УДК 004.912

С.Ф. Липницкий

ИНДЕКСИРОВАНИЕ ТЕКСТОВОЙ ИНФОРМАЦИИ НА ОСНОВЕ МОДЕЛИРОВАНИЯ ВЕРБАЛЬНЫХ АССОЦИАЦИЙ

Предлагаются алгоритмы индексирования текстовых документов на основе математической модели представления знаний в виде вербально-ассоциативных сетей. Рассматриваются структурные особенности информационной базы, построенной с учетом результатов моделирования. Приводятся формулы для вычисления информативности слов и их вербально-ассоциативных пар при индексировании текстовой информации.

Введение

Под вербальными ассоциациями в лингвистике и психолингвистике понимают семантические связи между словами в языке (тексте, речи), которые соответствуют ассоциативным отношениям между обозначаемыми ими сущностями в реальном мире. В результате формализации вербальных ассоциаций в работе [1] построена модель представления знаний в информационных системах в виде вербально-ассоциативной сети, т. е. графа отношения вербально-ассоциативной связи слов в тематическом корпусе текстов, представляющем некоторую предметную область. Каждая вершина этого графа помечена значением информативности соответствующего слова, а каждое ребро – значением информативности вербально-ассоциативной связи между инцидентными ему вершинами (словами).

Как показано в статье [1], предложенная модель представления знаний может быть использована при автоматическом индексировании текстовых документов и запросов на поиск информации, реферировании и аннотировании текста, а также при информационном поиске.

На уровне моделирования индексирование текстов и запросов сводится к построению их вербально-ассоциативных сетей. При программной реализации информационной системы поисковые образы текстовых документов и запросов будем представлять в виде совокупностей вербально-ассоциативных пар слов, каждому из которых, а также самим парам соответствуют значения их информативности.

Рассмотрим алгоритмы индексирования тематических корпусов текстов, политематических и монотематических текстов, а также запросов пользователей на поиск информации с учетом следующей структуры информационной базы.

1. Архитектура информационной базы

Информационная база включает базу данных и базу знаний (рис. 1). В базе данных содержатся тематические корпусы текстов и архив различных полнотекстовых документов и их рефератов, а в базе знаний – лингвистические словари и вербально-ассоциативные сети тематических корпусов текстов и отдельных документов (поисковые образы).

1.1. База данных

В базе данных представлены тематические и полные корпусы текстов. Тематический корпус – это совокупность текстовых документов по конкретной тематике, характеризующей соответствующую предметную область. Полный корпус представляет собой объединение всех тематических корпусов. Кроме корпусов текстов в базе данных могут храниться различные полнотекстовые документы, их рефераты и прочая актуальная информация.

Корпусы текстов. Формально текст T – это любое непустое подмножество входного языка информационной системы, если на этом подмножестве определена редукция $\prec^r = \prec \setminus \prec^2$ линейного порядка $\prec$ (транзитивного и антисимметричного бинарного отношения на множестве

T, которое связанно на T, т. е. для любых $a, b \in T$ или $a \prec b$, или $b \prec a$, или $a = b$). Цепочки текста T – суть предложения этого текста. Обозначим через Ct_i i-й тематический корпус текстов, т. е. множество текстов, соответствующих некоторой i-й предметной области. Тогда совокупность всех тематических корпусов $Cf = \{Ct_i \mid i = \overline{1, n}\}$ – это полный корпус текстов.

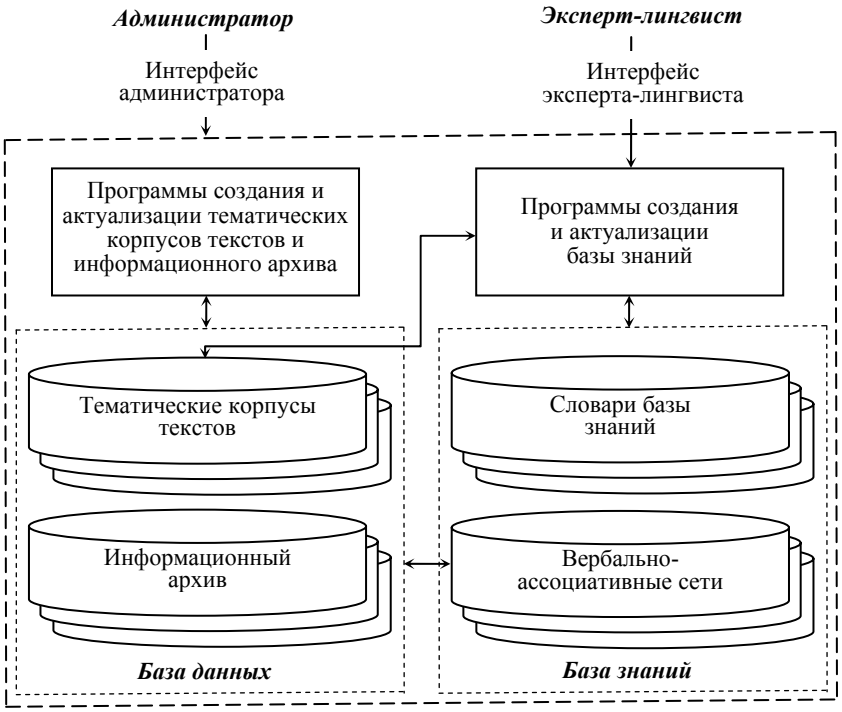

Рис. 1. Состав информационной базы

1.2. База знаний

Под базой знаний будем понимать совокупность лингвистических словарей и вербально-ассоциативных сетей предметных областей. Для реализации функций индексирования текстовой информации будем использовать частотные словари словоформ и вербально-ассоциативных пар слов, словарь словоизменительных парадигм и словарь синонимичных словоформ.

Частотный словарь словоформ. Пусть a – некоторая словоформа, n^a_{Cf} и $n^a_{Ct_i}$ ($i = \overline{1, n}$) – абсолютные частоты ее появления соответственно в полном и i-м тематическом корпусах текстов, W_{Cf} – множество всех словоформ полного корпуса текстов Cf. Тогда совокупность кортежей $Dic_a = \{\langle a,\ n^a_{Cf}\ ,\ n^a_{Ct_1}\ ,\ n^a_{Ct_2}\ , \ldots\ ,\ n^a_{Ct_n}\rangle \mid a \in W_{Cf}\}$ будем называть частотным словарем словоформ (табл. 1).

Таблица 1

Фрагмент частотного словаря словоформ

Словоформа	n^a_{Cf}	$n^a_{Ct_1}$	...	$n^a_{Ct_n}$	Код (номер) парадигмы
...					
стол	0204055	0056534	...	0014445	00000094
стола	0401657	0074526	...	0023747	00000094
...					

Частотный словарь вербально-ассоциативных пар слов. Структура этого словаря аналогична структуре словаря Dic_a. Рассмотрим пару произвольных слов $a, b \in W_{Cf}$. Обозначим через n^{ab}_{Cf} и $n^{ab}_{Ct_i}$ абсолютные частоты совместной встречаемости слов a и b в одном и том же предложении полного Cf и тематического Ct_i ($i = \overline{1, n}$) корпусов текстов. Тогда множество Dic_{ab} =

$= \{\langle (a, b),\ n_{Cf}^{ab},\ n_{Ct_1}^{ab},\ n_{Ct_2}^{ab}, \ldots,\ n_{Ct_n}^{ab} \rangle |\ a, b \in W_{Cf},\ n_{Cf}^{ab} \neq 0,\ n_{Ct_i}^{ab} \neq 0,\ i = \overline{1, n}\}$ назовем частотным словарем вербально-ассоциативных пар слов (табл. 2).

Таблица 2

Фрагмент частотного словаря вербально-ассоциативных пар слов

Пара словоформ	n_{Cf}^{a}	$n_{Ct_1}^{a}$	...	$n_{Ct_n}^{a}$
...				
(стол, комнате)	03020	00543	...	00121
(стола, поверхности)	04023	00623	...	00242
...				

Словарь словоизменительных парадигм. Пусть a – произвольное слово, а Par_a – множество всех словоизменений слова a (включая a). Тогда словарь словоизменительных парадигм представим в виде множества $Dic_{par} = \{(a, Par_a) \mid a \in W_{Cf}, a \in Par_a\}$ (табл. 3).

Таблица 3

Состав и структура словаря парадигм (конечное состояние)

Код (номер) парадигмы	Парадигма
...	
00000094	стол
	стола
	...
...	

Словарь парадигм создается и актуализируется в человеко-машинном режиме с использованием соответствующего инструментария эксперта-лингвиста.

Словарь синонимичных словоформ. Словарь состоит из групп синонимичных словоформ, которые могут быть использованы при определении информативности слов и их вербально-ассоциативных пар.

Обозначим через Syn_a множество всех синонимов слова $a \in W_{Cf}$. Тогда словарь синонимичных словоформ – это множество $Dic_{syn} = \{(a, Syn_a) \mid a \in W_{Cf}, a \in Syn_a\}$ (табл. 4).

Таблица 4

Фрагмент словаря синонимичных словоформ

Словоформа	Синонимичные словоформы
...	
языкознание	лингвистика
	языковедение
...	

Словарь синонимов создается экспертом-лингвистом с использованием специальных программных средств.

2. Индексирование тематических корпусов текстов

Условно объединим все тексты полного корпуса и всех тематических корпусов и будем рассматривать эти корпусы как отдельные текстовые документы.

2.1. Информативность слов и вербально-ассоциативных пар слов в тематическом корпусе текстов

Рассмотрим полный корпус текстов Cf и тематический корпус Ct_i, соответствующий некоторой i-й предметной области. Информативность всякой словоформы a из корпуса Ct_i при его индексировании будем вычислять как отношение абсолютной частоты встречаемости сло-

воформы a в тематическом корпусе текстов Ct_i к абсолютной частоте ее появления в полном корпусе текстов Cf [2]. На практике при вычислении информативности $I^a_{Ct_i}$ слова a необходимо учитывать его словоизменения и синонимию, зафиксированные в словарях Dic_{par} и Dic_{syn} соответственно. В этом случае указанное отношение частот примет вид

$$I^a_{Ct_i} = \frac{n^a_{Ct_i} + \sum\limits_{b\in Par_a,\, b\neq a} n^b_{Ct_i} + \sum\limits_{c\in Syn_a,\, c\neq a} (n^c_{Ct_i} + \sum\limits_{d\in Par_c,\, d\neq c} n^d_{Ct_i})}{n^a_{Cf} + \sum\limits_{b\in Par_a,\, b\neq a} n^b_{Cf} + \sum\limits_{c\in Syn_a,\, c\neq a} (n^c_{Cf} + \sum\limits_{d\in Par_c,\, d\neq c} n^d_{Cf})}, \quad (1)$$

где $n^a_{Ct_i}$, $n^b_{Ct_i}$, $n^c_{Ct_i}$ и $n^d_{Ct_i}$ – абсолютные частоты встречаемости слов a, b, c и d в тематическом корпусе текстов Ct_i; n^a_{Cf}, n^b_{Cf}, n^c_{Cf} и n^d_{Cf} – абсолютные частоты их появления в полном корпусе текстов Cf.

По аналогии с формулой (1) информативность вербально-ассоциативной связи слов a и b будем вычислять по формуле

$$I^{ab}_{Ct_i} = \frac{n^{ab}_{Ct_i} + \sum\limits_{\substack{c\in Par_a,\, c\neq a \\ d\in Par_b,\, d\neq b}} n^{cd}_{Ct_i} + \sum\limits_{\substack{r\in Syn_a,\, r\neq a \\ s\in Syn_b,\, s\neq b}} (n^{rs}_{Ct_i} + \sum\limits_{\substack{p\in Par_r,\, p\neq r \\ q\in Par_s,\, q\neq s}} n^{pq}_{Ct_i})}{n^{ab}_{Cf} + \sum\limits_{\substack{c\in Par_a,\, c\neq a \\ d\in Par_b,\, d\neq b}} n^{cd}_{Cf} + \sum\limits_{\substack{r\in Syn_a,\, r\neq a \\ s\in Syn_b,\, s\neq b}} (n^{rs}_{Cf} + \sum\limits_{\substack{p\in Par_r,\, p\neq r \\ q\in Par_s,\, q\neq s}} n^{pq}_{Cf})}, \quad (2)$$

где $n^{ab}_{Ct_i}$, $n^{cd}_{Ct_i}$, $n^{rs}_{Ct_i}$ и $n^{pq}_{Ct_i}$ – абсолютные частоты совместной встречаемости пар слов (a, b), (c, d), (r, s) и (p, q) в одном и том же предложении тематического корпуса текстов Ct_i; n^{ab}_{Cf}, n^{cd}_{Cf}, n^{rs}_{Cf} и n^{pq}_{Cf} – абсолютные частоты их совместного появления в полном корпусе Cf.

Моделью поискового образа тематического корпуса текстов Ct_i является вербально-ассоциативная сеть предметной области, определяемой корпусом Ct_i.

Построим этот поисковый образ в виде множества вербально-ассоциативных пар слов, в котором словам каждой пары приписаны значения их информативности, а парам – значения информативности вербально-ассоциативных связей между их словами:

$$\omega(Ct_i) = \{\langle (a, I^a_{Ct_i}); (b, I^b_{Ct_i}); I^{ab}_{Ct_i}\rangle \mid a\in Ct_i, b\in Ct_i, I^a_{Ct_i} > I^0_{Ct_i}, I^{ab}_{Ct_i} > I^{00}_{Ct_i}\}, \quad (3)$$

где $\omega : Cf \to PO$ – инъективное отображение полного корпуса текстов Cf в множество PO их поисковых образов.

2.2. Алгоритм индексирования тематических корпусов текстов

Алгоритм 1. На входе алгоритма – тематический корпус текстов Ct_i, на выходе – поисковый образ корпуса Ct_i в виде выражения (3).

1. Сформировать множество $W^{ab}_{Ct_i}$ всех вербально-ассоциативных пар слов вида (a, b) тематического корпуса текстов Ct_i.

2. Вычислить информативность $I^a_{Ct_i}$ и $I^b_{Ct_i}$ слов a и b всех вербально-ассоциативных пар слов из множества $W^{ab}_{Ct_i}$ по формуле (1).

3. Вычислить информативность $I^{ab}_{Ct_i}$ всех вербально-ассоциативных пар слов из множества $W^{ab}_{Ct_i}$ по формуле (2).

4. Сформировать поисковый образ тематического корпуса текстов Ct_i в виде множества (3). Конец.

3. Индексирование политематических текстов

Различают монотематические и политематические тексты. Монотематический текст – это текст небольшого объема (статья, доклад), содержание которого отображает некоторую единую тематику. В политематическом тексте (например, книге) обычно представлено несколько монотематических текстов (субтекстов).

3.1. Информативность слов и вербально-ассоциативных пар слов в политематическом тексте

Рассмотрим политематический текст T ($T \in Cf$). Моделью поискового образа текста T является его вербально-ассоциативная сеть [1]. Представим эту сеть в виде совокупности вербально-ассоциативных пар слов

$$\omega(T) = \{\langle (a, I_T^a); (b, I_T^b); I_T^{ab} \rangle \mid a \in T, b \in T, I_T^a > I_T^0, I_T^{ab} > I_T^{00}\}. \quad (4)$$

В этой совокупности словам a и b каждой пары соответствуют значения их информативности I_T^a и I_T^b в тексте T, а самой паре – значения информативности I_T^{ab} вербально-ассоциативных связей между этими словами.

Для вычисления информативности I_T^a произвольного слова a текста T будем использовать формулу, аналогичную формуле (1):

$$I_T^a = \frac{n_T^a + \sum\limits_{b \in Par_a,\, b \neq a} n_T^b + \sum\limits_{c \in Syn_a,\, c \neq a} (n_T^c + \sum\limits_{d \in Par_c,\, d \neq c} n_T^d)}{n_{Cf}^a + \sum\limits_{b \in Par_a,\, b \neq a} n_{Cf}^b + \sum\limits_{c \in Syn_a,\, c \neq a} (n_{Cf}^c + \sum\limits_{d \in Par_c,\, d \neq c} n_{Cf}^d)}, \quad (5)$$

где n_T^a, n_T^b, n_T^c и n_T^d – абсолютные частоты встречаемости слов a, b, c и d в тексте T.

По аналогии с формулой (2) информативность вербально-ассоциативной связи слов a и b в тексте T вычисляется по формуле

$$I_T^{ab} = \frac{n_T^{ab} + \sum\limits_{\substack{c \in Par_a,\, c \neq a \\ d \in Par_b,\, d \neq b}} n_T^{cd} + \sum\limits_{\substack{r \in Syn_a,\, r \neq a \\ s \in Syn_b,\, s \neq b}} (n_T^{rs} + \sum\limits_{\substack{p \in Par_r,\, p \neq r \\ q \in Par_s,\, q \neq s}} n_T^{pq})}{n_{Cf}^{ab} + \sum\limits_{\substack{c \in Par_a,\, c \neq a \\ d \in Par_b,\, d \neq b}} n_{Cf}^{cd} + \sum\limits_{\substack{r \in Syn_a,\, r \neq a \\ s \in Syn_b,\, s \neq b}} (n_{Cf}^{rs} + \sum\limits_{\substack{p \in Par_r,\, p \neq r \\ q \in Par_s,\, q \neq s}} n_{Cf}^{pq})}, \quad (6)$$

где n_T^{ab}, n_T^{cd}, n_T^{rs} и n_T^{pq} – абсолютные частоты совместной встречаемости пар слов (a, b), (c, d), (r, s) и (p, q) в одном и том же предложении текста T.

3.2. Алгоритм индексирования политематических текстов

Алгоритм 2. На входе алгоритма – политематический текст T, на выходе – поисковый образ текста T в виде множества (4).

1. Сформировать множество W_T^{ab} всех вербально-ассоциативных пар слов вида (a, b) текста T.

2. Вычислить информативность I_T^a и I_T^b слов a и b всех вербально-ассоциативных пар слов из множества W_T^{ab} по формуле (5).

3. Вычислить информативность I_T^{ab} всех вербально-ассоциативных пар слов из множества W_T^{ab} по формуле (6).

4. Сформировать поисковый образ текста T в виде множества (4). Конец.

4. Индексирование монотематических текстов

Моделью поискового образа монотематического текста является его вербально-ассоциативная сеть, т. е. подграф сети некоторой предметной области, которой соответствует тематический (или динамический [2]) корпус текстов, релевантный индексируемому тексту. В связи с этим процесс его индексирования сводится к поиску тематического корпуса текстов или созданию динамического (рис. 2).

Рис. 2. Структурно-функциональная схема индексирования монотематического текста

Пусть Q – монотематический текст, который нужно проиндексировать, т. е. создать его поисковый образ в виде совокупности кортежей $\{\langle (a, I_Q^a); (b, I_Q^b); I_Q^{ab} \rangle \mid a \in Q, b \in Q, I_Q^a > I_Q^0, I_Q^{ab} > I_Q^{00}\}$, где (a, b) – вербально-ассоциативная пара слов; I_Q^a – информативность слова a текста Q; I_Q^b – информативность слова b; I_Q^{ab} – информативность вербально-ассоциативной связи между словами a и b; I_Q^0 и I_Q^{00} – пороговые значения информативности.

Для выявления статистических характеристик I_Q^a, I_Q^b и I_Q^{ab} текста Q возможны две стратегии. Первая – поиск релевантного тематического корпуса текстов и вторая (в случае отрицательных результатов поиска) – создание динамического корпуса текстов путем отыскания релевантных текстов в полном корпусе.

4.1. Поиск релевантного тематического корпуса текстов

Текст Q будем рассматривать как запрос на поиск релевантного ему тематического корпуса текстов. Исключим из всех поисковых образов документов полного корпуса текстов значения информативности слов и вербально-ассоциативных пар слов, т. е. преобразуем выражение (3) к виду

$$\omega(Ct_i) = \{(a, b) \mid a \in Ct_i, b \in Ct_i\}. \quad (7)$$

Аналогичным образом запишем поисковое предписание, т. е. поисковый образ текста Q:

$$\omega(Q) = \{(c, d) \mid c \in Q, d \in Q\}. \quad (8)$$

Поиск релевантного тематического корпуса текстов в полном корпусе Cf реализуется в два этапа. На первом этапе проводится поиск по поисковому предписанию (8). При этом используется оптимальная по релевантности (как показано в статье [2]) поисковая функция

$$\pi(Q) = \{Ct \mid \eta(\omega(Ct), \omega(Q)) < 0,\ Ct \in Cf\},\ \eta(\omega(Ct), \omega(Q)) = \begin{cases} -1, \text{если } \omega(Q) \cap \omega(Ct) \neq \varnothing; \\ 0, \text{если } \omega(Q) \cap \omega(Ct) = \varnothing. \end{cases} \quad (9)$$

Поисковая функция – это частичное мультиотображение $\pi : Z \to Cf \cup S_{\text{веб}}$ множества Z запросов в множество Cf документов полного корпуса текстов и веб-страниц Интернета $S_{\text{веб}}$. Отображение $\eta : \omega(Cf \cup S_{\text{веб}}) \times \omega(Z) \to R$ в выражении (9) – это критерий выдачи (R – множество действительных чисел) [3], т. е. мера близости запросов и текстовых документов.

На втором этапе поиска релевантного тематического корпуса текстов из множества найденных выбирается корпус, которому соответствует наибольшее значение критерия выдачи. В большинстве известных информационных систем в качестве такого критерия применяется косинус угла между векторами поискового предписания и поискового образа документа в евклидовом пространстве E. Рассмотрим этот критерий, используя принятые выше обозначения.

Обозначим через W множество всех различных вербально-ассоциативных пар слов, входящих в поисковые образы текстов из множества $Cf \cup S_{\text{веб}}$, а также поисковых предписаний из множества $\omega(Z)$. Пусть их количество равно n. Лексикографически упорядочим все вербально-ассоциативные пары из множества W, т. е. представим W в виде кортежа $W = \langle (a_1, b_1), (a_2, b_2), \ldots, (a_n, b_n) \rangle$. Для каждого проиндексированного тематического корпуса текстов $Ct \in Cf$ построим вектор его поискового образа в пространстве E: $\mathbf{F}_{Ct} = (p_1, p_2, \ldots, p_n)$, где $p_i = 1$, если вербально-ассоциативная пара (a_i, b_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 $\mathbf{F}_{Ct}$ и $\mathbf{F}_Q$ воспользуемся критерием выдачи

$$\cos\varphi = \frac{\mathbf{F}_{Ct}\mathbf{F}_Q}{|\mathbf{F}_{Ct}||\mathbf{F}_Q|} = \frac{\sum_{i=1}^{n} p_i q_i}{\sqrt{\sum_{i=1}^{n} p_i^{\,2}} \sqrt{\sum_{i=1}^{n} q_i^{\,2}}}. \quad (10)$$

При реализации информационной системы критерий (10) целесообразно преобразовать следующим образом. Пусть l – количество совпавших вербально-ассоциативных пар поискового образа $\omega(Ct)$ и поискового предписания $\omega(Q)$, n_{Ct} – количество пар в множестве $\omega(Ct)$, а n_Q – их количество в множестве $\omega(Q)$. Тогда критерий (10) примет вид

$$\cos\varphi = \frac{l}{\sqrt{n_{Ct} n_Q}}. \quad (11)$$

Результатом поиска является тематический корпус текстов $\pi(Q) = Ct$, для которого значение $\cos\varphi$ является наибольшим из всех значений, таких, что $\cos\varphi \geq \eta_0$. Если такой тематический корпус текстов не найден, то оперативно формируется динамический корпус.

4.2. Создание динамического корпуса текстов

При формировании динамического корпуса текстов, релевантного запросу Q, в полном корпусе текстов Cf нужно найти все документы, релевантные тексту Q.

Пусть $D \in Cf$ – произвольный документ из полного корпуса текстов. Построим вектор $\mathbf{F}_D$ поискового образа $\omega(D)$ документа D по аналогии с вектором $\mathbf{F}_{Ct}$: $\mathbf{F}_D = (d_1, d_2, \ldots, d_n)$. Компонента вектора $d_k = 1$, если соответствующая вербально-ассоциативная пара слов имеется в поисковом образе текста D, иначе $d_k = 0$. В качестве критерия выдачи будем использовать аналог критерия (10):

$$\cos\psi = \frac{\mathbf{F}_D\mathbf{F}_Q}{|\mathbf{F}_D||\mathbf{F}_Q|} = \frac{\sum_{k=1}^{n} d_k q_k}{\sqrt{\sum_{k=1}^{n} d_k^2} \sqrt{\sum_{k=1}^{n} q_k^2}}. \quad (12)$$

Упростим критерий (12) и запишем его в виде, аналогичном выражению (11). Обозначим через r количество совпавших вербально-ассоциативных пар поискового предписания $\omega(Q)$ и поискового образа $\omega(D)$ документа D. Пусть также m_Q – количество пар в множестве $\omega(Q)$, а m_D – их количество в множестве $\omega(D)$. Тогда критерий (12) можно представить в виде

$$\cos\psi = \frac{r}{\sqrt{m_D m_Q}}. \qquad (13)$$

Будем считать, что документ $D \in Cf$ релевантен запросу Q и принадлежит создаваемому динамическому корпусу текстов Dt, если критерий (13) не меньше порогового значения η_0', т. е. $Dt = \pi(Q) = \{D \mid D \in Cf_i, \cos\psi \ge \eta_0'\}$.

4.3. Алгоритм индексирования монотематических текстов

Обозначим через Kt ($Kt \in \{Ct, Dt\}$) релевантный тексту Q тематический (Ct) или динамический (Dt) корпус текстов, а через W_Q – множество всех вербально-ассоциативных пар слов текста Q. Тогда поисковый образ текста Q будем строить в виде

$$\omega(Q) = \{\langle (c, I_{Kt}^c); (d, I_{Kt}^d); I_{Kt}^{cd}\rangle \mid c \in W_Q, d \in W_Q, I_{Kt}^c > I_{Kt}^0, I_{Kt}^{cd} > I_{Kt}^{00}\}. \qquad (14)$$

Алгоритм 3. На входе алгоритма – монотематический текст Q, на выходе – поисковый образ текста Q в виде множества (14).

1. $Kt := \varnothing$.
2. Преобразовать поисковое предписание $\omega(Q)$ к виду (8).
3. Искать в полном корпусе текстов релевантные тематические корпусы в соответствии с поисковой функцией (9). Если хотя бы один тематический корпус найден, то перейти к п. 4, иначе – к п. 6.
4. Если в п. 3 найдено более одного тематического корпуса текстов, то выбрать один корпус с наибольшим значением критерия выдачи (11) и перейти к п. 5. Иначе перейти к п. 5.
5. Поместить найденный тематический корпус текстов в множество Kt.
6. Искать в полном корпусе текстов документы, релевантные запросу Q, в соответствии с критерием выдачи (13). Если хотя бы один документ найден, то перейти к п. 7, иначе – к п. 9.
7. Поместить все найденные в п. 6 документы в множество Kt.
8. Индексировать корпус текстов Kt (алгоритм 1). Представить поисковый образ корпуса Kt в виде выражения (14), полагая, что $\omega(Q) = \omega(Kt)$. Конец (поисковый образ монотематического текста Q сформирован).
9. Конец (текст Q не проиндексирован из-за недостаточной представительности полного корпуса текстов Cf, пополнить корпус текстов Cf).

Заключение

Предложенные в статье алгоритмы могут быть использованы при индексировании, поиске и реферировании текстовой информации в Интернете, корпоративных сетях и в локальных базах данных. При соответствующем подборе тематики и языка представления корпусов текстов возможны поиск и реферирование документов на различных входных языках. Реализация этой функции сводится к формированию корпусов текстов и созданию словарей базы знаний (без коррекции программного обеспечения системы). При наличии персональных тематических корпусов текстов обеспечивается адаптация процессов поиска и реферирования к информационным потребностям соответствующих индивидуальных и корпоративных пользователей.

Список литературы

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

2. Липницкий, С.Ф. Моделирование информационного поиска на основе динамических корпусов текстов / С.Ф. Липницкий, А.А. Мамчич // Весці НАН Беларусі. Сер. фіз.-тэхн. навук. – 2011. – № 1. – С. 72–81.

3. Липницкий, С.Ф. Моделирование информационного мониторинга Интернета на основе тематических корпусов текстов / С.Ф. Липницкий // Весці НАН Беларусі. Сер. фіз.-тэхн. навук. – 2011. – № 3. – С. 92–99.

Поступила 18.05.12

Объединенный институт проблем
информатики НАН Беларуси,
Минск, Сурганова, 6
e-mail: lipn@newman.bas-net.by

S.F. Lipnitsky

TEXT INFORMATION INDEXING ON THE BASIS OF VERBAL ASSOCIATIONS MODELING

Algorithms of indexing text documents based on a mathematical model of knowledge representation in the form of verbal-associative networks are developed. Structural features of the knowledge base built by taking into account the simulation results are considered. Formulas for calculating words informativity and their verbal-associative pairs are proposed.