УДК 004.85; 004.91

# МЕТОДЫ ВЫЧИСЛЕНИЯ РЕЛЕВАНТНОСТИ ФРАГМЕНТОВ ТЕКСТА НА ОСНОВЕ ТЕМАТИЧЕСКИХ МОДЕЛЕЙ В ЗАДАЧЕ АВТОМАТИЧЕСКОГО АННОТИРОВАНИЯ

**И. В. Машечкин[1], М. И. Петровский[1], Д. В. Царёв[1]**

Рассмотрены наиболее актуальные методы вычисления релевантности (значимости) фрагментов текста на основе анализа тематических моделей для последующего построения аннотаций в форме выдержек, т.е. аннотаций, полностью состоящих из последовательности фрагментов исходного текста. В качестве тематических моделей выбраны популярные модели семантики документов и коллекции документов, используемые в задачах анализа текстовой информации: модели, основанные на латентно-семантическом анализе, модель вероятностного латентно-семантического анализа и модель скрытого распределения Дирихле. Предложен новый метод вычисления релевантности фрагментов текста, основанный на оценке весов тематик в нормализованном пространстве тематик, получаемом с помощью факторизации неотрицательных матриц, которая используется в качестве матричного разложения в модели латентно-семантического анализа. Эксперименты, проведенные с использованием методов автоматического аннотирования на эталонных тестовых наборах DUC 2001 и DUC 2002 на основе стандартных метрик оценки качества аннотаций ROUGE, показали превосходство методов вычисления релевантности фрагментов текста, основанных на латентно-семантическом анализе, по сравнению с методами, основанными на вероятностных тематических моделях, по качеству получаемых аннотаций. Приведены результаты тестирования, показывающие, что предложенный метод вычисления релевантности фрагментов текста, использующий факторизацию неотрицательных матриц для тематического моделирования, дает лучшие результаты по сравнению со всеми рассмотренными методами. Работа выполнена в рамках государственного контракта № 14.514.11.4016 и при поддержке РФФИ (проекты 11–07–00616 и 12–07–00585).

**Ключевые слова:** релевантность фрагментов текста, автоматическое аннотирование, семантические модели текста, тематические модели, латентно-семантический анализ, сингулярное разложение, факторизация неотрицательных матриц, вероятностные тематические модели, вероятностный латентно-семантический анализ, скрытое распределение Дирихле.

**1. Введение.** Настоящая статья посвящена актуальной на сегодняшний день задаче, решаемой методами интеллектуального анализа текстовых данных (англ. text mining), — задаче автоматического аннотирования текста [1, 2]. Это одна из основных алгоритмических задач, возникающая при реализации многих прикладных систем: в базах данных для анализа текстовых данных (Oracle Text), в web-поисковых системах (еще в 1998 г. система Inxight Summarizer использовалась для построения аннотаций в AltaVista), в текстовых редакторах (AutoSummarize в Microsoft Office) и др.

Наиболее популярные методы автоматического аннотирования основаны на анализе семантических моделей текста [1–7]. Модели семантики текста описывают текстовые единицы и их взаимосвязи между собой, основываясь на семантических значениях используемых текстовых единиц. В качестве текстовых единиц, в зависимости от модели, могут быть слова, фрагменты текста (например, предложения), документы коллекции, синонимические ряды (“синсеты”) [13–15] и т.п. На основе анализа построенной модели текста происходит вычисление релевантности (значимости) его фрагментов и последующее включение фрагментов с наибольшей релевантностью в аннотацию.

В последнее время в различных задачах автоматического аннотирования активно используются тематические модели (англ. topic model), которые можно рассматривать как один из типов семантических моделей [5, 8]. В тематических моделях в качестве текстовых единиц помимо термов текста используются либо фрагменты текста при моделировании отдельных документов, либо отдельные документы при

[1] Московский государственный университет им. М. В. Ломоносова, факультет вычислительной математики и кибернетики, Ленинские горы, д. 1, стр. 52, 119991, Москва; И. В. Машечкин, профессор, e-mail: mash@cs.msu.su; М. И. Петровский, доцент, e-mail: michael@cs.msu.su; Д. В. Царёв, математик, e-mail: tsarev@mlab.cs.msu.su

моделировании коллекции документов. Тематические модели коллекции текстовых единиц определяют их тематики в виде семантических связей между термами текста и оценивают степени принадлежности каждой текстовой единицы к каждой тематике.

В данной работе рассматриваются методы вычисления релевантности фрагментов текста на основе анализа тематических моделей для последующего построения аннотаций в форме выдержек, т.е. аннотаций, полностью состоящих из последовательности фрагментов исходного текста. Кроме того, предполагается, что аннотации строятся для широкого круга читателей, т.е. освещаются все главные темы исходного текста, а не делается акцент на определенные темы, связанные с интересами конкретных читателей.

Для того чтобы определять семантические связи между словами с помощью тематического моделирования коллекции фрагментов текста, фрагменты должны представлять собой семантически связные последовательности слов (например, предложения, параграфы и т.п.). Кроме того, разбиение текста на фрагменты должно опираться на тип анализируемого контента и длину требуемой аннотации: так, для переписки, состоящей из коротких текстовых сообщений, в качестве фрагментов могут выступать сами сообщения, для новостных или научных статей — предложения или параграфы (при построении длинной аннотации большого текста) и т.п.

Настоящая статья имеет следующую структуру. В разделе 2 дан краткий обзор тематических моделей текста, которые широко применяются в задачах автоматического аннотирования для вычисления релевантности фрагментов текста. В разделе 3 описываются методы вычисления релевантности фрагментов текста, основанные на тематических моделях. Кроме того, в данном разделе представлен собственный разработанный метод вычисления релевантности фрагментов текста, использующий факторизацию неотрицательных матриц для тематического моделирования (пункт 3.1.2). Раздел 4 посвящен экспериментальному исследованию рассмотренных методов вычисления релевантности фрагментов текста в задаче автоматического аннотирования на эталонных тестовых наборах данных DUC 2001 и DUC 2002.

**2. Тематические модели.** Анализ моделей семантики как отдельных документов, так и коллекций документов широко применяется при решении различных задач интеллектуального анализа текстовой информации: информационный поиск (англ. information retrieval) [9, 16], классификация (англ. classification) [10], кластеризация (англ. clustering) [11], определение тематики (англ. topic extraction) [12], автоматическое аннотирование (англ. text summarization) [2–7]. Обычно анализ моделей семантики используется для устранения полисемии и омонимии (наличие у слова двух и более значений: омонимия — это случайное совпадение слов; полисемия — наличие у слова разных исторически связанных значений), а также для определения наиболее значимых частей документа, например слов или предложений.

Тематические модели предназначены для описания текстов с точки зрения их тематик. Модели данного типа обычно используют в качестве текстовых единиц либо отдельные документы при анализе коллекции документов, либо текстовые фрагменты документа при анализе отдельного документа. В литературе [7, 9, 17–19] среди тематических моделей, как правило, выделяют три основных подхода: *латентно-семантический анализ*, *вероятностный латентно-семантический анализ* и *скрытое распределение Дирихле*.

*Модели, основанные на латентно-семантическом анализе.* Эти модели описывают семантическую взаимосвязь текстовых единиц и встречающихся в них слов путем объединения семантически схожих слов в тематики и сопоставления тематик текстовым единицам [20].

Для выделения семантических связей между текстовыми единицами и их словами используется метод латентно-семантического анализа (англ. Latent Semantic Analysis, `LSA`). Латентно-семантический анализ работает с векторным представлением типа “мешка слов” (англ. “bag-of-words”) текстовых единиц [21]. Таким образом, коллекция текстовых единиц представляется в виде числовой матрицы, строки которой соответствуют словам, входящим в коллекцию, а столбцы — текстовым единицам. Объединение термов в тематики и представление текстовых единиц в пространстве тематик осуществляется путем применения к данной матрице одного из матричных разложений. В настоящее время наиболее популярными матричными разложениями являются сингулярное разложение (англ. Singular Value Decomposition, `SVD`) и факторизация неотрицательных матриц (англ. Non-negative Matrix Factorization, `NMF`) [17, 28, 46]. После применения к текстовой матрице одного из матричных разложений формируется семантическая модель совокупности текстовых единиц, состоящая из

1) матрицы отображения пространства слов в пространство тематик;

2) вектора, элементы которого соответствуют весам выделенных тематик из коллекции текстовых единиц (или диагональной матрицы, диагональные элементы которой соответствуют весам тематик);

3) матрицы представления текстовых единиц в пространстве тематик.

*Модель вероятностного латентно-семантического анализа.* Данная семантическая модель совпадает с моделями, основанными на латентно-семантическом анализе [9]. Отличие между ними заключается в способе построения самих моделей. По сравнению с обычным латентно-семантическим анализом, который основан на линейной алгебре и является способом снижения размерности матрицы (с помощью одного из матричных разложений), вероятностный латентно-семантический анализ основан на предположении, что совместное появление пар [текстовая единица ($d$), слово ($t$)] обусловлено скрытыми (латентными) переменными — тематиками ($z$). Вероятностная модель появления пары [ текстовая единица, слово] может быть записана в виде: $P(d_i, t_j) = \sum_{n=1}^{k} P(z_n) P\big(d_i \mid z_n\big) P\big(t_j \mid z_n\big)$.

Зафиксировав число скрытых тематик $k$, с помощью вероятностного латентно-семантического анализа можно оценить следующие величины:

— $P(z_n)$: вероятность того, что случайно выбранное предложение $d_i$ соответствует тематике $z_n$;

— $P\big(d_i \mid z_n\big)$: вероятность того, что предложение $d_i$ попадет в группу предложений, соответствующих тематике $z_n$;

— $P\big(t_j \mid z_n\big)$: вероятность того, что терм $t_j$ попадет в группу термов, связанных с тематикой $z_n$.

Для нахождения параметров модели используется `EM-алгоритм` (англ. Expectation Maximization) [9] — стандартная итерационная процедура идентификации скрытых переменных путем максимизации функции правдоподобия.

Для представления полученных вероятностей в матричной форме определим следующие матрицы:

— матрица $\widetilde{U}$, элементы которой $\widetilde{u}_{jn}$ соответствуют условным вероятностям $P\big(t_j \mid z_n\big)$;

— матрица $\widetilde{V}$, элементы которой $\widetilde{v}_{in}$ соответствуют условным вероятностям $P(d_i \mid z_n)$;

— диагональная матрица $\widetilde{\Sigma}$, диагональные элементы которой $\widetilde{\sigma}_n$ соответствуют вероятностям $P(z_n)$.

Тогда модель совместной вероятности может быть записана в виде, аналогичном латентно-семантическому анализу: $P = \widetilde{U}\widetilde{\Sigma}\widetilde{V}^{\mathrm{T}}$.

Вероятностный латентно-семантический анализ (англ. Probabilistic Latent Semantic Analysis, `PLSA`) был предложен в 1999 г. и является первой вероятностной тематической моделью с латентными переменными. Данная модель также называется *моделью аспектов* (англ. aspect model) [9].

*Модель скрытого распределения Дирихле.* Скрытое распределение Дирихле (англ. Latent Dirichlet Allocation, `LDA`) — вероятностная модель порождения набора текстовых единиц. Основная идея данной вероятностной модели состоит в том, что текстовые единицы представлены как случайные смеси скрытых тематик; в свою очередь, каждая тематика характеризуется распределением вероятностей в пространстве слов из общего словаря [18]. Процесс порождения текста состоит из двух шагов. На первом шаге для каждой текстовой единицы $d$ выбирается случайный вектор $\theta_d$ из распределения Дирихле с параметром $\alpha$ (обычно $\alpha$ принимается равным $50/T$, где $T$ — число тематик [45]), соответствующий распределению вероятностей в пространстве тематик. На втором шаге для порождения каждого слова текстовой единицы $d$ сначала выбирается тематика $z_{di}$ из мультиномиального распределения с параметром $\theta_d$; затем для выбранной тематики $z_{di}$ выбирается слово $t_{di}$ из распределения $\Phi_{z_{di}}$, которое является распределением Дирихле с параметром $\beta$ (обычно параметр $\beta = 0.01$, увеличение $\beta$ приводит к большей разреженности тематик) [22, 45].

Для оценки параметров $\Phi = \{\phi_{zt}\}$, $\Theta = \{\theta_{dz}\}$, $Z = \{z_{dt}\}$ скрытого распределения Дирихле используются семплирование Гиббса, простой вариационный EM-алгоритм и метод распространения ожидания (англ. Expectation–Propagation) [18, 22, 36, 45].

**3. Методы вычисления релевантности фрагментов текста.** В этом разделе приводятся наиболее актуальные методы вычисления релевантности фрагментов текста, основанные на тематических моделях, для последующего построения аннотаций в форме выдержек, т.е. аннотаций, полностью состоящих из последовательности фрагментов исходного текста. Таким образом, в качестве текстовых единиц далее будем рассматривать фрагменты текста. Кроме того, предполагается, что аннотация строится для широкого круга читателей, т.е. освещаются все главные темы исходного текста, а не делается акцент на определенные темы, связанные с интересами конкретных читателей.

Описанные в разделе 2 тематические модели, как правило, используют модель “мешка слов” для представления своих текстовых единиц, т.е. каждый из $n$ фрагментов текста отображается в числовой вектор $A_j$ фиксированной размерности $m$, где $m$ — число признаков текста, а $i$-я компонента вектора определяет вес $i$-го признака. В качестве признаков в модели “мешка слов” используются лексемы, входящие в текст, а размерность признакового пространства равна размерности словаря лексем. Под лексемами в общем случае понимаются все различные слова текста. Однако обычно применяются некоторые меры по

предварительной обработке текста с целью получения более “информативного” признакового пространства.

Цель предварительной обработки текста — оставить только те признаки, которые наиболее информативны, т.е. наиболее сильно характеризуют аннотируемый текст. К тому же уменьшение анализируемых признаков приводит к уменьшению использования вычислительных ресурсов. В интеллектуальном анализе текстовых данных для обозначения признака текста принято использовать термин “терм”.

Используемые в данной работе эталонные тестовые наборы документов DUC 2001 и DUC 2002 являются полностью англоязычными и состоят из новостных статей, поэтому для формирования списка термов использовались такие методы предварительной обработки текста, как удаление стоп-слов и приведение слов к нормализованной форме (стемминг). Для проведения сравнений методов однодокументного автоматического аннотирования на наборах DUC 2001 и DUC 2002 в качестве фрагментов текста обычно выбираются предложения, что связано, в том числе, с требуемой длиной результирующих аннотаций в 100 слов.

Исходный текст представляется в виде числовой матрицы $A = [A_1, A_2, \ldots, A_n]$, строки которой соответствуют термам текста, а столбцы — его предложениям. Каждое предложение исходного текста представляется в виде вектора в пространстве термов, координатами которого являются весовые коэффициенты соответствующих термов. Формально $j$-е предложение текста отображается в вектор $A_j = [a_{1j}, a_{2j}, \ldots, a_{mj}]^{\mathrm{T}}$, где $m$ — число термов текста, $a_{ij} = L_{ij}\, G_i$, $L_{ij}$ — локальный вес терма $i$ в предложении $j$ $(1 \leqslant j \leqslant n)$ и $G_i$ — глобальный вес терма $i$ в совокупности предложений исходного текста.

В вероятностных тематических моделях, таких как модель вероятностного латентно-семантического анализа и модель скрытого распределения Дирихле, используется частотный локальный вес (TF, табл. 1), а глобальный вес не учитывается [9, 18, 25]. Для моделей, основанных на латентно-семантическом анализе, в задаче однодокументного аннотирования обычно используют сочетание бинарного локального веса и энтропии в качестве глобального веса (весовая схема BI*EN, табл. 1) [5, 23, 25].

Таблица 1

Веса для модели представления “мешок слов”

| Название | Описание |
|---|---|
| Частотный вес (TF) | $L_{ij} = t_{ij}$, где $t_{ij}$ — число появлений терма $i$ в предложении $j$. |
| Бинарный вес (BI) | $L_{ij} = \chi(t_{ij}) = \begin{cases} 1, & \text{если } t_{ij} > 0, \\ 0, & \text{если } t_{ij} = 0. \end{cases}$ |
| Энтропия (EN) | $G_i = 1 - \sum\limits_{j=1}^{N} \left( \dfrac{p_{ij} \log p_{ij}}{\log N} \right)$, где $p_{ij} = \dfrac{t_{ij}}{F_i}$, $F_i = \sum\limits_{k=1}^{N} t_{ik}$. |

Далее приводятся описания методов вычисления релевантности предложений текста, основанных на тематических моделях. Кроме того, в пункте 3.1.2 представлен разработанный нами метод вычисления релевантности предложений текста, использующий факторизацию неотрицательных матриц для тематического моделирования.

**3.1. Методы, основанные на латентно-семантическом анализе.** Латентно-семантический анализ широко используется в различных областях интеллектуального анализа текстовых данных, в том числе в информационном поиске [21], классификации документов [10, 24], автоматическом аннотировании [5, 23] и т.д. Построение представлений предложений текста в пространстве его тематик осуществляется применением к матрице текста одного из матричных разложений. Первым и являющимся до сих пор наиболее популярным является сингулярное разложение (англ. Singular Value Decomposition, `SVD`) [21, 26, 27]. В данном подразделе также рассматривается применение факторизации неотрицательных матриц (англ. Non-negative Matrix Factorization, `NMF`) [11, 12, 17, 28] в качестве разложения для латентно-семантического анализа и представляется наш собственный разработанный метод вычисления релевантности фрагментов текста, основанный на оценке весов тематик в нормализованном пространстве тематик, получаемом с помощью факторизации неотрицательных матриц.

Для определения числа тематик в методах однодокументного аннотирования, основанных на латентно-семантическом анализе, используется следующий подход. Для документа подсчитывается число его слов и определяется, какой процент $p\%$ от всего текста должна составлять аннотация. Тогда если исходная матрица документа имеет размерность $m \times n$, то число тематик подсчитывается по формуле $k =$

$\frac{p}{100}\min(m,n)$ [29].

**3.1.1. Методы на основе сингулярного разложения.** При сингулярном разложении аппроксимацию исходной матрицы текста $A \in \mathbb{R}^{m\times n}$ в пространстве, состоящем из $k$ тематик, можно записать в виде $A \approx A_k = U_k\Sigma_k V_k^{\mathrm{T}}$, где $k \ll \min(m,n)$, $A_k \in \mathbb{R}^{m\times n}$, $U_k \in \mathbb{R}^{m\times k}$, $V_k \in \mathbb{R}^{n\times k}$, $\Sigma_k = \operatorname{diag}(\sigma_1,\ldots,\sigma_k) \in \mathbb{R}^{k\times k}$ [20].

Каждый столбец $j$ матрицы $A$, соответствующий вектору предложения $j$ в аннотируемом тексте, отображается в столбец $j$ матрицы $V_k^{\mathrm{T}}$, который соответствует предложению $j$ в пространстве $k$ тематик. Матрица $U_k$ задает отображение между пространством $k$ тематик и пространством $m$ термов. Каждое сингулярное число $\sigma_l$, где $1 \leqslant l \leqslant k$, соотносится с весом каждой из выделенных тематик в тексте [27].

Для выделения ключевых предложений на основе такого представления используется один из следующих методов.

1. Тематики рассматриваются в порядке от 1 до $k$. Для каждой тематики $l$, где $1 \leqslant l \leqslant k$, рассматривается строка $v_l^{\mathrm{T}} = [v_{1l}, v_{2l}, \ldots, v_{nl}]$ матрицы $V_k^{\mathrm{T}} \in \mathbb{R}^{k\times n}$, элементы которой характеризуют вес тематики $l$ в каждом из $n$ предложений. Находится максимальное по модулю $v_{il}$. Это означает, что предложение $i$ лучше всего соответствует тематике $l$. Если предложение $i$ уже входит в аннотацию, то выбирается следующая максимальная по модулю компонента вектора $v_l^{\mathrm{T}}$ и т.д. [27].

2. Основной недостаток первого метода заключается в том, что в аннотацию могут попасть предложения, которые соответствуют тематикам с небольшим весом. Веса тематик оцениваются соответствующими сингулярными числами. Поэтому для устранения этого недостатка число предложений, соответствующих тематикам, выбирается исходя из процентного соотношения веса тематики к сумме весов всех построенных тематик [27].

3. Наиболее новым является метод [5, 29], в котором каждому предложению документа сопоставляется некоторая числовая оценка — релевантность предложения. Далее выбирается необходимое количество предложений в порядке убывания их релевантности. Релевантность предложения $j$, где $1 \leqslant j \leqslant n$, — это длина $j$-го вектор-столбца матрицы $\Sigma_k^2 V_k^{\mathrm{T}}$ (далее этот метод будем обозначать `SVD`). В основе данного метода лежит предположение о том, что значимость каждой тематики в документе аппроксимируется квадратом соответствующего сингулярного числа.

**3.1.2. Методы на основе факторизации неотрицательных матриц.** У матрицы $A \in \mathbb{R}^{m\times n}$ элементы принимают неотрицательные значения, поскольку они являются весами соответствующих термов в предложениях, поэтому для аппроксимации матрицы $A$ можно применить факторизацию неотрицательных матриц: $A \approx A_k = W_k H_k$, где $k \ll \min(m,n)$, матрицы $A_k \in \mathbb{R}^{m\times n}$, $W_k \in \mathbb{R}^{m\times k}$ и $H_k \in \mathbb{R}^{k\times n}$ имеют неотрицательные элементы [ 28, 46].

Матрица $W_k$ задает отображение между пространством $k$ тематик и пространством $m$ термов, матрица $H_k$ соответствует представлению предложений в пространстве тематик. В связи с тем, что элементы матрицы $W_k$ неотрицательны, можно установить, какие термы текста наилучшим образом характеризуют каждую из выделенных тематик, которым соответствуют столбцы матрицы $W_k$. Аналогично можно установить, какие из выделенных тематик наилучшим образом характеризуют каждое предложение текста. Таким образом, факторизация неотрицательных матриц, в отличие от сингулярного разложения, предоставляет хорошо интерпретируемое семантическое пространство (пространство тематик).

В данной работе факторизация неотрицательных матриц была реализована мультипликативным алгоритмом, описанным в работе [28]. Однако авторами также проводились исследования по применению и других алгоритмов, реализующих факторизацию неотрицательных матриц, в частности:

— метод наименьших квадратов, использующий проекцию градиента [40];

— вероятностная факторизация неотрицательных матриц [41];

— альтернативный метод наименьших квадратов [12];

— различные ортогональные факторизации неотрицательных матриц [42, 43].

Наиболее популярный метод вычисления релевантности предложений текста на основе факторизации неотрицательных матриц описан в [30]. Кроме того, его модификации используются и в задаче построения аннотаций на основе запроса [31], и в задаче многодокументного аннотирования [32].

Основная идея этого метода заключается в подсчете общей релевантности (англ. Generic Relevance, GR) для предложений текста.

Общая релевантность $j$-го предложения вычисляется по формуле $\mathrm{GR}_j = \sum\limits_{i=1}^{k}\bigl(h_{ij}\,\mathrm{weight}(H_{i*})\bigr)$, где $\mathrm{weight}(H_{i*}) = \left(\sum\limits_{q=1}^{n} h_{iq}\right)\left(\sum\limits_{p=1}^{k}\sum\limits_{q=1}^{n} h_{pq}\right)^{-1}$ — относительная релевантность $i$-й тематики среди всех выде-

ленных тематик. Далее выбирается необходимое число предложений с наибольшими значениями общей релевантности.

В отличие от сингулярного разложения, при факторизации неотрицательных матриц отсутствует диагональная матрица, элементы которой оценивают веса выделенных тематик. Поэтому в методе общей релевантности в качестве оценки весов используется относительная релевантность. В свою очередь, коллективом авторов настоящей статьи был предложен собственный метод вычисления релевантности предложений текста [23], основанный на оценке весов тематик в нормализованном пространстве тематик, получаемом с помощью факторизации неотрицательных матриц.

В работах, посвященных исследованию сходимости мультипликативного алгоритма факторизации неотрицательных матриц [44] и кластеризации документов на основе мультипликативного алгоритма [11], рекомендуется нормировать столбцы матрицы $W_k$.

Поэтому первый этап предложенного метода состоит в нормировке пространства $k$ тематик, т.е. в приведении длин вектор-столбцов матрицы $W_k$ к единице: $A_k = W_k H_k = \operatorname{Norm} W_k \operatorname{Norm} H_k$, где $\operatorname{Norm} W_k = W_k \operatorname{diag}\left(\dfrac{1}{\|w^1\|}, \ldots, \dfrac{1}{\|w^k\|}\right)$, $\operatorname{Norm} H_k = \operatorname{diag}\left(\|w^1\|, \ldots, \|w^k\|\right) H_k$, $\|w^l\| = \sqrt{\sum\limits_{p=1}^{m} w_{p,l}^2}$, $1 \leqslant l \leqslant k$.

Столбцы матрицы $\operatorname{Norm} H_k = [\operatorname{norm} h_{ij}]$ соответствуют $n$ предложениям в нормированном пространстве $k$ тематик. Каждая из $k$ строк $\operatorname{Norm} H_k$ соответствует вектору, показывающему, насколько подробно представлена соответствующая тематика в каждом из $n$ предложений. Тем самым, чем больше длина вектор-строки матрицы $\operatorname{Norm} H_k$, тем соответствующая тематика "больше" представлена во всем документе. Исходя из этого, второй этап оценки весов тематик состоит в вычислении веса тематики $l$ как длины $l$-й вектор-строки матрицы $\operatorname{Norm} H_k$: $\|\operatorname{norm} h_l\| = \sqrt{\sum\limits_{q=1}^{n} \operatorname{norm} h_{lq}^2} = \|w^l\| \sqrt{\sum\limits_{q=1}^{n} h_{lq}^2} = \|w^l\| \, \|h_l\|$, $1 \leqslant l \leqslant k$. Отметим, что авторами исследовалось применение различных норм для нормировки столбцов матрицы $W_k$ и для вычисления длин строк матрицы $\operatorname{Norm} H_k$. Лучшие результаты показало использование евклидовой нормы. Таким образом, взвешенному представлению $n$ предложений в нормированном пространстве $k$ тематик соответствует матрица

$$\operatorname{Weighted} H_k = \operatorname{diag}\left(\|w^1\| \, \|h_1\|, \ldots, \|w^k\| \, \|h_k\|\right) \operatorname{Norm} H_k = \operatorname{diag}\left(\|w^1\|^2 \, \|h_1\|, \ldots, \|w^k\|^2 \, \|h_k\|\right) H_k.$$

Тогда релевантность каждого предложения вычисляется как сумма элементов соответствующего вектор-столбца в матрице $\operatorname{Weighted} H_k$. Таким образом, релевантность $j$-го предложения, где $1 \leqslant j \leqslant n$, вычисляется по формуле

$$R_j(\operatorname{Weighted} H_k) = \sum_{i=1}^{k} (\operatorname{weighted} h_{ij}) = \sum_{i=1}^{k} \left(\|w^i\|^2 \, \|h_i\| \, h_{ij}\right).$$

Далее выбирается необходимое число предложений с наибольшими значениями полученной релевантности. Предложенный метод будем далее обозначать через `NMF`.

**3.2. Методы, основанные на вероятностном латентно-семантическом анализе.** В вероятностном латентно-семантическом анализе (англ. Probabilistic Latent Semantic Analysis, `PLSA`), как и в латентно-семантическом анализе, считаем, что существует $k$ скрытых тематик $z \in Z = \{z_1, \ldots, z_k\}$, число которых $k$ задается заранее. При фиксированном числе скрытых тематик $k$ с помощью вероятностного латентно-семантического анализа оцениваются следующие величины: $P(z_n)$, $P\big(d_i \mid z_n\big)$, $P\big(t_j \mid z_n\big)$.

В работе [6] проводится сравнение ряда подходов выбора предложений текста для аннотации с использованием вероятностного латентно-семантического анализа. В основном эти подходы аналогичны методам на основе сингулярного разложения, описанного в разделе 3.1.1.

По итогам экспериментов в [6] лучшие результаты показал метод, в котором релевантность предложения $i$ вычисляется по формуле $R_i = \sum\limits_{z \in Z} P\big(d_i \mid z\big) P(z) = P(d_i)$. Далее выбирается необходимое число предложений с наибольшими значениями релевантности [6].

Данный метод будем обозначать как `PLSA`. В нем число тематик подбиралось эмпирически, и наилучшие результаты для набора DUC 2002 были получены для двух тематик [6]. Кроме того, в `EM-алгоритме`

использовался модифицированный E-шаг: $P\big(z \mid d,t\big) = \dfrac{P(z)\Big[P\big(d \mid z\big)P\big(t \mid z\big)\Big]^{\beta}}{\sum\limits_{z' \in Z} P(z')\Big[P\big(d \mid z'\big)P\big(t \mid z'\big)\Big]^{\beta}}$, $\beta = 0.75$.

**3.3. Методы, основанные на скрытом распределении Дирихле.** В настоящее время скрытое распределение Дирихле в области автоматического аннотирования в основном применяется для задачи многодокументного аннотирования, т.е. аннотация строится не к каждому документу из коллекции, а к самой коллекции документов, написанных на одну тему [7]. Кроме того, скрытое распределение Дирихле используется для визуализации изменения важности основных тематик в больших коллекциях документов с течением времени [33]. Таким образом, скрытое распределение Дирихле на практике применяют для тематического моделирования коллекций документов, а не отдельных документов. Однако учитывая, что скрытое распределение Дирихле является дальнейшим развитием вероятностного латентно-семантического анализа [34], авторами настоящей статьи были проведены сравнительные эксперименты с методом вычисления релевантности предложений, аналогичным методу PLSA, описанному в подразделе 3.2. Иными словами, релевантность предложения $i$ вычисляется по формуле $R_i = \sum\limits_{z \in Z} P\big(d_i \mid z\big)P(z) = P(d_i)$. Отличие заключается лишь в том, что соответствующие вероятности $P\big(d \mid z\big)$ и $P(z)$ вычислялись путем применения к матрице текста скрытого распределения Дирихле. Данный метод будем далее обозначать как LDA.

Скрытое распределение Дирихле было реализовано с помощью семплирования Гиббса [22, 45], для этого использовалась библиотека Matlab Topic Modeling Toolbox 1.4 [35]. Результатом семплирования Гиббса являются

1) $WZ \in \mathbb{N}_0^{m \times k}$ — матрица, элементы $WZ_{ij}$ которой показывают, сколько раз терм $i$ был отнесен к тематике $j$;

2) $DZ \in \mathbb{N}_0^{n \times k}$ — матрица, элементы $DZ_{ij}$ которой показывают, сколько раз термы из документа $i$ были отнесены к тематике $j$;

3) $Z \in \mathbb{N}_1^{N_{\text{tot}}}$ — вектор, каждая компонента $Z_i$ которого соответствует номеру тематики $i$-го терма документа, $N_{\text{tot}}$ — число всех термов в документе.

Отображение всех термов $T$ в тематики, т.е. формирование вектора $Z$, осуществляется исходя из оценок вероятностей принадлежности $i$-го терма документа к каждой из тематик $c$ [45]:

$$P\big(z_i = c \mid z_{-i}, t_i, d_i, \alpha, \beta\big) \approx \frac{WZ_{t_i c} + \beta}{\sum\limits_t WZ_{tc} + m\beta} \frac{DZ_{d_i c} + \alpha}{\sum\limits_z DZ_{d_i z} + k\alpha}.$$

Здесь $z_i$ — тематика $i$-го терма в документе, $1 \leqslant i \leqslant N_{\text{tot}}$; $c \in \{1, \ldots, k\}$ — номер тематики; $t_i \in \{1, \ldots, m\}$ — номер $i$-го терма в словаре термов; $d_i \in \{1, \ldots, n\}$ — номер предложения $i$-го терма в документе.

Тогда параметры модели скрытого распределения Дирихле вычисляются по формулам [22, 45]:

1) $P\big(z_i \mid d_j\big) = \dfrac{DZ_{ji} + \alpha}{\sum\limits_z DZ_{jz} + k\alpha}$ — вероятность появления тематики с номером $z_i$ в $j$-ом предложении документа (или $\theta_{ji}$);

2) $P\big(t_i \mid z_j\big) = \dfrac{WZ_{ij} + \beta}{\sum\limits_t WZ_{tj} + m\beta}$ — вероятность появления терма с номером $t_i$ в $j$-ой тематике (или $\phi_{ji}$).

Вероятности $P(d \mid z)$ и $P(z)$ вычислялись аналогично М-шагу в вероятностном латентно-семантическом анализе [9]:

1) $P\big(d \mid z\big) = \dfrac{\sum\limits_t TF(d,t)P\big(z \mid d,t\big)}{\sum\limits_{d'}\sum\limits_t TF(d',t)P\big(z \mid d',t\big)}$ — вероятность того, что предложение $d$ попадет в группу предложений, соответствующих тематике $z$;

2) $P(z) = \dfrac{1}{\sum\limits_d\sum\limits_t TF(d,t)} \sum\limits_d\sum\limits_t TF(d,t)P\big(z \mid d,t\big)$ — вероятность появления тематики $z$ в случайно выбранном предложении документа.

**4. Результаты экспериментального исследования.** Для оценки алгоритмов вычисления релевантности фрагментов текста использовались аннотации документов, получаемые из наиболее релевант-

ных фрагментов. Таким образом, оценка алгоритмов вычисления релевантности фрагментов текста сводилась к задаче оценки качества соответствующих алгоритмов автоматического аннотирования.

В настоящее время самым распространенным средством для оценки качества алгоритмов аннотирования является полностью автоматизированная утилита ROUGE (аббревиатура для Recall-Oriented Understudy for Gisting Evaluation) [2, 4–6, 30, 37]. На вход подается множество построенных алгоритмом аннотаций и на каждую такую аннотацию одна (или более) модельная аннотация. Далее с помощью специальных метрик аннотации сравниваются, и алгоритму присваивается некоторая оценка (средняя по всем аннотациям) [38]. Впоследствии для оценки качества различных алгоритмов автоматического аннотирования утилита ROUGE стала использоваться на крупнейшей конференции, посвященной задачам автоматического построения аннотаций Document Understanding Conference (DUC) [39].

Эталонные наборы данных включают в себя текстовые документы и модельные аннотации к ним, которые обычно строятся человеком, причем для одного документа может быть построено несколько модельных аннотаций, так как, вообще говоря, одна аннотация может быть хорошей для одного человека и, в то же время, быть плохой для другого. Наборы данных могут различаться в зависимости от вида задачи аннотирования. Для оценки методов однодокументного аннотирования, в которых аннотации строятся к каждому документу из коллекции, используются наборы данных DUC 2001 и DUC 2002 [39].

В работе [38] проводится анализ применимости различных метрик для различных задач аннотирования; в частности, для оценки алгоритмов однодокументного аннотирования на наборах DUC 2001 и DUC 2002 рекомендуется использовать метрики ROUGE-2, ROUGE-L, ROUGE-S, ROUGE-W. Кроме того, одним из результатов работы [38] является вывод, что наборы DUC 2001 и DUC 2002 обладают достаточным количеством модельных аннотаций для объективной оценки качества рассматриваемых алгоритмов (табл. 2).

Таблица 2

Эталонные наборы данных DUC

| Название набора данных | Число документов | Число модельных аннотаций |
|---|---|---|
| DUC 2001 | 297 | 925 |
| DUC 2002 | 533 | 1112 |

Для оценки качества алгоритмов аннотирования в данной работе использовались наборы DUC 2001 и DUC 2002. Из-за схожести получаемых результатов на данных наборах основные результаты тестирования алгоритмов приводятся для набора DUC 2002, поскольку он обладает бóльшим числом документов и модельных аннотаций, чем набор DUC 2001. Однако результирующие данные также приводятся для набора DUC 2001. Отметим, что набор DUC 2001 (в отличие от DUC 2002) включает в себя специальный набор модельных аннотаций (всего 147 аннотаций), в котором аннотации полностью состоят из предложений исходного текста документа (данный набор будем обозначать как DUC 2001 Extracts). Поэтому в дополнение к результатам тестирования на наборах DUC 2001 и DUC 2002 приводятся результаты и для набора DUC 2001 Extracts. Для сравнения аннотаций использовалась утилита ROUGE с метриками ROUGE-2, ROUGE-L, ROUGE-S и ROUGE-W.

В наборах DUC 2001 (в том числе DUC 2001 Extracts) и DUC 2002 модельные аннотации состоят из 100 слов, поэтому и аннотации, получаемые алгоритмами, должны состоять из 100 слов. Все рассматриваемые алгоритмы последовательно выбирали предложения текста для аннотации в порядке убывания их релевантности до тех пор, пока суммарное количество слов в выбранных предложениях не превысит 100. Затем выбранные предложения упорядочивались в порядке появления их в тексте, а получаемая таким образом аннотация подавалась на вход утилите ROUGE, которая, в свою очередь, оставляла только первые 100 слов [29] (параметры ROUGE: ROUGE-1.5.5.pl -e duc2002 -m **-l 100** -2 4 - n 2 -w 1.2 -s -a duc2002.xml).

Полное сравнение методов автоматического аннотирования, основанных на латентно-семантическом анализе (подраздел 3.1), проводилось авторами в работе [23]. Поэтому для сравнения с методами, основанными на вероятностных тематических моделях, были выбраны только методы SVD и NMF как показавшие наилучшие результаты, соответственно, для сингулярного разложения и факторизации неотрицательных матриц.

В тематических моделях число тематик $k$ выбирается существенно меньшим, чем размерности исходной матрицы текста $A \in \mathbb{R}^{m \times n}$, т.е. $k \ll \min(m, n)$ [20, 45]. Для набора DUC 2002 среднее число строк матриц документов составляло 239, а среднее число столбцов — 37. Поэтому в табл. 3, показывающей изменение качества получаемых аннотаций в зависимости от числа выбираемых тематик, приводятся

данные для числа тематик от 1 до 10. В модели скрытого распределения Дирихле параметр $\beta$ обычно задают равным 0.01, а параметр $\alpha$ равным $50/T$, где $T$ — число тематик [45]. Так как в рассматриваемой задаче анализируются документы по отдельности, то число выбираемых тематик получается гораздо меньшим, чем при анализе коллекции документов, поэтому в табл. 3 также приводятся данные для случая $\alpha = 0.01$. Из-за схожести результатов при использовании различных метрик сравнения аннотаций, в табл. 3 приводятся данные для метрики ROUGE-2. Соответственно, чем выше значение ROUGE-2-F, тем выше качество полученных алгоритмом аннотаций.

Таблица 3

Сравнение методов построения аннотаций

| Число тематик | SVD | NMF | PLSA | LDA ($\alpha = 50/T$) | LDA ($\alpha = 0.01$) |
|---|---|---|---|---|---|
| 1 | 0.18608 | 0.18607 | 0.16996 | 0.18802 | 0.18802 |
| 2 | 0.18973 | 0.19105 | **0.17019** | 0.18768 | 0.1756 |
| 3 | 0.18868 | 0.18999 | 0.16963 | 0.18655 | 0.16658 |
| 4 | 0.19027 | 0.19159 | 0.16983 | 0.18641 | 0.16227 |
| 5 | 0.18941 | 0.19022 | 0.16934 | **0.1888** | 0.15769 |
| 6 | 0.18981 | 0.18985 | 0.16964 | 0.18705 | 0.15438 |
| 7 | 0.1897 | 0.19131 | 0.16992 | 0.18512 | 0.15351 |
| 8 | 0.18925 | 0.1916 | 0.16996 | 0.1871 | 0.15391 |
| 9 | 0.18829 | 0.19121 | 0.16911 | 0.18495 | 0.14534 |
| 10 | 0.18823 | 0.19172 | 0.16992 | 0.18606 | 0.14489 |

В табл. 4, 5 и 6 (соответственно, для наборов DUC 2002, DUC 2001 и DUC 2001 Extracts) показаны значения различных метрик `ROUGE` для рассмотренных методов аннотирования, основанных на тематических моделях. Значение, следующее за "±", показывает стандартное отклонение, рассчитанное по результатам 10 запусков алгоритма. В методах `SVD` и `NMF` число тематик выбиралось исходя из соотношения длины аннотации (100 слов) к длине самого текста (подраздел 3.1). Для метода `LDA` параметр $\alpha$ рассчитывался как $50/T$. В методах `PLSA` и `LDA` число тематик фиксировалось и выбиралось, соответственно, равным 2 и 5 (подразделы 3.2, 3.3), поскольку именно при этом значении данные методы показывали наилучшее качество получаемых аннотаций (табл. 3), хотя также исследовался подход выбора числа тематик, аналогичный используемому в методах `SVD` и `NMF`.

Таблица 4

Сравнение методов аннотирования на наборе DUC 2002

| Метрики ROUGE | SVD | NMF | PLSA | LDA |
|---|---|---|---|---|
| ROUGE-2-F | 0.19052 | **0.19209**±0.0007 | 0.17019±0.0003 | 0.1888 ±0.001 |
| ROUGE-L-F | 0.37035 | **0.37126**±0.0008 | 0.33901±0.0003 | 0.36856±0.001 |
| ROUGE-S4-F | 0.15157 | **0.15316**±0.0007 | 0.1337 ±0.0002 | 0.15214±0.0008 |
| ROUGE-W-F | 0.20956 | **0.21011**±0.0006 | 0.19184±0.0002 | 0.20823±0.0007 |

В дополнение к табл. 4, 5 и 6 приводится табл. 7, в которой показаны значения этих же метрик ROUGE:

— при случайном выборе предложений для аннотации (`RANDOM`);

— при выборе самых длинных по числу слов предложений (`WORD COUNT`);

— при выборе первых предложений документа (`FIRST K`);

— при выборе последних предложений документа (`LAST K`).

Полученные данные, приведенные в табл. 3–6, показывают, что методы вычисления релевантности фрагментов текста, основанные на латентно-семантическом анализе, превосходят методы, основанные

Таблица 5

Сравнение методов аннотирования на наборе DUC 2001

| Метрики ROUGE | SVD | NMF | PLSA | LDA |
|---|---|---|---|---|
| ROUGE-2-F | 0.15932 | **0.161353**±0.0014 | 0.139259±0.0004 | 0.155089±0.0012 |
| ROUGE-L-F | 0.32521 | **0.329423**±0.0015 | 0.293128±0.0002 | 0.324411±0.0015 |
| ROUGE-S4-F | 0.1272 | **0.128735**±0.0013 | 0.109984±0.0003 | 0.125513±0.001 |
| ROUGE-W-F | 0.18397 | **0.18611** ±0.0009 | 0.166496±0.0002 | 0.182989±0.0009 |

Таблица 6

Сравнение методов аннотирования на наборе DUC 2001 Extracts

| Метрики ROUGE | SVD | NMF | PLSA | LDA |
|---|---|---|---|---|
| ROUGE-2-F | 0.33737 | **0.34228**±0.006 | 0.280724±0.002 | 0.300527±0.008 |
| ROUGE-L-F | 0.43287 | **0.4372** ±0.005 | 0.374622±0.002 | 0.399901±0.006 |
| ROUGE-S4-F | 0.3117 | **0.31595**±0.006 | 0.256316±0.002 | 0.275368±0.008 |
| ROUGE-W-F | 0.26584 | **0.26744**±0.003 | 0.221477±0.0014 | 0.241775±0.005 |

Таблица 7

Сравнение методов аннотирования на наборе DUC 2002

| Метрики | RANDOM | WORD COUNT | FIRST K | LAST K |
|---|---|---|---|---|
| ROUGE-2-F | 0.12364 | 0.16996 | 0.16308 | 0.08562 |
| ROUGE-L-F | 0.29483 | 0.33853 | 0.33034 | 0.23202 |
| ROUGE-S4-F | 0.09503 | 0.13341 | 0.12923 | 0.06480 |
| ROUGE-W-F | 0.16227 | 0.19166 | 0.19312 | 0.12694 |

на вероятностных тематических моделях, по качеству получаемых аннотаций. Кроме того, из приведенных результатов тестирования следует, что предложенный метод вычисления релевантности фрагментов текста, использующий факторизацию неотрицательных матриц для тематического моделирования, показывает лучшие результаты по сравнению со всеми рассмотренными методами (на эталонных тестовых наборах DUC 2001 и DUC 2002 с использованием стандартных метрик оценки качества аннотаций `ROUGE`).

**5. Заключение.** В настоящей статье были рассмотрены наиболее актуальные методы вычисления релевантности (значимости) фрагментов текста на основе анализа тематических моделей для последующего построения аннотаций в форме выдержек, т.е. аннотаций, полностью состоящих из последовательности фрагментов исходного текста. В качестве тематических моделей были выбраны популярные модели семантики документов и коллекции документов, используемые в задачах анализа текстовой информации: модели, основанные на латентно-семантическом анализе, модель вероятностного латентно-семантического анализа и модель скрытого распределения Дирихле. Кроме того, в статье был представлен собственный разработанный метод вычисления релевантности фрагментов текста, основанный на оценке весов тематик в нормализованном пространстве тематик, получаемом с помощью факторизации неотрицательных матриц.

Эксперименты, проведенные с методами автоматического аннотирования на эталонных тестовых наборах DUC 2001 и DUC 2002 с использованием стандартных метрик оценки качества аннотаций `ROUGE`, показали превосходство методов вычисления релевантности фрагментов текста, основанных на латентно-семантическом анализе, по сравнению с методами, основанными на вероятностных тематических моделях, по качеству получаемых аннотаций. Кроме того, приведенные результаты тестирования показывают, что предложенный метод вычисления релевантности фрагментов текста, использующий факторизацию неотрицательных матриц для тематического моделирования, показывает лучшие результаты по сравнению со всеми рассмотренными методами.

По итогам проведенного обзора литературы можно сделать вывод, что сейчас в методах автоматиче-

ского однодокументного аннотирования, в основном, используется латентно-семантический анализ, применение вероятностных тематических моделей не нашло в них широкого применения. Однако в последнее время именно вероятностное тематическое моделирование применяется в различных задачах, связанных с анализом больших коллекций документов, в том числе в задаче многодокументного аннотирования.

СПИСОК ЛИТЕРАТУРЫ


1. *Mani I., Maybury M. (Eds.)* Advances in automatic text summarization. Cambridge: MIT Press, 1999.
2. *Jezek K., Steinberger J.* Automatic text summarization (the state of the art 2007 and new challenges) // Proc. of Znalosti-2008. Bratislava, 2008. 1–12.
3. *Barzilay R., Elhadad M.* Using lexical chains for text summarization // Proc. of the ACL/EACL-97 Workshop on Intelligent Scalable Text Summarization. Madrid, 1997. 10–17.
4. *Mihalcea R., Tarau P.* Textrank: bringing order into texts // Proc. of the Conference on Empirical Methods in Natural Language Processing. Barcelona, 2004. 404-411.
5. *Steinberger J., Jezek K.* Text summarization and singular value decomposition // Lecture Notes on Computer Science. Vol. 3261 Heidelberg: Springer, 2004. 245–254.
6. *Bhandari H., Shimbo M., Ito T., Matsumoto Y.* Generic text summarization using probabilistic latent semantic indexing // The Third International Joint Conference on Natural Language Processing. Hyderabad, 2008. 133–140.
7. *Arora R., Ravindran B.* Latent Dirichlet allocation based multi-document summarization // Proc. of the Second Workshop on Analytics for Noisy Unstructured Text Data. New York: ACM Press, 2008. 91–97.
8. *Blei D.* Probabilistic topic models // Communications of the ACM. 2012. **55**, N 4. 77–84.
9. *Hofmann T.* Probabilistic latent semantic indexing // Proc. of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 1999. 50–57.
10. *Arora S., Ge R., Moitra A.* Learning topic models — going beyond SVD // Proc. of the 53rd Annual IEEE Symposium on Foundations of Computer Science (FOCS-2012). New Brunswick: IEEE Press, 2012. 1–10.
11. *Xu W., Liu X., Gong Y.* Document clustering based on non-negative matrix factorization // Proc. of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2003. 267–273.
12. *Berry M., Browne M., Langville A., Pauca V., Plemmons R.* Algorithms and applications for approximate nonnegative matrix factorization // Computational Statistics and Data Analysis. 2007. **52**, N 1. 155–173.
13. WordNet [Электронный ресурс] : A lexical database for English (http://wordnet.princeton.edu/).
14. Русский WordNet [Электронный ресурс] (http://wordnet.ru/).
15. RussNet: тезаурус русского языка [Электронный ресурс] (http://project.phil.spbu.ru/RussNet/index_ru.shtml).
16. *Mandala R., Takenobu T., Hozumi T.* The use of WordNet in information retrieval // Proc. of ACL Workshop on Usage of WordNet in Natural Language Processing Systems. Montreal, 1998. 31–37.
17. *Rakesh P., Shivapratap G., Divya G., Soman K.* Evaluation of SVD and NMF methods for latent semantic analysis // International Journal of Recent Trends in Engineering. 2009. **1**, N 3. 308–310.
18. *Blei D., Ng A., Jordan M.* Latent Dirichlet allocation // Journal of Machine Learning Research. 2003. **3**. 993–1022.
19. *Blei D., Lafferty J.* Topic models // Text Mining: Classification, Clustering, and Applications. Boca Raton: CRC Press, 2009. 71–94.
20. *Berry M.W., Dumais S.T., O'Brien G.W.* Using linear algebra for intelligent information retrieval // SIAM Review. 1995. **37**, N 4. 573–595.
21. *Ye Y.* Comparing matrix methods in text-based information retrieval. Tech Rep. School of Mathematical Sciences, Peking University. Beijing, 2000.
22. *Griffiths T.L., Steyvers M.* Finding scientific topics // Proc. of the National Academy of Sciences. 2004. **101**, N 1. 5228–5235.
23. *Mashechkin I., Petrovskiy M., Popov D., Tsarev D.* Automatic text summarization using latent semantic analysis // Programming and Computer Software. 2011. **37**, N 6. 299–305.
24. *Tsarev D., Petrovskiy M., Mashechkin I.* Using NMF-based text summarization to improve supervised and unsupervised classification // Proc. of the 11th International Conference on Hybrid Intelligent Systems (HIS). Malacca: IEEE Press, 2011. 185–189.
25. *Chisholm E., Kolda T.* New term weighting formulas for the vector space method in information retrieval // Technical Report Number ORNL-TM-13756. Oak Ridge National Laboratory. Oak Ridge, 1999.
26. *Landauer T., Dumais S.* A solution to Plato's problem: the latent semantic analysis theory of the acquisition, induction and representation of knowledge // Psychological Review. 1997. **104**. 211–240.
27. *Gong Y., Liu X.* Generic text summarization using relevance measure and latent semantic analysis // Proc. of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2001. 19–25.
28. *Lee D., Seung H.* Learning the parts of objects by non-negative matrix factorization // Nature. 1999. **401**. 788–791.
29. *Steinberger J.* Text summarization within the LSA framework. PhD Thesis. University of West Bohemia in Pilsen. Pilsen, 2007.

30. *Lee J.-H., Park S., Ahn C.-M., Kim D.* Automatic generic document summarization based on non-negative matrix factorization // Information Processing and Management: an International Journal. 2009. **45**, N 1. 20–34.
31. *Park S.* Personalized summarization agent using non-negative matrix factorization // PRICAI 2008: Trends in Artificial Intelligence. Vol. 5351. Heidelberg: Springer, 2008. 1034–1038.
32. *Park S., Lee J.-H., Kim D.-H., Ahn C.-M.* Multi-document summarization using weighted similarity between topic and clustering-based non-negative semantic feature // Advances in Data and Web Management. Vol. 4505. Heidelberg: Springer, 2007. 108–115.
33. *Liu S., Zhou M., Pan S., Qian W., Cai W., Lian X.* Interactive, topic-based visual text summarization and analysis // Proc. of the 18th ACM Conference on Information and Knowledge Management. New York: ACM Press, 2009. 543–552.
34. *Daud A., Li J., Zhou L., Muhammad F.* Knowledge discovery through directed probabilistic topic models: a survey // Frontiers of Computer Science in China. 2010. **4**, N 2. 280–301.
35. Matlab Topic Modeling Toolbox 1.4 [Электронный ресурс] (http://psiexp.ss.uci.edu/).
36. *Minka T., Lafferty J.* Expectation-propagation for the generative aspect model // UAI'02 Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers, 2002. 352–359.
37. ROUGE: Recall-Oriented Understudy of Gisting Evaluation [Электронный ресурс] (http://www.berouge.com/).
38. *Lin C.-Y.* Looking for a few good metrics: automatic summarization evaluation — how many samples are enough? // Proc. of the NTCIR-5 Workshop. Tokyo: National Institute of Informatics, 2004. 1765–1776.
39. Document Understanding Conferences [Электронный ресурс] (http://duc.nist.gov/).
40. *Lin C.-J.* Projected gradient methods for non-negative matrix factorization // Neural Computation. 2007. **19**, N 10. 2756–2779.
41. *Ding C., Li T., Peng W.* On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing // Computational Statistics & Data Analysis. 2008. **52**, N 8. 3913–3927.
42. *Mirzal A.* Converged algorithms for orthogonal nonnegative matrix factorizations // Computing Research Repository. Vol. 1010. 2010.
43. *Ding C., Li T., Peng W., Park H.* Orthogonal nonnegative matrix tri-factorizations for clustering // KDD'06 Proc. of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2006. 126–135.
44. *Lin C.-J.* On the convergence of multiplicative update algorithms for nonnegative matrix factorization // IEEE Transactions on Neural Networks. 2007. **18**, N 6. 1589–1596.
45. *Steyvers M., Griffiths T.* Probabilistic topic models // Handbook of Latent Semantic Analysis. Vol. 427. Philadelphia: Psychology Press, 2007. 414–440.
46. *Воеводин В.В., Кузнецов Ю.А.* Матрицы и вычисления. М.: Наука, 1984.

Поступила в редакцию
24.10.2012