УДК 004.912 **DOI** 10.15622/sp.54.8

С.А. Краснов, А.С. Илатовский, А.Д. Хомоненко, В.Н. Арсеньев

ОЦЕНКА СЕМАНТИЧЕСКОЙ БЛИЗОСТИ ДОКУМЕНТОВ НА ОСНОВЕ ЛАТЕНТНО-СЕМАНТИЧЕСКОГО АНАЛИЗА С АВТОМАТИЧЕСКИМ ВЫБОРОМ РАНГОВЫХ ЗНАЧЕНИЙ

Краснов С.А., Илатовский А.С., Хомоненко А.Д., Арсеньев В.Н. **Оценка семантической близости документов на основе латентно-семантического анализа с автоматическим выбором ранговых значений.**

Аннотация. Предлагается метод оценки семантической близости документов на основе латентно-семантического анализа, учета динамики изменения сингулярных значений матрицы «терм-документ» и автоматического определения диапазона ранговых значений. Оценка семантической близости документов рассматривается применительно к решению задач выявления дублирования и противоречий в базах данных.

Приводится краткий обзор подходов, используемых при оценке семантической близости документов, выявлении дублирования и противоречий в базах данных и хранилищах данных. Приводятся результаты численных примеров оценки семантических зависимостей между термами документов в интересах выявления дублирования и противоречий в базах данных. При этом в качестве результирующей характеристики рассчитывается степень соответствия λ сравниваемых документов.

Приведены сравнительные оценки расчета степени соответствия λ документов с помощью основных методов (косинусной меры близости, векторной модели, коэффициента ранговой корреляции Спирмена, статической меры tf-idf — частота термина — обратная документная частота).

Показано, что использование предложенного метода анализа динамики изменения сингулярных чисел матрицы «терм-документ» с автоматическим выбором диапазона используемых ранговых значений позволяет устранить зависимость метода латентно-семантического анализа от выбора оптимального ранга.

Ключевые слова: оценка семантической близости документов, выявление дублирования и противоречий, базы данных, латентно-семантический анализ, статистический анализ, косинусная мера близости, векторная модель.

1. **Введение.** В настоящее время наблюдается большой рост объемов информации, хранимой и обрабатываемой в базах данных (БД) и информационных системах, сетях Интернет и Интранет. При этом достаточно важной и распространенной является задача выявления и устранения дублированной и противоречивой информации. В основе ее решения лежит теория интеллектуального анализа данных [1, 2].

При решении указанной задачи и других задач интеллектуальной обработки рассматриваются различные варианты прикладных областей и разные подходы. В частности, в [3] предлагаются методы выявления дублированной и противоречивой информации в Википедии. В [4, 5] для устранения конфликтов на

уровне семантики используется онтологический подход. В [6, 7] предлагаются решения задач разрешения противоречий в электронной библиотеке eLIBRARY.ru и автоматизированной классификации документов с использованием нечеткого логического вывода по алгоритму Мамдани. В [8] решается задача автоматического рустранения конфликтов в БД реляционного типа. Предлагаемые подходы служат для обнаружения конфликтов именования и структурных конфликтов или для предотвращения конфликтов путем введения более детализированных описаний связей между объектами, запрещения определенных действий. В [9] рассматривается задача стилевой и жанровой классификации на основе глубокого парсинга (разбора) текста.

В статье [10] представлено разрешение конфликтов и управление репликацией при интеграции данных из нескольких источников с помощью объектной модели. Разрешение конфликтов выполняется с помощью операций переключения и вспомогательной алгебры объекта. В работе [11] выявляются противоречия в тексте путем отыскания конфликтных операторов как основы информационного анализа текста. Вводится определение противоречия и предлагается их типология. Отмечается достижение хороших характеристик производительности анализа текста на основе предложенного подхода с использованием отрицания и антонимии.

В числе современных работ в области интеллектуального анализа данных отметим следующие. В [12] затрагиваются вопросы применения библиотеки BigARTM с открытым кодом для тематического моделирования больших текстовых коллекций. В [13] дается математическое обоснование метода категориальных векторов, обеспечивающего повышение эффективности поисковых информационных запросов. В [14] систематически изложены вопросы информационного поиска с подробными запросами.

Решение задачи выявления и устранения дублирования и противоречий в БД и многих других задач, например информационного поиска, кластеризации, классификации (рубрикации) и фильтрации, часто требует оценки семантической близости документов. При этом широко используется метод латентно-семантического анализа (ЛСА) [15-19].

Целью настоящей статьи является повышение *точности* оценки СБД с помощью метода ЛСА для выявления дублирования и противоречий в БД и информационных системах. Дублирования и противоречия в БД могут возникать из-за ошибок ввода и

редактирования данных, выполняемых человеком; ошибок, возникающих в процессе передачи данных между БД в распределенных информационных системах и так далее.

Современные системы управления базами данных (СУБД) реляционного типа имеют механизмы защиты от возникновения избыточности данных и соответствующих аномалий с помощью нормализации отношений [17]. Средства СУБД также позволяют обеспечить целостность БД.

Противоречия (конфликты) в БД разделяют в основном на три группы:

1. Конфликты наименований — из-за использования различной терминологии в процессе описания некоторой предметной области, приводят к омонимии и синонимии.

2. Структурные конфликты — связаны с различными представлениями одного и того же объекта в различных источниках.

3. Семантические конфликты — возникают, когда данные противоречивы.

Выявляемые в статье противоречия охватывают все три группы конфликтов в той или иной мере. Под такими противоречиями понимаем наличие близких, но не совпадающих значений одних и тех же полей БД или документов, что может приводить к неоднозначности их интерпретации.

Подход к выявлению и устранению дублированной и противоречивой информации описан в работе [20], где показана необходимость провести оценку смысловой близости документов с помощью метода ЛСА. Отметим, что в статье [21] приводится обзор работ по оценке близости текстов. В ней выделяются три подхода: основанные на строках, основанные на теле документа и основанные на знаниях. В статье [22] рассматривается метод сравнения текстов на русском языке для определения семантической близости. Дается обзор существующих методов сравнения. Предложен метод определения степени подобия между текстовыми пассажами в пределах семантического класса.

Для достижения поставленной нами цели требуется комплексное применение исходных и разработанных методов в области семантического анализа информации [23-26].

Метод ЛСА позволяет получить более точные результаты по отношению к простой векторной модели, однако без предварительной обработки данных метод ЛСА обладает

недостаточной точностью для выявления дублирования и противоречий в БД. Кроме того, у метода ЛСА отсутствует универсальный способ автоматической аппроксимации исходной матрицы матрицей, содержащей только k значимых сингулярных значений, отражающих основную структуру ассоциативных зависимостей исходной матрицы. Отсюда следует необходимость решения задачи выбора оптимального ранга аппроксимирующей матрицы [7, 17, 19, 20].

2. Этапы выявления дублирования и противоречий в БД. Как отмечалось, для выявления и устранения дублирования и противоречий в текстовых данных требуется провести оценку их смысловой близости с помощью метода ЛСА. Кроме того, ЛСА измеряет корреляционные зависимости типа «терм-терм», «терм-вектор» и «вектор-вектор». Результативность данного метода зависит не только от частот использования слов (термов) в документах, но и от выявления более глубоких (скрытых) связей [7, 23].

Для устранения факторов, затрудняющих выявления дублирования и противоречий в БД с высокой точностью, предлагается комплексное решение по подбору параметров метода, опирающееся на результаты [20] и основанное на комбинации:

– предложенных методов предварительной обработки данных, повышающих качество формирования векторной модели данных;

– метода ЛСА с анализом динамики изменения сингулярных значений матрицы «терм-документ», дающего возможность автоматически выделить диапазон ранговых значений сингулярных чисел;

– косинусной меры близости (КМБ) на основе скалярного произведения векторов, позволяющей провести оценку смысловой близости сравниваемых данных.

В результате анализа лингвистических и статистических методов предварительной обработки данных определены факторы, влияющие на точность выявления дублирования и противоречий в БД с помощью метода ЛСА:

– оценка важности слова в контексте документов (статистические меры);

– нормирование векторов (НВ);

Представлением документа (в векторном пространстве термов) будет вектор вещественных чисел, $\vec{d}_i = (d_{i1},...,d_{i|m|})^T$, в котором каждое вещественное число будет координатой вектора, соответствовать

конкретному термину и иметь значение, равное весу термина в соответствующем документе. При оценке СБД с помощью метода ЛСА исследовано влияние статической меры *tf-idf* и процедуры НВ:

$$d_{ij} = \frac{w_{ij}}{\left|\overrightarrow{w_{ij}}\right|}, \quad w_{ij} = tf_{ij} \times \log \frac{|D|}{df_j}, \quad tf_{ij} = \frac{n_i}{\sum_k n_k}, \tag{1}$$

где d_{ij} — нормированные веса по *tf-idf*, (частота встречаемости термина —— обратная документная частота), $0 \le d_{ij} \le 1$;

$|D|$ — число анализируемых документов;

n_i — количество употреблений слова в документе;

$\sum_k n_k$ — общее количество слов, содержащихся в документе;

tf_{ij} — частота встречаемости слова в документе (число раз, которое *j*-е слово встретилось в *i*-м документе);

df_j — документная частота (число документов, в которых встретилось *j*-ое слово);

$\left|\overrightarrow{w_{ij}}\right|$ — нормированный вектор w_{ij} в евклидовом пространстве.

В *tf-idf* наибольший вес получают слова с высокой частотой в пределах документа и с небольшой частотой встречаемости в других документах. При НВ матрицы «терм-документ» скалярное произведение не зависит от нормы векторов. Это позволяет упростить сравнение результатов скалярных произведений. Операция нормирования производится перед расчетом λ — степени соответствия документов.

Существенным фактором, влияющим на точность выявления дублирования и противоречий в БД при использовании метода ЛСА, является определение оптимального ранга аппроксимирующей матрицы [7, 12, 20]. Поэтому анализируется влияние сингулярных значений на точность выявления дублирования и противоречий в БД и предлагается метод анализа динамики изменения сингулярных значений матрицы «терм-документ» с автоматическим определением диапазона используемых ранговых значений. Этапы процесса выявления дублирования и противоречий представлены в таблице 1.

Таблица 1. Этапы выявления дублирования и противоречий в БД

Этап	Описание
1. Лингвистический анализ	
Лексический анализ; морфологический анализ.	Разделение текста документа на лексемы и выделение термов (для проведения экспериментов использовался алгоритм стемминга М. Портера).
2. Статистический анализ	
Частотный анализ и вычисление весовых коэффициентов термов; нормализация векторов.	Подсчет числа повторений термов; получение весовых характеристик; формирование матрицы «терм-документ».
3. ЛСА	
Сингулярное разложение матрицы «терм-документ»; определение значимого диапазона значений рангов рассматриваемой матрицы; получение аппроксимированных матриц с меньшими рангами.	Выбор оптимального диапазона ненулевых сингулярных значений на основе метода анализа динамики изменения сингулярных чисел анализируемой матрицы с автоматическим выбором диапазона ранговых значений.
3. Анализ соответствия векторов	
Оценка косинусной меры близости (λ векторов); вычисление результирующего λ полученных значений; задание порогового значения λ; сравнение полученного значения λ с пороговым значением.	Вычисление скалярных произведений векторов для каждого значимого ранга; вычисление среднего арифметического для каждого значимого ранга; выявление дублирования и противоречий в БД.

Для экспериментальной части работы использован язык программирования Python 3.4 и его библиотеки с открытым исходным кодом: NumPy и SciPy. Такой выбор обусловлен тем, что в них имеется поддержка больших многомерных массивов, матриц и высокоуровневых математических функций для операций с ними.

3. Метод анализа динамики изменения сингулярных чисел матрицы «терм-документ» с автоматическим выбором диапазона используемых ранговых значений. На первом шаге необходимо сформировать матрицу A «терм-документ», которая опишет анализируемые документы и будет содержать исходные данные для метода ЛСА. Ее элементы будут содержать веса термов, полученные

после применения статистической меры *tf-idf* (отношение частоты слов к инверсной частоте документа) $0 \le d_{ij} \le 1$ и НВ матрицы. Далее необходимо выполнить сингулярное разложение матрицы A в произведение трех матриц:

$$A = UWV^T, \tag{2}$$

где U и V — унитарные матрицы, которые состоят из левых и правых сингулярных векторов, а W — матрица с неотрицательными элементами на диагонали, которые называются сингулярными значениями матрицы A.

Согласно теореме Эккарта-Янга, если в матрице W оставить только наибольшие сингулярные значения σ, а в матрицах U и V — соответствующие этим значениям столбцы, то матрицы U_σ и V_σ будут лучшими их приближениями, отражающими ассоциативные зависимости представления термов и документов в пространстве размерности σ [17, 20].

Следующим шагом является оптимальный выбор ненулевых сингулярных значений $\sigma_1 \ge \sigma_2 \ge ... \ge \sigma_n > 0$ матрицы A, которые влияют на результат выявления дублирования и противоречий в БД. Результатом приведения матрицы A к рангам, имеющим близкие к нулю сингулярные значения, являются равные матрицы, учет которых ведет к увеличению вычислительной сложности метода ЛСА и тем самым снижается оперативность выявления и устранения дублирования и противоречий в БД. Чтобы решить задачу оптимального выбора сингулярных значений, предлагается описанный ниже эвристический метод выбора значимых рангов, который является сущностью метода анализа динамики изменения сингулярных чисел матрицы «терм-документ» с автоматическим выбором диапазона используемых ранговых значений.

Определим функцию $f(i) = \sigma_i, \quad i \in N, \quad i < P,$ где P — количество документов. Значимыми рангами являются только ранги $r_p, r_{p+1}, ..., r_{m-1}, r_m, p \le m,$ заключенные между соответствующими сингулярными значениями $\sigma_p \ge \sigma_m \ge 0,$ претерпевающими резкое изменение $\Delta\sigma_i = \sigma_i - \sigma_{i-1}, i \in \{\mathrm{p}; \mathrm{m}\}$ относительно предыдущих сингулярных значений $\sigma_i \ge \sigma_p, i \le p; \sigma_i \ge \sigma_m, i \le \mathrm{m}.$

Определение границ значимых рангов осуществляется с помощью понятия производной функции сингулярных значений $f'(\mathrm{i}) = \sigma_i - \sigma_{i-1}, i \in N, i < n.$

Далее осуществляется поиск максимального значения производной функции $f'_{\max}$, достижимого при σ_{max}, $1 < \max \le \frac{n}{2}$.

Затем определяется первый локальный минимум σ_p, следующий за σ_{max}. Ранги $r_1, r_2, ..., r_i, ..., r_{p-1}, r_p, i \le p$, соответствующие сингулярным значениям, большим σ_p, признаются незначимыми.

В качестве правой границы значимых рангов выбирается ранг r_n, соответствующий последнему ненулевому сингулярному значению σ_n.

На заключительном этапе необходимо рассчитать λ документов, используя КМБ:

$$\cos(\vec{\mathrm{X}}_j, \vec{\mathrm{X}}_i) = \sum_{i=1}^{M} x_j^{(\mathrm{i})} x_i^{(\mathrm{i})}, \tag{3}$$

где $x_j^{(\mathrm{i})} x_i^{(\mathrm{i})}$ — элементы разных векторов, между которыми вычисляется мера близости; M — размерность пространства векторов.

Значения КМБ ограничены промежутком [-1;1] при использовании операции НВ. Степень соответствия $\lambda_{j,i}^{l}$ векторов $\vec{X}_j, \vec{X}_i (\mathrm{i} < \mathrm{j} \le \mathrm{P})$ вычисляется для каждого значимого ранга $r_l, p \le l \le m$. Далее необходимо вычислить результирующую $\overline{\lambda}_{j,l}$ векторов $\vec{X}_j, \vec{X}_i$:

$$\overline{\lambda}_{j,l} = \frac{\sum_{l=p}^{m} \lambda_{j,i}^{l}}{m - p + 1}. \tag{4}$$

Таким образом, получаем результирующую степень соответствия для конкретной пары документов по всему значимому диапазону ранговых значений. Полученное значение необходимо сравнить, например, с пороговым значением для автоматического или автоматизированного принятия решения по устранению дублирования или противоречия в БД.

Отметим, что метод автоматического определения диапазона используемых ранговых значений позволяет с большей точностью гарантировать, что данные действительно дублированные или противоречивые, потому что значения λ всех пар векторов

оцениваются на каждом значимом ранге. При этом случайные всплески полученных значений λ при неправильных ранговых значениях сглаживаются, а значения λ явно дублированных или противоречивых данных стремятся к единице, так как находится среднее арифметическое по всем значениям λ одной пары векторов матриц, полученных из диапазона используемых ранговых значений.

Реализуемость рассмотренных нами этапов выявления дублирования и противоречий в БД подтверждается представленными результатами экспериментов, а также результатами, изложенными в других источниках, например [20, 27].

4. Сравнительная оценка точности выявления дублирования и противоречий в БД. Оценка точности выявления дублирования и противоречий в БД производится путем сравнения полученных результатов с эталоном по единой выборке. В качестве эталона используется выборка с заранее известной дублированной и противоречивой информацией, размеченной вручную. Для примера рассмотрим выборку *D* записей электронной библиотеки eLIBRARY.ru. Фамилии авторов изменены (таблицы 1, 2). Полная выборка содержит 60 документов и включает 7 пар дублированных и 3 пары схожих документов. Схожие документы, как и дублированные, имеют минимальные расхождения, поэтому в рамках этой модели рассматриваются как противоречивые.

Таблица 1. Образцы дублированных документов выборки *D*

1	*Модель функционирования системы автоматической рубрикации документов в нестационарном режиме. Иванов А.Д.,* Кубнов В.П., *Петров С.А*., Юремин А.С. Проблемы информационной безопасности. Компьютерные системы. 2011. № 4. С. 16.
2	*Модель функционирования системы автоматической рубрикации документов в нестационарном режиме. Петров С.А.,* Юремин А.С., *Иванов А.Д*., Кубнов В.П.. Проблемы информационной безопасности. Компьютерные системы. 2011. № 4. С. 16-23.
3	*Иванов А.Д., Петров С.А. Применение метода латентно-семантического анализа для автоматической рубрикации текстов в системах электронного документооборота.* Сборник материалов первой международной научно-практической конференции «*Интеллектуальные системы на транспорте*» (ИнтеллектТранс-2011)/СПб.: ПГУПС, 2011. С. 291-294.
4	*Применение метода латентно-семантического анализа для автоматической рубрикации текстов в системах электронного документооборота: доклад Петров С.А., Иванов А.Д.* В сборнике: *Интеллектуальные системы на транспорте*. Сборник материалов Первой международной научно-практической конференции. Редактор А. А. Порниенко. 2011. С. 291-294.

Таблица 2. Образцы схожих документов выборки *D*

1.	Информационная система геодезического мониторинга транспортных сооружений. Дрынь М.Я., Пикитчин А.А., Ледянин В.В., Иванов А.Д. *В сборнике:* Интеллектуальные системы на транспорте. ИнтеллектТранс-2012 сборник материалов II Международной научно-практической конференции. Федеральное агентство железно-дорожного транспорта, ФГБОУ ВПО "Петербургский государственный университет путей сообщения"; Редактор: Порниенко А.А.. 2012. *С. 150-154.*
2.	Информационная система геодезического мониторинга транспортных сооружений. Дрынь М.Я., Пикитчин А.А., Ледянин В.В., Иванов А.Д. *В книге:* Интеллектуальные системы на транспорте. ИнтеллектТранс-2012 сборник материалов II Международной научно-практической конференции. Федеральное агентство железно-дорожного транспорта, ФГБОУ ВПО "Петербургский государственный университет путей сообщения"; Редактор: Порниенко А.А. 2012. *С. 36.*
3.	Анализ и интеграция диагностической информации АСУ в условиях катастрофических отказов: доклад. *Салиниченко С.В., Иванов А.Д. В сборнике:* Проблемы математической и естественно-научной подготовки в инженерном образовании сборник трудов *II* Международной научно-методической конференции. *Редактор: В. А. Родаковский.* 2013. *С. 99-103.*
4.	Анализ и интеграция диагностической информации АСУ в условиях катастрофических отказов. *Иванов А.Д., Салиниченко С.В. В книге:* Проблемы математической и естественно-научной подготовки в инженерном образовании Тезисы докладов *2-ой* международной научно-методической конференции. 2012. *С. 168-169.*

В качестве меры точности выявления дублирований и противоречий в БД нами используется оценка семантической близости сравниваемых пар документов.

4.1. О применении метода ЛСА. Результаты исследований показали низкую точность выявления дублирования и противоречий в БД (рисунок 1) с помощью простой векторной модели (ПВМ) без использования метода ЛСА.

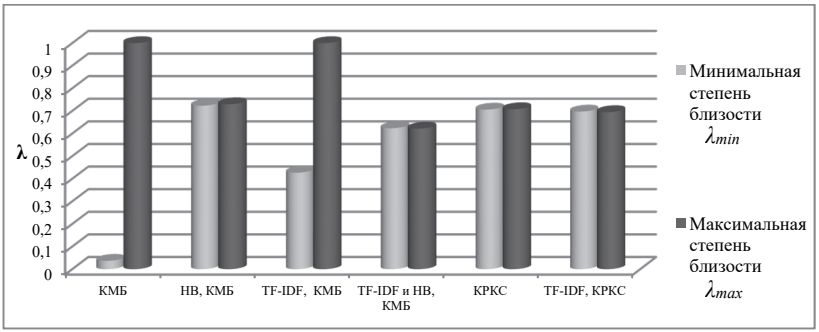

Рис. 1. Оценка семантической близости пар документов без метода ЛСА

То есть необходимо, чтобы левый цилиндр в каждой подгруппе сравниваемых методов был значительно выше правого (рисунок 1). Простая векторная модель не позволяет получить значение Δ, требуемое для выявления дублирования и противоречий в БД. Для наглядности результатов применения ПВМ без операции НВ в матрицы A проведено нормирование полученных результатов: каждое значение ряда заменено отношением к наибольшему по модулю значению ряда.

Операция нормирования при использовании коэффициента ранговой корреляции Спирмена (КРКС) [27] не меняет результатов применения метода, так как порядок расположения отсортированных элементов векторов матрицы A не изменяется.

4.2. Влияние предварительной обработки данных на точность метода ЛСА. На основе выборки D проведен анализ эффективности метода ЛСА в зависимости от различных комбинаций методов предварительной обработки данных (рисунок 2), из которого следует, что применение статической меры *tf-idf* в совокупности с НВ дает наилучший результат, так как удается достичь устойчивых значений Δ на достаточно продолжительном участке ранговых значений. На рисунке 3 представлены значения Δ для каждого ранга. В качестве меры близости выбрано скалярное произведение векторов. Для наглядности проведено нормирование результатов.

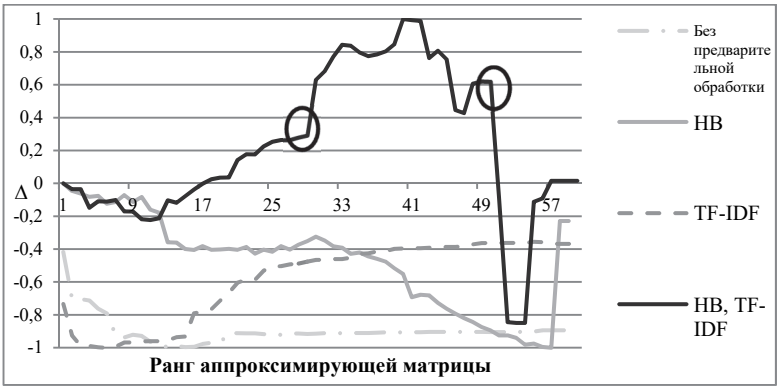

Рис. 2. Точность комбинаций методов предварительной обработки данных

Предположительно, точки изменения характера монотонности графика функции, отмеченные кругами черного цвета, ограничивают промежуток значений рангов, лучше всего подходящих для выявления дублированной и противоречивой информации. Далее выполняется обоснование метода выявления таких точек.

4.3. Влияние мер близости на точность метода ЛСА. Проведен анализ применения методов расчета λ документов при применении метода ЛСА (рисунок 3): с помощью КМБ и с помощью КРКС [27].

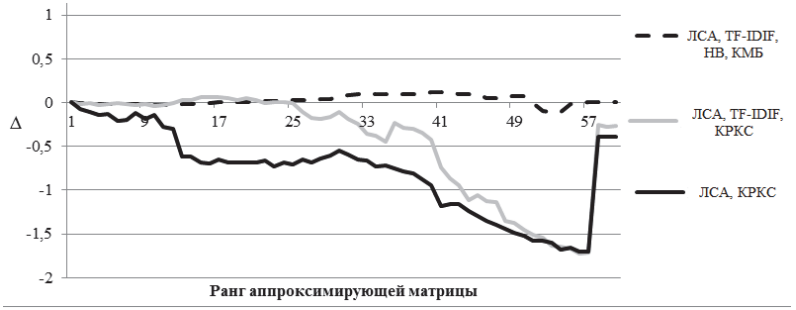

Рис. 3. Точность различных методов расчета λ документов

На рисунке 3 для каждого ранга приведено найденное значение Δ . Анализ показал, что применение КМБ дает более устойчивые к изменению ранга аппроксимирующей матрицы результаты. Устойчивость к изменению ранга позволяет выбрать более широкий диапазон сингулярных значений без критического падения точности метода.

4.4. Оценка точности применения ЛСА с анализом динамики изменения сингулярных чисел матрицы «терм-документ» и автоматическим выбором диапазона ранговых значений. Рассмотрим алгоритм выбора значимых рангов на примере выборки D (рисунок 4). В качестве правой границы значимых рангов примем ранг r_m, соответствующий последнему ненулевому значению σ_m функции $f(\mathrm{i})$.

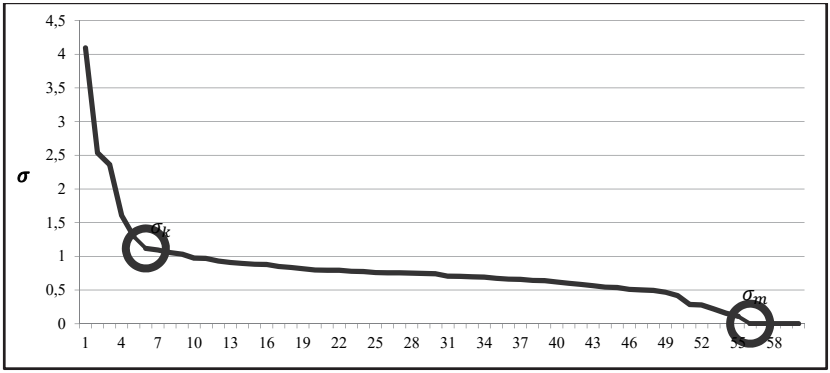

Рис. 4 . Сингулярные значения

Для определения левой границы r_k значимых рангов рассмотрим график производной функции сингулярных значений $f'(\mathrm{i})$ (рисунок 5). Резкая смена характера убывания функции сингулярных значений $f(\mathrm{i})$ при $i = k$ (рисунок 5) соответствует первому локальному минимуму $\Delta\sigma_k$ (рисунок 5), следующему за максимальным значением $\Delta\sigma_{\max}$ функции $f'(\mathrm{i})$ на интервале $(1;\frac{n}{2})$. Ранг r_k, соответствующий сингулярному значению σ_k, примем за левую границу значимых рангов.

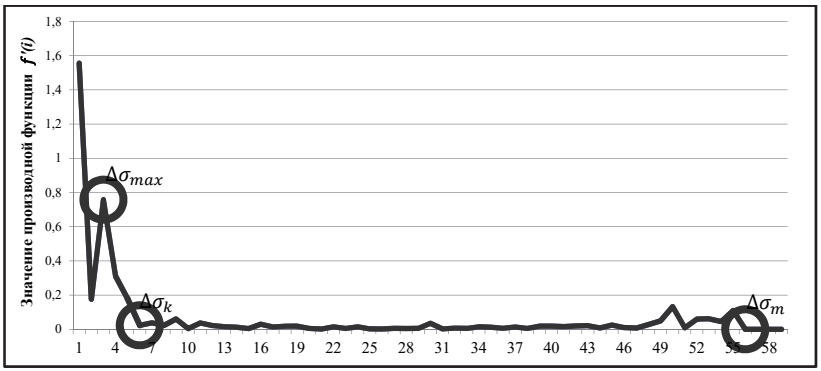

Рис. 5. Производная функция сингулярных значений

Оценим качество подобранных параметров метода ЛСА для выявления дублирования и противоречий в БД (рисунок 6) на основе оценки семантической близости документов. Применение предварительной обработки данных (*tf-idf*, НВ), ранга аппроксимирующей матрицы (ранг 40), методов определения λ документов (КМБ) позволяет получить значение $\Delta_{lsa,40} = 0.120$, которое незначительно превышает значение $\Delta_{mlsa} = 0.103$, полученное с помощью разработанного метода анализа динамики изменения сингулярных чисел матрицы «терм-документ» с автоматическим выбором диапазона используемых ранговых значений.

Оптимальный ранг аппроксимирующей матрицы найден с помощью графика значений Δ (рисунок 2). Значение Δ возможно получить только при заведомо известных дублированных и противоречивых парах документов. Кроме того, значение $\Delta_{lsa,1-60} = 0,086$, полученное при выборе всего рангового диапазона, менее показательно. Таким образом, разработанный метод позволяет

автоматически выбрать диапазон значимых рангов, уменьшить объем вычислений и сохранить точность метода ЛСА.

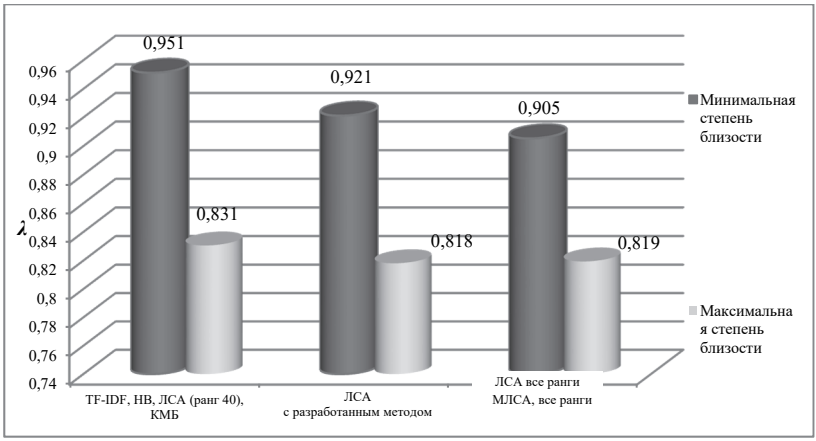

Рис. 6. Точность метода ЛСА с разработанным методом

Рассмотрим кривую распределения количества пар документов, превышающих определенное значение λ (рисунок 7).

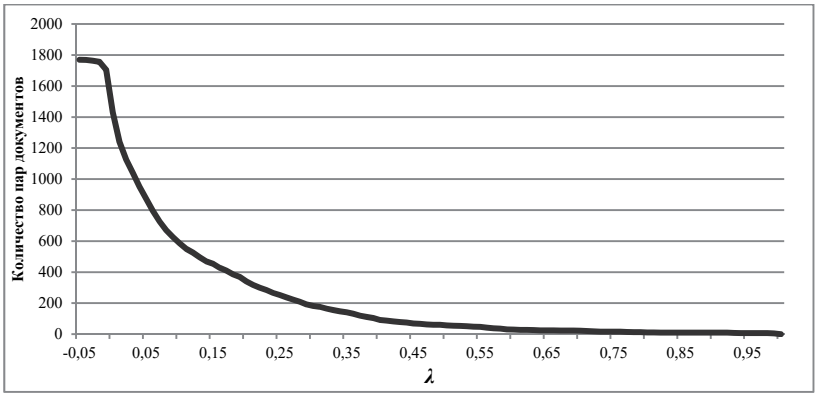

Рис. 7. Распределение пар документов по степени соответствия

Приведенная на рисунке 7 кривая имеет экспоненциальный вид, что дает возможность в автоматическом режиме значительно уменьшить область поиска дублирования и противоречий в БД: с 1770 пар документов до 10 пар документов при пороговом значении $\lambda_{\text{пор.}} = 0,85$. Найденные 10 пар документов являются дублированными или противоречивыми.

Большое значение величины $\Delta_{mlsa} = 0{,}103$ и экспоненциальный характер распределения позволяют выявить дублирования и противоречия в БД с меньшими временными затратами, так как обеспечивается максимальное значение разности между минимальным значением $\lambda_{\min}$ пары документов, содержащих дублирования и противоречия, и максимальным значением $\lambda_{\max}$ пары документов, не содержащих дублирования и противоречий.

5. Заключение. Проведенные эксперименты подтверждают, что на точность выявления дублирования и противоречий в БД на основе оценки семантической близости документов с помощью метода ЛСА существенное влияние оказывают следующие факторы: применение статистической меры tf-idf, нормирование векторов при формировании матрицы «терм-документ», определение оптимального ранга аппроксимирующей матрицы, применение косинусной меры близости.

Использование метода анализа динамики изменения сингулярных чисел матрицы «терм-документ» с автоматическим выбором диапазона используемых ранговых значений при латентно-семантическом анализе совместно с методами предварительной обработки данных позволяет обеспечить высокую точность поиска дублирования и противоречий в БД, а так же снизить временные затраты на их поиск и устранение. При проведении экспериментов использовалась небольшая выборка из 1770 документов. Вполне вероятно, при существенном изменении объема тестовой БД изменятся и рекомендуемые значения параметров. Решение этого вопроса требует отдельного исследования.

Литература

1. *Witten I.H., Frank E., Hall M.A.* Data Mining: Practical Machine Learning Tools and Techniques: 3rd edition // Morgan Kaufmann. 2011. 664 p.
2. *Паклин Н. Б., Орешков В. И.* Бизнес-аналитика: от данных к знаниям (+ CD) // СПб.: Изд. Питер. 2009. 624 с.
3. *Weissman S., Ayhan S.,* Bradley *J., Lin J.* Identifying Duplicate and Contradictory Information in Wikipedia // Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '15). 2015. pp. 57–60.
4. *Йоцов В.С., Сгурев В.С., Юсупов Р.М., Хомоненко А.Д.* Онтологии для разрешения семантических конфликтов // Труды СПИИРАН. 2008. Вып. 7. С. 26–40.
5. *Ram S., Park J.* Semantic Conflict Resolution Ontology (SCROL): An Ontology for Detecting and Resolving Data- and Schema-Level Semantic Conflicts // IEEE Transactions on Knowledge and Data engineering. 2004. vol. 16. no. 2. pp. 189–202.
6. *Хомоненко А.Д., Войцеховский С.В., Логашев С.В., Дашонок В.Л.* Устранение семантических противоречий в elibrary.ru на основе нечеткого вывода // Проблемы информационной безопасности. Компьютерные системы. 2015. № 1. С. 24–33.

7. *Хомоненко А.Д., Логашев С.В., Краснов С.А.* Автоматическая рубрикация документов с помощью латентно-семантического анализа и алгоритма нечеткого вывода Мамдани // Труды СПИИРАН. 2016. № 1(44) С. 5–19.
8. *Lawrence R.* Automatic Conflict Resolution to Integrate Relational Schema // Ph.D. Thesis. 2001. 165 p.
9. *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. pp. 171–181.
10. *Peng Z., Kambayashi Y.* Resolving Conflicts and Handling Replication during Integration of Multiple Databases by Object Deputy Model // Proceedings of the 20th International Conference on Conceptual Modeling: Conceptual Modeling. 2001. LNCS 2224. pp. 285–298.
11. *De Marneffe M.-C., Rafferty A. N., Manning C.D.* Finding Contradictions in Text // Proceedings of ACL-08: HLT. 2008. pp. 1039–1047.
12. *Воронцов К., Фрей А., Ромов П.* BigARTM: библиотека с открытым кодом для тематического моделирования больших текстовых коллекций // Аналитика и управление данными в областях с интенсивным использованием данных. 2015 С. 28–36.
13. *Бондарчук Д.В., Тимофеева Г.А.* Математические основы метода категориальных векторов в интеллектуальном анализе данных // Вестник Уральского государственного университета путей сообщения. 2015. № 4(28). С. 4–8.
14. *Gupta M., Bendersky M.* Information Retrieval with Verbose Queries // Foundations and Trends in Information Retrieval. 2015. vol. 9. no. 3–4. pp. 209–354.
15. *Landauer T., Foltz P., Laham D.* An introduction to Latent Semantic Analysis // Discourse processes. 1998. vol. 25. no. 2–3. pp. 259–284.
16. *Foltz P.W.* Using latent semantic indexing for information filtering // ACM Conference on Office Information Systems (COIS). 1990. pp. 40–47.
17. *Бубнов В.П. и др.* Модели информационных систем: учеб. пособие // М.: ФГБОУ «Учебно-методический центр по образованию на железнодорожном транспорте». 2015. 188 с.
18. *Dumais S.* Latent semantic indexing: TREC-3 report // Proc. of the Third Text REtrieval Conference. 1995. pp. 219–230.
19. *Соловьев А.Н.* Моделирование процессов понимания речи с использованием латентно-семантического анализа: диссертация на соискание степени к.ф-м.н. // С.-Петерб. гос. ун-т. Санкт-Петербург. 2008.
20. *Хомоненко А.Д., Дашонок В.Л., Краснов С.А.* Выявление противоречий в семантически близкой информации на основе латентно-семантического анализа // Проблемы информационной безопасности. Компьютерные системы. 2014. № 2. С. 73–84.
21. *Gomaa W.H., Fahmy A.A.* A Survey of Text Similarity Approaches. International. Journal of Computer Applications 2013. vol. 68. no. 13. pp. 13–18.
22. *Бермудес С.Х.Г., Керимова С.У.* О методе определения текстовой близости, основанном на семантических классах // Инженерный вестник Дона. 2016. № 4(43). URL: ivdon.ru/ru/magazine/archive/n4y2016/3832 (дата обращения: 01.08.2017).
23. *Kuznetsov S., Poelmans J.* Knowledge representation and processing with formal concept analysis // Wiley interdisciplinary views: Data mining and knowledge discovery. 2013. vol. 3. pp. 200–215.
24. *Jones K.S.* A statistical interpretation of term specificity and its application in retrieval // Journal of Documentation. 2004. vol. 60. no. 5. pp. 493–502.
25. *Manning C.D., Raghavan P., Schütze H.* An Introduction to Information Retrieval Draft // Online edition. Cambridge University Press. 2009. 544 p.

26. *Бондарчук Д.В.* Использование латентно-семантического анализа в задачах классификации текстов по эмоциональной окраске // Бюллетень результатов научных исследований. 2012. № 2(3). С. 146–152.
27. *Краснов С.А., Хомоненко А.Д., Яковлев Я.В.* 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.

Краснов Сергей Александрович — к-т техн. наук, старший преподаватель, Военно-космическая академия имени А.Ф. Можайского (ВКА им. А.Ф. Можайского). Область научных интересов: информационные технологии, защита информации, системы искусственного интеллекта. Число научных публикаций — 30. kras25@rambler.ru; ул. Ждановская 13, Санкт-Петербург, 197198; р.т.: +7(911)7346550.

Илатовский Анатолий Сергеевич — курсант, Военно-космическая академия имени А.Ф. Можайского (ВКА им. А.Ф. Можайского). Область научных интересов: информационные технологии, защита информации, системы искусственного интеллекта. Число научных публикаций — 2. letsbrainup@gmail.com; ул. Ждановская 13, Санкт-Петербург, 197198; р.т.: +7(967)968-20-63.

Хомоненко Анатолий Дмитриевич — д-р техн. наук, профессор, заведующий кафедрой информационных и вычислительных систем, ФГБОУ ВО Петербургский государственный университет путей сообщения Императора Александра I. Область научных интересов: численная теория массового обслуживания, программирование, операционные и информационные системы. Число научных публикаций — 150. khomon@mail.ru, http://www.pgups.ru; Московский пр., 9, Санкт-Петербург, 190031; р.т.: 8(812)457-80-23, Факс: 8(812)310-75-25.

Арсеньев Владимир Николаевич — д-р техн. наук, профессор, профессор кафедры бортовых информационных и измерительных комплексов, Военно-космическая академия имени А.Ф. Можайского (ВКА им. А.Ф. Можайского). Область научных интересов: методы анализа и синтеза сложных систем. Число научных публикаций — 100. vladar56@mail.ru; ул. Ждановская 13, Санкт-Петербург, 197198; р.т.: +79112620222.

DOI 10.15622/sp.54.8

S.A. KRASNOV, A.S. ILATOVSKY, A.D. KHOMONENKO, V.N. ARSENIEV

ASSESSMENT OF SEMANTIC SIMILARITY OF DOCUMENTS ON THE BASIS OF THE LATENT SEMANTIC ANALYSIS WITH THE AUTOMATIC CHOICE OF RANK VALUES

Krasnov S.A., Ilatovsky A.S., Khomonenko A.D., Arseniev V.N. **Assessment of Semantic Smilarity of Documents on the basis of the Latent Semantic Analysis with the Automatic Choice of Rank Values.**

Annotation. The method of assessment of semantic similarity of documents, which is based on the use of the latent and semantic analysis, dynamics of change of singular values of a term-document matrix and automatic determination of a range of rank values, is offered. Assessment of semantic similarity of documents is considered in relation to the solution of problems of identification of duplication and contradictions in databases and storages of data.

A short review of the approaches used at assessment of semantic similarity of documents, identification of duplication and contradictions in databases is provided. Results of numerical examples of assessment of semantic dependences between terms of documents for the benefit of identification of duplication and contradictions in databases and storages of data are given. In this case, the degree of correspondence between the compared documents as the resultant characteristic is calculated.

Comparative estimates of the accuracy of the calculation of the degree of correspondence of λ documents with the help of the main methods (cosine proximity measure, vector model, Spearman rank correlation coefficient, static measure tf-idf — frequency of the term — reverse document frequency) are given.

It is shown that application of the offered method of the latent and semantic analysis with automatic detection of a range of rank values allows eliminating dependence of results of application of a method of the latent semantic analysis on the chosen rank.

Keywords: assessment of semantic similarity of documents, identification of duplications and contradictions, databases, latent semantic analysis, statistical analysis, cosine measure of proximity, vector model.

Krasnov Sergey Aleksandrovich — senior lecturer, Mozhaisky Military Space Academy. Research interests: information technology, information security, artificial intelligence systems. The number of publications — 30. kras25@rambler.ru; 13, Zhdanovskaya street, St.-Petersburg, 197198, Russia; office phone: +7(911)7346550.

Ilatovsky Anatoly Sergeyevich — cadet, Mozhaiskii Military Space Academy. Research interests: information technology, information security, system of artificial intelligence. The number of publications — 2. letsbrainup@gmail.com; 13, Zhdanovskaya street, St.-Petersburg, 197198, Russia; office phone: +7(967)968-20-63.

Khomonenko Anatoly Dmitrievich — Ph.D., Dr. Sci., professor, head of information and computing systems department, Emperor Alexander I St. Petersburg State Transport University. Research interests: queuing systems, artificial intelligence, databases. The number of publications — 150. khomon@mail.ru, http://www.pgups.ru; 9, Moskovsky pr., Saint Petersburg, 190031; office phone: 8(812)457-80-23, Fax: 8(812)310-75-25.

Arseniev Vladimir Nikolaevich — Ph.D., Dr. Sci., professor, professor of the pulpit on-board (side) information and measuring complex, Mozhaisky Military Space Academy. Research interests: modeling and testing of complex systems, control of aircraft. The number of publications — 100. vladar56@mail.ru; 13, Zhdanovskaya street, St.-Petersburg, 197198, Russia; office phone: +79112620222.

References

1. Witten I.H., Frank E., Hall M.A. Data Mining: Practical Machine Learning Tools and Techniques: 3rd edition. Morgan Kaufmann. 2011. 664 p.
2. Paklin N.B., Oreshkov V.I. *Biznes-analitika: ot dannyh k znanijam* (+ *CD*) [Business analysis: from data to knowledge (+ CD)]. SPb: Izd. Peter. 2009. 624 p. (In Russ.).
3. Weissman S., Ayhan S., Bradley J., Lin J. Identifying Duplicate and Contradictory Information in Wikipedia. Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '15). 2015. pp. 57–60.
4. Jotsov V.S., Sgurev V.S., Yusupov R.M., Khomonenko A.D. [The Ontology for the Semantic Conflicts Resolution]. *Trudy SPIIRAN –SPIIRAS Proceedings*. 2008. vol. 7. pp. 26–40. (In Russ.).
5. Ram S., Park J. Semantic Conflict Resolution Ontology (SCROL): An Ontology for Detecting and Resolving Data- and Schema-Level Semantic Conflicts. *IEEE Transactions on Knowledge and Data engineering*. 2004. vol. 16. no. 2. pp. 189–202.
6. Khomonenko A.D., Voitsekhovskii S.V., Logashev S.V., Dashonok V.L. [Resolving Semantic Inconsistencies in ELIBRARY.RU Based on Fuzzy Inference]. *Problemy informacionnoj bezopasnosti. Komp'juternye sistemy – Automatic Control and Computer Sciences*. 2015. vol. 1. pp. 24–33. (In Russ.).
7. Khomonenko A.D., Logashev S.V., Krasnov S.A. [Automatic categorization of documents using latent semantic analysis and the Mamdani fuzzy inference algorithm]. *Trudy SPIIRAN –SPIIRAS Proceedings*. 2016. vol. 1(44). pp. 5–19. (In Russ.).
8. Lawrence R. Automatic Conflict Resolution to Integrate Relational Schema. Ph.D. Thesis. 2001. 165 p.
9. 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. pp. 171–181.
10. Peng Z., Kambayashi Y. Resolving Conflicts and Handling Replication during Integration of Multiple Databases by Object Deputy Model. Proceedings of the 20th International Conference on Conceptual Modeling: Conceptual Modeling. 2001. LNCS 2224. pp. 285–298.
11. De Marneffe M.C., Rafferty A. N., Manning C.D. Finding Contradictions in Text. Proceedings of ACL-08: HLT. Proceedings of ACL-08: HLT. 2008. pp. 1039–1047.
12. Vorontsov K., Frei A., Romov P. [BigARTM: open source library for thematic modeling of large text collections]. *Analitika i upravlenie dannymi v oblastjah s intensivnym ispol'zovaniem dannyh – Analytics and data management in areas with intensive data use*. 2015. pp. 28–36. (In Russ.).
13. Bondarchuk D.V., Timofeeva G.A. [Mathematical foundations of the method of categorical vectors in the intellectual analysis of data]. *Vestnik Ural'skogo gosudarstvennogo universiteta putej soobshhenija – Newsletter Ural State University of Communications*. 2015. vol. 4(28). pp. 4–8. (In Russ.).
14. Gupta M., Bendersky M. Information Retrieval with Verbose Queries. Foundations and Trends in Information Retrieval. vol. 9. no. 3–4. 2015. pp. 209–354.
15. Landauer T., Foltz P., Laham D. An Introduction to Latent Semantic Analysis. *Discourse processes.* 1998. vol. 25. no. 2–3. pp. 259–284.
16. Foltz P.W. Using the latent semantic indexing for information filtering. ACM Conference on Office Information Systems (COIS). 1990. pp. 40–47.
17. Bubnov V.P. et al. *Modeli informacionnyh sistem* [Models of information systems]. Moscow: Izd.FGBOU «The Educational-methodical center on formation on a railway transportation». 2015. 188 p. (In Russ.).

18. Dumais S. Latent semantic indexing: TREC-3 report. Proc. Of the Third Text REtrieval Conference. 1995. pp. 219–230.
19. Soloviev A.N. *Modelirovanie processov ponimanija rechi s ispol'zovaniem latentno-semanticheskogo analiza: dissertacija na soiskanie stepeni k.f-m.n.* [Modeling the processes of understanding speech using latent-semantic analysis. Ph.D. thesis. S.-Peterb. gos. un-t. Sankt-Peterburg. 2008. (In Russ.).
20. Khomonenko A.D., Dashonok V.L., Krasnov S.A. [Revealing of contradictions in semantically close information on the basis of latent-semantic analysis]. *Problemy informacionnoj bezopasnosti. Komp'juternye sistemy – Automatic Control and Computer Sciences*. 2014. vol. 2. pp. 73–84. (In Russ.).
21. Gomaa W.H., Fahmy A.A. A Survey of Text Similarity Approaches. *International Journal of Computer Applications*. 2013. vol. 68. no. 13. pp. 13–18.
22. Bermudes S.H.G., Kerimova S.U. [About the method of determination of text closeness based on semantic classes]. *Inzhenernyj vestnik Dona – Engineering journal of Don*. 2016. vol. 4(43). Available at: ivdon.ru/ru/magazine/archive/n4y2016/3832 (accessed: 01.08.2017). (In Russ.).
23. Kuznetsov S., Poelmans J. Knowledge representation and processing with formal concept analysis. *Wiley inter disciplinary views: Data mining and knowledge discovery*. 2013. vol. 3. pp. 200–215.
24. Jones K.S. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*. 2004. vol. 60. no. 5. pp. 493–502.
25. Manning C.D., Raghavan P., Schütze H. An Introduction to Information Retrieval Draft. Cambridge University Press. 2009. 544 p.
26. Bondarchuk D.V. [The use of latent-semantic analysis in problems of classification of texts on emotional coloring]. *Bjulleten' rezul'tatov nauchnyh issledovanij – Bulletin of the results of scientific research*. 2012. vol. 2(3). pp. 146–152. (In Russ.).
27. Krasnov S.A., Khomonenko A.D., Yakovlev Ya.V. [Estimation of the effectiveness of the algorithm for calculating the Spearman's rank correlation coefficient in the method of latent-semantic analysis in the automatic classification of documents]. *Bjulleten' rezul'tatov nauchnyh issledovanij – Bulletin of the results of scientific research*. 2012. vol. 2(3). pp. 153–162. (In Russ.).