Раздел I. Алгоритмы обработки информации

УДК 004.021 DOI 10.18522/2311-3103-2021-3-6-17

М.С. Анферова, А.М. Белевцев

РАЗРАБОТКА АЛГОРИТМОВ ИНТЕЛЛЕКТУАЛЬНОГО СЕРВИСА ПОИСКА И МОНИТОРИНГА ИНФОРМАЦИИ

Описана проблема стратегического анализа и выбора направлений развития инновационного предприятия в условиях перехода к 6 технологическому укладу и индустрии 4.0. В данных условиях поисково-аналитическая обработка информации не может быть полноценно выполнена без применения автоматизированных информационно-аналитических систем, в том числе и на базе искусственного интеллекта. В ходе анализа были определены основные приоритетные функции, которые должны обеспечивать разрабатываемые сервисы. Обозначены основные трудности при разработке данных сервисов, такие как: предварительная обработка данных и автоматизированная проверка актуальности баз данных. Для эффективного решения поставленных задач сервис интеллектуального мониторинга и поиска информации должен использовать комплексный подход с учетом эффективности применения методов для отдельных подзадач, обеспечивать высокую эффективность реализации всех этапов процедуры интеллектуального мониторинга. В связи с этим в данной работе описывается не только разработка общего интеллектуального поискового алгоритма, но и отдельные блок-алгоритмы, необходимые для обеспечения приоритетных функций разрабатываемого сервиса. В работе представлены следующие алгоритмы: алгоритм информационного поиска, необходимый для решения задачи полнотекстового поиска документов в пределах базы информационных ресурсов информационно-аналитического комплекса; алгоритм процедуры внесения новых документов; алгоритм предварительной обработки данных, выключающий в себя стемминг и удаление знаков препинания для последующего анализа текста; алгоритм оценки ранжирования и релевантности информации, включающий в себя векторизацию документов; алгоритм кластеризации результатов поиска информации на основе нейронной сети Кохонена; алгоритм проверки актуальности информации -проверка соответствия локальной копии документа актуальной версии на веб-ресурсе источника. Предложен и обоснован язык программирования Python для реализации представленного алгоритма. Система обеспечивает автоматизированный непрерывный мониторинг с высокой периодичностью отправки запроса без участия оператора, что повысит качество и эффективность информационного поиска в условиях большого объёма неструктурированной информации.

Технологические тренды; мониторинг; искусственный интеллект; Big Data; алгоритм; распознавание текста; кластеризация.

M.S. Anferova, A.M. Belevtsev

DEVELOPMENT OF ALGORITHMS OF INTELLIGENT SERVICE FOR INFORMATION SEARCH AND MONITORING

This paper describes the problem of strategic analysis and the choice of directions for the development of an innovative enterprise in the conditions of transition to the 6th technological order and industry 4.0. In these conditions, search and analytical processing of information cannot be fully performed without the use of automated information and analytical systems, including those based on artificial intelligence. During the analysis, the main priority functions that the developed services should provide were identified. The main difficulties in the development of these services are identified, such as: pre-processing of data and automated checking of the relevance of databases. To effectively solve the

tasks set, the intelligent monitoring and information retrieval service should use an integrated approach, taking into account the effectiveness of applying methods for individual subtasks, and ensure high efficiency of implementing all stages of the intelligent monitoring procedure. In this regard, this paper describes not only the development of a general intelligent search algorithm, but also individual block algorithms necessary to ensure the priority functions of the service being developed. The paper presents the following algorithms: an information search algorithm necessary to solve the problem of full-text search of documents within the database of information resources of the information and analytical complex; an algorithm for the procedure for entering new documents; an algorithm for pre-processing data that includes stemming and removing punctuation marks for subsequent text analysis; an algorithm for evaluating the ranking and relevance of information, including vectorization of documents; an algorithm for clustering information search results based on the Kohonen neural network; the algorithm for checking the relevance of information is to check whether the local copy of the document corresponds to the current version on the source's web resource. The Python programming language for the implementation of the presented algorithm is proposed and justified. The system provides automated continuous monitoring with a high frequency of sending a request without the participation of an operator, which will increase the quality and efficiency of information search in conditions of a large volume of unstructured information.

Technological trends; monitoring; search robot; artificial intelligence; Big Data; algorithm; text recognition; clustering.

Введение. Проблема мониторинга и стратегического анализа информации в сетях общего и специального назначения постоянно усложняется.

Это обусловлено экспоненциальным ростом объемов информации (Big Data) [1], ее разнородностью, в том числе и по форме представления, отсутствием структурированности и высокой динамикой обновления.

В этой связи создание новых технологий мониторинга, обеспечивающих эффективность, полноту и высокий уровень релевантности информационного поиска является актуальным [2, 3].

Решение данной задачи может быть получено на основе создания на основе создания интеллектуальных поисковых роботов для распознавания и кластеризации неструктурированной информации в информационно-аналитических комплексах [4].

При этом общая концепция технологии подобного сервиса должна быть основана на комплексном подходе, обеспечивающего как синергетический эффект, так и высокую эффективность реализации отдельных этапов процедуры интеллектуального мониторинга.

Проведенный анализ исследовательских проектов [5] позволил выделить следующие приоритетные функции, которые должны обеспечивать разрабатываемые сервисы:

Ф1. Формирование поискового образа на основе методов теории исчисления предикатов с использованием ключевых слов и метаданных.

Ф2. Полнотекстовый поиск документов в пределах базы информационных ресурсов информационно-аналитического комплекса.

Ф3. Поиск документов, семантически близких к заданным эталонам в пределах рассматриваемой предметной области.

Ф4. Кластеризация найденных документов с целью упрощения их дальнейшего восприятия и фильтрации аналитиком.

Ф5. Автоматизированная проверка актуальности базы информационных ресурсов с целью реализации задачи непрерывного мониторинга.

При этом набольшую сложность представляют собой следующие взаимосвязанные проблемы:

- предварительная обработка данных и формирование первичной коллекции (Ф1-Ф4);
- автоматизированная проверка актуальности базы информационных ресурсов с целью реализации задачи непрерывного мониторинга (Ф5).

Основная часть. Общая процедура алгоритмов интеллектуального сервиса поиска и мониторинга информации будет представлять собой совокупность нескольких отдельных блоков-алгоритмов (рис. 1).

Рис. 1. Общая процедура алгоритмов интеллектуального сервиса поиска и мониторинга информации

Для реализации представленного алгоритма предлагается использовать язык программирования Python. Это обусловлено следующими факторами:

- простым и удобным интерфейсом;
- большим количеством специализированных библиотек для работы с сайтами, текстами и анализа, такие как ScikitLearn,NLTK, Gensim, spaCy, NetworkX и Yellowbrick.

1. Разработка алгоритма информационного поиска. Для решения задачи полнотекстового поиска документов в пределах базы информационных ресурсов информационно-аналитического комплекса в первую очередь необходимо составить полнотекстовый поисковый индекс для разрабатываемого сервиса.

Для поиска некоторого элемента в хранилище данных по заданному запросу потребуется время, пропорциональное количеству элементов в данном хранилище. Индекс, представляя собой структурированный, а не хаотический набор данных, позволяет осуществлять доступ на порядок быстрее.

Существует два основных типа поискового индекса [6]:

1. Прямой индекс.
2. Инвертированный индекс.

Для решения поставленной задачи удобно использовать инвертированный индекс, дополнив его весом слова в документе, так как инвертированный индекс хранит список документов, содержащих каждое слово, поисковая система может использовать прямой доступ, чтобы найти документы, связанные с каждым словом в запросе, и быстро получить их.

В качестве меры взвешивания слова можно воспользоваться F-мерой, а именно precision (точность) и recall (полнота) – это метрики, которые используются при оценке большей части алгоритмов извлечения информации [7]. Суть точности и полноты таких мер очень проста.

Точность системы в пределах класса – это доля документов, действительно принадлежащих данному классу относительно всех документов, которые система отнесла к этому классу. Полнота системы – это доля найденных классификатором документов, принадлежащих классу относительно всех документов этого класса в тестовой выборке.

Эти значения легко рассчитать на основании таблицы контингентности (табл. 1), которая составляется для каждого класса отдельно [8].

Таблица 1

Таблица контингентности

Категория i		Экспертная оценка	
		положительная	отрицательная
Оценка системы	Положительная	TP	FP
	Отрицательная	FN	TN

В таблице содержится информация, сколько раз система приняла верное и сколько раз неверное решение по документам заданного класса.

Точность и полнота определяются следующим образом [9]:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (1)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}, \quad (2)$$

где TP – истинно-положительное решение; TN – истинно-отрицательное решение; FP – ложно-положительное решение; FN – ложно-отрицательное решение.

В практическом применении значения точности и полноты удобнее рассчитывать с использованием матрицы неточностей (confusion matrix). В случае если количество классов относительно невелико (не более 100–150 классов), этот подход позволяет наглядно представить результаты работы классификатора.

Матрица неточностей (рис. 2) – это матрица размера N на N, где N – это количество классов. Столбцы этой матрицы резервируются за экспертными решениями, а строки за решениями классификатора. Когда мы классифицируем документ из тестовой выборки мы инкрементируем число, стоящее на пересечении строки класса, который вернул классификатор и столбца класса, к которому действительно относится документ.

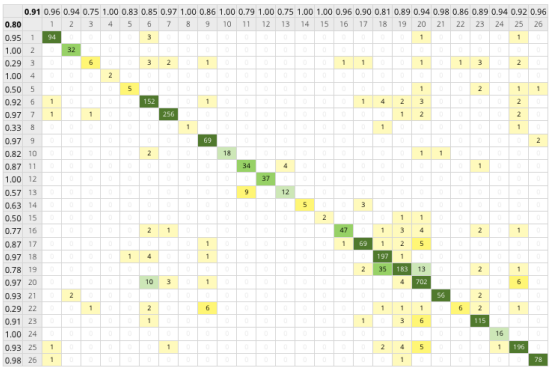

Рис. 2. Матрица неточностей (26 классов, результирующая точность – 0.8, результирующая полнота – 0.91)

Проведенный анализ показал, что классификатор определяет верно большинство документов. Диагональные элементы матрицы явно выражены. Тем не менее, в рамках некоторых классов (3, 5, 8, 22) классификатор показывает низкую точность.

Получив такую матрицу точности и полноты для каждого класса, последующие расчеты упрощаются. Точность равняется отношению соответствующего диагонального элемента матрицы и суммы всей строки класса. Полнота – отношение диагонального элемента матрицы и суммы всего столбца класса может быть представлена в следующем виде:

$$\text{Precision}_c = \frac{A_{c,c}}{\sum_{i=1}^{n} A_{c,i}}. \quad (3)$$

Тогда результирующая точность классификатора рассчитывается как арифметическое среднее его точности по всем классам. Аналогично с полнотой. Технически этот подход называется macro-averaging [10].

$$Recall_{\text{c}} = \frac{A_{c,c}}{\sum_{i=1}^{n} A_{i,c}} \quad (4)$$

Возникает проблема поиска баланса между максимальной точностью и полнотой. Для решения данной проблемы вводится метрика F-мера. С ее помощью будет проще принимать решение о том, какую реализацию запускать дальше (production).

F-мера представляет собой гармоническое среднее между точностью и полнотой. Она стремится к нулю, если точность или полнота стремится к нулю [11].

$$\text{F} = 2\frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recal}}. \quad (5)$$

Данная формула придает одинаковый вес точности и полноте, поэтому F-мера будет падать одинаково при уменьшении и точности и полноты. Возможно рассчитать F-меру придав различный вес точности и полноте.

$$\text{F} = (\beta^2 + 1)\frac{\text{Precision}\times\text{Recall}}{\beta^2\text{Precision}+\text{Recal}}, \quad (6)$$

где β принимает значения в диапазоне 0<β<1 если вы хотите отдать приоритет точности, а при β>1 приоритет отдается полноте. При β=1 формула сводится к предыдущей, и получается сбалансированная F-мера (рис. 3–5).

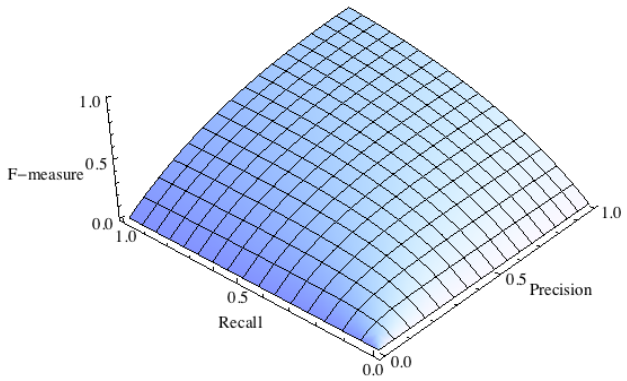

Рис. 3. Сбалансированная F-мера

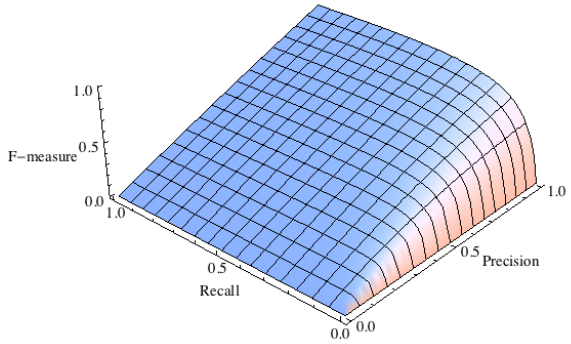

Рис. 4. F-мера с приоритетом точности $(\beta^2 - \frac{1}{4})$

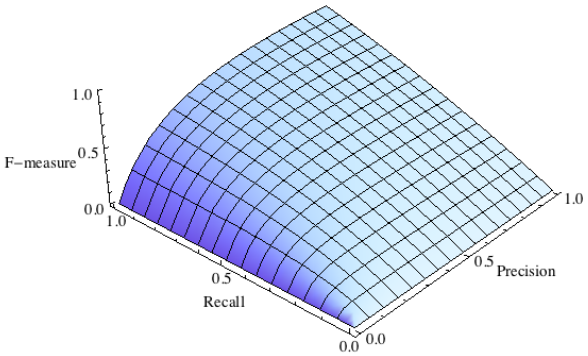

Рис. 5. F-мера с приоритетом полноты $(\beta^2 - 2)$

Проведенный анализ показал, что F-мера сводит к одному числу две других основополагающих метрики: точность и полноту, поэтому ее можно использовать как формальную метрику оценки качества классификатора.

2. Алгоритм процедуры внесения новых документов. Все документы, хранящиеся в базе данных информационно-аналитического комплекса, должны быть проиндексированы с использованием алгоритмов нормализации текста. Полученный таким образом индекс будет основой информационного поиска.

Принципиальный алгоритм процедуры внесения новых документов представлен на рис. 6.

Рис. 6. Алгоритм процедуры внесения новых документов

3. Предварительная обработка данных. Предварительная обработка текста включает в себя такие следующие операции:

- Стемминг (stemming).
- Удаление знаков препинания.

Стемминг – это нахождение основы слова (стеммы), передающей его лексическое значение [12].

Проведенный анализ показал, что в базах данных преимущественно информация хранится в виде HTML документов.

Разметка HTML, которая сама по себе структурирована, может производиться и отображаться множеством иногда беспорядочных способов. Это связано с тем, что в сети Интернет веб-страницы не обязаны структурироваться в строгом соответствии с каким-то набором стандартов. В связи с данной непредсказуемостью возникает проблема извлечения данных из HTML документов методичным и предсказуемым способом.

Тогда алгоритм предварительной обработки текстов может быть представлен в виде следующего алгоритма (рис. 7).

Рис. 7. Алгоритм предварительной обработки текстов

4. Разработка алгоритма оценки ранжирования и релевантности. Для сравнения документов между собой математическими методами, необходимо провести их векторизацию.

Такую матрицу можно получить при помощи класса CountVectorizer из библиотеки scikit-learn.

Вес терма («важность» слова для идентификации данного текста) в документе можно определить разными способами. Если в разрабатываемом поисковом сервисе за меру веса терма берем меру TF-IDF, то это позволит нам использовать поисковый индекс в качестве кэша для хранения векторов документов.

TF (term frequency – частота слова) – отношение числа вхождений некоторого слова к общему числу слов документа. Таким образом, оценивается важность слова ti в пределах отдельного документа.

$$TF(t,d) = \frac{n_t}{\sum_k n_k}, \quad (7)$$

где n_t – число вхождений слова t в документ, а в знаменателе – общее число слов в данном документе.

Релевантность в информационном поиске – это семантическое соответствие поискового запроса полученному документу [13].

Введем понятие формальной релевантности – соответствие, определяемое путём сравнения образа поискового запроса с поисковым образом документа.

Одним из методов для оценки релевантности является TF-IDF-метод [14].

Проведенный анализ показал, что целесообразно использовать данный метод в основе алгоритма оценки релевантности и ранжирования предлагаемого интеллектуального сервиса поиска и мониторинга информации, так как обеспечивается функция ранжирования соответствующих поисковому запросу документов по увеличению метрики TF-IDF.

5. Разработка алгоритма кластеризации результатов поиска. Следующим этапом разрабатываемого сервиса является кластеризация–процесс объединения в группы объектов, обладающих схожими признаками [15].

Применительно к задаче анализа данных, метод кластеризации используется для кластерного анализа – многомерной статистической процедуры, выполняющей сбор данных, содержащих информацию о выборке объектов, и затем упорядочивающей объекты в сравнительно однородные группы [16].

Проведенный мониторинг показал, что существует несколько основных подходов разбиения групп объектов на кластеры:

1. Вероятностный подход.
2. Подходы на основе систем искусственного интеллекта:
3. Логический подход. Построение дендрограммы осуществляется с помощью дерева решений.
4. Иерархический подход. Предполагается наличие вложенных групп (кластеров различного порядка).

Проведенный анализ показал, что в настоящее время наиболее высокой интенсивностью исследований обладают методы искусственного интеллекта [17, 18]. Так же кластеризация с помощью методов искусственного интеллекта основана на подобии образов: нейронная сеть размещает близкие образы в один кластер [19].

В этой связи для решения поставленных задач предлагается использовать подходы на основе систем искусственного интеллекта. В качестве простейшего примера можно привести алгоритм на основе нейронной сети Кохонена (рис. 8).

6. Разработка алгоритма оценки актуальности. Организация мониторинга инновационного потенциала играет особо значимую роль в процессе обеспечения устойчивого развития предприятий. Это предусматривает многоаспектное исследование по выявлению тенденций и перспектив дальнейшего развития предприятия. Мониторинг инновационного потенциала позволяет провести анализ и прогнозирование доходности изменений структуры объектов интеллектуальной собственности, инновационного развития, а также платежеспособности, ликвидности, финансовой устойчивости и деловой активности предприятия.

Под оценкой актуальности в данной работе понимается проверка соответствия локальной копии документа актуальной версии на веб-ресурсе источника. Предлагаемый алгоритм работает посредством программы-робота, автоматически в фоновом режиме проходящей по коллекции документов и проверяющей первоисточники документов на предмет обновления (рис. 9).

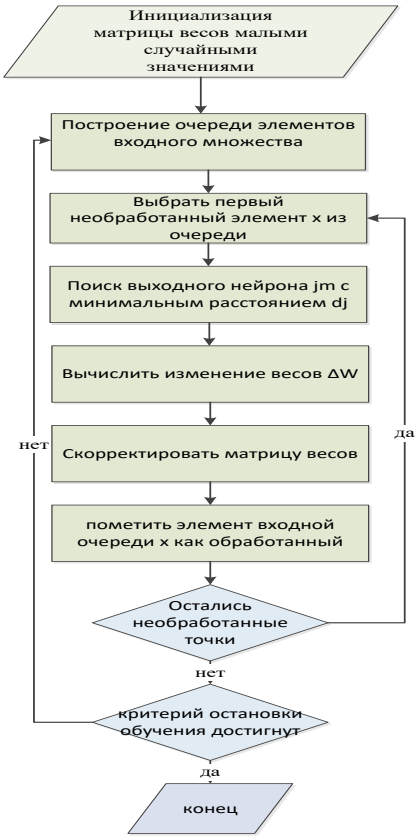

Рис. 8. Алгоритм кластеризации

Необходимость данного функционала обусловлена высокой степенью важности своевременного обнаружения обновлений при работе стратегическом анализе инновационных направлений развития предприятий.

Рис. 9. Алгоритм проверки актуальности информации

Выводы. Предлагаемый интеллектуальный сервис поиска и мониторинга информации реализован в информационно-аналитическом комплексе АРМ Аналитика [20]. Внедрение данного алгоритма позволило сократить общее время поиска информации в 5–6 раз по отношению к запросам, формируемым в поисковых системах общего назначения.

Предложенная процедура формирует первичную коллекцию на основе интеллектуального поиска информации в заданной предметной области, что позволяет существенно повысить релевантность поиска и обработки информации, в отличие от систем глобального мониторинга, поиск которых основан на индикаторах: рост использования ключевых слов, увеличение численности новых авторов, цитирование работ из смежных областей.

Системное программное обеспечение на основе предложенного алгоритма может быть использовано при создании интеллектуальных сервисов и существенно повысить релевантность поиска и обработки информации в информационно-аналитическом комплексе стратегического анализа инновационных направлений развития предприятия.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Белевцев А.М., Садреев Ф.Г., Белевцев А.А., Балыбердин В.А.* Разработка интеллектуальных сервисов мониторинга технологических трендов в информационно-аналитических комплексах // Наукоемкие технологии. – 2019. – Т. 20, № 3. – С. 24-29.
2. *Белевцев А.М., Балыбердин В.А., Бендерский Г.П., Белевцев А.А.* Анализ направлений развития нано- и IT-технологий для построения специализированных сетевых коммуникационных систем нового поколения // Известия ЮФУ. Технические науки. – 2015. – № 3 (164). – С. 35-45.
3. *Микова Н.С., Соколова А.В.* Мониторинг глобальных технологических трендов: теоретические основы и лучшие практики // Форсайт. – 2014. – Т. 8, № 4.
4. *Анферова М.С., Белевцев А.М.* Анализ направлений создания алгоритмов эффективного поиска информации в сетях общего и специального назначения // Матер. III Всероссийской научно-технической конференции «Актуальные проблемы современной науки и производства». – Рязань: РГРТУ, 2018.
5. *Анферова М.С., Белевцев А.М*. Поисковые роботы для автоматизированного мониторинга информации в сетях общего и специального назначения // 18-я Международная научно-практическая конференция «Управление качеством». – 2019.
6. *Jacob Devlin and Ming-Wei Chang*. Research Scientists, Google AI Language: Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing (англ.). – Google, Inc, 2018.
7. *Charles L. Clarke A., Gordon V.* Cormack Dynamic Inverted Indexes for a Distributed Full-Text Retrieval System // MultiText Pro ject Technical Report MT-95-01. – University of Waterloo, Waterloo, Ontario N2L 3G1, Canada, 1995.
8. *Павлов Ю.Н., Майструк К.А*. Сравнение методов оценки тональности текста // Молодой ученый. – 2016. – № 12 (116). – С. 59-64.
9. *Olson David L, and Delen, Dursun.* Advanced Data Mining Techniques. – Springer, 1st edition (February 1, 2008). – 2008. – 138 p.
10. *Manning C., Raghavan P., Schütze H.* Introduction to Information Retrieval. – Cambridge University Press, 2008.
11. *Powers, David M.W.* Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation // Journal of Machine Learning Technologies. – 2011. – No. 2 (1). – P. 37-63.
12. *Lovins Julie Beth.* Development of a Stemming Algorithm // Mechanical Translation and Computational Linguistics. – 1968. – Vol. 11.
13. Словарь по кибернетике / под ред. академика В.С. Михалевича. – 2-е. изд. – Киев: Гл. ред. Украинской советской энциклопедии им. М.П. Бажана, 1989. – 751 с.
14. *Salton G. and Buckley C.* Term-weighting approaches in automatic text retrieval // Information Processing & Management. – 1988. – Vol. 24 (5). – P. 513-523.

15. *Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д.* Прикладная статистика: Классификация и снижение размерности. – М.: Финансы и статистика, 1989. – 607 с.
16. *Бериков В.С., Лбов Г.*С. Современные тенденции в кластерном анализе // Всероссийский конкурсный отбор обзорно-аналитических статей по приоритетному направлению «Информационно-телекоммуникационные системы». – 2008. – 26 с.
17. *Анферова М.С., Белевцев А.М.* Анализ направлений создания алгоритмов эффективного поиска информации в сетях общего и специального назначения // Матер. III Всероссийской научно-технической конференции «Актуальные проблемы современной науки и производства». – Рязань: РГРТУ, 2018.
18. *Анферова М.С., Белевцев А.М.* Анализ направлений развития технологий мониторинга в условиях большого объёма неструктурированной информации // XXIV Всероссийская научно-техническая конференция с международным участием им. профессора О.Н. Пьявченко ”Компьютерные и информационные технологии в науке, инженерии и управлении” «КомТех-2020».
19. *Эндрю М.* Реальная жизнь и искусственный интеллект // Новости искусственного интеллекта. РАИИ, 2000.
20. *Белевцев А.М., Балыбердин В.А., Белевцев А.А., Садреев Ф.Г.* О разработке интеллектуальных сервисов мониторинга технологических трендов в информационно-аналитических комплексах // Наукоемкие технологии. – 2019. – № 3.

REFERENCES

1. *Belevtsev A.M., Sadreev F.G., Belevtsev A.A., Balyberdin V.A.* Razrabotka intellektual'nykh servisov monitoringa tekhnologicheskikh trendov v informatsionno-analiticheskikh kompleksakh [Development of intelligent services for monitoring technological trends in information and analytical complexes], *Naukoemkie tekhnologii* [High-tech technologies], 2019, Vol. 20, No. 3, pp. 24-29.
2. *Belevtsev A.M., Balyberdin V.A., Benderskiy G.P., Belevtsev A.A.* Analiz napravleniy razvitiya nano- i IT-tekhnologiy dlya postroeniya spetsializirovannykh setevykh kommunikatsionnykh sistem novogo pokoleniya [Analysis of the directions of development of nano-and IT-technologies for the construction of specialized network communication systems of a new generation], *Izvestiya YuFU. Tekhnicheskie nauki* [Izvestiya SFedU. Engineering Sciences], 2015, No. 3 (164), pp. 35-45.
3. *Mikova N.S., Sokolova A.V.* Monitoring global'nykh tekhnologicheskikh trendov: teoreticheskie osnovy i luchshie praktiki [Monitoring global technological trends: theoretical foundations and best practices], *Forsayt* [Foresight], 2014, Vol. 8, No. 4.
4. *Anferova M.S., Belevtsev A.M.* Analiz napravleniy sozdaniya algoritmov effektivnogo poiska informatsii v setyakh obshchego i spetsial'nogo naznacheniya [Analysis of the directions of creating algorithms for effective information search in general and special purpose networks], *Mater. III Vserossiyskoy nauchno-tekhnicheskoy konferentsii «Aktual'nye problemy sovremennoy nauki i proizvodstva»* [Materials of the III All-Russian Scientific and Technical Conference "Actual problems of modern science and production"]. Ryazan': RGRTU, 2018.
5. *Anferova M.S., Belevtsev A.M.* Poiskovye roboty dlya avtomatizirovannogo monitoringa informatsii v setyakh obshchego i spetsial'nogo naznacheniya [Search robots for automated monitoring of information in general and special purpose networks], *18-ya Mezhdunarodnaya nauchno-prakticheskaya konferentsiya «Upravlenie kachestvom»* [18th International Scientific and Practical Conference "Quality Management"], 2019.
6. *Jacob Devlin and Ming-Wei Chang*. Research Scientists, Google AI Language: Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing (англ.). Google, Inc, 2018.
7. *Charles L. Clarke A., Gordon V.* Cormack Dynamic Inverted Indexes for a Distributed Full-Text Retrieval System, *MultiText Pro ject Technical Report MT-95-01. University of Waterloo, Waterloo, Ontario N2L 3G1, Canada, 1995*.
8. *Pavlov Yu.N., Maystruk K.A.* Sravnenie metodov otsenki tonal'nosti teksta [Comparison of methods for assessing the tonality of the text], *Molodoy uchenyy* [Young scientist], 2016, No. 12 (116), pp. 59-64.
9. *Olson David L, and Delen, Dursun.* Advanced Data Mining Techniques. Springer, 1st edition (February 1, 2008), 2008, 138 p.

10. *Manning C., Raghavan P., Schütze H.* Introduction to Information Retrieval. Cambridge University Press, 2008.
11. *Powers, David M.W.* Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation, *Journal of Machine Learning Technologies*, 2011, No. 2 (1), pp. 37-63.
12. *Lovins Julie Beth.* Development of a Stemming Algorithm, *Mechanical Translation and Computational Linguistics*, 1968, Vol. 11.
13. Slovar' po kibernetike [Dictionary of Cybernetics], ed. by akad. V.S. Mikhalevicha. 2nd. ed. Kiev: Gl. red. Ukrainskoy sovetskoy entsiklopedii im. M.P. Bazhana, 1989, 751 p.
14. *Salton G. and Buckley C.* Term-weighting approaches in automatic text retrieval, *Information Processing & Management*, 1988, Vol. 24 (5), pp. 513-523.
15. *Ayvazyan S.A., Bukhshtaber V.M., Enyukov I.S., Meshalkin L.D.* Prikladnaya statistika: Klassifikatsiya i snizhenie razmernosti [Applied statistics: Classification and dimension reduction]. Mscow: Finansy i statistika, 1989, 607 p.
16. *Berikov V.S., Lbov G.S.* Sovremennye tendentsii v klasternom analize [Modern trends in cluster analysis], *Vserossiyskiy konkursnyy otbor obzorno-analiticheskikh statey po prioritetnomu napravleniyu «Informatsionno-telekommunikatsionnye sistemy»* [All-Russian competitive selection of review and analytical articles in the priority direction "Information and telecommunications systems"], 2008, 26 p.
17. *Anferova M.S., Belevtsev A.M.* Analiz napravleniy sozdaniya algoritmov effektivnogo poiska informatsii v setyakh obshchego i spetsial'nogo naznacheniya [Analysis of the directions of creating algorithms for effective information search in general and special purpose networks] *Mater. III Vseros-siyskoy nauchno-tekhnicheskoy konferentsii «Aktual'nye problemy sovremennoy nauki i proizvodstva»* [Materials of the III All-Russian Scientific and Technical Conference "Actual problems of modern science and production"]. Ryazan': RGRTU, 2018.
18. *Anferova M.S., Belevtsev A.M.* Analiz napravleniy razvitiya tekhnologiy monitoringa v usloviyakh bol'shogo ob"ema nestrukturirovannoy informatsii [Analysis of trends in the development of monitoring technologies in the conditions of a large volume of unstructured information], *XXIV Vserossiyskaya nauchno-tekhnicheskaya konferentsiya s mezhdunarodnym uchastiem im. professora O.N. P'yavchenko "Komp'yuternye i informatsionnye tekhnologii v nauke, inzhenerii i upravlenii" «KomTekh-2020»* [XXIV All-Russian Scientific and Technical Conference with international participation named after Professor O. N. Piavchenko "Computer and information technologies in science, engineering and management ""Comtech-2020"].
19. *Endryu M.* Real'naya zhizn' i iskusstvennyy intellekt [Real life and artificial intelligence], *Novosti iskusstvennogo intellekta*, RAII, 2000.
20. *Belevtsev A.M., Balyberdin V.A., Belevtsev A.A., Sadreev F.G.* O razrabotke intellektual'nykh servisov monitoringa tekhnologicheskikh trendov v informatsionno-analiticheskikh kompleksakh [On the development of intelligent services for monitoring technological trends in information and analytical complexes], *Naukoemkie tekhnologii* [High-tech technologies], 2019, No. 3.

Статью рекомендовал к опубликованию д.т.н., профессор В.А. Балыбердин.

Анферова Маргарита Сергеевна – Московский авиационный институт (национальный исследовательский университет); e-mail: gludkina@yandex.ru; г. Москва, Россия; тел.: +79055220749; старший преподаватель.

Белевцев Андрей Михайлович – e-mail: ambelevtsev@yandex.ru; тел.: +79037691788; д.т.н.; профессор.

Anferova Margarita Sergeevna – Moscow Aviation Institute (National Research University); e-mail: gludkina@yandex.ru; Moscow, Russia; phone: +79055220749; senior lecturer.

Belevtsev Andrey Michailovitch – e-mail: ambelevtsev@yandex.ru; phone: +79037691788; dr. of eng. science; professor.