УДК 681.3.064 DOI: 10.34759/trd-2020-115-16

Исследование влияния искажения алфавита поиска на идентификацию сущностей на основе частотного анализа данных

Малышаков Г.В.

Московский авиационный институт (национальный исследовательский университет), Волоколамское шоссе, 4, Москва, А-80, ГСП-3, 125993, Россия

e-mail: malshakov@mail.ru

Статья поступила 13.11.2020

Аннотация

В статье исследовано качество частотной идентификации сущностей предметной области при различных стратегиях исключения лексем из алфавита поиска. Показаны зависимости автокорреляции данных сущности предметной области от используемой при частотной идентификации длине лексем при различных объёмах исходных данных, степени вхождения лексем алфавита в данные объектов сущности предметной области от используемой длины лексем, распознавания сущностей предметной области от степени вхождения лексем алфавита в её данные, частотной идентификации сущностей предметной области при благоприятной и неблагоприятной стратегиях исключения её объектов из анализируемых данных при различных длинах лексем алфавита.

На основе выполненных исследований идентификации при благоприятной и неблагоприятной стратегиях исключения её данных по объектам из анализируемых данных установлено влияние степени вхождения лексем алфавита в анализируемые данные на качество частотной идентификации сущностей предметной области.

Ключевые слова: ошибка, частотная идентификация, сущность, прикладное программное обеспечение, корреляция.

При проектировании, производстве и эксплуатации сложных наукоёмких изделий авиационной техники основополагающим является создание единого информационного пространства для всех участников их жизненного цикла [1 - 5]. При этом от способности обмениваться данными (интероперабельности) используемых информационных компонент информационного пространства зависит эффективность взаимодействия участников жизненного цикла [6 - 8].

Чтобы обеспечить обмен данными между прикладными программами необходимо определить местоположение требуемых данных в их базах данных. Для этого автором предложен метод частотной идентификации сущности предметной области[9], в котором принадлежность анализируемых данных к данным сущности выполняется на основе вычисления коэффициента корреляции Пирсона [10, 11], между частотами встреч лексем алфавита построенного на основе данных сущности и частотами их встреч в анализируемых данных.

$$r_{XA} = \frac{\sum_{i=1}^{n}(X_i - \overline{X}) * (A_i - \overline{A})}{\sqrt{\sum_{j=1}^{n}(X_j - \overline{X})^2 * \sum_{j=1}^{n}(A_j - \overline{A})^2}}$$

где $A(i) = \{A_1, A_2, \ldots A_n\}$ - частоты встреч лексем алфавита построенного на основе данных сущности; $X(j) = \{X_1, X_2, \ldots X_n\}$ - частоты встречи в анализируемых данных лексем алфавита.

Если вычисленное значение коэффициента корреляции Пирсона находится в диапазоне от **0,7** до **1** то между анализируемыми данными и данными сущности

имеется высокая степень взаимосвязи [12 - 15]. В этом случае принимается решение о том, что анализируемые данные принадлежат данным поисковой сущности.

В исследованиях использовались данные радиокомпонент диодов [16] и транзисторов [17 - 19] располагающихся в отдельных таблицах реляционной базы данных [20, 21].

Для ускорения идентификации (уменьшения вычислительных затрат) из поискового алфавита сущности исключается часть лексем, из-за чего алфавит перестаёт полностью покрывать объекты сущности предметной области, что приводит к увеличению количества ошибок пропуска обнаружения (рис. 1). Это возникает из-за того что в силу исключения лексем у части данных поисковой сущности отсутствуют лексемы идентификации в алфавите и они перестают идентифицироваться. Степень вхождения лексем в данные сущности после введённых ограничений для каждой сущности индивидуальна.

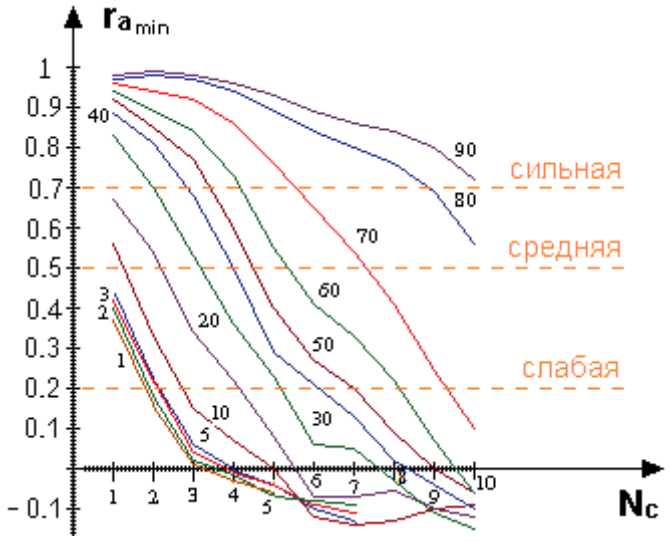

Рисунок 1. Зависимости автокорреляции данных сущности предметной области от используемой при идентификации длины лексем

При увеличении ограничения на длину лексем ($\mathbf{N_c}$) используемых при идентификации значение допустимой автокорреляции ($\mathbf{ra_{min}}$) по которой принимается решение о том что данные принадлежат поисковой сущности снижается, что приводит к увеличению количества ложных срабатываний.

В таблице 1 и рис. 1 приведены результаты исследования изменения вхождения в данные сущности лексем алфавита в зависимости от используемой длины (**Nc**).

Таблица 1. Степень вхождения лексем алфавита сущности предметной области от используемой длины

		Nc									
	Количество записей	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
лучшее (для сущности с №12)	12256	100	100	100	100	100	100	100	100	100	100
усреднённое значение для всех рассматриваемых сущностей	-----	100	99	93	90	87	77	71	66	62	59
худшее (для сущности с №119)	10	100	90	40	20	20	0	0	0	0	0

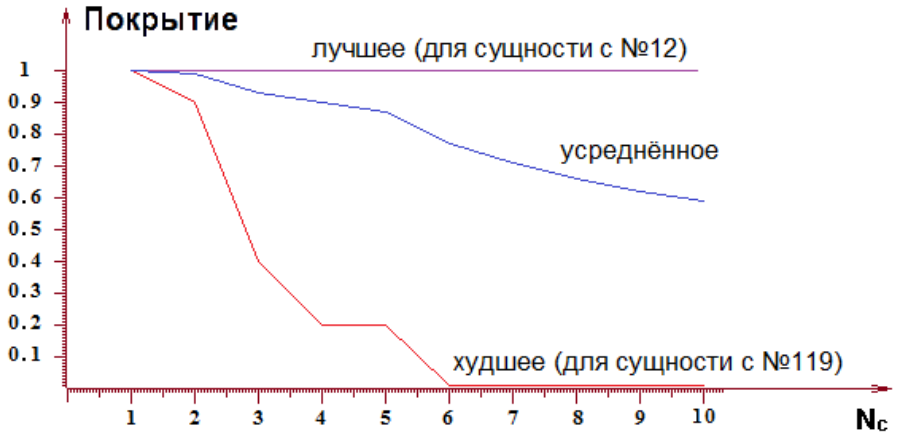

Рисунок 2. Зависимости степени вхождения лексем алфавита в объекты сущности предметной области от используемой длины

При уменьшении степени вхождения лексем алфавита в анализируемые данные сущности увеличивается количество ошибок пропуска обнаружения сущности

(таблица 2, рис. 3). Полученные величины были нормированы относительно общего количества вычислений для рассматриваемого диапазона покрытия.

Таблица 2. Распознавание сущностей предметной области в зависимости от степени вхождения лексем алфавита в её данные

Диапазон покрытия	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
нормированное количество ошибок пропуска обнаружения	0	0	0	0,08	0,10	0,10	0,13	0,12	0,25	0,25
нормированное количество правильно не обнаруженных сущностей	0,57	0,517	0,97	0,62	2,13	1,47	2,97	3,13	13,2	25,9
нормированное количество правильно обнаруженных сущностей	0	0,034	0	0,01	0,06	0,12	0,12	0,14	0,21	0,20

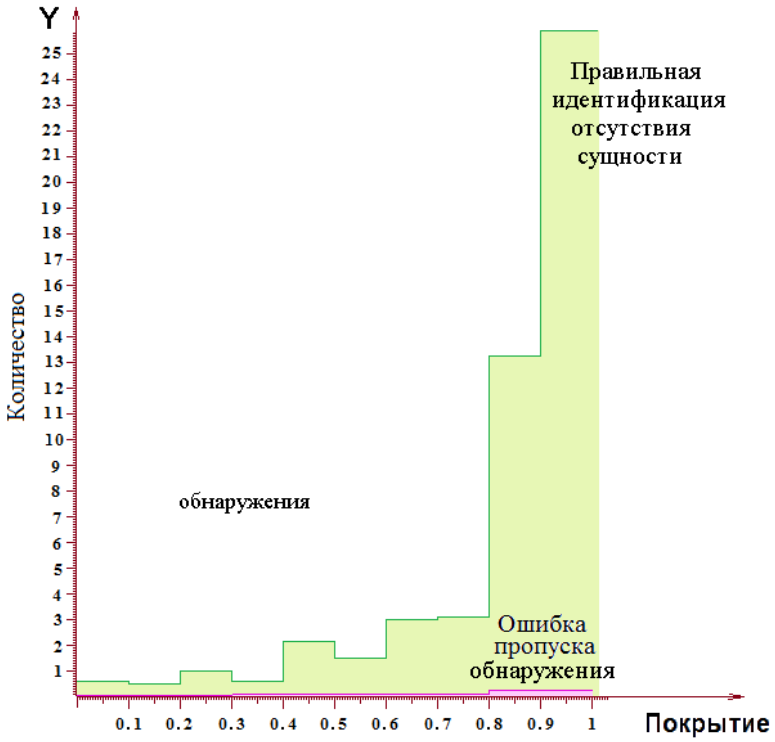

Рисунок 3. Распознавание сущностей предметной области в зависимости от степени вхождения лексем алфавита в её данные

Видно, что скорость нарастания ошибок пропуска обнаружения сущностей предметной области при максимальном покрытии значительно меньше, чем правильное не-обнаружение сущности предметной области, поэтому покрытие необходимо брать по возможности максимальным (из последнего диапазона).

Ограничение на количество символов в лексемах **Nc** можно увеличивать, пока покрытие алфавита частотной идентификации сущности предметной области будет не меньше 90%. В этом случае правильное не обнаружение сущности предметной области будет максимальным.

В зависимости от вхождения лексем алфавита в данные объектов сущности предметной области удалять объекты можно в порядке минимального вхождения в него лексем алфавита (стратегия ***благоприятного исключения***) либо в порядке максимального вхождения в него лексем алфавита (стратегия ***неблагоприятного исключения***).

В первом случае в силу того, что удаляемые объекты не входят (либо минимально входят) в лексемы алфавита частотной идентификации сущности предметной области, их исключение минимально влияет на расчёт коэффициента корреляции Пирсона. Во втором случае влияние искажения данных является максимальным.

Полученные экспериментальным путём данные по количеству правильно идентифицируемых сущностей предметной области при неблагоприятной стратегии

исключения её объектов из анализируемых данных для различных длин лексем алфавита сведены в таблицу 3 (рис. 4).

Таблица 3. Количество правильно идентифицируемых сущностей предметной области при неблагоприятной стратегии исключения её объектов из анализируемых данных для различных длин лексем алфавита

Величина выборки (количество используемых объектов от эталона)	Nc									
	1	2	3	4	5	6	7	8	9	10
0% - 20%	1	0	0	0	0	0	0	0	0	0
21% - 45%	88	47	32	23	24	21	18	21	22	24
46% - 85%	75	108	109	102	103	86	69	75	73	60
86% - 100%	11	19	34	48	48	68	87	79	80	90

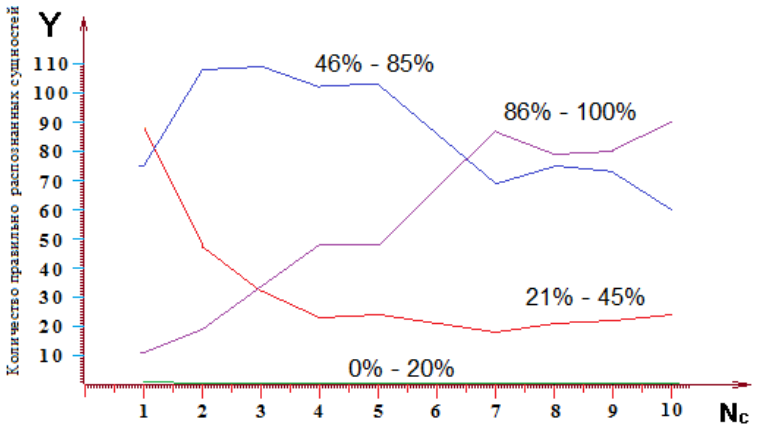

Рисунок 4. Зависимости количества правильно идентифицируемых сущностей предметной области при неблагоприятной стратегии исключения её объектов из анализируемых данных для различных длин лексем алфавита

С учётом полученных экспериментальных данных исключения объектов из анализируемых данных для различных длин лексем по неблагоприятной стратегии сделаны следующие выводы:

- чтобы выполнялось правильное распознавание, величина выборки должна быть больше **20%**;

- правильное распознавание сущностей предметной области легче обеспечить при большей выборке (**V**) благодаря возможности увеличения длины используемых лексем (**Nc**) в алфавите;

- при выборке от 21% до 45% от эталонных данных происходит уменьшение количества правильно идентифицируемых сущностей из-за уменьшения степени вхождения лексем алфавита в анализируемые данные в силу увеличения длины (**Nc**) используемых лексем;

- при выборке от 86% до 100% для хорошей частотной идентификации требуется большой объём анализируемых данных в силу того что увеличение длины используемых лексем в алфавите приводит к уменьшению степени вхождения лексем алфавита в анализируемые данные.

Данные по количеству правильно идентифицируемых сущностей предметной области при благоприятной стратегии исключения её объектов из анализируемых данных для различных длин лексем алфавита сведены в таблицу 4 (рис. 5).

Таблица 4. Количество правильно идентифицируемых сущностей предметной области при благоприятной стратегии исключения её объектов из анализируемых данных для различных длин лексем алфавита

Величина выборки (количество используемых объектов от эталона)	**Nc**									
	1	2	3	4	5	6	7	8	9	10
0% - 20%	47	34	33	30	33	29	32	32	32	34
21% - 45%	73	70	64	63	60	58	50	53	55	53
46% - 85%	50	61	66	58	62	54	52	45	45	42

86% - 100%	5	9	12	22	20	34	40	45	43	45

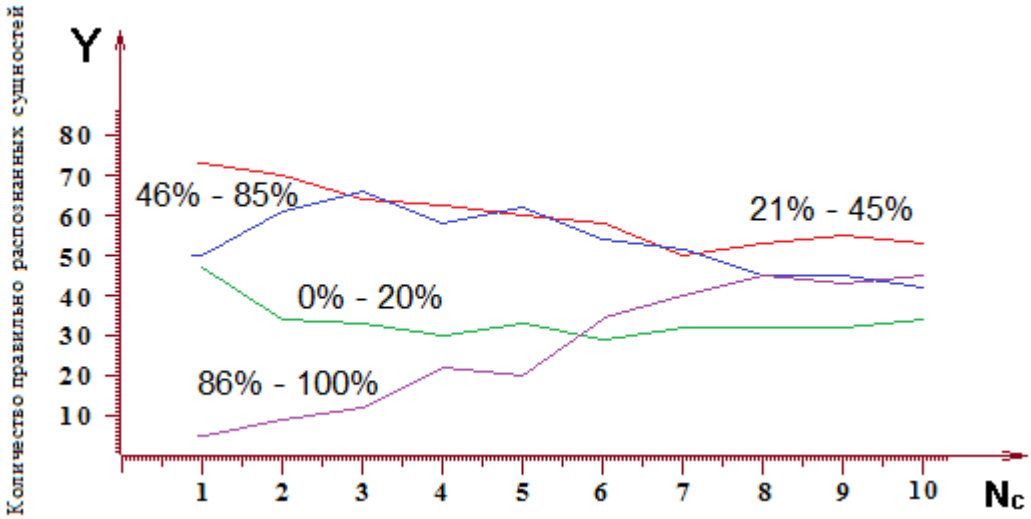

Рисунок 5. Зависимости количества правильно идентифицируемых сущностей предметной области при благоприятной стратегии исключения её объектов из анализируемых данных для различных длин лексем алфавита

С учётом полученных экспериментальных данных исключения объектов из анализируемых данных для различных длин лексем по благоприятной стратегии сделаны следующие выводы:

- при благоприятной стратегии исключения для выборки 0 до 20% правильная идентификация становится возможной;

- при благоприятной стратегии исключения объектов сущности предметной области зависимости количества правильно распознанных сущностей предметной области для различных объёмов выборок от ограничения на количество символов в лексемах алфавита частотной идентификации сущности предметной области более равномерны;

- правильная идентификация зависит от индивидуальных особенностей анализируемых данных сущности: для одних сущностей она возможна при объёме

данных менее 20% от эталонных, для других она начинается с объёма 86% от эталонных.

Как показали исследования на качество частотной идентификации сущностей предметной области влияет степень вхождения лексем алфавита в анализируемые данные сущности, уникальные особенности данных сущностей, используемый при идентификации объём данных относительно использованного при создании алфавита поиска.

Несмотря на то, что в ходе идентификации данных сущности предметной области возможны ошибки, использование частотной идентификации при поиске требуемых данных в базе данных прикладных программ приводит к заметному облегчению и ускорению этого процесса, обеспечивая обмен данными с ней.

Полученные результаты исследования влияния искажения алфавита поиска на идентификацию сущностей на основе частотного анализа данных имеет практическое значение для снижение вычислительных затрат при обеспечении заданного качества идентификации. Частотный анализ данных в отличие от существующих подходов достижения интероперабельности позволяет в автоматическом режиме находить сходные сущности предметной области в различных прикладных программах, обеспечивая их интероперабельность.

Библиографический список

1. Заковряшин А.И. ИПИ технология создания наукоемких изделий // Труды МАИ. 2011. № 49. URL: http://trudymai.ru/published.php?ID=28072&PAGEN_2=2

2. Соколов В.П., Завалишин И.В., Милюков И.А. Технологическая среда проектирования сложных технических объектов // Труды МАИ. 2011. № 49. URL: http://trudymai.ru/published.php?ID=28199&PAGEN_2=2

3. Беккер Д.А. Роль информационных технологий в управлении современным промышленным предприятием // Актуальные проблемы гуманитарных и естественных наук. 2013. № 7-1. С. 141 - 144.

4. Воробьев Е.Б., Лисина О.В. Использование информационных технологий в государственном и муниципальном управлении // Научные горизонты. 2018. № 2 (6). С. 45 – 49.

5. Елисеев Ю.С., Поклад В.А., Елисеев Д.Н. Применение информационных технологий при проектировании газотурбинных установок // Труды МАИ. 2012. № 56. URL: http://trudymai.ru/published.php?ID=30155

6. Арженко А.Ю., Волков П.А., Вестяк В.А. Повышение эффективности функционирования ERP-систем посредством создания взаимосвязанных баз данных // Труды МАИ. 2010. № 37. URL: http://trudymai.ru/published.php?ID=13438

7. Буряков А.А. Методы и технологии обмена данными систем автоматизированного проектирования // Труды МАИ. 2005. № 20. URL: http://trudymai.ru/published.php?ID=34133

8. Арженко А.Ю., Волков П.А., Вестяк В.А. Организация взаимодействия между базами данных предприятий аэрокосмической отрасли в рамках автоматизированного учета с использованием ERP-системы // Труды МАИ. 2010. № 41. URL: http://trudymai.ru/published.php?ID=23766

9. Мальшаков Г.В. Методы, алгоритмы и программные инструменты достижения интероперабельности прикладного программного обеспечения на основе частотного анализа данных: дисс. ... канд. тех. наук. – М.: МАИ, 2017. – 150 с.

10. Лунев И.С., Некруткин В.В. Замечание о некоторых классических критериях математической статистики // Вестник Санкт-Петербургского университета. Математика. Механика. Астрономия. 2019. Т. 6. № 2. С. 221 - 231.

11. Шашков В.Б. Автоматизированный расчет критерия Пирсона (математическая статистика без статистических таблиц) // Вестник Оренбургского государственного университета. 2005. № 9 (47). С. 172 - 174.

12. Мещерякова Т.В., Фирюлин М.Е. Статистическая обработка эмпирических данных функционирования автоматизированных систем // Охрана, безопасность, связь. 2019. Т. 3. № 4 (4). С. 110 - 116.

13. Гржибовский А.М., Иванов С.В., Горбатова М.А. Корреляционный анализ данных с использованием программного обеспечения STATISTICA и SPSS // Наука и Здравоохранение. 2017. № 1. С. 7 - 36.

14. Унгуряну Т.Н., Гржибовский А.М. Корреляционный анализ с использованием пакета статистических программ STATA // Экология человека. 2014. № 9. С. 60 - 64.

15. Савина Л.Н., Попырин А.В. Вычисление коэффициентов взаимной сопряженности Пирсона и Чупрова средствами EXCEL // Сборник научных трудов SWorld. 2013. Т. 2. № 2. С. 53 - 56.

16. Гаврилова М.А. Полупроводниковые приборы. Диоды. Выпрямительные диоды, магнитодиоды КД 102Б... АД 425А, Б: Справочник. - СПб.: Изд-во РНИИ "Электронстандарт", 1994. - 256 с.

17. Воробьев М.Д. Элементная база радиоэлектроники. Биполярные и полевые транзисторы. Интегральные схемы. – М.: Изд-во МЭИ, 1991. - 81 с.

18. Овсянников Н.И. Кремниевые биполярные транзисторы: Справочное пособие. - Минск: Вышэйшая школа, 1989. - 301 с.

19. Лысенко А.П. Биполярные транзисторы. – М.: Московский государственный институт электроники и математики, 2006. - 78 с.

20. Брешенков А.В. Методика проектирования реляционных баз данных // Инженерный журнал: наука и инновации. 2013. № 11 (23). С. 29.

21. Ковтун И.И. Принципиальные решения функционально-реляционной методологии проектирования автоматизированных систем // Информатизация и связь. 2013. № 3. С. 47 - 54.

Studying search alphabet distortion impact on entities identification based on data frequency analysis

Malshakov G.V.

Moscow Aviation Institute (National Research University), Volokolamskoe shosse, 4, Moscow, A-80, GSP-3, 125993, Russia

e-mail: malshakov@mail.ru

Abstract

The article recounts the results of studying the impact of the alphabet tokens degree of occurrence in the entity objects of the subject domain on the identification of entities based on data frequency analysis with various strategies for tokens exclusion.

The following dependences are shown:

- Dependence of autocorrelation on minimum allowable number of characters in tokens for various samplings (as a percentage of the initial data);

- Dependence of covering the entity objects alphabet of the subject domain on limitation on the number of characters in the alphabet tokens;

- Dependence of the number of correctly recognized entities of the subject domain for various samplings volumes on the limitation of the number of characters in lexemes of the alphabet of the subject domain entity's frequency identification at unfavorable strategy for the subject domain entity objects exclusion;

- Dependence of the number of correctly recognized subject domain entities for various samplings sizes on the limitation on the number of characters in the lexemes of the alphabet of the frequency identification of the subject domain entity at a favorable strategy of excluding objects of the subject domain entities.

The impact of the degree of occurrence of alphabet lexemes in the entity objects of the subject domain and the strategy of excluding alphabet lexemes on the entities identification based on data frequency analysis was established.

The above said studies allow reducing the number of errors in identifying domain entities based on the data frequency analysis while ensuring the interoperability of the applied software.

Keywords: error, identification, entity, application software, data frequency analysis, correlation.

References

1. Zakovryashin A.I. *Trudy MAI*, 2011, no. 49. URL: http://trudymai.ru/eng/published.php?ID=28072&PAGEN_2=2

2. Sokolov V.P., Zavalishin I.V., Milyukov I.A. *Trudy MAI*, 2011, no. 49. URL: http://trudymai.ru/eng/published.php?ID=28199&PAGEN_2=2

3. Bekker D.A. *Aktual'nye problemy gumanitarnykh i estestvennykh nauk*, 2013, no. 7-1, pp. 141 - 144.

4. Vorob'ev E.B., Lisina O.V. *Nauchnye gorizonty*, 2018, no. 2 (6), pp. 45 – 49.

5. Eliseev Yu.S., Poklad V.A., Eliseev D.N. *Trudy MAI*, 2012, no. 56. URL: http://trudymai.ru/eng/published.php?ID=30155

6. Arzhenenko A.Yu., Volkov P.A., Vestyak V.A. *Trudy MAI*, 2010, no. 37. URL: http://trudymai.ru/eng/published.php?ID=13438

7. Buryakov A.A. *Trudy MAI*, 2005, no. 20. URL: http://trudymai.ru/eng/published.php?ID=34133

8. Arzhenenko A.Yu., Volkov P.A., Vestyak V.A. *Trudy MAI*, 2010, no. 41. URL: http://trudymai.ru/eng/published.php?ID=23766

9. Mal'shakov G.V. *Metody, algoritmy i programmnye instrumenty dostizheniya interoperabel'nosti prikladnogo programmnogo obespecheniya na osnove chastotnogo analiza dannykh* (Methods, algorithms, and software tools for achieving interoperability of software application based on frequency data analysis), Doctor's thesis, Moscow, MAI, 2017, 150 p.

10. Lunev I.S., Nekrutkin V.V. *Vestnik Sankt-Peterburgskogo universiteta. Matematika. Mekhanika. Astronomiya*, 2019, vol. 6, no. 2, pp. 221 - 231.

11. Shashkov V.B. *Vestnik Orenburgskogo gosudarstvennogo universiteta*, 2005, no. 9 (47), pp. 172 - 174.

12. Meshcheryakova T.V., Firyulin M.E. *Okhrana, bezopasnost', svyaz'*, 2019, vol. 3, no. 4 (4), pp. 110 – 116.

13. Grzhibovskii A.M., Ivanov S.V., Gorbatova M.A. *Nauka i Zdravookhranenie*, 2017, no. 1, pp. 7 - 36.

14. Unguryanu T.N., Grzhibovskii A.M. *Ekologiya cheloveka*, 2014, no. 9, pp. 60 - 64.

15. Savina L.N., Popyrin A.V. *Sbornik nauchnykh trudov SWorld*, 2013, vol. 2, no. 2, pp. 53 - 56.

16. Gavrilova M.A. *Poluprovodnikovye pribory. Diody. Vypryamitel'nye diody, magnitodiody KD 102B... AD 425A, B* (Semiconductor devices. Diodes: Rectifyiing diodes,

magneto-diodes KD 102B ... AD 425A, B), Saint Petersburg, Izd-vo RNII “Elektronstandart”, 1994, 256 p.

17. Vorob'ev M.D. *Elementnaya baza radioelektroniki. Bipolyarnye i polevye tranzistory. Integral'nye skhemy* (Elemental base of radio electronics. Bipolar and field effect transistors.

18. Ovsyannikov N.I. *Kremnievye bipolyarnye tranzistory* (Silicon Bipolar Transistors), Minsk, Vysheishaya shkola, 1989, 301 p.

19. Lysenko A.P. *Bipolyarnye tranzistory* (Bipolar transistors), Moscow, Moskovskii gosudarstvennyi institut elektroniki i matematiki, 2006, 78 p.

20. Breshenkov A.V. *Inzhenernyi zhurnal: nauka i innovatsii*, 2013, no. 11 (23), pp. 29.

21. Kovtun I.I. *Informatizatsiya i svyaz'*, 2013, no. 3, pp. 47 - 54.