УДК 004.912 + 004.738.52

И.С. МИСУНО, Д.А. РАЧКОВСКИЙ, С.В. СЛИПЧЕНКО

ВЕКТОРНЫЕ И РАСПРЕДЕЛЕННЫЕ ПРЕДСТАВЛЕНИЯ, ОТРАЖАЮЩИЕ МЕРУ СЕМАНТИЧЕСКОЙ СВЯЗИ СЛОВ

Abstract: *Methods for formation of multidimensional vector representation of words reflecting their semantic similarity are considered. The methods are based on statistics of co-occurrence of words and contexts that is extracted from large text corpuses. Prototypes of software systems for processing of textual information, formation of semantic representations and text search are implemented. Results of experimental investigation of the developed representations in a number of tests are provided.*
Key words: *semantic similarity of words, vector representations, distributed representations.*

Анотація: *Розглянуто методи формування багатовимірних векторних представлень слів, що відображають їх семантичну близькість. Використовується статистика сумісного вживання слів і контекстів у великих корпусах текстів. Реалізовано прототип програмних засобів обробки текстової інформації, формування семантичних представлень та текстового пошуку. Приведено результати експериментальних досліджень адекватності розроблених представлень у ряді тестових задач.*
Ключові слова: *семантична близькість слів, векторні представлення, розподілені представлення.*

Аннотация: *Рассматриваются методы формирования многомерных векторных представлений слов, отражающих их семантическую близость. Используется статистика совместной встречаемости слов и контекстов, извлекаемая из больших корпусов текстов. Реализован прототип программных средств обработки текстовой информации, формирования семантических представлений и текстового поиска. Приводятся результаты экспериментальных исследований адекватности разработанных представлений в ряде тестовых задач.*
Ключевые слова: *семантическая близость слов, векторные распределенные представления.*

1. Введение

Интеллектуализация систем обработки текстовой информации требует учета ее смыслового содержания. Классическими проблемами поиска являются синонимия и полисемия. При автоматическом переводе и проверке правописания также возникают проблемы выбора подходящего слова из нескольких (разных по смыслу) вариантов, предлагаемых системой; реферирование требует отбора предложений, отражающих тему и смысл текста, и т.п.

Для решения подобного рода проблем предлагается ряд подходов. Одни основаны на использовании лингвистических ресурсов, конструируемых людьми: списков синонимов, тезаурусов или онтологий (например, WordNet [1]). Такие ресурсы требуются для разных предметных областей и языков. Их создание очень трудоемко и поэтому большое внимание привлекают автоматические методы получения и представления семантической информации. В подходах, основанных на "гипотезе распределений", используются интерпретации идеи Zellig Harris [2] о том, что семантическая близость слов отражает их совместную встречаемость в сходных контекстах. Часто такие слова могут заменяться одно другим без существенного изменения смысла высказывания.

В частности, в "моделях векторных пространств" [3, 4] формируются векторные представления слов и других компонент текстов путем автоматического извлечения статистики их совместной встречаемости из больших массивов текстовой информации. Эта информация фиксируется в так называемых семантических или контекстных векторах, сходство которых отражает меру семантической близости слов [5 – 9].

Векторные модели близки к методам распределенного представления данных, развиваемым в рамках парадигмы ассоциативно-проективных нейронных сетей (АПНС) [10, 11]. Однако ранее в АПНС не исследовались представления, отражающие семантику. Это обусловило

следующие задачи данной работы: исследовать ряд подходов к формированию представлений, отражающих семантическую близость слов, а также разные меры нахождения их сходства; разработать и исследовать ряд модификаций таких методов, укладывающихся в схему АПНС и позволяющих эффективную обработку представлений с точки зрения скорости и использования памяти; испытать полученные контекстные векторы для оценки как адекватности отображения семантической близости, так и перспективности практического использования разработанных методов.

2. Векторные и распределенные представления текстовой информации

Современные методы векторного представления текстовой информации являются развитием моделей векторных пространств VSM (Vector Space Models) и GVSM (Generalized Vector Space Model – обобщенная модель векторного пространства) [3]. В таких моделях компоненты текстов (слова, словосочетания, фрагменты, целые документы) представляются многомерными векторами. Значения элементов векторов вычисляются на основании частот встречаемости компонент текстов в контексте других компонент.

Пусть компонентами текста являются слова, а контекстами – документы. Построим матрицу A встречаемости слов в документах коллекции размерностью $w \times d$, где w – размер словаря коллекции, d – число документов. Элементы матрицы a_{ij} – "веса" слов – например, частота их встречаемости в документе и др. [12]. Строки A отражают распределение слов по документам и их можно считать "контекстными векторами" соответствующих слов. Степень сходства векторов обычно оценивают как косинус угла между ними [12]. В методе LSI (Latent Semantic Indexing – латентное семантическое индексирование [4–6] предлагается сокращать размерность контекстных векторов с помощью метода SVD (разложение по сингулярным значениям).

Для больших коллекций документов полноразмерная матрица A оказывается слишком большой для обработки и даже хранения в оперативной памяти компьютера [5, 6, 13]. Один из подходов к решению этой проблемы основан на применении распределенных представлений. В отличие от локальных представлений, где каждый информационный компонент данных представлен элементом кодвектора, то есть одним из ортогональных измерений входного пространства, в распределенных представлениях каждому компоненту данных соответствует N-мерный вектор со случайно сгенерированными элементами [10]. В многомерном пространстве почти ортогональных измерений гораздо больше, чем строго ортогональных. Для ряда распределений элементов случайные векторы близки к ортогональным и достаточно хорошо аппроксимируют исходный базис. Если в качестве текстовых компонентов рассматривать документы и $N < d$, то размерность $w \times N$ новой матрицы S меньше исходной $w \times d$:

$$S = AR. \qquad (1)$$

Здесь R – матрица случайных векторов $d \times N$. Углы между парами строк S (представлениями слов) аппроксимируют углы между соответствующими парами строк A (см. также анализ в [13–17]). Матрицу S можно сформировать без A в процессе последовательного пословного прохода корпуса текстов (метод случайного индексирования RI – Random Indexing [7]). При этом к строкам

$\bar{s}_i$ матрицы S прибавляются индексные векторы документов $\bar{r}_j$ (строки матрицы R), когда слово i встречается в документе j с весом a_{ij}:

$$\bar{s}_i = \sum_{j:w(i)\in d(j)} \bar{r}_j a_{ij} \ . \tag{2}$$

Вычислительная сложность формирования S сравнима со сложностью формирования A и равна $O(LM)$, где L – число слов в корпусе; M – число ненулевых элементов в $\bar{r}$ (обычно M постоянно и $M << N$).

Метод RL (Random Labels, случайные метки) можно рассматривать как аналог RI, где в качестве контекстов используются не документы, а слова в некоторой окрестности целевого слова (например, в окне размером до 10 слов [8, 9, 18]). Здесь "исходным материалом" может быть набор матриц B_k частот совместной встречаемости целевого слова w с контекстом k (например, со словами, находящимися на расстоянии 1 от целевого). Размерность B_k равна $w \times w_k$, где w_k – размер словаря слов контекста k. Каждому слову в корпусе присваивается индексный разреженный вектор размерностью N. Такие векторы формируют строки $\bar{r}_j$ матрицы R. Затем вычисляется матрица C:

$$C = (f_1 B_1 + f_2 B_2 + ... + f_k B_k) R, \tag{3}$$

где f_k – "вес" контекста; C имеет размерность $w \times N$ и содержит в столбцах N-мерные контекстные векторы, которые отражают семантическую близость слов.

Построение C может осуществляться без B_k, а путем прохода текстового массива “контекстным окном” и прибавления к контекстному вектору целевого слова индексных векторов слов контекстов. Можно взвешивать элементы этой суммы, например, таким образом, чтобы индексные векторы слов, находящиеся ближе к целевому, получали большие веса:

$$\bar{c}_i = \sum_k \sum_{j:w(i)\in w(j,k)} f_k \bar{r}_j a_{ij} \ . \tag{4}$$

При формировании контекстных векторов по схемам (1), (2), в них фиксируется только информация о частоте встречаемости слова и контекста. Каждый контекст является уникальным и независимым от других. Совместная встречаемость слов в одном контексте увеличивает сходство их представлений (контекстных векторов). В схемах (3), (4) слово-контекст может встречаться много раз. Эта схема в варианте [18] и др. непосредственно не учитывает информацию о совместной встречаемости двух слов в контексте, так как индексный вектор слова не участвует в формировании его контекстного вектора. Однако такая информация учитывается косвенно, за счет фиксации в контекстных векторах информации о совместной встречаемости слов с одинаковыми словами контекста. Для обеих схем в качестве индексных векторов можно использовать уже полученные контекстные [9]. Это позволяет учитывать в контекстных векторах информацию о совместной встречаемости более высоких порядков [19].

В методах RI, RL [7, 8] в качестве "индексных векторов" $\bar{r}_i$ документов и слов используются разреженные тернарные векторы с малым числом ненулевых элементов $\{-1,1\}$, а итоговые контекстные векторы действительные. Обработка векторов с дискретными (бинарными или тернарными) элементами намного более эффективна, чем с действительными. Поэтому предлагается использовать дискретные контекстные векторы, которые могут конструироваться применением пороговых операций к элементам действительных векторов, а также путем непосредственного формирования из подмножеств индексных векторов с использованием процедур контекстно-зависимого прореживания [11]. Для пороговых операций результирующий вектор формируется как $\bar{x}^* = t(\bar{x}, \bar{\theta})$, где t определена следующим образом:

$$t_i = 1 \text{ при } x_i > \theta_i^+ \geq 0; \quad t_i = -1 \text{ при } x_i < \theta_i^- \leq 0; \quad \text{иначе } t_i = 1. \tag{5}$$

Значения θ_i могут быть фиксированными (например, 0 при тернаризации с нулевым порогом) или подобраны для обеспечения заданного числа ненулевых элементов, или зависеть, например, от среднего значения $E(\bar{x}_i)$. Процедура может применяться к матрицам A и B, C и S. Для векторов с $x_i \geq 0$ может применяться центрирование путем вычитания среднего контекстного вектора по всем словам или вычитания вектора с переставленными элементами.

Альтернативным способом является "непосредственное" конструирование дискретных контекстных векторов. Пусть индексные векторы $\bar{r}_j$ имеют по M ненулевых элементов. Для каждого вектора $\bar{r}_j$ сформируем случайную перестановку $\bar{r}_j^{\sim}$. Поместим в вектор $\bar{q}_j = \bar{q}_j(\bar{r}_j^{\sim}, n)$ n первых ненулевых элементов из $\bar{r}_j^{\sim}$, разместив их по номерам, которые они имели в $\bar{r}_j$. Например, при $n = M$ $\bar{q}_j(\bar{r}_j, M) \equiv \bar{r}_j$, а при $n_1 < n_2 < M$ в $\bar{q}_j(\bar{r}_j, n_1)$ и $\bar{q}_j(\bar{r}_j, n_2)$ будет содержаться подмножество ненулевых элементов r_j, причем в $\bar{q}_j(\bar{r}_j, n_2)$ будут все ненулевые элементы $\bar{q}_j(\bar{r}_j, n_1)$. Контекстный кодвектор $\bar{z}_i$ получим как

$$\bar{z}_i = \sum_{j:w(i) \in k(j)} \bar{q}_i(\bar{r}_j, f(j)). \tag{6}$$

Здесь $f(j) \in (0, M)$ – функция, регулирующая представительство $\bar{r}_j$ в $\bar{z}_i$ и учитывающая важность контекста $k(j)$. Затем подвергнем результат пороговой операции:

$$\bar{z}_i^* = t(\bar{z}_i, \bar{\theta}). \tag{7}$$

Для векторов $\bar{z}$ и $\bar{r}$ с элементами $\{0,1\}$ $\theta = 0$ (7) эквивалентно замене суммирования в (6) дизъюнкцией. $f(j)$ можно определять как

$$f(j) = \begin{cases} ent(\varphi(j)) + 1, y < \varphi(j) - ent(\varphi(j)); \\ ent(\varphi(j)), \textit{иначе}. \end{cases} \tag{8}$$

Здесь $\varphi(j)$ – нецелочисленная функция важности; y – псевдослучайное число из диапазона (0,1), фиксированное для данного j. При числе контекстов $L > M$ эта процедура случайно отбирает

("сэмплирует") некоторые документы в качестве контекстов, игнорируя другие (так называемые *скетчи* [20]). Сходные результаты (для документов, расположенных в случайном порядке) можно получить, положив $y = 1/j$.

Если все контексты имеют одинаковую важность и не учитывается частота их встречаемости, то можно принять $\varphi(j) = \varphi = M/L$. Если слово встретилось в $L = 10$ контекстах, в его контекстный вектор войдет по $M/10$ единиц от индексного кодвектора каждого из них. Если индексные векторы имеют малое пересечение, для всех векторов $\bar{z}_i$ получаем $|\bar{z}_i| \approx M$. Для нелинейного учета частот встречаемости слова в контексте

$$\varphi(j) = M(1 + \log a_{ij}) idf_j / F_j \text{ при } a_{ij} \neq 0;\ \varphi(j) = 0 \text{ при } a_{ij} = 0, \quad (9)$$

где $idf_j = \log(d/d_j)$ – обратная частота встречаемости документа; d_j – число документов, в которых встретился контекст j, $F_j = \sum_j \varphi_j$.

3. Формирование и тестирование семантических представлений

Сравнительные экспериментальные исследования характеристик семантических векторов проводились для разных вариантов их формирования на ряде тестовых задач. В качестве обучающего материала для формирования векторов использовались англоязычные корпусы текстов TASA, BNC и корпус страниц Интернет.

3.1. Корпусы текстов

TASA (Touchstone Applied Sciences Association [21]) – коллекция текстов по литературе, биологии, экономике, промышленности, науке, искусству, социологии и бизнесу, рекомендованных для чтения в колледжах США. Более чем 37600 текстов, каждый из которых состоит примерно из 166 слов из словаря в 94000 слов, в сумме содержат более чем 10 млн. слов.

BNC (British National Corpus [22]) – коллекция текстов современного английского языка (разговорного и письменного), содержащая порядка 100 млн. слов. Письменная часть (порядка 90% коллекции) состоит из газетного материала, специальной периодики для аудитории разных возрастов, академической и популярной литературы, писем и меморандумов, школьных и университетских сочинений и прочих текстов. Разговорная часть состоит из записей разговоров добровольцев разных возрастов, регионов и социального положения.

Постоянно растущим текстовым корпусом является Интернет. Например, летом 2005 число web-страниц, проиндексированных Google, составляло порядка 9 млрд., Alta-Vista – 3,5 млрд.

3.2. Меры сходства

Для исследования адекватности отражения контекстными векторами семантической близости слов проводилось сравнение величины или ранга сходства контекстных векторов с эталонными результатами. В качестве контекстных векторов использовались как непосредственно строки матриц A и B, C и S, так и их дискретизированные (бинарные и тернарные) варианты.

Величина сходства контекстных векторов оценивалась по ряду мер (всего испытывалось около 90 мер, здесь приводятся только некоторые из них). Обозначим $\bar{u}$ и $\bar{v}$ векторы, между которыми оценивается сходство; $\bar{u}^{\sim}$ – случайная перестановка $\bar{u}$; $|\bar{u}|_p$ – норма l_p вектора $\bar{u}$; $(\bar{u},\bar{v})$ – скалярное произведение векторов.

Меры на основе скалярного произведения (x –показатель степени, обычно x =1, p =0, 1, 2):

$$COS_p^x(\bar{u},\bar{v}) = (\bar{u},\bar{v})^x / (|\bar{u}|_p |\bar{v}|_p) = (\bar{u}^x / |\bar{u}|_p, \bar{v}^x / |\bar{v}|_p),$$
$$COX_p^x(\bar{u},\bar{v}) = \left(\bar{u}-\bar{u}^{\sim}, \bar{v}-\bar{v}^{\sim}\right)^x / \left(|\bar{u}-\bar{u}^{\sim}|_p |\bar{v}-\bar{v}^{\sim}|_p\right). \quad (10)$$

Сходство по Жаккару:

$$Jaccard = (\bar{u},\bar{v}) / \left[(\bar{u},\bar{u}) + (\bar{v},\bar{v}) - (\bar{u},\bar{v})\right]. \quad (11)$$

Взаимная информация:

$$MI(\mathbf{u},\mathbf{v}) = \sum_{\mathbf{u}=\{0,1\}} \sum_{\mathbf{v}\in=\{0,1\}} P(\mathbf{u},\mathbf{v}) \log_2 \frac{P(\mathbf{u},\mathbf{v})}{P(\mathbf{u})P(\mathbf{v})} \approx \sum_{\mathbf{u}=\{0,1\}} \sum_{\mathbf{v}\in=\{0,1\}} \frac{(\mathbf{u},\mathbf{v})}{N} \log_2 \frac{(\mathbf{u},\mathbf{v})}{(\mathbf{u},\mathbf{u})(\mathbf{v},\mathbf{v})} N. \quad (12)$$

При вычислении сходства по частотам совместной встречаемости слов u и v в некотором контексте также использовались $MI(u,v)$ и точечная взаимная информация *PMI*:

$$PMI(u,v) = p(u,v) / (p(u)p(v)). \quad (13)$$

Меры на основе расстояний Минковского ($p=1$ – манхэттенова, $p=2$ – евклидова):

$$L_p(\bar{u},\bar{v}) = \left(\left\|\bar{u}-\bar{v}\right\|^p\right)^{1/p}. \quad (14)$$

Меры на основе "информационного отклонения" D_δ [23]. Пусть $p_i = u_i / \|\bar{u}\|_1$, $q_i = v_i / \|\bar{v}\|_1$,

$$D_\delta = \sum\nolimits_i \left[\delta p_i + (1-\delta) q_i - p_i^\delta q_i^{1-\delta}\right] / \left[\delta(1-\delta)\right], \quad (15)$$

где δ – отклонение, применяемое в байесовской теории распознавания и определяющее интегральную ошибку отклонения распределений p и q. При δ=0 и 1 (15) дает расстояние Кульбака-Лейблера $\mathbf{p}\log(\mathbf{p}/\mathbf{q})$ и $\mathbf{q}\log(\mathbf{q}/\mathbf{p})$, а при δ=0,5 – расстояние Хеллингера $(p^{0.5}-q^{0.5})^2$.

Производилось также усреднение результатов, полученных на разных реализациях контекстных векторов. Обозначим величину сходства для z-й реализации векторов как SIM_z. Усредненные величины вычислялись как

$$E(SIM_z) = 1/Z \sum\nolimits_{z=1,Z} SIM_z. \quad (16)$$

Далее результаты определялись по сравнению величин $E(SIM_z)$. В другой версии по каждой SIM_z определялся ближайший сосед, и результат определялся голосованием.

В другом варианте производилось усреднение путем суммирования контекстных векторов, полученных по каждой их реализации, и сходство находилось по этим векторам:

$$u_z = \sum\nolimits_{z=1,Z} u_z. \quad (17)$$

3.3. Программная реализация

Для формирования контекстных векторов и проведения экспериментов по определению семантического сходства и текстового поиска создан ряд вариантов программного обеспечения. Один из вариантов представляет собой пакет модульных консольных программ, реализованных на C++. Последовательность вызовов программ при обработке коллекции, входные и выходные данные приведены на рис. 1. Программы пакета выполняют следующие операции:

а) индексирование корпуса – преобразование текстового формата корпуса в последовательность индексов слов для унификации и ускорения работы с корпусом;

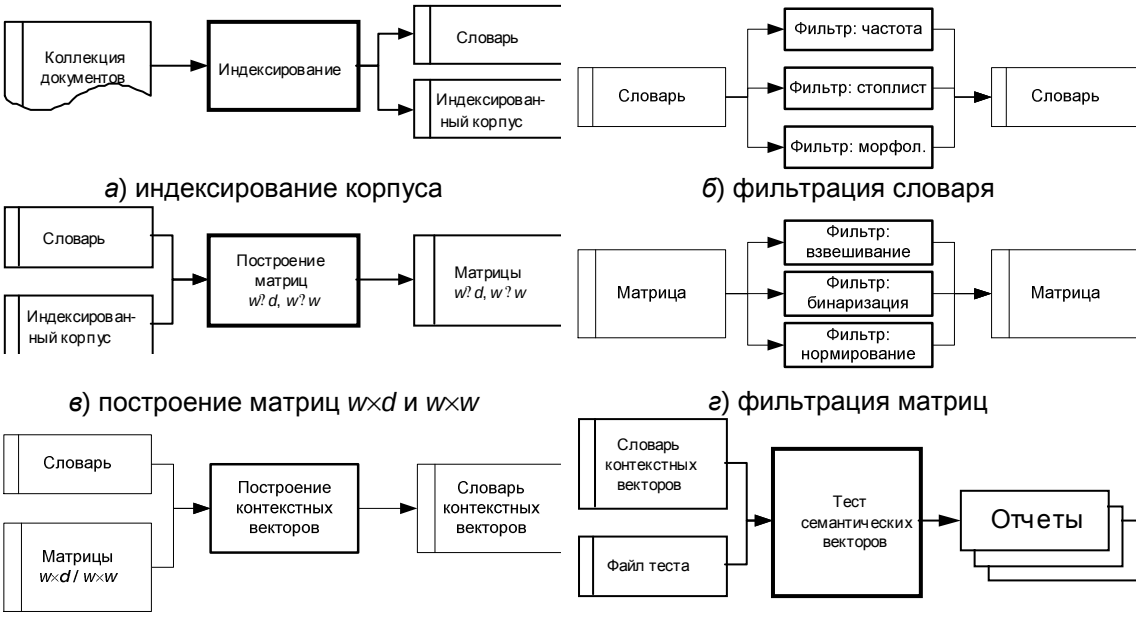
б) фильтрация словаря – включает фильтрацию слов по частоте, исключение стоп-слов, морфологию (приведение к базовой словоформе, или *обрезку*);

в) построение матриц частот совместной встречаемости слов в контекстах корпуса (коллекции). Например, $w \times d$ и $w \times w$ для коллекции – это подсчет частот совместной встречаемости слов в документах и в контекстном окне;

г) преобразование, взвешивание и нормирование матриц;

д) построение контекстных векторов слов – на основе матриц совместной встречаемости и словаря формируются контекстные векторы слов. Возможно применение опций и вариантов алгоритмов, а также последующее преобразование файлов контекстных векторов: дискретизация, взвешивание, нормирование и т.д. Словарь контекстных векторов записывается в файл для последующего использования;

е) *применение контекстных векторов для формирования представлений документов*. Например, тестирование сформированных контекстных векторов. Возможно циклическое выполнение программ с разными начальными значениями генераторов случайных чисел и т.д. По результатам формируются отчеты (в формате XML и др).



а) индексирование корпуса *б*) фильтрация словаря

в) построение матриц *w*×*d* и *w*×*w* *г*) фильтрация матриц

д) построение контекстных векторов *е*) тестирование

Рис. 1. Последовательность вызовов программ построения контекстных векторов

Для осуществления поиска в текстовых коллекциях, а также для оценки характеристик алгоритмов поиска разработан макет поисковой системы. Система работает под управлением веб-сервера MS IIS (Internet Information Services), веб-интерфейс пользователя реализован на Dynamic HTML с отображением в веб-браузере.

Работа с контекстными векторами на серверной стороне осуществляется с помощью технологии программирования ASP (Active Server Pages), JScript и COM-объектов, реализованных на языке C++. COM-объекты реализуют чтение словаря, индексированный поиск вектора по слову, вычисление контекстного вектора нескольких слов, вычисление коэффициента сходства между двумя произвольными контекстами (как отдельного слова, так и вычисленного контекста), поиск наиболее близких документов и произвольный доступ к документам в коллекции. Диаграмма последовательности вызовов, осуществляемая при взаимодействии пользователя с системой поиска, приведена на рис. 2.

Пользователь вводит запрос в окно веб-браузера, связанного с веб-сервером MS IIS. Сервер выполняет программу на языке ASP/JScript, которая производит обработку запроса с помощью вызовов COM-объектов. Результаты поиска в виде HTML передаются пользователю для просмотра в браузере. Внешний вид окна веб-браузера при работе с системой поиска представлен на рис. 3.

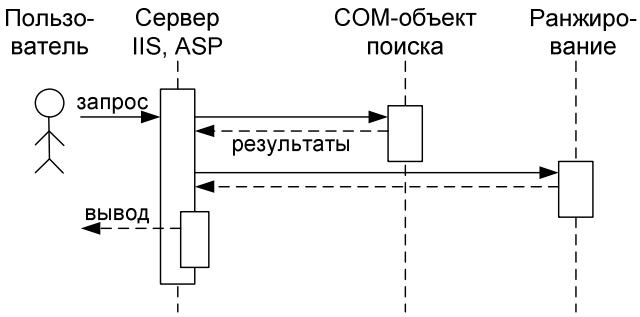


Рис. 2. Диаграмма последовательности вызовов при работе системы поиска

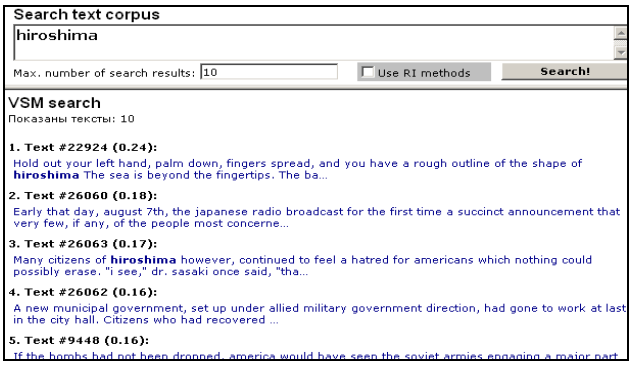


Рис. 3. Пример работы макета программы поиска

Для определения семантической меры близости слов с помощью корпуса текстов Интернет использовались возможности поисковой системы AltaVista (идея [24]). Для пары слов *u* и *v* величины сходства вычислялись по информационным мерам. Например, для PMI (13):

$$score(u, v) = hits(u \text{ AND } v) / (hits(u)\, hits(v)); \tag{18}$$

$$score(u, v) = hits(u \text{ NEAR } v) / (hits(u)\, hits(v)). \tag{19}$$

Здесь $hits(q)$ – количество найденных документов по запросу q. Ключевое слово AND позволяет найти документы, содержащие оба слова, а NEAR – документы, в которых слова находятся в окне шириной 10 друг от друга.

Для экспериментов была разработана система, состоящая из двух программ на языке C#, предоставляющем удобные средства выполнения сетевых операций. Первая программа выполняет запросы к публичной поисковой системе и разбор возвращенного HTML-кода веб-страниц с помощью регулярных выражений и сохраняет результаты в файле в формате XML. Этот

файл анализируется второй программой, которая подсчитывает величины сходства (а также формирует при необходимости список ответов на вопросы тестов и протокол результатов).

4. Результаты экспериментов

Для исследования характеристик контекстных векторов применялся ряд тестов. Исследовалась зависимость результатов тестирования от способов учета контекстов (схемы с матрицей A слова-документы и B слова-слова); различных преобразований, взвешиваний и нормировки элементов; размерности N индексных векторов и числа их положительных p и отрицательных n единичных элементов; алгоритмов формирования и типов контекстных векторов (действительные и/или дискретные с разными параметрами); меры сходства; типов усреднения, текстовых корпусов и способов их предобработки (фильтрация слов по частоте, учет морфологии и др.); разных реализаций генераторов случайных чисел. В ходе экспериментального исследования проводились массовые серийные эксперименты с разными реализациями случайных кодвекторов. В данной статье приводятся некоторые результаты и их краткое обсуждение.

Синонимический тест TOEFL. Тест TOEFL (Test Of English as a Foreign Language [25]) – распространенный тест, применяемый для тестирования знания английского языка абитуриентами университетов США. Способность выполнить этот тест с результатами выше уровня догадок требует владения некоторыми знаниями о семантике слов. Средние результаты абитуриентов составляют 64,5% правильных ответов [6]. Синонимическая часть этого теста состоит в выборе замены тестового слова одним из 4 вариантов, наиболее близким ему по смыслу.

При предварительной обработке коллекции документов исключались слова стоп-листа SMART [3], отбрасывались слова, встречающиеся менее 3 и более 14000 раз, применялась обрезка (truncation) слов по длине – *tr*:*n*, выделение базовой части слова – *morph*. Исследовалось, в основном, влияние учета разных типов контекста и способов нормировки контекстных векторов. Для метода RL (4) контекстом k является окно размером $2k$ (по k слов с обеих сторон целевого), веса слов $f_k = 2^{1-k}$ убывают по мере удаления от центра окна. Контекстные векторы RL получены при $k = 1-4$. Размерность индексных векторов RI, RL N=1800, количество ненулевых элементов $p = n = 8$. Если не указано иначе, результаты получены усреднением по E=10 реализациям векторов.

В табл. 1 приведены результаты испытаний семантических представлений, полученных на корпусах BNC, TASA. Столбец mean – среднее значение, dev – среднеквадратичное отклонение, sum – усреднение индивидуальных сходств по (16), Magnus – результаты из [8]. На TASA наилучший результат 67,8% наблюдается при отсутствии обрезки слов и контексте $k = 3,4$. В целом полученные результаты сравнимы с [8]. При нахождении результатов по суммарному сходству (16) результаты до 10% лучше среднего значения и достигают 72,5%. На большей коллекции BNC метод RL демонстрирует 72,3% при $k = 1$.

Таблица 1. Результаты на TOEFL % (контекстные векторы RL)

tr	TASA													BNC
	mean				*dev*				*sum*				*Magnus*	*mean*
k	1	2	3	4	1	2	3	4	1	2	3	4	*сред.* ±0,73	1
6	54,3	56,9	56,0	53,9	3,02	3,27	3,15	1,42	53,8	57,5	57,5	57,5	56,3	61,6
8	59,8	61,1	61,5	62,4	3,53	2,87	2,55	2,34	61,3	63,8	63,8	62,5	62,8	69,9
10	67,4	66,3	67,0	66,6	2,27	2,56	3,92	2,44	70,0	71,3	**72,5**	68,8	66,8	70,5
12	66,1	65,0	65,8	63,5	1,97	2,44	2,03	2,73	67,5	67,5	71,3	65,0	64,6	**72,3**
—	65,9	65,6	*67,8*	*67,8*	2,44	2,32	3,39	2,67	65,0	66,3	71,3	73,8	65,6	71,3

Ранее не имелось надежных результатов на этом же экспериментальном материале для локальных матриц A и B. В [6] приводится результат 36,8%. Мы провели эксперименты для различных нормировок матриц A и B и мер сходства. В результатах исследований использована нотация SMART [3]: l означает преобразование частоты $(1+\log tf)$; b – бинаризацию элементов матрицы по (5) с порогом 0; t – умножение на idf; c – нормирование векторов слов к l_2; n – отсутствие соответствующего взвешивания или нормировки.

Результаты, приведенные в табл.2, по-разному зависят от выбранной нормировки и меры. Например, для матрицы слова-документы A результаты слабо зависят от нормировок, и даже при бинаризации A с порогом 0 (*bnn*) результаты не ухудшаются, а достигают 61% для меры *cox*. Для матрицы слова-слова B с единичным окном $k=1$ лучший результат составляет 72,5% (*ntn*), бинаризация также дает хороший результат – до 65%. Таким образом, бинаризация исходных матриц заслуживает внимания как альтернатива исходным частотным представлениям слов *tf*, а выбор подходящего нормирования и меры сходства позволяет значительно улучшить результат, который превышает 64,5%, показанные методом LSI в [6].

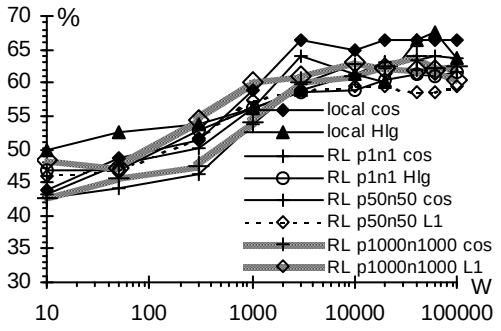


Рис. 4. Зависимость результатов от W

Для исследования влияния размера словаря на уровень результатов в качестве контекстов отбиралось W слов с наибольшей частотой встречаемости в словаре корпуса TASA. На рис. 4 показана зависимость результатов от W для локальных (local) и распределенных (RL) представлений ($N=2000$, контекст $k=1$, нормировка $B-nnn$). Даже для $W=10$ слов наблюдаются результаты на уровне 50%, которые увеличиваются с ростом W. При $W=3000$ результаты уже близки к максимальным, что открывает возможности значительного снижения размерности строк матриц B_k (и соответствующей вычислительной сложности нахождения сходства и затрат памяти) при сохранении «качества» контекстных векторов.

Результаты исследования влияния нормировок, применяемых при конструировании распределенных контекстных векторов, приведены в табл. 3, 4. Для мер сходства cos, L_2, *cox*,

$H\lg$ результаты отличаются в пределах 10%, в среднем лучшие результаты получены для $\cos$. Как и в табл. 1, наблюдается сильный разброс индивидуальных результатов (*min–max*). Средние значения при усреднении по E реализациям приведены в столбце avg, а результаты при суммировании сходств по (16) – в столбце sum. В столбце N100K приведены результаты, соответствующие «развернутому» представлению векторов RI $\left(N^*L\right)$ с количеством ненулевых элементов E. Как увеличение N, так и усреднение по (16) стабильно и сравнимо улучшает результаты – до 25%.

Таблица 2. Результаты локальных методов на TOEFL, %

	local (A, RI)					local (B, RL, *k*=1)				
Нор	*cos*	*cox*	L_1	*Hlg*	$D_{0,3}$	*cos*	*cox*	L_1	*Hlg*	$D_{0,3}$
nnn	55,0	60,0	46,3	53,8	52,5	67,5	66,3	65,0	66,3	66,3
lnn	55,0	*61,3*	45,0	53,8	55,0	63,8	63,8	58,8	62,5	61,3
bnn	56,3	*61,3*	47,5	53,8	50,0	65,0	62,5	53,8	61,3	60,0
ltn	55,0	*61,3*	50,0	56,3	51,3	67,5	66,3	57,5	62,5	66,3
ntn	55,0	60,0	47,5	53,8	51,3	**72,5**	70,0	65,0	67,5	71,3

Таблица 3. Результаты на TOEFL, % (RL, разные нормировки)

RL2000	p1n1					p50n50					p1000n1000				
Норм \ Мера	*cos*	*cox*	L_2	*Hlg*	$D_{0,3}$	*cos*	*cox*	L_2	*Hlg*	$D_{0,3}$	*cos*	*cox*	L_2	*Hlg*	$D_{0,3}$
nnn	63,6	62,3	61,9	61,9	60,6	64,5	63,8	64,5	64,4	64,5	63,9	63,4	63,9	61,6	62,3
lnn	61,3	59,8	60,5	58,8	57,3	61,3	60,8	61,3	60,8	61,0	61,1	61,5	61,1	60,0	60,5
bnn	58,6	56,6	57,4	56,8	53,5	56,4	56,6	56,4	56,3	56,6	56,8	54,9	56,8	57,0	57,9
ltn	62,3	60,4	61,9	59,3	60,4	60,5	59,4	60,5	60,3	61,0	61,1	59,8	61,1	61,6	61,1
ntn	*66,6*	63,9	66,3	62,4	62,5	64,0	63,5	64,0	64,5	63,1	64,3	64,6	64,3	63,5	63,0

Таблица 4. Результаты на TOEFL по методу RI, %

p=n=	1	1	1	1	N100K	
N	*min*	*max*	*avg*	*sum*	*n=p=E*	*n*=0
4000p1n1 cos *E*25	36,3	50,0	43,4	*53,8*	53,8	38,8
4000p1n1 cox *E*25	35,0	48,8	40,1	52,5	53,8	**55,0**
2000p1n1 cos *E*50	32,5	48,8	38,9	50,0	–	–
2000p1n1 cox *E*50	27,5	46,3	36,6	51,3	–	–
1000p1n1 cos *E*100	20,0	47,5	36,1	52,5	48,8	30,0
1000p1n1 cox *E*100	21,3	43,8	34,0	51,3	47,5	47,5

p=n=	1	50	1000
N	*avg*	*avg*	*avg*
10000 cos	*48,3*	45,1	45,9
4000 cos	41,8	41,8	43,3
2000 cos	39,5	38,6	40,6
1000 cos	38,9	35,0	—

Исследовалось применение контекстных векторов с дискретными элементами для оценки меры семантической близости слов. Результаты, полученные на тесте TOEFL, приведены на рис. 5. Для дискретизации по фиксированному порогу наилучшие результаты достигаются при значениях порога 0,1,2. Для бинарных кодвекторов наилучшие результаты дает мера сходств MI, для тернарных не наблюдается явно "лучшей" меры. Для дискретизации с переменным порогом, обеспечивающей фиксированное число ненулевых элементов M, для мер MI, cos, QA максимальные результаты наблюдаются при малых M и $p=n=1$. Для больших $p=n$ результаты растут с ростом M.

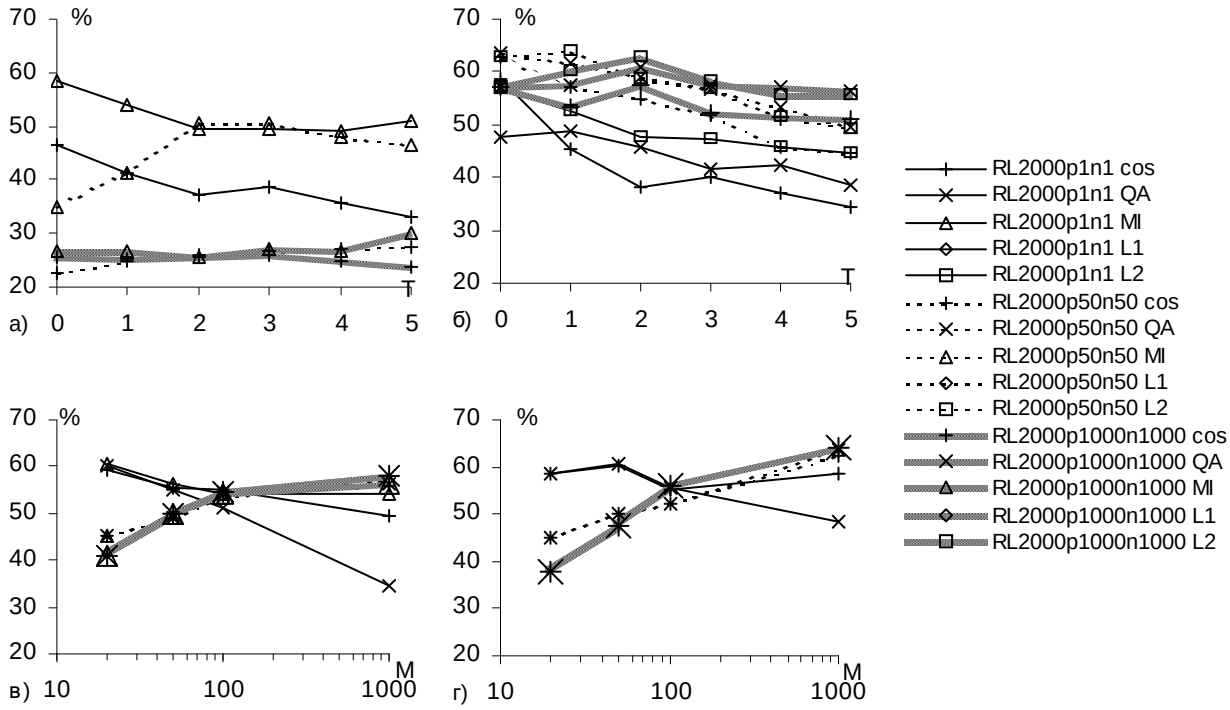


Рис. 5. Результат на TOEFL для дискретных контекстных векторов: а, б – бинаризация и тернаризация по порогу; в, г – бинаризация и тернаризация по количеству ненулевых элементов

Таким образом, существуют параметры формирования контекстных векторов и меры сходства, при которых результаты для дискретных контекстных векторов близки к результатам, полученным для действительных контекстных векторов. Результаты, полученные в протестированном диапазоне параметров для бинарных векторов (~60%), несколько ниже, чем для тернарных (~63%).

Крайне интересны результаты, полученные на TOEFL при использовании алгоритма, работающего с запросами к поисковым машинам Интернет (см. конец разд. 3.3). В работе [24] приводится результат 62,5-73,75% (объем проиндексированных Altavista страниц составлял порядка 350 млн.). В нашей работе 2003 года (примерно 2 млрд. страниц Altavista) был зафиксирован результат 81,25% В июне 2005 число страниц достигло 3,5 млрд, однако был изменен алгоритм поиска Altavista, что не дало возможности получить сопоставимые результаты.

Синонимический тест ESL. Синонимическая секция теста ESL (Test of English as a Second Language [26]) содержит набор предложений, в каждом из которых выделено целевое слово и даны 4 варианта его замены. Тест состоит из 40 заданий. Этот тест сложнее TOEFL, но в нем присутствует контекстная информация в виде слов предложения. Результаты выполнения теста ESL с использованием контекстных векторов RI, RL, сгенерированных с $N = 1800/8$ и параметрами $tr = 0, k = 3$, представлены в табл. 5, столбец “ESL”.

В экспериментах были использованы различные методики учета контекста предложения. Для *NoContext* информация о контексте предложения не используется. *Clos*: n – учет контекста как

суммы близости слова с n словами предложения. Sent: n – учет контекста определением семантической близости с вектором суммы n соседних слов предложения. Результаты демонстрируют недостоверное улучшение результатов при учете одного слова контекста (*Clos* :2 – 52%) по сравнению с отсутствием учета контекста (*NoCtx* – 50%).

Многозначность при автоматическом переводе. На основе многовариантных результатов автоматического перевода с помощью системы PROMPT С. Лебедевым был составлен тест из 40 примеров. Каждый пример содержит исходное предложение на русском языке и его перевод PROMPT на английский с двумя вариантами перевода целевого слова, один из которых надо выбрать. Например, (а) "*Простирать руки к небу*" (*To extend* [*wash*] *hands to the sky*) – "*Простирать грязное белье*" (*To wash* [*extend*] *dirty linen*); (б) "*Он медленно бродил по улицам*" (*He slowly wandered* [*fermented*] *along the streets*) – "*Молодое вино еще бродило*" (*The young wine still fermented* [*wandered*]).

Результаты экспериментов по выбору вариантов перевода неоднозначных слов с помощью семантических представлений представлены в табл. 5, столбец "Homonyms". Параметры формирования представлений: *truncation*:*morph*, $k = 3$. Лучший результат составил ~70%.

Оценка меры близости слов людьми (Wordpairs). В исследовании Rubenstein, Goodenough [27] с помощью опроса испытуемых определены меры синонимии для 65 пар слов. В этих парах встречались слова как весьма похожие, например, *gem–jewel (драгоценность)*, так и совершенно несвязанные, например, *noon–string (луна–струна)*. Miller и Charles [28] выделили из первоначального набора 30 пар слов с наиболее стабильными оценками сходства разной степени. Мы провели эксперименты с определением степени сходства этих слов по контекстным векторам. Результаты представлены в табл. 6. Наибольший коэффициент корреляции составил 0,47 для контекстных векторов, полученных на корпусе BNC методом RL.

Извлечение семантического ряда. По сформированным контекстным векторам можно вычислять сходство любых слов словаря и использовать эту информацию, например, при расширении поискового запроса, при нахождении синонимов или вариантов замены слов и др. Примеры "семантических рядов", полученных путем выбора из словаря слов с наибольшим значением сходства контекстных векторов со словами, заданными в левом столбце, приведены в табл. 7.

Таблица 5. Результаты на тестах ESL, Homonyms, %

	ESL		Homonyms			
	TASA	BNC	TASA		BNC	
Context	RL	RL	RI	RL	RI	RL
NoCtx	36	50	—	—	—	—
Clos:1	34	48	60	65,5	56,4	62,7
Clos:2	32	*52*	59,1	64,6	58,3	60,1
Clos:3	28	48	60,1	*68,6*	56,6	60,5
Clos:*	18	28	59,6	67,9	56,6	60
Sent:*	22	36	64,5	**70,7**	54,9	58,7

Поиск текстовой информации. Исследованные нами методы поиска включали как традиционный поиск VSM, так и методы на базе контекстных векторов – GVSM, RI, RL (разд. 2). Поиск по семантическим представлениям учитывает "смысл" слов и текстов. Например, поиск "hiroshima" в TASA по методам контекстных векторов возвращает некоторые тексты, не содержащие слова "hiroshima", хотя в них идет речь об атомной бомбардировке Японии в 1945 году и ее последствиях. Запрос "vessel" (судно, посудина, кровеносный сосуд) возвращал некоторые тексты, не содержащие этого слова, в которых говорилось о кровеносной системе и давлении крови.

Таблица 6. Результаты на WordPairs

RL, 1800/8	TASA	BNC
Rubenstein-Goodenough		
Корр.	0,232	0,470
Ошибка	0,121	0,109
Стьюд.	1,926	4,290
Miller-Charles		
Корр.	0,338	0,469
Ошибка	0,172	0,161
Стьюд.	1,970	2,911

Для численной оценки качества работы системы поиска применялась оценка средней (интерполированной) точности по 11-точечной характеристике recall-precision [3]. Исследования проводились на стандартных коллекциях Medlars, Cranfield, Time Magazine [29].

Векторы документов и запросов формировались суммированием контекстных векторов слов с соответствующим взвешиванием. Сходство между запросом и документом для VSM и GVSM вычислялось как скалярное произведение их векторов, для RI сходство находилось по *cos*. Конфигурации экспериментов и результаты приведены в табл. 8 . Применялись нормировки контекстных векторов, указанные в нотации SMART [3] (см. разд. “TOEFL”). Нотация $lt3$ означает, что к контекстным векторам слов применялась тернаризация с порогом 0.

Таблица 7. Примеры семантических рядов

Слово	Ближайшие слова
oxford	0,79 freiburg; 0,79 cerling; 0,79 grahamstow; 0,74 yale;
fly	0,65 flew; 0,55 walk; 0,54 move; 0,53 swim; 0,52 run
bank	0,55 banks; 0,41 savings; 0,39 checking; 0,38 park;
climate	0,53 weather; 0,39 moist; 0,36 dry; 0,36 winters
leg	0,54 wing; 0,53 hand; 0,52 arm; 0,50 shoulder;

Результаты поиска по GVSM превосходят VSM (улучшение до 25%) для коллекции MED, на TIME и CRAN оба метода дают сравнимые результаты. Применение представлений, учитывающих семантику посредством контекстных векторов, показывает результаты на уровне (или на несколько процентов выше) GVSM на всех трех коллекциях. Применение контекстных представлений с дискретными элементами дает результаты, сравнимые с контекстными представлениями с действительными элементами, но обеспечивает большие возможности повышения эффективности обработки вследствие сокращения объема памяти и упрощения вычислительных операций.

На рис. 6 приведены графики зависимости средней точности поиска от полноты. Кривые средней точности поиска по методам GVSM-ltc и RI-ltc с действительными векторами находятся рядом друг с другом и выше других методов, а кривая для контекстных векторов RI с дискретизацией проходит немного ниже, но также над кривыми других методов при полноте больше 15%.

Таблица 8. Средняя точность поиска

Метод	Med	Time	Cran
VSM nnn	0,45	0,42	0,26
VSM ltc	0,55	**0,68**	**0,42**
GVSM nnn	0,54	0,39	0,21
GVSM ltc	0,70	0,66	**0,42**
RI2000p5n5 nnn	0,54	0,48	0,28
RI2000p5n5 ltc	***0,69***	***0,69***	0,45
RI2000p5n5 lt3c	**0,63**	0,70	0,45

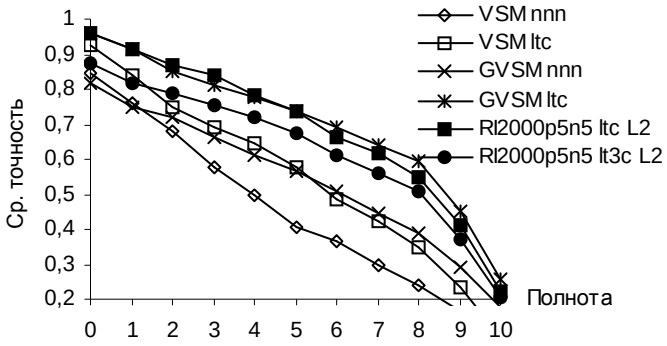


Рис. 6. Средняя точность поиска (Medlars)

5. Заключение

Проведено исследование векторных и распределенных представлений слов, отражающих их семантическую близость и моделирующих когнитивные способности людей по учету семантики и контекста. Изучался ряд вариантов автоматического формирования таких представлений путем оценки частот совместной встречаемости слов и контекстов разного типа в коллекциях документов. Тестовые исследования продемонстрировали дееспособность использованных подходов и их потенциал в приложениях, требующих информации о семантическом сходстве слов.

Предложенные и исследованные представления и методы превосходят аналоги по своей эффективности и масштабируемости, показывая сравнимые результаты. Применение распределенных представлений исключает необходимость конструирования и вычислительно сложной обработки полной матрицы слова-контексты, объем которой может быть очень велик для больших текстовых массивов. Использование бинарных и тернарных контекстных векторов позволяет сократить затраты памяти и вычислительных ресурсов на их хранение и обработку на обычных компьютерах (использование 1-2 битов для хранения элемента вектора вместо 64, применение логических операций вместо операций с плавающей запятой), а также обеспечить эффективную аппаратную поддержку в виде специализированных вычислителей и нейрокомпьютеров [30]. Такие представления совместимы с форматом данных АПНС [10], что открывает возможность использования арсенала методов информационной обработки, разработанного в АПНС, а также расширить парадигму АПНС семантическими представлениями. Разреженные варианты представлений с малым числом ненулевых элементов дополнительно повышают эффективность обработки (например, путем использования процедур, построенных на обратном индексе) и являются нейробиологически релевантными.

Разработаны и реализованы программные средства, позволяющие формировать и использовать распределенные представления в ряде практически важных задач. Проведенные исследования и полученные результаты создают предпосылки и открывают перспективы для создания новых информационных технологий и прикладных систем анализа и обобщения информации с элементами понимания семантики. Это системы автоматической проверки правописания с учетом контекста и темы текстов; составление многоязычных тезаурусов; поиск текстов; определение смысла многозначных слов; извлечение релевантной информации из текстов; создание профиля интересов пользователя и селекции текстов по интересам и др.

Авторы выражают благодарность Л. М. Касаткиной и А. М. Соколову за плодотворные обсуждения, Prof. T. Landauer за корпус TASA и вопросы TOEFL, J. Glick за доступ к API AltaVista, А. Курсину за программу morph, С. Лебедеву за тест по автоматическому переводу.

СПИСОК ЛИТЕРАТУРЫ

1. Miller G. Wordnet: a lexical database for english // Communications of the ACM. – 1995.– Vol. 38 (11).– P. 39 – 41.
2. Harris Z. Mathematical Structures of Language. – New York: John Wiley & Sons, 1968. – 230 p.
3. Salton G. Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer. – Addison-Wesley, Reading, MA. – 1989. – 530 p.
4. Yang Y., Carbonell J., Brown R., Frederking R. Translingual information retrieval: learning from bilingual corpora // Artificial Intelligence. – 1998. – N 103. – P. 323 – 345.
5. Deerwester S., Dumais S., Furnas G., Landauer T., Harshman R. Indexing by latent semantic analisys // Journal of American Society for Information Sciences. – 1990. – N 1 (6). – P. 391 – 407.
6. Landauer T., Dumais S. A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition induction and representation of knowledge // Psychological review. – 1997. – N 104(2). – P. 211 – 240.
7. Kanerva P., Kristoferson J., Holst A. Random indexing of text samples for Latent Semantic Analysis // 22nd Annual Conference of the Cognitive Science Society. – U Pennsylvania: Erlbaum, 2000. – P. 1036/
8. Karlgren J. and Sahlgren M. From Words to Understanding // In Uesaka Y., Kanerva P., Asoh H. (eds.) Foundations of Real-World Intelligence. – Stanford: CSLI Publications, 2001. – P. 294 – 308.
9. Caid W., Dumais S., Gallant S. Learned vector-space models for document retrieval // Information Processing and Management. – 1995. – Vol. 31, 3. – P. 419 – 429.
10. Куссуль Э.М. Ассоциативные нейроподобные структуры. – Киев: Наукова думка, 1991. – 144 с.
11. Рачковский Д.А., Слипченко С.В., Куссуль Э.М., Байдык Т.Н. Процедура связывания для бинарного распределенного представления данных // Кибернетика и системный анализ. – 2005. – № 3. – С. 3 – 18.
12. Grossman D., Frieder O. Information Retrieval: Algorithms and Heuristics. – Boston: Kluwer, 1998. – 255 p.
13. Papadimitriou C.H., Raghavan P., Tamaki H., Vempala S. Latent Semantic Indexing: A Probabilistic Analysis // 17th ACM Symposium on the Principles of Database Systems. – ACM Press, 1998. – P. 159 – 168.
14. Kaski S. Dimensionality Reduction by Random Mapping: Fast Similarity Computation for Clustering // IJCNN'98, Int. Joint Conference on Neural Networks. – Piscataway, NJ: IEEE Service Center. – 1998. – Vol. 1. – P. 413 – 418.
15. Johnson W., Lindenstrauss J. Extensions of Lipshitz mapping into Hilbert space // Contemporary Mathematics. – 1984. – N 26. – P. 189 – 206.
16. Indyk P. Algorithmic Applications of Geometric Embeddings // FOCS-01. – IEEE. – 2001. – Vol. 1. – P. 10 – 35.
17. Charikar M. Similarity estimation techniques from rounding algorithms // ACM Symposium on Theory of Computing. – 2002. – P. 380 – 388.
18. Burgess C., Lund K. The dynamics of meaning in memory // In Dietrich E., Markman A.B. (eds.) Cognitive Dynamics. – Mahwah, NJ: Erlbaum. – 2000. – P. 117 – 156.
19. Kontostathis A., Pottenger W. A framework for understanding latent semantic indexing (LSI) performance // Information Processing and Management. – Preprint. – Elsevier Science. – 2004. – 24 p.
20. Broder A.Z. On the resemblance and containment of documents // Compression and Complexity of Sequences. – 1997. – Vol. 1. – P. 21.
21. TASA (Touchstone Applied Sciences Association) text collection. http://lsa.colorado.edu/spaces.html.
22. Oxford University Computing Services. The British National Corpus. http://www.natcorp.ox.ac.uk/.
23. Zhu H. Bayesian geometric theory of learning algorithms // Int. conf. on Neural Networks. – 1997. – Vol. 2. –P. 1041 – 1044.
24. Turney P.D. PMI-IR versus LSA on TOEFL // ECML 2001. 12th European Conf. on Machine Learning. – Freiburg, Germany. – 2001. – September 5–7. – 2167. – P. 491 – 502.
25. TOEFL (Test of English as a Foreign Language). http://www.ets.org/.
26. Tatsuki D. Basic 2000 Words – Synonym Match 1 // In: Interactive JavaScript Quizzes for ESL Students. http://www.aitech.ac.jp/~iteslj/quizzes/js/dt/mc-2000-01syn.html. – 1998.

27. Rubenstein H., Goodenough J. Contextual correlates of synonymy // Computational Linguistics. –1965. – N 8. – P. 627 – 633.
28. Miller G., Charles W. Contextual correlates of semantic similarity // Language and Cognitive Processes. – 1991. – Vol. 6. – 1. – P. 1 – 28.
29. Cornell University SMART text collections. ftp://ftp.cs.cornell.edu/pub/smart/.
30. Амосов Н., Байдык Т., Гольцев А., Касаткин А., Касаткина Л., Куссуль Э., Рачковский Д. Нейрокомпьютеры и интеллектуальные роботы. – Киев: Наукова думка, 1991. – 272 с.