УДК 004.912

А. Г. Яшина, ассистент, e-mail: ayashina.pmi@gmail.com,
Д. Е. Прозоров, д-р техн. наук, профессор, e-mail: prozorov.de@gmail.com
Вятский государственный университет, г. Киров

Взвешенная косинусная мера векторной модели информационного поиска речевых документов

Рассмотрена задача поиска речевых документов по текстовому запросу. Предложено использование взвешенной косинусной меры сходства в качестве функции релевантности векторной модели информационного поиска. Веса вычисляются посредством нечеткого сравнения слов на основе длины наибольшей общей подстроки, либо расстояния Левенштейна. Проведен анализ эффективности поиска речевых документов в зависимости от используемых алгоритмов сравнения слов.

Ключевые слова: *информационный поиск, речевые документы, векторная модель информационного поиска, косинусная мера*

Введение

Развитие информационных технологий привело к увеличению количества электронных документов различного типа. Существующие системы информационного поиска ориентированы в основном на обработку текстовых документов [1, 2]. Однако к настоящему времени широкое распространение получили мультимедийные документы, в которых существенная часть информации передается посредством речи. Примерами таких документов являются радио- и видеоновости, аудиокниги, записи лекций, доклады конференций и т. п. Документы, содержащие только речь (без музыки), называются речевыми [3].

Поиск речевых документов по текстовому или устному запросу относится к области *"Spoken Document Retrieval"* (SDR), которая находится на стыке распознавания речи и информационного поиска. Основное отличие поиска речевых документов от текстового поиска заключается в наличии ошибок распознавания, которые искажают изначальное содержание речевых документов, что отражается на эффективности поиска. Поэтому актуальной задачей является разработка методов вычисления оценки релевантности речевых документов запросу, в которых учитываются ошибки распознавания.

Методы поиска речевых документов по текстовому запросу с учетом ошибок распознавания можно разделить на две группы [4]. К первой группе относятся методы, основанные на распознавании слитной речи в тексте и применении алгоритмов поиска по тексту [5, 6]. Методы второй группы используют фонемное распознавание или транскрибирование речевых документов [7—10]. Существуют комбинированные методы, объединяющие оба подхода [11, 12].

Выделяют [1] следующие модели информационного поиска: булевы, векторные, вероятностные. Модель информационного поиска описывает "представление" содержания документов и запросов пользователя, а также метод вычисления оценки релевантности документа запросу. Векторную модель широко применяют при информационном поиске по тексту. Основные достоинства данной модели заключаются в простоте и наличии различных методов взвешивания слов и вычисления оценки релевантности. Документы и запросы в векторной модели представляются векторами пространства, базисные векторы которого соответствуют словам, входящим в содержание документов. Оценка релевантности документа запросу вычисляется как значение меры близости двух векторов.

В качестве меры близости широко используют косинусную меру близости (косинус угла между векторами) [1] на основе точного сравнения слов.

Существуют также модифицированные меры близости на основе нечеткого сравнения строк. Например, в работе [13] вводится "мягкая" косинусная

мера (*soft cosine measure*), применяемая для классификации текстов:

$$\cos(\boldsymbol{a}, \boldsymbol{b}) = \frac{\sum_{i=0}^{N}\sum_{j=0}^{N} s_{i,j}a_i b_j}{\sum_{i=0}^{N}\sum_{j=0}^{N} s_{i,j}a_i a_j \cdot \sum_{i=0}^{N}\sum_{j=0}^{N} s_{i,j}b_i b_j},$$

где $s_{i,j}$ — мера сходства слов сравниваемых текстов. "Мягкая" косинусная мера близости позволяет сравнивать содержание текстов в векторном пространстве, базис которого не является ортогональным.

Метод, описанный в работе [13], при сравнении текстов учитывает сходство слов, которое может быть "по смыслу", полученное на основе словаря синонимов, или "по написанию", вычисленное в результате нечеткого сравнения. Сходство слов "по написанию" в работе [13] вычисляется как расстояние Левенштейна между n-граммами слов. Сравнение строк на основе расстояния Левенштейна (расстояния редактирования) [1, 14] и метрики Джаро-Винклера [14] используется также для выявления нечетких дубликатов в текстах [14].

В данной работе описана векторная модель поиска речевых документов по текстовому запросу. В качестве функции релевантности предложена модифицированная взвешенная косинусная мера, веса которой вычисляются на основе нечеткого сравнения слов распознанного речевого документа и текстового запроса. Нечеткое сравнение слов выполняется посредством нахождения длины наибольшей общей подстроки, либо расстояния Левенштейна. Проведен анализ эффективности поиска речевых документов в зависимости от используемых алгоритмов точного и нечеткого сравнения слов.

Постановка SDR-задачи

Речевой документ d_k коллекции D является набором независимых распознанных слов $\{w_i^{d_k}|$. Текстовый запрос Q задается набором ключевых слов $\{q_j\}$, введенных пользователем.

SDR-задача [1] заключается в вычислении оценки релевантности r_k документов $d_k \in D$ запросу Q:

$$r_k = F(d_k, Q), \quad (1)$$

где F — функция релевантности d_k запросу Q. Требуется сформировать список документов, ранжированный по значению r_k.

Векторная модель информационного поиска

Пусть $T = \{t_i\}$ является множеством слов, встречающихся в коллекции D. Каждому документу $d_k \in D$ поставим в соответствие вектор $\mathbf{d}_k = (v_0, v_1, ..., v_n)$ n-мерного пространства ($n = |T|$), где v_i — вес слова $t_i \in T$ в документе $d_k \in D$. Широко используемым методом взвешивания слов документа является *tf-idf* [1, 15]:

$$v_i = tf_i \cdot idf_{d_k, i}, \quad (2)$$

где tf_i — частота вхождения слова t_i в документ d_k, а $idf_{d_k, i}$ — обратная частота документа, вычисляемая как

$$idf_{d_k, i} = \log\left(\frac{|D|}{|(d_k \supset t_i)|}\right),$$

где $|D|$ — число документов в коллекции D; $|(d_k \supset t_i)|$ — число документов, содержащих слово t_i.

Текстовому запросу Q поставим в соответствие вектор $\mathbf{q} = (u_0, u_1, ..., u_n)$, определяемый в результате точного сравнения со словами $t_i \in T$:

$$u_i = \begin{cases} 0, & t_i \notin Q, \\ 1, & t_i \in Q. \end{cases} \quad (3)$$

Значение функции релевантности $F(d_k, Q)$ документа $d_k \in D$ запросу Q вычисляется как косинусная мера близости между векторами $\mathbf{d}_k$ и $\mathbf{q}$ [1]:

$$F(d_k, Q) = \cos(\mathbf{d}_k; \mathbf{q}) = \frac{\sum_{i=1}^{n} v_i u_i}{\sqrt{\sum_{i=1}^{n} v_i^2} \cdot \sqrt{\sum_{i=1}^{n} u_i^2}}. \quad (4)$$

Взвешенная косинусная мера близости

При поиске по коллекции текстовых документов используется метод точного сравнения строк (2) [1]. Однако указанный метод менее эффективен при поиске по коллекции распознанных речевых документов, так как в тексте присутствуют слова, распознанные с ошибками. Частично учесть при поиске искажения слов, полученные в результате распознавания речевых документов, позволяют методы нечеткого сравнения строк. "Мягкая" косинусная мера [13] имеет вычислительную сложность $O(|T|^2)$, что существенно влияет на время обработки запроса.

Введем взвешенную косинусную меру близости:

$$\cos'(\mathbf{d}_k; \mathbf{q}) = \frac{\sum_{i=1}^{n} s_i v_i u_i}{\sqrt{\sum_{i=1}^{n} s_i v_i^2} \cdot \sqrt{\sum_{i=1}^{n} s_i u_i^2}}, \quad (5)$$

где s_i — вес слова $t_i \in T$, $u_i = 1$, $i = 1...n$.

Для определения веса s_i необходимо найти меру сходства между словом $t_i \in T$ и словами запроса $q_j \in Q$.

В данной работе для решения рассматриваемой задачи предлагается вычислять значение меры сходства на основе нечеткого сравнения слов. В совре-

менных системах, в которых используют алгоритмы нечеткого сравнения строк, например, для исправления опечаток при вводе запроса пользователем, применяют алгоритмы поиска длины наибольшей общей подстроки и расстояния Левенштейна.

Мера сходства g_{t_i, q_j} на основе нормированной длины наибольшей общей подстроки равна

$$g_{t_i, q_j} = \frac{|s_{q_j, t_i}|}{\max(|q_j|; |t_i|)}, \quad (6)$$

где $|s_{q_j, t_i}|$ — длина наибольшей общей подстроки строк q_j и t_i, а $|s|$ — длина строки s.

Расстояние Левенштейна [1, 14] позволяет вычислять сходство двух строк посредством определения минимального числа операций редактирования, требуемых для преобразования одной строки во вторую. Операциями редактирования являются "вставка", "удаление" и "замена" символа. Для вычисления расстояния Левенштейна применяют алгоритм динамического программирования Вагнера—Фишера [16].

Мера сходства на основе расстояния Левенштейна:

$$g_{t_i, q_j} = 1 - \frac{D(q_j, t_i)}{|q_j| + |t_i|}, \quad (7)$$

где $D(q_j, t_i)$ — значение расстояния Левенштейна между строками q_j и t_i.

Тогда веса $\{s_i\}$, $i = \overline{1, n}$, можно вычислить, определив слова $t_i \in T$, для которых g_{t_i, q_j} максимально:

$$s_i = \begin{cases} g_{t_i, q_j}, \exists\, q_j \in Q\colon g_{t_i, q_j} = \max\limits_{i = 1, \ldots, n} (g_{t_i, q_j}) \\ 0, \text{ в противном случае.} \end{cases} \quad (8)$$

Рассмотрим в качестве примера коллекцию из трех документов. В табл. 1 представлено верное содержание документов и содержание, полученное в результате распознавания.

Множество T всех слов, встречающихся в коллекции, содержит девять слов. В табл. 2 показаны *tf-idf* веса (2) слов множества T; для вычисления значения *idf* использовали десятичный логарифм.

Векторы соответствующих документов равны

$$\mathbf{d}_1 = (0{,}48;\ 0{,}48;\ 0{,}48;\ 0;\ 0;\ 0;\ 0;\ 0;\ 0),$$

$$\mathbf{d}_2 = (0;\ 0;\ 0;\ 0{,}48;\ 0{,}18;\ 0{,}48;\ 0;\ 0;\ 0),$$

$$\mathbf{d}_3 = (0;\ 0;\ 0;\ 0;\ 0{,}18;\ 0;\ 0{,}48;\ 0{,}48;\ 0{,}48).$$

Пусть требуется вычислить значения релевантности документов относительно запроса "заметали следы". Вычислим значения меры сходства для каждого слова $t_i \in T$ относительно слов запроса. В табл. 3 представлены значения весов s_i, которые получены на основе расчета мер сходства (6) и (7).

В соответствии с правилом (8) вектор весовых коэффициентов косинусной меры (5) формируется на основе максимальных значений мер сходства (табл. 3):

$$\mathbf{s}_{\text{sub}} = (0;\ 0;\ 0;\ 0;\ 0{,}5;\ 1;\ 0;\ 0;\ 0),$$

$$\mathbf{s}_{\text{Lev}} = \{0;\ 0;\ 0;\ 0;\ 0{,}8;\ 1;\ 0;\ 0;\ 0),$$

где $\mathbf{s}_{\text{sub}}$ — вектор весовых коэффициентов, вычисленных как мера сходства на основе длины наибольшей общей подстроки (6), а $\mathbf{s}_{\text{Lev}}$ — вектор весовых коэффициентов, вычисленных как мера сходства на основе расстояния Левенштейна (7).

Таблица 1

Содержание речевых документов коллекции

Документ	Верное содержание	Распознанное содержание
Документ 1	Торжественно гарцевавших	Торжественно гонцы ваших
Документ 2	Заметали следы	За мечтали следы
Документ 3	Мечтали о золотом веке	Мечтали по золотому веки

Таблица 2

tf-idf* веса слов множества *T

Слова	*idf*	Документ 1		Документ 2		Документ 3	
		tf	*tf-idf*	*tf*	*tf-idf*	*tf*	*tf-idf*
торжественно	0,48	1	0,48	0	0	0	0
горцы	0,48	1	0,48	0	0	0	0
ваших	0,48	1	0,48	0	0	0	0
за	0,48	0	0	1	0,48	0	0
мечтали	0,18	0	0	1	0,18	1	0,18
следы	0,48	0	0	1	0,48	0	0
по	0,48	0	0	0	0	1	0,48
золотому	0,48	0	0	0	0	1	0,48
веки	0,48	0	0	0	0	1	0,48

Таблица 3

Значения мер сходства

$g^{substr}_{t_i, q_j}$			$g^{Levenshtein}_{t_i, q_j}$		
слова	заметали	следы	слова	заметали	следы
торжественно	0,083	0,083	торжественно	0,5	0,412
горцы	0	0,2	горцы	0,385	0,6
ваших	0,125	0	ваших	0,462	0,5
за	0,25	0	за	0,4	0,286
мечтали	**0,5**	0,143	мечтали	**0,8**	0,417
следы	0,125	**1**	следы	0,462	**1**
по	0	0	по	0,2	0,286
золотому	0,125	0,125	золотому	0,625	0,462
веки	0,125	0,2	веки	0,5	0,556

Примечание: $g^{substr}_{t_i, q_j}$ — значение меры сходства, вычисленные на основе длины наибольшей общей подстроки (6), $g^{Levenshtein}_{t_i, q_j}$ — значение меры сходства на основе расстояния Левенштейна (7).

Таблица 4

Значения взвешенного косинуса и ранг документов

Документ	Мера сходства для вычисления s_i		Ранг
	Длина наибольшей общей подстроки	Расстояние Левенштейна	
Документ 1	0	0	3
Документ 2	0,936	0,917	1
Документ 3	0,577	0,667	2

Используя (5), получим значения взвешенной косинусной меры (табл. 4).

В результате (табл. 4) речевые документы будут ранжированы по значению релевантности: документ 2, документ 3 и документ 1.

Эксперимент

В ходе эксперимента оценивали эффективность векторной модели поиска с использованием точного и нечеткого сравнения слов запроса со словами распознанных речевых документов. Анализировали следующие методы.

Метод 1 — определение релевантности документа запросу на основе точного совпадения слов (3) и косинусной меры (4).

Метод 2 — определение релевантности документа запросу на основе нечеткого сравнения слов и взвешенной косинусной меры (5), где веса s_i вычисляются как нормированная длина наибольшей общей подстроки (6).

Метод 3 — определение релевантности документа запросу на основе нечеткого сравнения слов и взвешенной косинусной меры (5), где веса s_i определяются на основе расстояния Левенштейна (7).

В различных системах, включающих поиск по текстовым документам, широко распространено использование свободной библиотеки высокоскоростного полнотекстового поиска *Lucene* [17, 18]. Поэтому для сравнения результатов дополнительно выполнялся поиск на основе данной библиотеки.

Метод 4 — полнотекстовый поиск.

Метод 5 — полнотекстовый поиск с неточным соответствием слов.

Эффективность поиска оценивали показателями полноты R и точности P, полученными в результате эксперимента на разработанной SDR-системе. В соответствии с работой [1] полнота определяется как доля найденных релевантных документов среди всех релевантных, а точность — доля релевантных документов среди найденных.

Оценку указанных показателей выполняли по 253 поисковым запросам на коллекции из 100 речевых документов. Для распознавания речи использована библиотека *pocketsphinx* [19], акустическая и языковая модели [20].

Полученные результаты представлены в табл. 5. Запись "> 0" означает, что релевантными документами считались те, для которых оценка релевантности запросу r_k (1) положительна. Запись "$t = N$" означает, что релевантными считались документы, имеющие ранг не более N.

Совместной оценкой полноты и точности является F_1-мера [1]:

$$F_1 = \frac{2RP}{R+P}.$$

Значения F_1-меры приведены в табл. 6.

Графики полноты/точности, построенные по полученным значениям, представлены на рисунке. Цифрами обозначены номера исследуемых методов.

Таблица 5

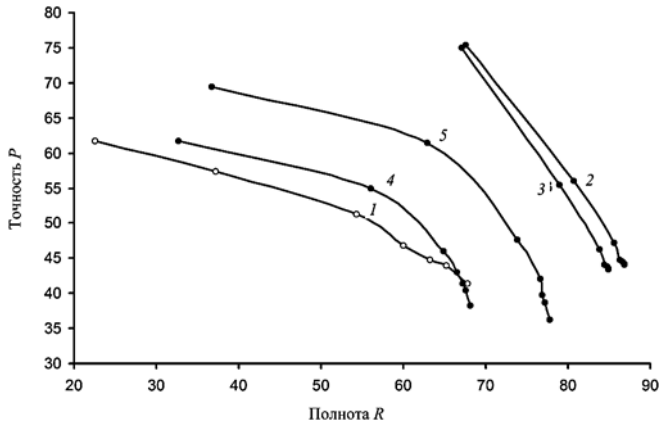
Показатели полноты и точности поиска

Метод	Значения полноты						
	>0	$t = 1$	$t = 2$	$t = 4$	$t = 6$	$t = 8$	$t = 10$
1	67,8	22,5	37,2	54,3	**60,0**	63,3	65,3
2	86,9	**67,5**	80,7	85,6	86,2	86,7	86,9
3	84,9	**67,1**	79,0	83,9	84,5	84,8	84,9
4	68,2	32,7	**56,0**	64,8	66,5	67,3	67,6
5	77,8	36,8	**62,9**	73,8	76,6	76,9	77,2
Метод	Значения точности						
	>0	$t = 1$	$t = 2$	$t = 4$	$t = 6$	$t = 8$	$t = 10$
1	41,3	61,7	57,3	51,3	**46,8**	44,8	43,9
2	44,1	**75,4**	56,0	47,2	44,7	44,5	44,2
3	43,4	**74,9**	55,4	46,2	44,1	43,8	43,5
4	38,3	61,7	**54,9**	46,0	43,0	41,4	40,4
5	36,3	69,4	**61,4**	47,7	42,1	39,8	38,7

Таблица 6

Значения F_1-меры (F_1)

Метод	>0	$t = 1$	$t = 2$	$t = 4$	$t = 6$	$t = 8$	$t = 10$
	F_1	F_1	F_1	F_1	F_1	F_1	F_1
1	51,4	33,0	45,1	52,8	**52,6**	52,5	52,5
2	58,5	**71,2**	66,1	60,9	58,9	58,8	58,6
3	57,4	**70,8**	65,1	59,6	57,9	57,8	57,5
4	49,1	42,7	**55,5**	53,8	52,2	51,2	50,6
5	49,5	48,1	**62,2**	57,9	54,3	52,4	51,5



Графики полноты/точности исследуемых методов (цифрами обозначены номера методов)

Заключение

Полученные результаты (табл. 5, 6 и рисунок) позволяют сделать следующие выводы:

1) наилучшее качество поиска речевых документов по текстовому запросу показали методы 2 и 3, включающие нечеткое сравнение слов и вычисление оценки релевантности посредством взвешенной косинусной меры (5);

2) наихудшее качество поиска из рассмотренных методов показали методы точного сравнения 1 и 4;

3) среднее качество поиска показал метод 5 (полнотекстовый поиск с неточным соответствием слов, реализованный в библиотеке *Lucene*);

4) использование длины наибольшей общей подстроки (6) при вычислении оценки релевантности документа запросу как взвешенной косинусной меры (5) в отличие от расстояния Левенштейна (7) в среднем улучшает полноту поиска на 2 % и точность поиска на 1 %;

5) использование предложенной взвешенной косинусной меры (5) для вычисления оценки релевантности повышает показатель F_1-меры поиска на 15 % относительно метода полнотекстового поиска с неточным соответствием слов, реализованного в библиотеке *Lucene* [16];

6) как видно из табл. 5 и 6, при использовании векторной модели поиска на основе предложенной взвешенной косинусной меры (5) ложно-определенные системой релевантные документы реже попадают в верхние ранги ранжированного результирующего списка по сравнению с традиционной векторной моделью и полнотекстовым поиском, реализованном в *Lucene*.

Список литературы

1. **Маннинг К. Д., Рагхаван П., Шютце Х.** Введение в информационный поиск. М.-СПб.-К.: Вильямс, 2011. 520 с.
2. **Лакаев А. С.** Контекстные технологии — новое направление развития информационных технологий анализа текстовой информации // Информационные технологии. 2013. № 12. С. 10—16.
3. **Larson M., Jones G. J. F.** Spoken Content Retrieval: A Survey of Techniques and Technologies // Information Retrieval. 2011. V. 5, N 4—5. P. 235—422.
4. **Wechsler M.** New Approaches to Spoken Document Retrieval, Information Retrieval // Information Retrieval. 2000. V. 3. P. 173—188.
5. **Jones G., Foote J., Jones K. S., Young S.** Video mail retrieval using voice: An overview of the stage-2 system // In: van Rijsbergen C., Ed., Proc. of the Final Workshop on Multimedia Information Retrieval (MIRO'95), Electronic Workshops in Computing. Glasgow: Springer, 1995.
6. **Wactlar H., Hauptmann A., Witbrock M.** Informedia: News-on-demand experiments in speech recognition // In: Proceedings of DARPA Speech Recognition Workshop. Arden House, Harriman, NY. 1996.
7. **Wechsler M.** Spoken Document Retrieval Based on Phoneme Recognition: PhD Thesis. Diss. No. 12879. ETH Zurich. 1998.
8. **Ng K., Zue V. W.** Phonetic Recognition for spoken document retrieval // Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing. 1998. V. 1, pp. 325—328.
9. **Яшина А. Г.** Алгоритм контекстного поиска речевых аудио-файлов на основе фонемного сравнения слов // Advanced Science, 2012. № 1. С. 73—85. URL: http://www.vyatsu.ru/uploads/file/1210/1_(2).pdf (дата обращения: 12.03.2015).
10. **Прозоров Д. Е., Яшина А. Г.** Анализ алгоритмов фонемного транскрибирования в задачах контекстного поиска речевых документов // Инфокоммуникационные технологии. 2013. Т. 12, № 4. С. 62—65.
11. **Witbrock M., Hauptmann A. G.** Speech recognition and information retrieval: Experiments in retrieving spoken documents // In: Proceedings of the DARPA Speech Recognition Workshop. Chantilly Virginia. 1997.
12. **Brown M., Foote J., Jones G., Jones K. S., Young S.** Open-vocabulary speech indexing for voice and video mail retrieval // In: ACM Multimedia Conference. Boston, MA. 1996.
13. **Sidorov G., Gelbukh A., Gomez-Adorno H., Pinto D.** Soft similarity and soft cosine measure: similarity of features in vector space model // Computaciony Sistemas. 2014. V. 18, N 3.
14. **Bilenko M., Mooney R., Cohen W., Ravikumar P., Fienberg S.** Adaptive name matching in information integration // IEEE Intelligent Systems. 2003. P. 16—23.
15. **Salton G., Buckley C.** Term-weighting approaches in automatic text retrieval // Information Processing and Management. 1988. V. 24, N 5, pp. 513—523.
16. **Wagner R., Fischer M.** The String-to-String Correction Problem // Journal of the Association for Computing Machinery. 1974. V. 21, N. 1, pp. 168—173.
17. **Hatcher E., Gospodnetic O., McCandless M.** Lucene in Action // Manning — 2009.
18. **Lucene.NET.** URL: https://www.nuget.org/packages/Lucene.Net (дата обращения: 12.03.2015).
19. **CMU** Sphinx. Open Source Toolkit For Speech Recognition // URL: http://cmusphinx.sourceforge.net (дата обращения: 12.03.2015).
20. **Voxforge-ru-0.2.** URL: http://sourceforge.net/projects/cmusphinx/files/Acoustic and Language Models/Russian Voxforge (дата обращения: 12.03.2015).

A. G. Yashina, Assistant, ayashina.pmi@gmail.com, **D. E. Prozorov,** PhD, Professor, prozorov.de@gmail.com
Vyatka State University, Kirov

Spoken Document Retrieval Vector Model Based on Weighted Cosine Measure

This paper presents a weighted cosine measure as a relevant function in a spoken document retrieval vector model. Measure weights are based on an approximate string matching. A length of the longest common substring and a Levenshtein distance are used in the weighted cosine measure. Search techniques have been evaluated using the collection which is consisted of 100 spoken documents on the russian language. Experiment results are analyzed. Results show that the weighted cosine measure improves a spoken document retrieval based on a vector model in comparison with a standard cosine measure and text retrieval methods of a Lucene search engine library.

Keywords: *information retrieval, spoken document retrieval, information retrieval vector model, cosine measure*

References

1. **Manning K. D., Raghavan P., Schutze X.** *Introduction to information search*, M.-SPb.-K.: Williams, 2011, 520 p.

2. **Lakayev A. S.** Contextual technologies — the new direction of development of the text information analysis information technologies, *Information technologies*, 2013, no. 12, pp. 10—16.

3. **Larson M., Jones G. J. F.** Spoken Content Retrieval: A Survey of Techniques and Technologies, *Information Retrieval*, 2011, vol. 5, no. 4—5, pp. 235—422.

4. **Wechsler M.** New Approaches to Spoken Document Retrieval, Information Retrieval, *Information Retrieval*, 2000, vol. 3, pp. 173—188.

5. **Jones G., Foote J., Jones K. S., Young S.** Video mail retrieval using voice: An overview of the stage-2 system, Ed. van Rijsbergen C., *Proc. of the Final Workshop on Multimedia Information Retrieval (MIRO'95), Electronic Workshops in Computing*, Glasgow, Springer, 1995.

6. **Wactlar H., Hauptmann A., Witbrock M.** Informedia: News-on-demand experiments in speech recognition, *Proc. of DARPA Speech Recognition Workshop*, NY., Arden House, Harriman, 1996.

7. **Wechsler M.** Spoken Document Retrieval Based on Phoneme Recognition: PhD Thesis. Diss. 1998, no. 12879, ETH Zurich.

8. **Ng K., Zue V. W.** Phonetic Recognition for spoken document retrieval, *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1998, vol. 1, pp. 325—328.

9. **Yashina A. G.** The spoken doccument retrieval algorithm based on a phonemic word matching, *Advanced Science*, 2012, no. 1, pp. 73—85, available at: http://www.vyatsu.ru/uploads/file/1210/1_(2).pdf

10. **Prozorov D. E., Yashina A. G.** The analysis of phonemic transcription algorithms in a spoken document retrieval, *Infocommunication technologies*, Samara, 2013, vol. 12, no. 4, pp. 62—65.

11. **Witbrock M., Hauptmann A. G.** Speech recognition and information retrieval: Experiments in retrieving spoken documents, *Proc. of the DARPA Speech Recognition Workshop*, Chantilly Virginia, 1997.

12. **Brown M., Foote J., Jones G., Jones K. S., Young S.** Open-vocabulary speech indexing for voice and video mail retrieval, *ACM Multimedia Conference*, Boston, MA., 1996.

13. **Sidorov G., Gelbukh A., Gomez-Adorno H., Pinto D.** Soft similarity and soft cosine measure: similarity of features in vector space model, *Computaciony Sistemas*, 2014, no. 18 (3).

14. **Bilenko M., Mooney R., Cohen W., Ravikumar P., Fienberg S.** Adaptive name matching in information integration, *IEEE Intelligent Systems*, 2003, pp. 16— 23.

15. **Salton G., Buckley C.** Term-weighting approaches in automatic text retrieval, *Information Processing & Management*, 1988, no. 24 (5), pp. 513—523.

16. **Wagner R., Fischer M.** The String-to-String Correction Problem, *Journal of the Association for Computing Machinery*, 1974, vol. 21, no. 1, pp. 168—173.

17. **Hatcher E., Gospodnetic O., McCandless M.** *Lucene in Action*, Manning, 2009.

18. **Lucene.NET,** available at: https://www.nuget.org/packages/Lucene.Net

19. **CMU** Sphinx. Open Source Toolkit For Speech Recognition, available at: http://cmusphinx.sourceforge.net

20. **Voxforge-ru-0.2,** available at: http://sourceforge.net/projects/cmusphinx/files/ Acoustic and Language Models/Russian Voxforge

Адрес редакции:

107076, Москва, Стромынский пер., 4

Телефон редакции журнала **(499) 269-5510**

E-mail: it@novtex.ru

Технический редактор *Е. В. Конова.*

Корректор *Е. В. Комиссарова.*

Сдано в набор 07.07.2015. Подписано в печать 20.08.2015. Формат 60×88 1/8. Бумага офсетная.

Усл. печ. л. 8,86. Заказ IT915. Цена договорная.

Журнал зарегистрирован в Министерстве Российской Федерации по делам печати, телерадиовещания и средств массовых коммуникаций.

Свидетельство о регистрации ПИ № 77-15565 от 02 июня 2003 г.

Оригинал-макет ООО "Адвансед солюшнз". Отпечатано в ООО "Адвансед солюшнз".

119071, г. Москва, Ленинский пр-т, д. 19, стр. 1.