УДК: 004.912 DOI: 10.53816/23061456_2022_3–4_24

МОДЕЛЬ ПРЕДСТАВЛЕНИЯ ДАННЫХ ПРИ АВТОМАТИЧЕСКОЙ РУБРИКАЦИИ РАЗНОРОДНОЙ ИНФОРМАЦИИ В ЗАЩИЩЕННЫХ ИНФОРМАЦИОННО-АНАЛИТИЧЕСКИХ СИСТЕМАХ

MODEL OF DATA REPRESENTATION BY THE AUTOMATIC RUBRICATION OF DIFFERENT INFORMATION IN PROTECTED INFORMATION-ANALYTICAL SYSTEMS

Канд. техн. наук С.А. Краснов, канд. техн. наук В.А. Лохвицкий, Р.С. Хабаров

Ph.D. S.A. Krasnov, Ph.D. V.A. Lokhvickii, R.S. Khabarov

ВКА им. А.Ф. Можайского

В статье рассмотрена задача автоматической рубрикации разнородной информации в защищенных информационно-аналитических системах (ИАС), позволяющая снизить временные затраты при поиске и отборе требуемых данных. Предложена математическая модель представления данных, в основе которой лежит модернизированный метод латентно-семантического анализа (МЛСА) и словарь ключевых словосочетаний с повышающим коэффициентом. МЛСА используется для выявления семантических зависимостей между термами документов и получения коэффициента соответствия сравниваемых векторов. Показано, что использование предложенного модернизированного метода латентно-семантического анализа в совокупности со словарем ключевых словосочетаний с повышающим коэффициентом позволяет повысить эффективность автоматической рубрикации разнородной информации в ИАС.
Ключевые слова: латентно-семантический анализ, автоматическая рубрикация, информационно-аналитические системы, интеллектуальная обработка информации.

The article deals with the problem of automatic rubrication of heterogeneous information in secure information and analytical systems (IAS), which allows to reduce the time spent in the search and selection of the required data. A mathematical model of data presentation is proposed, which is based on the modernized method of latent semantic analysis (MLSA) and a dictionary of key phrases with an increasing coefficient. The MLSA is used to identify the semantic dependencies between the terms of documents and to obtain the correspondence coefficient of the compared vectors.It is shown that the use of the proposed modernized method of latent-semantic analysis in conjunction with a dictionary of key phrases with an increasing coefficient makes it possible to increase the efficiency of automatic rubrication of heterogeneous information in the IAS.
Keywords: latent-semantic analysis, automatic rubrication, information-analytical systems, intelligent information processing.

Введение

Важной задачей государства и Министерства обороны Российской Федерации в области информационных технологий является расширение функциональных возможностей центров обработки и хранения данных в процессе анализа больших массивов разнородной

информации [1]. Достаточно много научных и практических работ посвящены разработке новых моделей, методов и алгоритмов интеллектуального анализа больших данных [2, 3]. При этом вопрос автоматической рубрикации разнородной информации остается одной из важнейших задач семантической интеграции данных в условиях постоянного возрастания объемов информации. Методы лингвистического анализа, хотя и позволяют точнее анализировать текстовую информацию, выделяя его структурные особенности, но являются более трудоемкими и сложными в использовании. Связано это, прежде всего, с богатством семантики и морфологии естественных языков. Формальное описание правил естественного языка и их реализация — весьма трудоемкий процесс, требующий привлечения специалистов из области лингвистики. Работы в данном направлении идут [1, 4, 5, 7], и существует множество практических реализаций, но на сегодняшний день лингвистический анализ по части анализа семантики весьма проблематичен. Поэтому в данной работе для выявления семантических взаимосвязей между разнородными документами предлагается расширить модель, основанную на МЛСА [6]. Формирование математической модели занимает особое место в математической лингвистике, где на ее основе синтезируются различные алгоритмы обработки текстов: автоматический перевод, автоматическая рубрикация, анализ языковых особенностей и др. [9, 11, 14, 17]. Надо отметить, что полноценной модели порождения текстовой разнородной информации в настоящее время не существует. Имеются лишь отдельные примеры, отражающие те или иные стороны языковых особенностей [1, 11].

Предлагаемая модель расширяется на этапе предварительной обработки разнородной информации при выделении корневых основ и подсчете частот их употребления. В ней учитывается повышающий коэффициент для ключевых словосочетаний при формировании семантического описания как обучающей коллекции документов, так и документов пришедших на рубрикацию. Актуальность проведения исследований заключается в необходимости развития защищенных ИАС, отвечающих за сбор и обработку специальной информации, достоверность которой играет важную роль.

Этапы автоматической рубрикации разнородной информации в ИАС на основе МЛСА

В современных условиях постоянного увеличения объема и ценности обрабатываемой разнородной информации в защищенных ИАС возникает проблема решения задачи ее сбора, обработки и автоматической рубрикации с целью оперативного доступа к ней соответствующих должностных лиц [8, 15, 16, 18]. Для достижения поставленной цели требуется комплексное применение существующих и модернизированных методов интеллектуального анализа текстовых данных [1, 3, 15].

Автоматическая рубрикация разнородной информации в ИАС разделяется на следующие основные этапы:

1. Формирование представления содержания и связей разнородной информации из множества Ω (формирование тематических рубрик);

2. Автоматическая рубрикация разнородной информации из множества D по сформированным тематическим рубрикам.

На этапе формирования тематической рубрики происходит создание матрицы информационных образов из набора эталонной информации (далее документы). При этом исходные эталонные документы и входной документ, пришедший для автоматической рубрикации, проходят предварительную обработку, которая состоит из следующих этапов:

1. Определение количественной и качественной структуры тематических рубрик;

2. Формирование словарей ключевых словосочетаний для каждой тематической рубрики с повышающим весовым коэффициентом (g = 1,2; 1,4; 1,6; … 3) (например: Российская Федерация, Вооруженные силы, автоматизированная система и т.д.). Повышающий коэффициент для остальных слов будет равен g = 1;

3. Подбор документов и формирование обучающей выборки для обучения тематического рубрикатора;

4. Разделение текста документа на лексемы и выделение термов (корневых основ) с учетом заранее сформированных словарей ключевых словосочетаний;

5. Устранение общих шумовых слов, символов, цифр и т.д.;

6. Расчет весовых коэффициентов термов с учетом всей коллекции информационных образов документов с помощью алгоритма tfc, снижающего размерность итоговой матрицы «терм-документ» путем исключения термов, попадающих во все тематические рубрики;

7. Формирование нормализованной матрицы «терм-документ» из информационных образов эталонных документов;

8. Разложение полученной матрицы «терм-документ» с помощью МЛСА и автоматическое определение диапазона ранговых значений [6];

9. Автоматизированное формирование информационного представления каждой тематический рубрики.

Этап автоматической рубрикации информации по тематическим рубрикам включает в себя:

1. Корреляционный анализ вектора входного документа и векторов матрицы терм-документ;

2. Выбор необходимого критерия для принятия решения о принадлежности документа к тематической рубрике.

Расширенная модель представления содержания и связей разнородной информации, основанная на МЛСА

Смоделируем процесс автоматической рубрикации разнородной информации, основанного на выделении слов и частотах их появления для формирования информационных образов документов в векторно-матричном представлении. При моделировании процесса автоматической рубрикации ключевым моментом является разработка модели представления содержания и связей разнородной информации.

Пусть **A** — матрица, где элемент (i, j) отображает употребление слова i в документе j. Отображение матрицы **A** имеет следующий вид:

$$t_i^T \rightarrow \begin{pmatrix} gx_{1,1} & \cdots & gx_{1,n} \\ \vdots & \ddots & \vdots \\ gx_{m,1} & \cdots & gx_{m,n} \end{pmatrix}, \quad \begin{matrix} d_j \\ \downarrow \end{matrix}$$

где элемент $gx_{i,j}$ представляет собой количество употреблений i-го слова в j-том документе с учетом значения повышающего коэффициента. Каждая строка в этой матрице — это вектор, соответствующий слову и отражающий его наличие в каждом документе $\mathbf{t}_\mathbf{i}^\mathbf{T}$.

$$\mathbf{t}_\mathbf{i}^\mathbf{T} = \left(gx_{i,1} \ \cdots \ gx_{i,n} \right).$$

Аналогично, каждый столбец представляет собой вектор, соответствующий документу и отражающий употребление слов в этом документе $\mathbf{d}_\mathbf{j}$:

$$\mathbf{d}_\mathbf{j} = \begin{pmatrix} gx_{1,j} \\ \vdots \\ gx_{m,j} \end{pmatrix}.$$

Для введения семантической составляющей в процессе рубрикации разнородной информации использован метод ЛСА.

Особенность разложения вида

$$\mathbf{A} = \mathbf{U}\mathbf{W}\mathbf{V}^\mathbf{T},$$

состоит в том, что если в W оставить только k наибольших сингулярных значений (другими словами, количество факторов), а в матрицах **V** и **U** только соответствующие этим значениям столбцы, то произведение получившихся матриц $\mathbf{V}_k$, $\mathbf{W}_k$ и $\mathbf{U}_k$ будет наилучшим приближением исходной матрицы **A** матрицей ранга k.

Если в качестве матрицы **A** использовать матрицу «терм-документ», то матрица $\tilde{\mathbf{A}}$, содержащая только k автоматически отобранных сингулярных значений матрицы **A**, отражает основную структуру ассоциативных зависимостей разнородной информации, присутствующих в исходной матрице [10, 12, 13].

Разложение имеет следующий вид:

$$\mathbf{t}_\mathbf{i}^\mathbf{T} \rightarrow \begin{pmatrix} gx_{1,1} & \cdots & gx_{1n} \\ \vdots & \ddots & \vdots \\ gx_{m,1} & \cdots & gx_{m,n} \end{pmatrix} = \left(\tilde{t}_i^T\right) \rightarrow$$

$$\rightarrow \left[(u_1) \cdots (u_l) \right] * \left[\begin{pmatrix} w_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_l \end{pmatrix} \right] * \left[\begin{matrix} (v_1) \\ \vdots \\ (v_l) \end{matrix} \right],$$

где числа $w_1, \cdots, w_l$ называются сингулярными числами, а $u_1, \cdots, u_l$ и $v_1, \cdots, v_l$ правыми и левыми сингулярными векторами соответственно. Следует заметить, что единственная часть матри-

цы **V**, которая влияет на элементы вектора $\mathbf{t_i}$ — это ее *i*-я строка. Обозначим этот вектор $\tilde{\mathbf{t}}_i$. Аналогично, на $\tilde{d}_j$ влияет только *j*-й столбец $\mathbf{V^T}$, $\tilde{\mathbf{d}}_\mathbf{j}$.

Кроме того, если из всех сингулярных значений отобрать *k* наибольших, то мы получим аппроксимацию исходной матрицы, матрицей ранга *k*. Понижение ранга матрицы «терм-документ» позволяет добиться следующих результатов:

– снизить отрицательное влияние синонимии и полисемии;

– сократить объем вычислений, в результате преобразования размерность матрицы «терм-документ» уменьшается.

Операция уменьшает влияние синонимии, так как понижение ранга «сливает» размерности, связанные со словами, имеющими близкие значения. Также уменьшается влияние полисемии: если многозначное слово имеет «правильное значение», то его элемент «сливается» с матричными элементами слов с таким же значением. Если же слово употреблено в «неправильном» значении, то частотная характеристика соответствующего элемента будет уменьшена.

Следует отметить, что у метода ЛСА существуют некоторые ограничения. В нем не используется информация о порядке слов, и, следовательно, метод не учитывает синтаксические отношения, логику или морфологию. Несмотря на это, результаты метода достаточно достоверно отображают смысловые корреляции между словами и документами [10].

Моделирование процесса автоматической рубрикации на основе МЛСА

Задача автоматической рубрикации разнородной информации определяется следующим образом. Имеется множество объектов *D,* не обязательно конечное, а также множество

$$T = \{t_i\}, i = 1 \ldots N_t,$$

состоящее из N_t рубрик объектов. Каждая рубрика t_i представлена некоторым описанием $\tilde{V}_i$, имеющим некоторую внутреннюю семантическую структуру. Процедура рубрикации f объектов $t \in T$ заключается в выполнении преобразований над ними (формирование информационных образов документов и представление их в виде векторов $\vec{\mathbf{b}}_i$), после которых либо делается вывод о соответствии *t* одной из семантических структур $\tilde{V}_i$, что означает отнесение *t* (т.е. вектора $\vec{\mathbf{b}}_i$) к рубрике t_i, либо вывод о невозможности рубрикации *t.* Элементами множества *T* являются информационные образы электронных версий разнородных документов. Множественная модель автоматического рубрикатора на основе МЛСА может быть представлена алгебраической системой следующего вида:

$$R = < D, T_{\text{с.руб.}}, \tilde{V}_l, R_t, f >,$$

где D — множество разнородных документов, подлежащих автоматическому рубрицированию; $T_{\text{с.руб.}}$ — множество тематик рубрикатора; $\tilde{V}_l$ — множество описаний; R_t — отношение $T_{\text{с.руб.}} \cdot \tilde{V}_l$, т.е. формирование семантического пространства; f — операция семантического рубрицирования вида $D \to T_{\text{с.руб.}}$.

Кроме сформулированной задачи автоматической рубрикации разнородных документов определяется задача обучения рубрикатора, под которой подразумевается частичное или полное формирование T, $\tilde{V}_l$, R_t и f на основе априорных данных.

Характерной чертой автоматической рубрикации на основе МЛСА является универсальность описаний матриц $\tilde{\mathbf{V}}$, которые с одной стороны используются для представления содержания рубрик, а с другой стороны — содержания анализируемых документов. Процедура рубрикации f использует, например, косинусную меру подобия вида F: $\tilde{V}_l * V_d \to [-1;1]$, позволяющую количественно оценивать тематическую близость описаний $\tilde{V}_d \in \tilde{V}$ и $\tilde{V}_l \in \tilde{V}$, где описание $\tilde{V}_d$ представляет содержание анализируемого документа, а $\tilde{V}_l$ — содержание некоторой рубрики. Для вычисления косинусной меры подобия используется скалярное произведение векторов:

$$F\left(\tilde{V}_d, \tilde{V}_l\right) = \cos\left(\tilde{V}_d, \tilde{V}_l\right).$$

Действия процедуры семантической рубрикации *f* сводятся к преобразованию анализируемого документа *d* в представление $\tilde{V}_d \in \tilde{V}$, оценке подобия описания $\tilde{V}_d$ с описаниями рубрик $\tilde{V}$ (вычисление $F\left(\tilde{V}_d, \tilde{V}_l\right)$ и заключению по результатам сопоставления о принадлежности документа только одной рубрике, т.е. из всех $F\left(\tilde{V}_d, \tilde{V}_l\right)$

выбирается максимальная величина, которая и указывает на результирующую рубрику. Такое ограничение введено с целью гарантированного отнесения документа в соответствующую тематическую рубрику ИАС.

Определение принадлежности документа рубрике

Для произвольного объекта u расположим объекты обучающей выборки $\tilde{d}_j$ в матрице описаний $\tilde{\mathbf{V}}$ в порядке возрастания расстояний до u:

$$p(u,\tilde{d}_{1;u}) \le p(u,\tilde{d}_{2;u}) \le \cdots \le p(u,\tilde{d}_{m;u}),$$

где через $\tilde{d}_{j;u}$ обозначается тот объект обучающей выборки, который является j-м соседом объекта u. Аналогичное обозначение введем и для ответа на j-м соседе: $\tilde{d}'_{j;u}$.

Таким образом, произвольный объект u порождает свою переномерацию выборки:

$$a(u) = \arg\max \sum_{i=1}^{m} \left[\tilde{d}_{j;u} = \tilde{d}'_{j;u}\right] w(j,u),$$

при $w(j,u) = [j \le K]$ — метод K-ближайших соседей, где $w(j,u)$ — заданная весовая функция, которая оценивает степень важности j-го соседа для рубрикации объекта u.

Процесс представления разнородной информации в модели МЛСА

Процесс семантической рубрикации вновь поступающих документов представляет собой представление документов в структуре разработанной модели. При этом, чтобы не производить многократные пересчеты сингулярного разложения в МЛСА, необходимо представить новый документ в векторном представлении, но уже после выполнения операции SVD. Опишем данный процесс:

Пусть

$$\mathbf{d}^{\mathbf{new\_doc}} = \begin{pmatrix} gx_i \\ \vdots \\ gx_m \end{pmatrix}$$

— вектор весов, с учетом взвешивания, термов нового информационного образа документа, пришедшего на рубрикацию, при этом $i \in A$, тогда его информационный образ в пространстве семантических признаков будет иметь вид:

$$\tilde{d} = (d^{new\_doc})^T V_k W_k^{-1}.$$

В этом случае мера близости документа $\tilde{d}$ оценивается скалярным произведением векторов $\tilde{\mathbf{V}}_\mathbf{d}, \tilde{\mathbf{V}}_\mathbf{l}$.

Процесс дообучения системы автоматической рубрикации на основе модели МЛСА

Представляет собой процесс, аналогичный самой рубрикации разнородных документов, но имеет отличие в виде записи информационного образа документа в уже сформированное ранее пространство семантических признаков всех документов из обучающей выборки.

Пусть документ

$$\mathbf{d}^{\mathbf{new\_doobuch}} = \begin{pmatrix} gx_i \\ \vdots \\ gx_m \end{pmatrix}$$

— вектор весов термов нового информационного образа документа, сформированного для дообучения (новый столбец матрицы $\tilde{\mathbf{A}}$), при этом $i \in A$, тогда его представление в пространстве семантических признаков можно вычислить по формуле

$$\tilde{d}_j = W_k^{-1} V_k^T d^{new\_doobuch}.$$

После представления документа в математическом виде, т.е. выделения семантических признаков, происходит процесс перенастройки параметров семантического рубрикатора.

Результаты моделирования, оценка эффективности предложенной модели

Результаты предыдущих исследований показали, что применение МЛСА в задачах поиска, рубрикации, выявления дублирующей информации позволяет эффективно выявлять смысловые взаимосвязи между термами. Это позволило повысить автоматизацию решения задачи устранения конфликтов и избыточности информации, повысить точность при поиске и рубрикации информации [1, 2, 6, 7, 9, 11, 13].

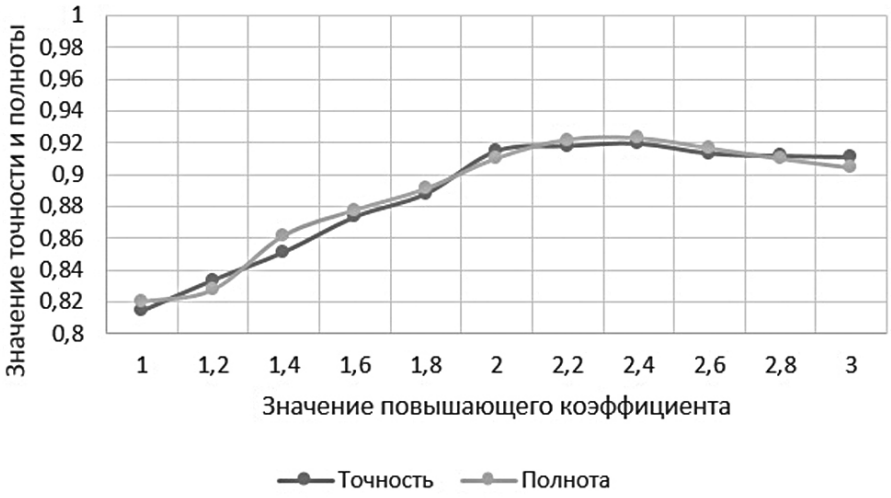

Рис. Зависимость полноты и точности от значения повышающего коэффициента

Используя ранее полученные результаты, была описана и оценена расширенная модель представления содержания и связей информации на основе МЛСА и проведена оценка влияния повышающего коэффициента *g* на общую эффективность системы автоматической рубрикации разнородной информации в распределенной ИАС. Для проведения экспериментальных расчетов использовались: сформированные словари с повышающими весовыми коэффициентами словосочетаний *g*; сформированная обучающая выборка, состоящая из 150 текстовых отобранных документов, для формирования десяти специальных тематических рубрик; сформированная обучающая выборка, состоящая из 1000 текстовых документов, подающихся на вход системы автоматической рубрикации разнородной информации. Результаты оценки системы автоматической рубрикации по полноте и точности представлены на рисунке, которые позволяют сделать вывод о положительном влиянии повышающего коэффициента *g* в диапазоне значений от 2 до 2,6.

Заключение

В работе исследована возможность использования сформированных словарей ключевых словосочетаний. Разработана и описана расширенная модель представления данных на основе модернизированного метода ЛСА с учетом повышающего коэффициента для ключевых словосочетаний.

Проведена оценка влияния значения повышающего коэффициента ключевых словосочетаний на полноту и точность автоматической рубрикации разнородной информации в ИАС. В результате проведенного исследования можно сделать вывод, что использование повышающего коэффициента *g* для решения задачи автоматической рубрикации разнородной информации в защищенных ИАС повысит ее эффективность.

При проведении экспериментов использовалась небольшая выборка из 1150 документов. Вполне вероятно, что при существенном изменении ее объема изменятся и рассматриваемые значения повышающего коэффициента. Проработка этого вопроса требует отдельного исследования.

Литература

1. Басыров А.Г., Бубнов В.П., Забродин А.В. и др. Модели и методы исследования информационных систем. Монография / под общ. ред. А.Д. Хомоненко. — СПб.: Лань. 2019. 204 с.

2. Бубнов В.П. и др. Модели информационных систем: учеб. пособие // — М.: ФГБОУ «Учебно-методический центр по образованию на железнодорожном транспорте». 2015. 188 с.

3. Galitsky B., Ilvovsky D., Kuznetsov S.O. Style and Genre Classification by Means of Deep Textual Parsing // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialogue 2016». 2016. P. 171–181.

4. Gupta M., Bendersky M. Information Retrieval with Verbose Queries // Foundations and Trends in Information Retrieval. 2015. Vol. 9. № 3–4. P. 209–354.

5. Witten I.H., Frank E., Hall M.A. Data Mining: Practical Machine Learning Tools and Techniques: 3rd edition // Morgan Kaufmann. 2011. 664 p.

6. Илатовский А.С., Хомоненко А.Д., Арсеньев В.Н. и др. Оценка семантической близости документов на основе латентно-семантического анализа с автоматическим выбором ранговых значений // Труды СПИИРАН. 2017. № 5 (54). С. 185–204.

7. Краснов С.А. О возможности смыслового анализа информации для выявления информационных интересов пользователей // Вестник Российского нового университета. Серия: Сложные системы: модели, анализ и управление. 2019. № 2. С. 157–163.

8. Краснов С.А., Еремин А.С. и др. Оценка оперативности автоматической рубрикации документов с помощью модели нестационарной системы обслуживания с эрланговским распределением длительности интервалов между запросами // Проблемы информационной безопасности. Компьютерные системы. 2012. № 3. С. 14–21.

9. Хомоненко А.Д., Логашев С.В. и др. Автоматическая рубрикация документов с помощью латентно-семантического анализа и алгоритма нечеткого вывода Мамдани // Труды СПИИРАН. 2016. № 1 (44). С. 5–19.

10. Dumais S. Latent semantic indexing: TREC-3 report // Proc. of the Third Text REtrieval Conference. 1995. P. 219–230.

11. Соловьев А.Н. Моделирование процессов понимания речи с использованием латентно-семантического анализа: диссертация на соискание степени к.ф-м.н. — Санкт-Петербург: СПбГУ. 2008. 165 с

12. Хомоненко А.Д., Дашонок В.Л. и др. Выявление противоречий в семантически близкой информации на основе латентно-семантического анализа // Проблемы информационной безопасности. Компьютерные системы. 2014. № 2. С. 73–84.

13. Хомоненко А.Д. Применение метода латентно-семантического анализа для автоматической рубрикации документов / А.Д. Хомоненко, С.А. Краснов // Известия Петербургского университета путей сообщения. 2012. Вып. 2 (31). С. 124–132.

14. Краснов С.А., Яковлев Я.В. и др. Оценка эффективности применения алгоритма вычисления коэффициента ранговой корреляции Спирмена в методе латентно-семантического анализа при автоматической рубрикации документов // Бюллетень результатов научных исследований. 2012. № 2 (3). С. 153–162.

15. Бубнов В.П., Краснов С.А., Еремин А.С. и др. Модель функционирования системы автоматической рубрикации документов в нестационарном режиме // Проблемы информационной безопасности. Компьютерные системы. 2011. № 4. С. 16–23.

16. Войцеховский С.В., Калиниченко С.В., Уланов А.В. и др. Модель оценивания оперативности обработки устаревающей информации // Научное обозрение. 2014. № 3. С. 155–157.

17. Хомоненко А.Д. и др. Системы искусственного интеллекта: Практикум. — СПб.: ВКА им. А.Ф. Можайского. 2014. 69 с.

18. Уланов А.В., Матвеев С.В. и др. Анализ оперативности обработки информации с ограниченным временем актуальности // Бюллетень результатов научных исследований. 2013. № 3 (8). С. 39–47.

References

1. Basyrov A.G., Bubnov V.P., Zabrodin A.V. et al. Models and methods of researching information systems. Monograph / under total. ed. HELL. Khomonenko. — SPb: Lan. 2019 . 204 p.

2. Bubnov V.P. and other Models of information systems: textbook. manual // — M.: FGBOU «Training and methodological center for education in railway transport». 2015.188 p.

3. Galitsky B., Ilvovsky D., Kuznetsov S.O. Style and Genre Classification by Means of Deep Textual Parsing // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialogue 2016». 2016. P. 171–181.

4. Gupta M., Bendersky M. Information Retrieval with Verbose Queries // Foundations and Trends in Information Retrieval. 2015. Vol. 9. № 3–4. Pp. 209–354.

5. Witten I.H., Frank E., Hall M.A. Data Mining: Practical Machine Learning Tools and

Techniques: 3rd edition // Morgan Kaufmann. 2011. 664 p.

6. Ilatovskiy A.S., Khomonenko A.D., Arsen'-ev V.N. et al. Assessment of the semantic proximity of documents based on latent-semantic analysis with automatic selection of rank values // Proceedings of SPIIRAS. 2017. № 5 (54). P. 185–204.

7. Krasnov S.A. On the possibility of semantic analysis of information to identify information interests of users // Bulletin of the Russian New University. Series: Complex systems: models, analysis and management. 2019. № 2. P. 157–163.

8. Krasnov S.A., Eremin A.S. et al. Evaluation of the efficiency of automatic rubrication of documents using a model of a non-stationary service system with Erlang distribution of the duration of intervals between requests // Problems of information security. Computer systems. 2012. № 3. P. 14–21.

9. Khomonenko A.D., Logashev S.V. et al. Automatic rubrication of documents using latent semantic analysis and the Mamdani fuzzy inference algorithm // Proceedings of SPIIRAS. 2016. № 1 (44). P. 5–19.

10. Dumais S. Latent semantic indexing: TREC-3 report // Proc. of the Third Text REtrieval Conference. 1995. Pp. 219–230.

11. Soloviev A.N. Modeling the processes of understanding speech using latent semantic analysis: dissertation for the degree of Ph.D. — St. Petersburg: St. Petersburg state un-t. 2008. 165 p.

12. Khomonenko A.D., Dashonok V.L. et al. Revealing contradictions in semantically close information on the basis of latent-semantic analysis // Problems of information security. Computer systems. 2014. № 2. P. 73–84.

13. Khomonenko A.D. Application of the method of latent-semantic analysis for automatic rubrication of documents / A.D. Khomonenko, S.A. Krasnov // Bulletin of the Petersburg University of Railways. 2012. Issue. 2 (31). P. 124–132.

14. Krasnov S.A., Yakovlev Ya.V. et al. Evaluation of the efficiency of using the algorithm for calculating the Spearman's rank correlation coefficient in the method of latent-semantic analysis for automatic rubrication of documents // Bulletin of scientific research results. 2012. № 2 (3). P. 153–162.

15. Bubnov V.P., Krasnov S.A., Eremin A.S. et al. Model of functioning of the system of automatic rubrication of documents in non-stationary mode // Problems of information security. Computer systems. 2011. № 4. P. 16–23.

16. Voitsekhovsky S.V., Kalinichenko S.V., Ulanov A.V. et al. Model for evaluating the efficiency of processing obsolete information // Scientific Review. 2014. № 3. P. 155–157.

17. Khomonenko A.D. et al. Artificial Intelligence Systems: Workshop. — SPb: VKA im. A.F. Mozhaisky. 2014 . 69 p.

18. Ulanov A.V., Matveev S.V. et al. Analysis of the efficiency of information processing with limited time of relevance // Bulletin of scientific research results. 2013. № 3 (8). P. 39–47.