ISSN 1816-0301 (Print)
ISSN 2617-6963 (Online)
УДК 004.912

Поступила в редакцию 03.06.2019
Received 03.06.2019

Принята к публикации 18.06.2019
Accepted 18.06.2019

Веб-поиск и адресное распространение информации на основе моделирования вербальных ассоциаций

С. Ф. Липницкий

Объединенный институт проблем информатики
Национальной академии наук Беларуси, Минск, Беларусь
E-mail: lipn@newman.bas-net.by

Аннотация. Предлагается математическая модель процессов сканирования веб-сайтов и адресного (избирательного) распространения найденной текстовой информации по запросам пользователей в виде их информационных профилей, т. е. накопленных архивов релевантных интернет-публикаций. Функциональными компонентами такой информационной системы являются подсистемы индексирования текстов, архивов пользователей и кратких сообщений, сканирования веб-страниц и адресной рассылки текстов и кратких сообщений пользователям. Индексирование текстов, архивов пользователей и кратких сообщений сводится к построению их вербально-ассоциативных сетей. В состав подсистемы индексирования входит совокупность лингвистических словарей для вычисления информативности слов и вербально-ассоциативных связей между ними. Словари формируются на основе использования публикаций из архивов пользователей. Сканирование веб-страниц осуществляется на основе программных решений в виде специализированных агентов, основная задача которых – систематическое получение и накопление новых данных из обновленных страниц. Сканирование реализуется в порядке, определяемом специальным упорядочивающим отношением, которое задается на множестве веб-страниц каждого сканируемого веб-сайта. Рассылка найденных публикаций происходит путем сравнения вербально-ассоциативных сетей этих публикаций и информационных профилей пользователей в виде поисковых образов.

Ключевые слова: веб-поиск, вербально-ассоциативные связи, информационный профиль, лингвистические словари, математическая модель, релевантность

Для цитирования. Липницкий, С. Ф. Веб-поиск и адресное распространение информации на основе моделирования вербальных ассоциаций / С. Ф. Липницкий // Информатика. – 2019. – Т. 16, № 3. – С. 79–88.

Web-search and address distribution of information on the basis of modeling of verbal associations

Stanislav F. Lipnitsky

The United Institute of Informatics Problems
of the National Academy of Sciences of Belarus, Minsk, Belarus
E-mail: lipn@newman.bas-net.by

Abstract. A mathematical model is proposed for the processes of websites scanning and addressing (selective) distribution of the text information found after the request of users in the form of their information profiles, i. e. accumulated archives of relevant Internet publications. The functional components of such an information system are three subsystems: the subsystem of texts indexing, user archives and short messages; web page scanning subsystem; the subsystem of address distribution of texts and short messages to users. Indexing the texts, archives of users and short messages is reduced to the construction of their verbal-associative networks. The indexing subsystem includes a set of linguistic dictionaries for calculating the informational content of words and verbal-associative links between them. Dictionaries are formed based on the use of publications from user's archives. Scanning the web pages is carried out on the basis of software solutions in the form of

specialized agents, whose main task is to obtain systematically and accumulate new data from updated pages. Scanning is implemented as the sequence determined by a special ordering relationship, which is fixed on the set of web pages of each web site scanned. Distribution of found publications occurs by comparing the verbal-associative networks of these publications and users' information profiles in the form of search images.

Keywords: web-search, verbal-associative network, information profile, linguistic dictionaries, mathematical model, relevance

For citation. Lipnitsky S. F. Web-search and address distribution of information on the basis of modeling of verbal associations. *Informatics*, 2019, vol. 16, no. 3, pp. 79–88 (in Russian).

Введение. Адресное (избирательное) распространение информации – это индивидуальное информирование о новых публикациях с учетом информационных потребностей пользователей. Первые информационные системы подобного назначения появились более полувека назад [1]. В них использовались главным образом ручные методы поиска и распространения информации. В настоящее время этот вид информационного обслуживания приобретает особую актуальность в связи с существованием большого количества интернет-сервисов, основанных на веб-технологиях [2].

К системам адресного распространения информации предъявляется ряд требований [3]. Назовем наиболее существенные из них:

– оперативность и регулярность рассылки новых публикаций;

– изложение краткого содержания каждой публикации в виде реферата, аннотации или набора ключевых слов;

– наличие обратной связи с пользователями рассылаемой информации для своевременной корректировки их информационных профилей.

В настоящей статье предлагается математическая модель процессов веб-поиска и адресного распространения текстовой информации. В отличие от подходов к интернет-мониторингу публикаций в других системах (см., например, [2]) разработанные в рамках модели алгоритмы основаны на использовании предложенных автором вербально-ассоциативных сетей в качестве знаний об информационных профилях пользователей [4]. Вершинами таких сетей являются словоформы, а ребра соответствуют вербально-ассоциативным связям между ними. Использование вербально-ассоциативных сетей обеспечивает адаптацию алгоритмов веб-поиска к информационным профилям, представленным в виде поисковых образов совокупностей релевантных публикаций.

Архитектура информационной системы. Функциональными компонентами системы веб-поиска и адресного распространения текстовой информации являются три подсистемы (рис. 1):

– индексирования текстов, архивов пользователей и кратких сообщений;

– сканирования веб-страниц;

– адресной рассылки текстов и кратких сообщений пользователям.

Индексирование текстов, совокупностей текстов и кратких сообщений сводится к построению их вербально-ассоциативных сетей. В состав подсистемы индексирования входит совокупность лингвистических словарей для вычисления информативности слов и вербально-ассоциативных связей между ними. Словари формируются на основе использования специальных наборов публикаций по каждой предметной области – тематических архивов пользователей. При программной реализации информационной системы поисковые образы текстов представляются в виде множеств слов и вербально-ассоциативных пар слов с соответствующими значениями информативности.

Для сканирования веб-страниц используются программные решения в виде специализированных агентов, основная задача которых – систематическое получение и накопление новых данных из обновленных страниц. Сканирование реализуется в порядке, определяемом специальным упорядочивающим отношением, которое задается на множестве веб-страниц каждого сканируемого веб-сайта.

Рассылка пользователям найденных публикаций осуществляется путем сравнения вербально-ассоциативных сетей этих публикаций и профилей пользователей в виде поисковых образов их архивов релевантных публикаций.

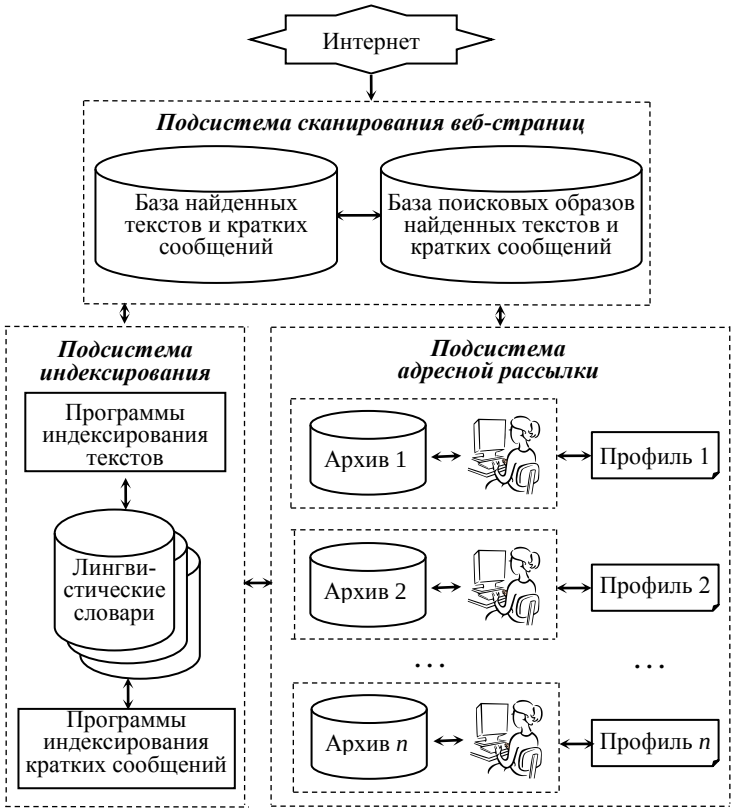

Рис. 1. Структурная схема информационной системы

Функциональная схема системы веб-поиска и адресного распространения найденной информации представлена на рис. 2. Полученные в результате сканирования веб-страниц тексты индексируются и накапливаются в базе данных. Одновременно формируется также база поисковых образов текстов и кратких сообщений. Далее тексты распределяются по архивам пользователей в соответствии с их информационными профилями.

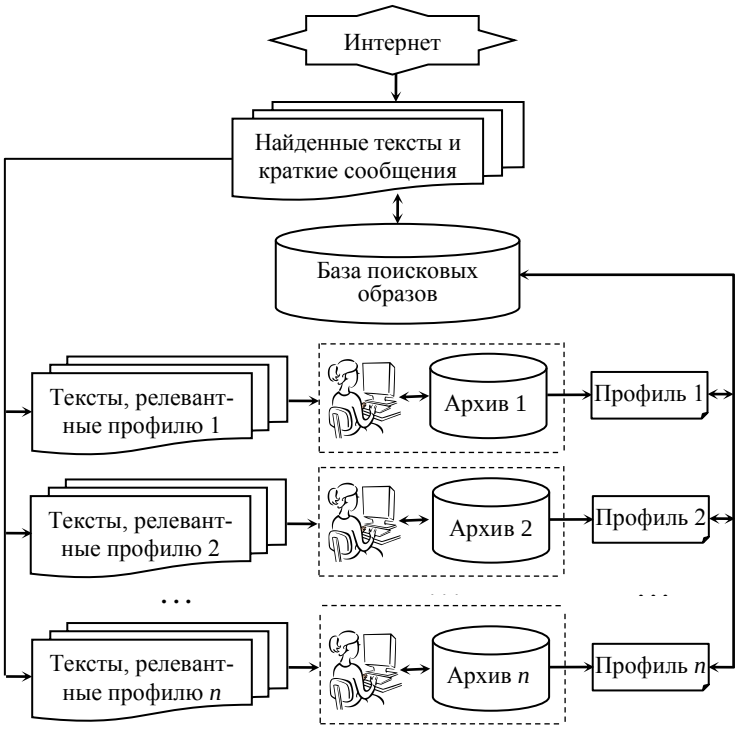

Рис. 2. Функциональная схема информационной системы

Индексирование текстов, архивов пользователей и кратких сообщений реализуется на основе использования семантических связей между словами. Промоделируем эти связи в виде вербально-ассоциативных сетей.

Индексирование текстов. Рассмотрим персональные архивы пользователей A_i $(i=\overline{1,m})$ и объединение всех архивов $U=\bigcup_{i=1}^{m} A_i$. Обозначим через W множество всех слов объединения U. Тогда отношение толерантности Θ (рефлексивное и симметричное бинарное отношение) на множестве W назовем отношением вербально-ассоциативной связи слов в объединенном архиве U, если любая упорядоченная пара слов (a, b) из множества W является элементом отношения Θ тогда и только тогда, когда слова a и b из этой пары содержатся хотя бы в одном предложении множества U.

Определим на множестве W антирефлексивное бинарное отношение Ω, такое, что для любых слов $a, b \in W$ соотношение $(a, b) \in \Omega$ выполняется тогда и только тогда, когда в архиве U существует предложение π, в котором слово a непосредственно предшествует слову b. Отношение Ω будем называть отношением дискурсивной сочетаемости слов в архиве U.

Формализуем понятие поискового образа текста.

Пусть T – произвольный текст, найденный в процессе сканирования веб-страниц. Обозначим через W_T множество всех слов текста T, а через Θ_T – сужение отношения Θ на множество W_T, т. е. $\Theta_T = \Theta \cap (W_T \times W_T)$. Отношение Θ_T назовем отношением вербально-ассоциативной связи слов в тексте T. Пару (a, b) любых слов из множества W_T, которая является элементом отношения Θ_T, т. е. $(a, b) \in \Theta_T$, будем называть вербально-ассоциативной парой текста T.

Построим также сужение Ω_T отношения Ω на множество W_T, т. е. $\Omega_T = \Omega \cap (W_T \times W_T)$. Отношение Ω_T назовем отношением дискурсивной сочетаемости слов в тексте T.

Обозначим через G_T граф отношения Θ_T. Пометим каждую вершину a графа G_T значением информативности I_T^a этого слова (с учетом синонимии и словоизменения), а каждое ребро (a, b) – значением информативности I_T^{ab} вербально-ассоциативной связи слов a и b в тексте T (также учитывая синонимию и словоизменения). Информативность I_T^a вычислим по формуле

$$I_T^a = n_T^a / n_U^a, \tag{1}$$

а информативность I_T^{ab} – по формуле

$$I_T^{ab} = n_T^{ab} / n_U^{ab} \tag{2}$$

из статьи [4]. В формуле (1) n_T^a и n_U^a – абсолютные частоты встречаемости слова a (с учетом синонимии и словоизменений) в тексте T и в объединении множеств U. В формуле (2) n_T^{ab}, n_U^{ab} – абсолютные частоты совместной встречаемости слов a и b (с учетом синонимии и словоизменений) в одном и том же предложении текста T и множества U.

Информация о частотах словоформ, а также о парадигматике и синонимии слов хранится в специальных лингвистических словарях [5, 6]:

– частотном словаре словоформ $Dic_a = \{\langle a,\ n_U^a,\ n_{A_1}^a,\ n_{A_2}^a,\ \ldots,\ n_{A_n}^a \rangle |\ a \in W\}$, в котором каждой словоформе приписаны частоты ее встречаемости n_U^a, $n_{A_1}^a$, $n_{A_2}^a$, …, $n_{A_n}^a$ в объединенном архиве U и во всех персональных архивах A_i $(i=\overline{1,m})$;

– частотном словаре вербально-ассоциативных пар слов $Dic_{ab} = \{\langle (a, b),\ n_U^{ab},\ n_{A_1}^{ab},\ n_{A_2}^{ab},\ \ldots,\ n_{A_n}^{ab} \rangle |\ a, b \in W,\ n_U^{ab} \neq 0,\ n_{A_i}^{ab} \neq 0,\ i=\overline{1,m}\}$, где n_U^{ab}, $n_{A_i}^{ab}$ – абсолютные частоты совместной встречаемости слов a и b в одном и том же предложении объединенного архива U и i-го персонального архива A_i $(i=\overline{1,m})$;

– словаре словоизменительных парадигм $Dic_{par} = \{(a, Par_a) \mid a \in W, a \in Par_a\}$, состоящем из пар ⟨*словоформа*, *парадигма*⟩. В позиции парадигмы Par_a представлены все словоизменения данной словоформы a;

– словаре синонимичных словоформ $Dic_{syn} = \{(a, Syn_a) \mid a \in W, a \in Syn_a\}$, включающем в себя пары ⟨*словоформа*, *синонимичные словоформы*⟩, в которых каждой словоформе a соответствует множество ее синонимов Syn_a.

Используя лингвистические словари, формулу (1) перепишем в виде

$$I_T^a = \frac{n_T^a + n_T^{Par_a} + n_T^{Syn_a}}{n_U^a + N_U^{Par_a} + N_U^{Syn_a}}, \tag{3}$$

где $n_T^{Par_a}$ – число вхождений всех словоформ текста T, являющихся словоизменениями словоформы a,

$$n_T^{Par_a} = \sum_{b \in Par_a,\, b \neq a} n_T^b;$$

$n_T^{Syn_a}$ – количество синонимов словоформы a в тексте T,

$$n_T^{Syn_a} = \sum_{c \in Syn_a,\, c \neq a} n_T^c,$$

аналогично

$$N_U^{Par_a} = \sum_{b \in Par_a,\, b \neq a} n_U^b,\ N_U^{Syn_a} = \sum_{c \in Syn_a,\, c \neq a} n_U^c.$$

Формулу (2) перепишем по аналогии с формулой (1):

$$I_T^{ab} = \frac{n_T^{ab} + n_T^{Par_{ab}} + n_T^{Syn_{ab}}}{n_U^{ab} + N_U^{Par_{ab}} + N_U^{Syn_{ab}}}, \tag{4}$$

где n_U^{ab}, n_T^{ab} – абсолютные частоты совместной встречаемости слов a и b в одном и том же предложении объединенного архива U и текста T.

Параметр $n_T^{Par_{ab}}$ в формуле (4) указывает на число вхождений всех пар словоформ, являющихся словоизменениями соответственно слов a и (или) b и входящих в одно и то же предложение текста T:

$$n_T^{Par_{ab}} = \sum_{\substack{c \in Par_a,\, d \in Par_b,\\ c \neq a \text{ и(или) } d \neq b\\ c,d \in \rho,\, \rho \in T}} n_T^{cd}.$$

Подобное выражение можно записать и для параметра $n_T^{Syn_{ab}}$:

$$n_T^{Syn_{ab}} = \sum_{\substack{c \in Syn_a,\, d \in Syn_b,\\ c \neq a \text{ и(или) } d \neq b\\ c,d \in \rho,\, \rho \in T}} n_T^{df}.$$

Для параметров $N_U^{Par_{ab}}$ и $N_U^{Syn_{ab}}$ верны аналогичные выражения, отличающиеся тем, что в каждом из них индекс T заменяется на U.

Пусть (a, b) – произвольное ребро графа G_T. Если $(a, b) \in \Omega_T$, то для всех таких пар (a, b) вершины a и b соединим дугой, направленной от a к b. Обозначим полученный смешанный граф Net_T и назовем его вербально-ассоциативной сетью текста T.

При практической реализации информационной системы сеть Net_T целесообразно представить в виде

$$Net_T = \{\langle (a, I_T^a); (b, I_T^b); (I_T^{ab}, Arc)\rangle \mid a \in T, b \in T\}, \quad (5)$$

где $Arc = 1$, если $(a, b) \in \Omega_T$; $Arc = -1$, если $(b, a) \in \Omega_T$, и $Arc = 0$, если $(a, b) \notin \Omega_T$ и $(b, a) \notin \Omega_T$.

С учетом изложенного выше индексирование каждого текстового документа, найденного в результате сканирования веб-сайтов, реализуется в три этапа:

– с использованием лингвистических словарей находятся абсолютные частоты каждого слова в тексте T и множестве текстов U;

– вычисляется информативность каждого слова текста T и вербально-ассоциативной связи между словами;

– формируется поисковый образ индексируемого текста T в виде вербально-ассоциативной сети.

Пример поискового образа текста: ⟨(*алгоритм*, 0,57); (*данные*, 0,2); (0,02, 0)⟩ ⟨(*словарь*, 0,32); (*лингвистический*, 0,27); (0,1, –1)⟩ ⟨(*алгоритм*, 0,57); (*данные*, 0,2); (0,02, 0)⟩ ⟨(*профиль*, 0,11); (*пользователя*, 0,21); (0,3, 1)⟩ ⟨(*алгоритм*, 0,57); (*релевантность*, 0,18); (0,01, 0)⟩.

Индексирование архивов пользователей. Каждый архив состоит из текстов, соответствующих информационным потребностям пользователя, сформировавшего данный архив в качестве своего информационного профиля. Если представить архив в виде последовательного объединения всех его текстов, то процесс индексирования архива сводится к индексированию текста.

Пусть A_i ($A_i \in \{A_1, A_2, ..., A_m\}$) – произвольный архив пользователя, включающий l текстовых документов $T_1, T_2, ..., T_l$, $A_i = \{T_1, T_2, ..., T_l\}$. Объединим все тексты $T_1, T_2, ..., T_l$ архива A_i в один текст путем их последовательной конкатенации («склеивания»). Тогда по аналогии с выражением (5) вербально-ассоциативная сеть архива пользователя A_i примет вид

$$Net_{A_i} = \{\langle (c, I_{A_i}^c); (b, I_{A_i}^d); (I_{A_i}^{cd}, Arc)\rangle \mid c \in A_i, d \in A_i\},$$

где $Arc = 1$, если $(c, d) \in \Omega_{A_i}$; $Arc = -1$, если $(d, c) \in \Omega_{A_i}$, и $Arc = 0$, если $(c, d) \notin \Omega_{A_i}$ и $(d, c) \notin \Omega_{A_i}$ (Ω_{A_i} – сужение отношения Ω на множество W_{A_i} всех слов архива A_i, $\Omega_{A_i} = \Omega \cap (W_{A_i} \times W_{A_i})$ – отношение дискурсивной сочетаемости в архиве A_i).

Индексирование кратких сообщений. Под кратким сообщением будем понимать текстовый документ, объем которого не позволяет выявить статистические характеристики его словоформ. Поэтому процессу индексирования краткого сообщения предшествует информационный поиск релевантного ему архива пользователя или создание нового релевантного архива.

Пусть Q – краткое сообщение, которое нужно проиндексировать, т. е. создать его поисковый образ $Net_Q = \{\langle (a, I_Q^a); (b, I_Q^b); (I_Q^{ab}, Arc)\rangle \mid a \in Q, b \in Q\}$ в виде вербально-ассоциативной сети, где (a, b) – вербально-ассоциативная пара слов; I_Q^a – информативность слова a сообщения Q; I_Q^{ab} – информативность вербально-ассоциативной связи между словами a и b. Для выявления статистических характеристик сообщения Q возможны две стратегии: поиск релевантного профиля пользователя и, в случае отрицательных результатов поиска, создание нового архива текстов путем отыскания релевантных документов в базе найденных текстов (рис. 3).

Краткое сообщение Q будем рассматривать как запрос на поиск релевантного ему профиля пользователя. Исключим из всех поисковых образов архивов A_i ($i = \overline{1, m}$) пользователей значения информативности слов: $\text{ПО}_{A_i} = \{a \mid a \in A_i, i = \overline{1, m}\}$. Аналогичным образом запишем поисковое предписание, т. е. поисковый образ сообщения Q: $\text{ПО}_Q = \{b \mid b \in Q\}$. При поиске релевантного профиля пользователя будем использовать векторную модель описания данных [7], а в качестве меры близости запросов и профилей пользователей – косинус угла между векторами поискового предписания и поискового образа архива пользователя. Рассмотрим эту меру близости.

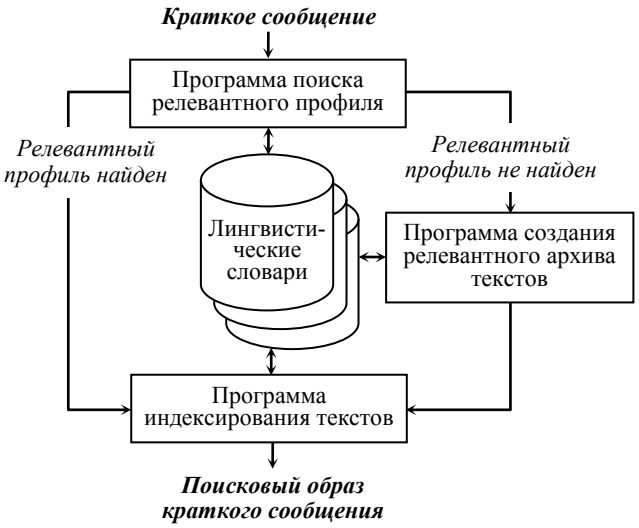

Рис. 3. Схема индексирования краткого сообщения

Обозначим через *Lex* множество всех различных слов, входящих в архивы пользователей и базу найденных текстов. Пусть их количество равно *n*. Введем в рассмотрение *n*-мерное евклидово пространство *E*. Для этого лексикографически упорядочим все слова из множества *Lex*, т. е. представим его в виде кортежа $Lex = \langle a_1, a_2, \ldots, a_n \rangle$. Для каждого проиндексированного архива $A \in \{A_1, A_2, ..., A_m\}$ пользователя построим вектор его поискового образа в пространстве *E*: $\mathbf{F}_A = (p_1, p_2, \ldots, p_n)$, где $p_i = 1$, если слово a_i входит в этот поисковый образ, в противном случае $p_i = 0$. Аналогично представим вектор поискового предписания, построенного для запроса *Q*: $\mathbf{F}_Q = (q_1, q_2, \ldots, q_n)$. Тогда для вычисления меры близости между векторами $\mathbf{F}_A$ и $\mathbf{F}_Q$ воспользуемся критерием выдачи

$$\cos\varphi = \frac{\mathbf{F}_A \mathbf{F}_Q}{|\mathbf{F}_A||\mathbf{F}_Q|} = \frac{\sum_{i=1}^{n} p_i q_i}{\sqrt{\sum_{i=1}^{n} p_i^{\,2}}\sqrt{\sum_{i=1}^{n} q_i^{\,2}}}. \quad (6)$$

Обозначим через *r* количество совпавших слов поискового образа архива *A* и поискового предписания *Q*. Пусть также m_A – количество слов в профиле *A*, а m_Q – их количество в предписании *Q*. Тогда критерий (6) можно представить в виде

$$\cos\varphi = \frac{r}{\sqrt{m_A m_Q}}. \quad (7)$$

Будем считать, что персональный архив текстов $A \in \{A_1, A_2, ..., A_m\}$ релевантен запросу *Q*, если критерий (7) не меньше некоторого порогового значения. Если это условие выполняется, то в поисковом образе $Net_Q = \{\langle (a, I_Q^a); (b, I_Q^b); (I_Q^{ab}, Arc) \rangle \mid a \in Q, b \in Q\}$ краткого сообщения *Q* информативность принимает следующие значения: $I_Q^a = I_A^a$, $I_Q^b = I_A^b$, $I_Q^{ab} = I_A^{ab}$. Если такой профиль не найден, то оперативно формируется новый релевантный архив текстов.

Обозначим через *Dat* множество найденных при сканировании Интернета текстов, а через *Im* – множество их поисковых образов. При формировании архива текстов, релевантных запросу *Q*, в базе найденных текстов *Dat* нужно найти все документы, релевантные тексту *Q*.

Пусть $D \in Im$ – поисковый образ произвольного текста из множества *Dat*. Построим вектор $\mathbf{F}_D$ поискового образа документа *D* по аналогии с вектором $\mathbf{F}_A$: $\mathbf{F}_D = (d_1, d_2, \ldots, d_n)$. При поиске текстов в множестве *Dat* в качестве критерия выдачи будем использовать аналог критерия (6):

$$\cos\psi = \frac{k}{\sqrt{m_D m_Q}}.$$

Обозначим через *Arel* сформированный релевантный архив текстов. Тогда в поисковом образе $Net_Q = \{\langle (a, I_Q^a); (b, I_Q^b); (I_Q^{ab}, Arc)\rangle \mid a \in Q, b \in Q\}$ короткого сообщения Q получим информативность $I_Q^a = I_{Arel}^a$, $I_Q^b = I_{Arel}^b$, $I_Q^{ab} = I_{Arel}^{ab}$.

Сканирование веб-страниц. Всякий веб-сайт в Интернете имеет гипертекстовую структуру и может быть представлен в виде орграфа, вершинами которого являются веб-страницы, а дугами – связи между ними. Среди разнообразия связей (ассоциативные, родо-видовые и др.) при решении задачи сканирования веб-сайтов нас будут интересовать только те из них, которые указывают на порядок следования страниц.

Рассмотрим последовательность шагов при сканировании веб-страниц путем их упорядочения. Пусть S_H – множество всех веб-страниц некоторого веб-сайта *H*. Определим на множестве S_H строгий порядок (транзитивное и антирефлексивное бинарное отношение) τ_H. Обозначим через ρ_H редукцию $\rho_H = \tau_H \setminus (\tau_H)^2$ строгого порядка τ_H. Редукция ρ_H означает, что для любых веб-страниц $a, b \in S_H$ отношение $(a, b) \in \rho_H$ выполняется тогда и только тогда, когда справедливо отношение $(a, b) \in \tau_H$, но не существует «промежуточной» веб-страницы *x*, такой, что $(a, x) \in \tau_H$ и $(x, b) \in \tau_H$. Таким образом, отношение ρ_H указывает на «непосредственное» следование страницы *b* за страницей *a* в веб-сайте *H*.

С учетом отношения ρ_H сканирование веб-страниц сайта *H* удобно реализовать в следующей последовательности: сканируются все веб-страницы, являющиеся висячими вершинами орграфа *H*; найденные тексты помещаются в специальную базу данных; отсканированные веб-страницы условно исключаются из орграфа *H*, далее процесс продолжается аналогичным образом. Порядок сканирования веб-страниц схематически показан на рис. 4.

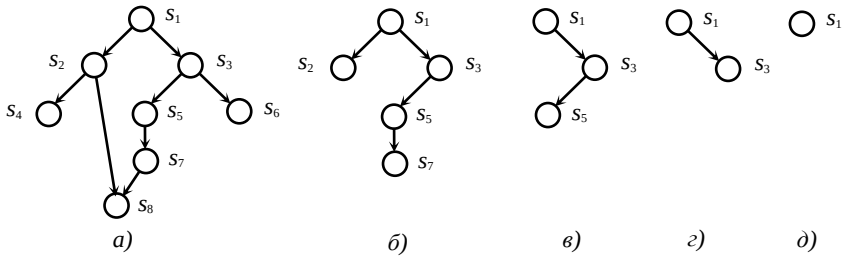

Рис. 4. Последовательное сканирование веб-страниц: *а)* s_4, s_8, s_6; *б)* s_2, s_7, s_3; *в)* s_5; *г)* s_3; *д)* s_1

Контроль новизны найденных текстов реализуется в два этапа: вначале проводится поиск в базе текстов и коротких сообщений по запросу, которым является текст, найденный при сканировании веб-сайтов, затем новизна отсканированного текста уточняется с использованием вербальных ассоциаций между его словами.

Пусть *T* – текст, поступивший в результате сканирования веб-сайтов, а *Q* – любой текст из множества *Dat*. Обозначим через $\mathbf{F}_T$ вектор поискового образа текста *T* (т. е. поисковое предписание), а через $\mathbf{F}_Q$ – вектор поискового образа текста *Q*. Воспользовавшись критерием выдачи, аналогичным критерию (6), на первом этапе контроля новизны текста *T* проведем поиск релевантных текстов в множестве *Dat* и выберем текст *Q′* с наибольшим значением критерия выдачи.

Рассмотрим второй этап контроля новизны найденных текстов. Обозначим через W_T и $W_{Q'}$ множества всех пар словоформ соответственно текстов *T* и *Q′*, через $W_{Dat} = \langle a_1b_1, a_2b_2, ..., a_lb_l\rangle$ – кортеж всех пар словоформ из базы данных *Dat*, а через *E* – *l*-мерное евклидово пространство. Рассмотрим объединение $W_T \cup W_{Q'}$ текстов W_T и $W_{Q'}$ и представим его вектором в пространстве *E*: $\mathbf{W}_{TQ'} = (I_{TQ'}^{a_1b_1}, I_{TQ'}^{a_2b_2}, \ldots, I_{TQ'}^{a_lb_l})$, где $I_{TQ'}^{a_1b_1}, I_{TQ'}^{a_2b_2}, \ldots, I_{TQ'}^{a_lb_l}$ – значения информативности вербально-ассоциативной связи в множестве $W_T \cup W_{Q'}$. При этом компонента вектора $\mathbf{W}_{TQ'}$

равна нулю, если соответствующей пары слов нет в множестве $W_T \cup W_{Q'}$. С учетом рассмотренных обозначений нормализованную информативность $I^{TQ}_{W_T \cup W_{Q'}}$ вербально-ассоциативной связи между текстами T и Q' можно интерпретировать как проекцию вектора $\mathbf{e} = (1, 1, ..., 1)$ на направление вектора $\mathbf{W}_{TQ'}$:

$$I^{TQ'}_{W_T \cup W_{Q'}} = \frac{\sum\limits_{a \in W_T,\, b \in W_{Q'}} I^{ab}_{W_T \cup W_{Q'}}}{\sqrt{\sum\limits_{a \in W_T,\, b \in W_{Q'}} (I^{ab}_{W_T \cup W_{Q'}})^2}}. \quad (8)$$

Нетрудно видеть, что тексты T и Q' совпадают, если $I^{TQ'}_{W_T \cup W_{Q'}} = I^{TQ'}_{W_T \cup W_T} = I^{TQ'}_{W_T}$, где

$$I^{TQ'}_{W_T} = \frac{\sum\limits_{a \in W_T,\, b \in W_T} I^{ab}_{W_T}}{\sqrt{\sum\limits_{a \in W_T,\, b \in W_T} (I^{ab}_{W_T})^2}}. \quad (9)$$

Таким образом, процесс контроля новизны найденных текстов реализуется следующим образом. Для каждого очередного найденного текста T вычисляются значения информативности вербально-ассоциативной связи между текстом T и каждым из текстов Q базы данных Dat по формулам (8) и (9). Если для текста T и некоторого текста $Q_+ \in Dat$ выполняется равенство $I^{TQ'}_{W_T \cup W_{Q'}} = I^{TQ'}_{W_T}$, то текст T не является новым. В противном случае он новый.

Адресная рассылка текстов. Адресная рассылка найденных при сканировании Интернета текстов сводится к поиску профиля пользователя с наибольшим значением критерия выдачи.

Рассылка реализуется в три этапа:

– ищутся все релевантные профили пользователей по поисковому предписанию, которым является поисковый образ очередного текста, найденного при сканировании веб-страниц;

– проверяется, является ли новым найденный текст;

– найденный новый текст помещается в архивы пользователей, для которых он оказался релевантным.

Заключение. Предложенная в статье математическая модель процессов сканирования вебсайтов и адресного распространения найденной информации может быть использована при индексировании, поиске и реферировании текстовой информации в Интернете, корпоративных сетях и локальных базах данных. При соответствующем подборе тематики и языка представления тематических корпусов текстов возможен веб-поиск текстовых документов на различных входных языках.

Список использованных источников

1. Ахремчик, Р. В. Система ИРИ в Центральной научной библиотеке Национальной академии наук Беларуси / Р. В. Ахремчик, Т. В. Пинчук // Научные и технические библиотеки. – 2014. – № 2. – С. 58–62.

2. Юдина, И. Г. Избирательное распространение информации на базе веб-сервисов: обзор интернет-ресурсов / И. Г. Юдина // Библиосфера. – 2008. – № 1. – С. 51–56.

3. Перегоедова, Н. В. Организация и методика библиографического информирования : конспект лекций / Н. В. Перегоедова. – Новосибирск : ГПНТБ СО РАН, 2008. – 36 с.

4. Липницкий, С. Ф. Модель представления знаний в информационных системах на основе вербальных ассоциаций / С. Ф. Липницкий // Информатика. – 2011. – № 4(32). – С. 21–28.

5. Липницкий, С. Ф. Моделирование анализа текстовых документов и кратких сообщений на основе вербальных ассоциаций / С. Ф. Липницкий // Информатика. – 2018. – Т. 15, № 1. – С. 70–80.

6. Липницкий, С. Ф. Индексирование текстовой информации на основе моделирования вербальных ассоциаций / С. Ф. Липницкий // Информатика. – 2012. – № 3(35). – С. 94–102.

References

1. Ahremchik R. V., Pinchuk T. V. Sistema IRI v Central'noj nauchnoj biblioteke Nacional'noj akademii nauk Belarusi [IRI system in the Central scientific library of the National academy of sciences of Belarus]. Nauchnye i tehnicheskie biblioteki [*Scientific and Technical Libraries*], 2014, no. 2, pp. 58–62 (in Russian).

2. Yudina I. G. Izbiratel'noe rasprostranenie informacii na baze veb-servisov: obzor internet-resursov [Selective dissemination of information based on web services: a review of Internet resources]. Bibliosfera [*Bibliosphere*], 2008, no. 1, pp. 51–56 (in Russian).

3. Peregoedova N. V. Organizacija i metodika bibliograficheskogo informirovanija. *Organization and Methods of Bibliographic Information*. Novosibirsk, Gosudarstvennaja publichnaja nauchno-tehnicheskaja biblioteka Sibirskogo otdelenija Rossijskoj akademii nauk, 2008, 36 p. (in Russian).

4. Lipnitsky S. F. Model' predstavlenija znanij v informacionnyh sistemah na osnove verbal'nyh associacij [Model of knowledge representation in information systems based on verbal associations]. Informatika [*Informatics*], 2011, no. 4(32), pp. 21–28 (in Russian).

5. Lipnitsky S. F. Modelirovanie analiza tekstovyh dokumentov i kratkih soobshhenij na osnove verbal'nyh associacij [Modeling analysis of text documents and short messages based on verbal associations]. Informatika [*Informatics*], 2018, vol. 15, no. 1, pp. 70–80 (in Russian).

6. Lipnitsky S. F. Indeksirovanie tekstovoj informacii na osnove modelirovanija verbal'nyh associacij [Indexing text information based on modeling verbal associations]. Informatika [*Informatics*], 2012, no. 3(35), pp. 94–102 (in Russian).

Информация об авторе

Липницкий Станислав Феликсович, доктор технических наук, главный научный сотрудник, Объединенный институт проблем информатики Национальной академии наук Беларуси, Минск, Беларусь.
E-mail: lipn@newman.bas-net.by

Information about the author

Stanislav F. Lipnitsky, Dr. Sci. (Eng.), Chief Researcher, The United Institute of Informatics Problems of the National Academy of Sciences of Belarus, Minsk, Belarus.
E-mail: lipn@newman.bas-net.by