorithm, TextRank, KMeans, LSA, ROUGE.

АВТОРСКИЙ УКАЗАТЕЛЬ

Научное издание
Прикладные информационные системы (ПИС-2019)
Сборник научных трудов
VI Всероссийской научно-практической конференции
с международным участием
(Россия, г. Ульяновск 27 мая – 09 июня, 2019 г.)

Ответственный за выпуск Е.Н. Эгов

ЛР №020640 от 22.10.97.

Подписано в печать 27.12.2019. Формат 60×84/16.
Усл. печ. л 6,74. Тираж 100 экз. Заказ № 36.

Ульяновский государственный технический университет
432027, г. Ульяновск, ул. Северный Венец, д. 32.
ИПК «Венец» УлГТУ, 432027, г. Ульяновск, ул. Северный Венец, д. 32.