УДК 004.912

С.Ф. Липницкий

МОДЕЛЬ ПРЕДСТАВЛЕНИЯ ЗНАНИЙ В ИНФОРМАЦИОННЫХ СИСТЕМАХ НА ОСНОВЕ ВЕРБАЛЬНЫХ АССОЦИАЦИЙ

Предлагается математическая модель представления знаний в системе поиска и обработки текстовой информации, основанная на статистическом исследовании внутритекстовых семантических связей. Формально определены понятия вербально-ассоциативных сетей предметных областей, монотематических и политематических текстов. Приводятся формулы для вычисления информативности вербально-ассоциативной связи слов, предложений и фрагментов текста.

Введение

Вербальные ассоциации – это семантические связи между словами в языке (тексте, речи), соответствующие ассоциативным отношениям между обозначаемыми ими сущностями в реальном мире. В настоящее время вербальные ассоциации изучаются главным образом в лингвистике и психолингвистике. Различают два типа таких ассоциаций – парадигматические и синтагматические [1, 2]. Парадигматические ассоциации существуют между словами языка независимо от контекста и объединяют понятия, обозначающие предметы или явления, между которыми имеется постоянная связь (например, пары слов *книга–знание*, *человек–дом*). В противоположность парадигматическим, синтагматические ассоциации возникают в тексте, т. е. между словами и словосочетаниями каждого конкретного его предложения (например, в парах слов *технология–информационная*, *текст–шрифт*).

В отличие от существующих моделей представления знаний о текстовых документах (см., например, [3, 4]) в предлагаемой модели исследуется статистика лексики не только анализируемого текста, но и предметных областей, представленных тематическими и полными корпусами текстов [5].

1. Вербально-ассоциативная сеть предметной области

Будем различать вербально-ассоциативные сети предметных областей и анализируемых текстов. В каждой предметной области ее сети соответствует тематический корпус текстов, на основе которого данная сеть формируется. Тематический корпус – это совокупность текстов по конкретной тематике. Все тематические корпусы образуют полный корпус текстов, который создается для каждого языка взаимодействия пользователей с информационной системой.

1.1. Отношение вербально-ассоциативной связи слов

Рассмотрим тематические корпусы текстов Ct_{ij} ($i = \overline{1, m}$, $j = \overline{1, n_i}$) и полные корпусы $Cf_i = \bigcup_{j=1}^{n_i} Ct_{ij}$. Полный корпус текстов Cf_i соответствует i-му входному языку (например, английскому), а тематический корпус Ct_{ij} – j-й предметной области для i-го языка (например, предметной области Mathematics, представленной текстами на английском языке).

Обозначим через W_i множество всех слов полного корпуса текстов Cf_i. Тогда отношение толерантности Θ_i (рефлексивное и симметричное бинарное отношение) на множестве W_i назовем *отношением вербально-ассоциативной связи слов в полном корпусе текстов* Cf_i, если любая упорядоченная пара слов (a, b) из множества W_i является элементом отношения Θ_i тогда и только тогда, когда слова a и b из этой пары содержатся хотя бы в одном предложении корпуса Cf_i. Если пара (a, b) любых слов из множества W_i является элементом отношения Θ_i, т. е. $(a, b) \in W_i$, то (a, b) (по аналогии с работой [6]) будем называть *вербально-ассоциативной парой*.

Обозначим через W_{ij} множество всех слов тематического корпуса текстов Ct_{ij}. Рассмотрим сужение Θ_{ij} отношения Θ_i на множество W_{ij}, т. е. $\Theta_{ij} = \Theta_i \cap (W_{ij} \times W_{ij})$. Отношение Θ_{ij} назовем *отношением вербально-ассоциативной связи слов в тематическом корпусе текстов* Ct_{ij}.

1.2. Информативность слов и вербально-ассоциативных пар предметной области

Информативность $I^a_{Ct_{ij}}$ слова a из тематического корпуса текстов Ct_{ij} – это вероятность нахождения слова a в данном корпусе при условии, что оно содержится в полном корпусе текстов. При достаточно больших объемах тематического и полного корпусов текстов формула для вычисления информативности слова имеет вид [7]

$$I^a_{Ct_{ij}} = n^a_{Ct_{ij}} / n^a_{Cf_i}, \quad (1)$$

где $n^a_{Ct_{ij}}$, $n^a_{Cf_i}$ – абсолютные частоты встречаемости слова a (с учетом синонимии и словоизменения) в тематическом Ct_{ij} и полном Cf_i корпусах текстов.

Понятие информативности вербально-ассоциативной пары предметной области определим по аналогии с понятием информативности слова.

Пусть имеется пара слов a, b входного языка информационной системы. Рассмотрим следующую совокупность событий (в теоретико-вероятностном смысле):

$S^{ab}_{Ct_{ij}}$ – слова a и b извлечены случайным образом из одного и того же предложения тематического корпуса текстов Ct_{ij};

$S^{ab}_{Cf_i}$ – слова a и b извлечены случайным образом из одного и того же предложения полного корпуса текстов Cf_i;

$H_{Ct_{ij}}$ – появление тематического корпуса текстов Ct_{ij}.

Обозначим через $P(S^{ab}_{Ct_{ij}} / S^{ab}_{Cf_i})$ вероятность извлечения слов a и b из одного и того же предложения множества Ct_{ij} при условии, что они уже извлечены из одного и того же предложения полного корпуса текстов Cf_i. Эта условная вероятность вычисляется следующим образом:

$$P(S^{ab}_{Ct_{ij}} / S^{ab}_{Cf_i}) = \frac{P(S^{ab}_{Ct_{ij}} \cdot S^{ab}_{Cf_i})}{P(S^{ab}_{Cf_i})} = \frac{P(S^{ab}_{Ct_{ij}}) \cdot P(S^{ab}_{Cf_i} / S^{ab}_{Ct_{ij}})}{P(S^{ab}_{Cf_i})}.$$

Вероятность $P(S^{ab}_{Ct_{ij}} / S^{ab}_{Cf_i})$ будем называть *информативностью* вербально-ассоциативной пары (a, b) в тематическом корпусе текстов Ct_{ij} (или в предметной области, определяемой корпусом Ct_{ij}).

По аналогии с формулой (1) информативность $I^{ab}_{Ct_{ij}}$ можно представить в виде

$$I^{ab}_{Ct_{ij}} = n^{ab}_{Ct_{ij}} / n^{ab}_{Cf_i}, \quad (2)$$

где $n^{ab}_{Ct_{ij}}$, $n^{ab}_{Cf_i}$ – абсолютные частоты совместной встречаемости слов a и b (с учетом синонимии и словоизменения) в одном и том же предложении тематического Ct_{ij} и полного Cf_i корпусов текстов.

1.3. Определение вербально-ассоциативной сети предметной области

Пусть S_{ij} – граф отношения Θ_{ij} вербально-ассоциативной связи слов в корпусе Ct_{ij}. Пометим каждую вершину a графа S_{ij} значением информативности $I^a_{Ct_{ij}}$ этого слова (с учетом синонимии и словоизменения), а каждое ребро (a, b) – значением информативности $I^{ab}_{Ct_{ij}}$ вербально-

ассоциативной связи слов *a* и *b* (также учитывая синонимию и словоизменения). Обозначим полученный граф через Net_{ij}.

Граф Net_{ij} назовем *вербально-ассоциативной сетью предметной области*, определяемой тематическим корпусом текстов Ct_{ij}.

При практической реализации информационной системы вербально-ассоциативная сеть Net_{ij} предметной области представляется в виде вербально-ассоциативного словаря (таблица).

Фрагмент вербально-ассоциативного словаря

Слово 1	Информативность слова 1	Слово 2	Информативность слова 2	Информативность вербально-ассоциативной пары
...				
текст	0,57	шрифт	0,29	0,28
...				
технология	0,43	информационная	0,51	0,78
...				

В таблице каждому слову ставится в соответствие его информативность $I^{a}_{Ct_{ij}}$, вычисляемая по формуле (1), а каждой вербально-ассоциативной паре слов – информативность вербально-ассоциативной связи между ними $I^{ab}_{Ct_{ij}}$, которая определяется в соответствии с формулой (2).

Пусть *a* – некоторая вершина сети Net_{ij}, а Net^{a}_{ij} – звездный подграф графа Net_{ij}, определяемый вершиной *a*, т. е. граф, образованный выделением одной вершины *a* и всех смежных с нею вершин.

Граф Net^{a}_{ij} будем называть *вербально-ассоциативным полем* слова *a*.

2. Вербально-ассоциативная сеть монотематического текста

Будем различать монотематические и политематические тексты. Под монотематическим будем понимать текст, посвященный единой тематике (например, статья, доклад). Политематический текст (например, книга) является конкатенацией («склейкой») нескольких монотематических текстов (например, разделов).

Вербально-ассоциативная сеть монотематического текста – это подграф сети некоторой предметной области, которой соответствует тематический (или динамический [8]) корпус текстов, релевантный анализируемому тексту. Динамический корпус создается в процессе информационного поиска из документов полного корпуса.

Рассмотрим процедуру поиска тематического корпуса текстов, релевантного анализируемому тексту.

2.1. Поиск релевантного тематического корпуса текстов

Пусть *T* – произвольный текст на *i*-м входном языке, который будем рассматривать как запрос на поиск релевантного ему тематического корпуса текстов Ct_{ij}. В результате индексирования текста *T*, т. е. реализации инъективного отображения $\omega : \{T \mid T \in Cf_i, i = \overline{1, m}\} \to PP$ множества входных текстов в множество *PP* текстов на внутреннем языке информационной системы, получим поисковое предписание $\omega(T) = \{b_1, b_2, ...\}$, где $b_1, b_2, ...$ – ключевые слова этого предписания. (В качестве ключевых будем использовать все слова текста *T*, имеющиеся в словаре *i*-го входного языка.)

Рассмотрим поисковый образ $\omega(Ct_{ij}) = \{a_1, a_2, ...\}$ некоторого тематического корпуса текстов $Ct_{ij} \in Cf_i$. Естественно считать, что корпус Ct_{ij} релевантен запросу *T*, если значение критерия выдачи для пары $(\omega(Ct_{ij}), \omega(T))$ не меньше некоторого порогового значения η_0. Под критерием выдачи будем понимать отображение $\eta : \{\omega(Ct_{ij}) \mid j = \overline{1, n_i}\} \times \{\omega(T) \mid T \in Cf_i, i = \overline{1, m}\} \to R$ де-

картова произведения множеств поисковых образов тематических корпусов текстов $\{\omega(Ct_{ij}) \mid j = \overline{1, n_i}\}$ и поисковых образов входных текстов (являющихся поисковыми предписаниями) $\{\omega(T) \mid T \in Cf_i, i = \overline{1, m}\}$ в множество действительных чисел R.

Введем в рассмотрение l-мерное евклидово пространство E. Для этого лексикографически упорядочим все слова полного корпуса текстов Cf_i, т. е. сформируем кортеж $W_{Cf_i} = \langle c_1, c_2, \ldots, c_l \rangle$. Для каждого тематического корпуса текста Ct_{ij} построим вектор его поискового образа в пространстве E: $\mathbf{F}_{Ct_{ij}} = (J_{c_1}, J_{c_2}, \ldots, J_{c_l})$, где $J_{c_1}, J_{c_2}, \ldots, J_{c_l}$ – значения информативности ключевых слов $c_1, c_2, \ldots, c_l$ соответственно. (Компонента вектора $J_{c_k} = 1$, если слово c_k присутствует в поисковом образе корпуса Ct_{ij}, и $J_{c_k} = 0$ в противном случае.) Аналогично представим вектор поискового предписания: $\mathbf{F}_T = (I_{c_1}, I_{c_2}, \ldots, I_{c_l})$.

Как показано в работе [8], в качестве критерия выдачи целесообразно использовать косинус угла между векторами $\mathbf{F}_T$ и $\mathbf{F}_{Ct_{ij}}$:

$$\cos\varphi = \frac{\mathbf{F}_T \mathbf{F}_{Ct_{ij}}}{|\mathbf{F}_T||\mathbf{F}_{Ct_{ij}}|} = \frac{\sum_{k=1}^{l} I_{c_k} J_{c_k}}{\sqrt{\sum_{k=1}^{l} I_{c_k}^2} \sqrt{\sum_{k=1}^{l} J_{c_k}^2}}. \qquad (3)$$

Соответствующая поисковая функция $\pi : \{T \mid T \in Cf_i, i = \overline{1, m}\} \to \{Ct_{ij} \mid j = \overline{1, n_i}\}$, т. е. частичное мультиотображение множества анализируемых текстов в множество тематических корпусов текстов, примет вид

$$\pi(T) = \{Ct_{ij} \mid j = \overline{1, n_i}, \cos\varphi \geq \eta_0\}.$$

Приведем описание алгоритма поиска тематического корпуса текстов Ct_{ij}, релевантного тексту T.

На вход алгоритма поступает текст T, в результате индексирования которого получаем вектор поискового предписания $\mathbf{F}_T = (I_{c_1}, I_{c_2}, \ldots, I_{c_l})$. Далее по предписанию $\mathbf{F}_T$ реализуется поиск поисковых образов тематических корпусов текстов в множестве $\{\omega(Ct_{ij}) \mid j = \overline{1, n_i}\}$. Результатом поиска считаем корпус Ct_{ij}, которому соответствует наибольшее из значений критерия выдачи (3), такое, что $\cos\varphi \geq \eta_0$. Если такой корпус не найден, для текста T необходимо сформировать динамический корпус текстов.

2.2. Формирование релевантного динамического корпуса текстов

Для создания динамического корпуса текстов, релевантного запросу T, в множестве Cf_i нужно найти все документы, релевантные тексту T.

Пусть $d \in Cf_i$ – произвольный документ из полного корпуса текстов. Построим вектор $\mathbf{F}_d$ поискового образа $\omega(d)$ документа d по аналогии с вектором $\mathbf{F}_{Ct_{ij}}$: $\mathbf{F}_d = (J_{c_1}, J_{c_2}, \ldots, J_{c_l})$. Компонента вектора $J_{c_k} = 1$, если слово c_k имеется в поисковом образе текста d, в противном случае $J_{c_k} = 0$. В качестве критерия выдачи будем использовать аналог критерия (3):

$$\cos\psi = \frac{\mathbf{F}_T \mathbf{F}_d}{|\mathbf{F}_T||\mathbf{F}_d|} = \frac{\sum_{k=1}^{l} I_{c_k} J_{c_k}}{\sqrt{\sum_{k=1}^{l} I_{c_k}^2} \sqrt{\sum_{k=1}^{l} J_{c_k}^2}}. \qquad (4)$$

Считаем, что документ d релевантен запросу T и принадлежит создаваемому динамическому корпусу текстов Ct, если значения критерия (4) не меньше порогового значения η_0', т. е. при поиске будем использовать поисковую функцию

$$\pi(T) = \{d \mid d \in Cf_i, \cos\psi \geq \eta_0'\}.$$

Алгоритм формирования динамического корпуса текстов, релевантных тексту T, работает следующим образом. На входе алгоритма – текст T (запрос на поиск). По полученному в результате индексирования запроса T поисковому предписанию $\mathbf{F}_T = (I_{c_1}, I_{c_2}, \ldots, I_{c_l})$ проводится поиск поисковых образов текстов в множестве $\{\omega(d) \mid d \in Cf_i\}$ в соответствии с критерием выдачи (4). Результатом поиска является множество текстовых документов $Ct = \{d \mid d \in Cf_i, \cos\psi \geq \eta_0'\}$, образующее динамический корпус текстов Ct.

2.3. Определение вербально-ассоциативной сети монотематического текста

Обозначим через $Kt \in \{Ct_{ij}, Ct\}$ релевантный тексту T тематический (Ct_{ij}) или динамический (Ct) корпус текстов, через Net_{Kt} – вербально-ассоциативную сеть предметной области, определяемой корпусом Kt, а через W_T – множество всех слов текста T. Исключим из сети Net_{Kt} все вершины, не принадлежащие множеству W_T, и все инцидентные им ребра. Полученный граф обозначим через Net_T. Граф Net_T назовем *вербально-ассоциативной сетью монотематического текста* T, определяемой корпусом текстов Kt.

Если ρ – любое предложение текста T, то, исключив из графа Net_{Kt} вершины и ребра (по аналогии с формированием сети Net_T), получим *вербально-ассоциативную сеть* Net_ρ *предложения* ρ.

3. Вербально-ассоциативная сеть политематического текста

Полнотекстовые документы характеризуются, как правило, политематичностью. В связи с этим при обработке таких текстов необходимо их разбиение на субтексты. (Субтекст – это «компонент связного текста, развивающий одну из его основных тем» [9].)

При вычислении информативности слов, предложений и субтекстов политематического текста, а также вербально-ассоциативных связей между ними будем использовать их статистические характеристики и характеристики полного корпуса текстов, не прибегая к анализу тематических и динамических корпусов текстов.

3.1. Информативность слов политематического текста

Пусть теперь Q ($Q \in Cf_i$) – политематический текст на i-м входном языке. Для вычисления информативности I_Q^a произвольного слова a этого текста будем использовать формулу, аналогичную формуле (1):

$$I_Q^a = n_Q^a / n_{Cf_i}^a,$$

где n_Q^a, $n_{Cf_i}^a$ – абсолютные частоты встречаемости слова a (с учетом синонимии и словоизменения) в тексте Q и полном корпусе текстов Cf_i.

3.2. Информативность вербально-ассоциативных пар политематического текста

Обозначим через W_Q множество всех слов текста Q, а через Θ_Q – сужение отношения Θ_i на множество W_Q, т. е. $\Theta_Q = \Theta_i \cap (W_Q \times W_Q)$. Отношение Θ_Q назовем *отношением вербально-ассоциативной связи слов* в тексте Q. Пару (a, b) любых слов из множества W_Q, которая является элементом отношения Θ_Q, т. е. $(a, b) \in W_Q$, будем называть *вербально-ассоциативной парой* текста Q. Формула для вычисления информативности I_Q^{ab} вербально-ассоциативных пар текста Q аналогична формуле (2):

$$I_Q^{ab} = n_Q^{ab} / n_{Cf_i}^{ab},$$

где n_Q^{ab}, $n_{Cf_i}^{ab}$ – абсолютные частоты совместной встречаемости слов *a* и *b* (с учетом синонимии и словоизменения) в одном и том же предложении текста *Q* и полного корпуса текстов Cf_i.

3.3. Определение вербально-ассоциативной сети политематического текста

Обозначим через S_Q граф отношения Θ_Q вербально-ассоциативной связи слов в политематическом тексте *Q*. Пометим каждую вершину *a* графа S_Q значением информативности I_Q^a этого слова (с учетом синонимии и словоизменения), а каждое ребро (*a*, *b*) – значением информативности I_Q^{ab} вербально-ассоциативной связи слов *a* и *b* в тексте *Q* (также учитывая синонимию и словоизменения).

Обозначим полученный граф через Net_Q и назовем его *вербально-ассоциативной сетью* политематического текста *Q*. Если *a* – некоторая вершина сети Net_Q, а Net_Q^a – звездный подграф графа Net_{ij}, определяемый вершиной *a*, граф Net_Q^a будем называть *вербально-ассоциативным полем* слова *a* в тексте *Q*.

3.4. Информативность предложений и субтекстов

Пусть, по-прежнему, $W_{Cf_i} = \langle c_1, c_2, \ldots, c_l \rangle$ – кортеж всех слов полного корпуса текстов Cf_i, *E* – *l*-мерное евклидово пространство, а $\pi = a_1 a_2 \ldots$ – произвольное предложение политематического текста *Q*. Представим это предложение вектором в пространстве *E*: $\mathbf{\Pi} = (I_\pi^{c_1}, I_\pi^{c_2}, \ldots, I_\pi^{c_l})$, где $I_\pi^{c_1}, I_\pi^{c_2}, \ldots, I_\pi^{c_l}$ – значения информативности слов предложения π. При этом компонента вектора $\mathbf{\Pi}$ равна нулю, если соответствующего слова нет в предложении π. С учетом рассмотренных обозначений нормализованную информативность I_Q^π предложения π можно интерпретировать как проекцию вектора $\mathbf{\Pi}$ на направление единичного вектора **e** в пространстве *E* [8], т. е. скалярное произведение векторов $\mathbf{\Pi}$ и **e**:

$$I_Q^\pi = \frac{\sum_{i=1}^{l} I_\pi^{c_i}}{\sqrt{\sum_{i=1}^{l} (I_\pi^{c_i})^2}}. \qquad (5)$$

Аналогично вектору $\mathbf{\Pi}$ представим вектор субтекста $Sub = \pi_1 \pi_2 \ldots$ текста *Q* с предложениями $\pi_1, \pi_2, \ldots$: $\mathbf{Sub} = (I_Q^{\pi_1}, I_Q^{\pi_2}, \ldots, I_Q^{\pi_l})$. Тогда информативность субтекста *Sub* имеет вид

$$I_Q^{Sub} = \frac{\sum_{i=1}^{l} I_Q^{\pi_i}}{\sqrt{\sum_{i=1}^{l} (I_Q^{\pi_i})^2}}. \qquad (6)$$

3.5. Информативность вербально-ассоциативной связи предложений и субтекстов

Пусть π и ρ – произвольные предложения или словосочетания текста *Q*, а I_Q^{ab} – информативность его вербально-ассоциативных пар (*a*, *b*). Тогда информативность $I_Q^{\pi\rho}$ вербально-ассоциативной связи этих предложений будем вычислять по аналогии с вычислением информативности предложений (см. (5)) по формуле

$$I_Q^{\pi\rho} = \frac{\sum_{a \in \pi,\, b \in \rho} I_Q^{ab}}{\sqrt{\sum_{a \in \pi,\, b \in \rho} (I_Q^{ab})^2}}.$$

Обозначим через Sub_1 и Sub_2 субтексты текста Q. Пусть, по-прежнему, $I_Q^{\pi\rho}$ – информативность вербально-ассоциативной связи между любыми предложениями π и ρ текста Q. Тогда для вычисления информативности вербально-ассоциативной связи между субтекстами Sub_1 и Sub_2 построим аналог формулы (6):

$$I_Q^{Sub_1, Sub_2} = \frac{\sum\limits_{\pi \in Sub_1,\, \rho \in Sub_2} I_Q^{\pi\rho}}{\sqrt{\sum\limits_{\pi \in Sub_1,\, \rho \in Sub_2} (I_Q^{\pi\rho})^2}}.$$

Заключение

Модель представления знаний о текстовых документах, основанная на статистическом исследовании вербально-ассоциативных отношений, может быть использована при решении следующих задач.

Автоматическое индексирование текстовых документов и запросов на поиск информации. В индексируемом тексте выявляются не только информативные (ключевые) слова, но и вербально-ассоциативные пары. Информативность слов и вербально-ассоциативных пар определяется в соответствии с моделями, представленными в разд. 2 (для монотематического текста) и в разд. 3 (для политематического текста). Поисковый образ проиндексированного текста включает два компонента: первый в виде пар *слово–информативность слова*, а второй как совокупность вербально-ассоциативных словосочетаний, каждому из которых соответствует значение их информативности. Запросы на поиск информации индексируются как монотематические тексты согласно модели, изложенной в разд. 2.

Автоматическое реферирование и аннотирование текста. Одним из этапов реализации этих процессов является построение в тексте сверхфразовых единств и субтекстов. Формирование этих фрагментов текста осуществляется путем вычисления информативности вербально-ассоциативных связей между его предложениями (см. модель в подразд. 3.5).

Информационный поиск. Документальный поиск проводится в два этапа. На первом этапе ищутся текстовые документы по ключевым словам, на втором результаты поиска ранжируются с использованием вербально-ассоциативных пар. Фактографическому поиску предшествует выявление в полнотекстовом документе релевантных запросу субтекстов.

Список литературы

1. Морковкин, В.В. Идеографические словари / В.В. Морковкин [Электронный ресурс]. – Режим доступа : http://rifmovnik.ru/ideog_book.htm. – Дата доступа : 07.09.2011.

2. Мартинович, Г.А. Вербальные ассоциации в ассоциативном эксперименте / Г.А. Мартинович. – СПб. : Изд-во СПбГУ, 1997. – 72 с.

3. Релевантность // Помощь Рамблер [Электронный ресурс]. – 2010. – Режим доступа : http://help.rambler.ru/article.html?s=154&id=567. – Дата доступа : 07.09.2011.

4. Brin, S. The Anatomy of a Large-Scale Hypertextual Web Search Engine / S. Brin, L. Page // The Stanford University InfoLab [Electronic resource]. – Mode of access : http://infolab. stanford.edu/~backrub/google.html. – Date of access : 07.09.2011.

5. Липницкий, С.Ф. Модели знаний о предметной области для решения задач поиска и обработки текстовой информации / С.Ф Липницкий // Информатика. – 2007. – № 2 (14). – С. 25–34.

6. Мартинович, Г.А. Вербальные ассоциации и организация лексикона человека / Г.А. Мартинович // Филологические науки [Электронный ресурс]. – 1989. – № 3. – С. 39–45. – Режим доступа : http://lit.lib.ru/m/martinowich_g_a/02assfilnauk.shtml. – Дата доступа : 07.09.2011.

7. Липницкий, С.Ф. Семантический анализ текста на основе ситуативно-синтагматической сети / С.Ф. Липницкий // Информатика. – 2005. – № 2 (6). – С. 102–110.

8. 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.

9. Вейзе, А.А. Чтение, реферирование и аннотирование иностранного текста / А.А. Вейзе. – М. : Высшая школа, 1985. – 127 с.

Поступила 05.10.11

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

S.F. Lipnitsky

A MODEL OF KNOWLEDGE REPRESENTATION IN INFORMATION SYSTEMS BASED ON VERBAL ASSOCIATIONS

A mathematical model of knowledge representation is suggested for the text searching and processing systems. The model is based on statistical analysis of intra-semantic relationships. The concepts of verbal and associative network of knowledge domain as well as monothematic and polythematic texts are formally defined. Formulas for calculating self-descriptiveness of verbal and associative relation of words, sentences and text fragments are presented.