УДК 004.912:519.7

МОДЕЛЬ И МЕТОД МНОГОАСПЕКТНОГО ПОИСКА ФАКТОГРАФИЧЕСКИХ ДАННЫХ ДЛЯ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ

В. А. Тертышный, И. В. Шевченко
Кременчугский национальный университет имени Михаила Остроградского
ул. Первомайская, 20, г. Кременчуг, 39600, Украина. E-mail: mainhousepost@hotmail.com

Построена формальная модель многоаспектного семантического пространства. Выявлены типы логических связей, которые могут быть использованы для многоаспектного фактографического поиска. Выбраны признаки и метрики для кластеризации контекстов. Сформирована специализированная логико-лингвистическая модель для выявления фактографических данных. Предложен метод многоаспектного анализа текстов, выявления релевантных контекстов и формирования фактографического отчета. При анализе запроса каждому аспекту присваивается весовой коэффициент. На этапе поиска релевантных документов используется нечеткий частотный анализ. На этапе фактографического поиска формируются матрицы связей термов в абзацах документов и матрицы частот пар термов по абзацам. Полученные матрицы сравниваются с эталонными матрицами по аспектам. Критерий сравнения – расстояние Кемени. На этапе определения степени принадлежности каждого абзаца к каждому аспекту используется пороговое преобразование. На последнем этапе осуществляется ранжирование аспектов и найденных абзацев по степени релевантности определенным аспектам.

Ключевые слова: фактографический поиск, логико-лингвистическая модель, многоаспектный анализ текстов.

МОДЕЛЬ І МЕТОД БАГАТОАСПЕКТНОГО ПОШУКУ ФАКТОГРАФІЧНИХ ДАНИХ ДЛЯ ПІДТРИМКИ ПРИЙНЯТТЯ РІШЕНЬ

В. А. Тертишний, І. В. Шевченко
Кременчуцький національний університет імені Михайла Остроградського
вул. Першотравнева, 20, м. Кременчук, 39600, Україна. E-mail: mainhousepost@hotmail.com

Побудовано формальну модель багатоаспектного семантичного простору. Виявлено типи логічних зв'язків, які можуть бути використані для багатоаспектного фактографічного пошуку. Обрано ознаки і метрики для кластеризації контекстів. Сформоване спеціалізовану логіко-лінгвістичну модель для виявлення фактографічних даних. Запропоновано метод багатоаспектного аналізу текстів, виявлення релевантних контекстів і формування фактографічного звіту. При аналізі запиту кожному аспекту присвоюється ваговий коефіцієнт. На етапі пошуку релевантних документів використовується нечіткий частотний аналіз. На етапі фактографічного пошуку формуються матриці зв'язків термів в абзацах документів і матриці частот пар термів по абзацах. Отримані матриці порівнюються з еталонними матрицями по аспектах. Критерій порівняння – відстань Кемені. На етапі визначення ступеня приналежності кожного абзацу до кожного аспекту використовується граничне перетворення. На останньому етапі здійснюється ранжування аспектів і знайдених абзаців за ступенем релевантності певним аспектам.

Ключові слова: фактографічний пошук, логіко-лінгвістична модель, багатоаспектний аналіз текстів.

АКТУАЛЬНОСТЬ РАБОТЫ. Компьютерная обработка информации, отображенной на естественном языке, является базовой проблемой в области построения интеллектуальных систем поддержки принятия решений на основе информационного поиска. Огромное количество информации, способной повлиять на принимаемые решения, представлено в виде текстов. Поэтому создание и развитие информационных технологий обработки текстовой информации является одной из важнейших задач на протяжении многих лет. Получено большое количество научных и практических результатов в осуществлении морфологического, синтаксического и частично семантического анализа [1], однако сложной задачей остается извлечение необходимой для принятия решений фактографической информации, особенно, в тех случаях, когда эта информация затрагивает несколько аспектов – различных сторон решения конкретной задачи. Поэтому современные средства автоматической обработки естественноязыковых данных не всегда удовлетворяют потребностям пользователя. Кроме того, полное математическое обеспечение процессов синтаксического и семантического анализа является сложным и его реализация в СППР, предназначенных для решения управленческих задач на предприятиях затруднена.

Таким образом, сегодня одной из актуальных задач является разработка методов и моделей фактографического многоаспектного анализа текстовых массивов с использованием статистических подходов и частотно-семантических методов [2].

Анализ литературных данных и постановка проблемы. Процесс обработки естественно-языковых текстов лежит в основе современных информационных технологий извлечения сведений из текстовых документов. При этом используются методы морфологического, синтаксического и частично семантического анализа. Тем не менее, извлечение нужных сведений, т.е. понимание содержания текста остается задачей, которая не решена в полном объеме [3–6].

Частотно-семантический анализ текстовых массивов является одним из перспективных направлений современных информационных технологий. Составляющими такого анализа является алгоритмы класси-

фикации и кластеризации текстовых документов. В этих алгоритмах используют векторную модель текстовых документов, основанную на представлении документов как векторов в некотором фазовом пространстве. Базис такого пространства часто образуют с помощью частотно-дистрибутивных характеристик лексем текстового словаря. Одна из проблем такого подхода обусловлена большой размерностью рассматриваемого векторного пространства. Считается, что такое пространство не позволяет выделить заданные семантические составляющие в интеллектуальном анализе текстов, несмотря на развитую теорию представления знаний. Проблеме представления знаний посвящено много литературных источников. Начиная с истоков проблемы формализации естественного языка [7, 8] и до настоящего этапа обработки текстовой информации [9–11], известны пять формальных моделей представления знаний: продукционные, фреймовые, логические, семантические сети и онтологии. Однако ни одна из этих моделей не может в полной мере отразить содержание текста на естественном языке. Методы интеллектуального анализа текстовой информации Data Mining помогают решить многие задачи, с которыми сталкивается аналитик, в частности, разработаны и усовершенствованы методы классификации и кластеризации описаны многими учеными [12, 13]. Именно поэтому в настоящее время в большинстве случаев используют для обработки текстов статистические методы без привлечения процедур синтаксического анализа. Но проблемой остается эффективное извлечение фактографических данных, – прецедентов и их параметров, в случае многоаспектного поиска.

Целью работы является упрощение процесса многоаспектного поиска фактографических данных путем модификации метода латентно-семантического анализа и соответствующей логико-лингвистической модели.

МАТЕРИАЛ И РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЙ. В продолжение исследований [14, 15] и в соответствии с поставленной целью необходимо решить следующие задачи:

– построить формальную модель многоаспектного семантического пространства;

– выявить возможные типы логических связей, которые могут быть использованы для многоаспектного фактографического поиска;

– выбрать признаки и метрики для кластеризации контекстов;

– сформировать специализированную логико-лингвистическую модель (ЛЛМ) для выявления фактографических данных;

– сформировать последовательность этапов многоаспектного анализа, выявления релевантных контекстов и формирования фактографического отчета для пользователя.

Формирование семантического пространства. Имеется множество релевантных документов, специально отобранных для данной онтологии и данного аспекта. Эти документы, в частности, содержат ключевые слова, обозначающие сущности ПрО (главный тезаурус) и лексемы, связывающие сущности с сущностями и сущности с признаками, действиями и параметрами (тезаурусы аспектов решаемых задач). Необходимо сформировать онтологию ПрО и соответствующее семантическое пространство.

Представим онтологию ПрО с учетом наличия аспектов в виде

$$O =< E(AT), ER, F, AS, AR >, \quad (1)$$

где E – набор сущностей; AT – множество атрибутов сущностей; ER – множество отношений сущностей; $F:E\times ER$– функции интерпретации отношений и сущностей; AS – множество аспектов, определяющих подмножества сущностей и связей; AR – пересечение аспектов.

Тогда многоаспектное семантическое пространство:

$$S =< O, M >, \quad (2)$$

где M – набор метрик для вычисления степени близости сущностей и релевантности результатов поиска. Пространство SS условно разделено на пересекающиеся подпространства, каждое из которых соответствует одному аспекту. На рис. 1 представлена схема связей между ПрО, аспектами и семантическими пространствами.

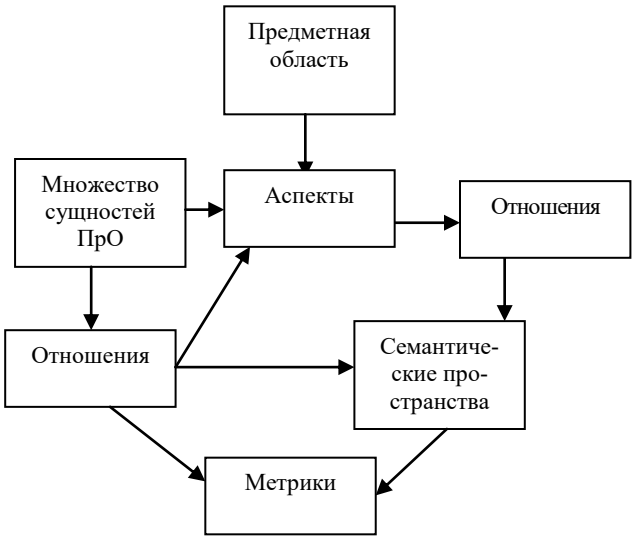

Рисунок 1 – Взаимосвязи между сущностями, аспектами и семантическими пространствами

Логические связи для многоаспектного фактографического поиска. Пусть T – входной текст, в котором имеются интерпретированные словоформы $X = \{x_1,...,x_m\}$ имеющие между собой скрытые отношения $R = \{r_1, r_2, ..., r_n\}$ Проблема извлечения фактографических данных (знаний) состоит в поиске интерпретации $\varphi: X \to TI$, где TI – суперпозиция отношений R, выраженная на ЕЯ.

Особенность предлагаемой онтологии и семантического пространства – учет того факта, что между сущностями имеется несколько ассоциаций по нескольким аспектам.

Введем следующие допущения:

а) семантическое пространство дискретно;

б) набор элементов пространства конечен и обозрим;

в) число комбинаций ключевых слов в контекстах велико, но конечно;

г) каждый смысловой аспект можно представить кластером, центр которого определен на множестве релевантных документов.

Многоаспектность задачи предполагает, что одна и та же сущность находится в разных отношениях с другими сущностями. Если две сущности связаны несколькими отношениями в разных аспектах, тогда для каждого аспекта следует сформировать отдельную матрицу связей, построить свой тезаурус и своё пространство признаков.

В одной задаче могут быть важны одни аспекты, в другой – другие. Тогда между двумя узлами семантической сети есть несколько дуг, каждая из которых относится к одному аспекту. Каждой подсети соответствует своя матрица связей. Это дает возможность вычислять силу связей между сущностями с учётом аспектов, которые важны в данной задаче.

В частотно-семантическом методе в качестве критерия оценки «силы связи» обычно используются две характеристики: общность элементов и частота встречаемости.

Выбор признаков и метрики для кластеризации. Используем два пространства признаков. В первом пространстве признаков каждой оси координат соответствует относительная частота встречаемости ключевых слов, находящихся в тезаурусе ПрО. Во втором пространстве признаков каждой оси координат соответствует относительная частота встречаемости словосочетаний, находящихся в тезаурусах аспектов. Кластеры образуют документы, наиболее близкие по своим матрицам частот к эталонной матрице конкретного аспекта.

Учитывая, что в контексте решаемой задачи значимыми элементами языка являются документ, абзац (контексты) и слово, обозначим:

$N_A, f_A, = N_A/N$ – количество и частота контекстов, где встретилось только слово А;

$N_B, f_B, = N_B/N$ – количество и частота контекстов, где встретилось только слово В;

$N_{AB}, f_{AB}, = N_{AB}/N$ – количество и частота контекстов, в которых наблюдалась совместная встречаемость слов А и В; N – общее количество контекстов.

В качестве основной метрики используем расстояние Кемени [16]. Как известно, бинарное отношение A на конечном множестве $Q = \{q_1, q_2, ..., q_k\}$ – это подмножество *декартова квадрата* $Q^2 = \{(q_m, q_n), m, n = 1,2,...,k\}$. При этом пара (q_m, q_n) входит в A тогда и только тогда, когда между q_m и q_n имеется рассматриваемое отношение. Каждую кластеризованную ранжировку, как и любое бинарное отношение, можно задать квадратной матрицей $\|x(a,b)\|$ из 0 и 1 порядка $k \times k$. При этом $x(a\ b) = 1$ тогда и только тогда, когда $a < b$ либо $a = b$. В первом случае $x(b\ a) = 0$, а во втором $x(b\ a) = 1$. При этом хотя бы одно из чисел $x(a\ b)$ и $x(b,a)$ равно 1.

Расстоянием Кемени между бинарными отношениями A и B, описываемыми матрицами $\|a(i,j)\|$ и $\|b(i,j)\|$ соответственно, называется величина

$$KD(A,B) = \sum |a(i,j) - b(i,j)|, \qquad (3)$$

где суммирование производится по всем i,j от 1 *до* k, т.е. расстояние Кемени между бинарными отношениями равно сумме модулей разностей элементов, стоящих на одних и тех же местах в соответствующих им матрицах.

При кластеризации с использованием расстояния Кемени центр кластера удобно определять как медиану Кемени – $Arg\ min \sum D(A_i, A)$ [16].

Логико-лингвистическая модель для выявления фактографических данных. Формальной моделью представления знаний, учитывающей содержание предложений естественного языка, является логико-лингвистическая модель [2]. Поэтому, если построить формальную содержательную модель текста любой тематики и структуры, то можно будет анализировать электронные текстовые документы по содержанию, извлекать из них знания, сравнивать их.

Логико-лингвистическая модель текстового документа – это теоретико-множественная модель, которая включает основные свойства текста и его составных частей и отражает основные взаимосвязи между интересующими исследователя структурными компонентами текста.

С учетом перечисленных выше логических связей ЛЛМ текста для многоаспектного поиска выглядит следующим образом:

$$T = < CT, AS, DB, A, RA > , \qquad (4)$$

где CT – маркер принадлежности текста к определенному классу, маркер присваивается тексту на предварительном этапе поиска релевантных текстов; AS – множество аспектов в тексте; DB – база данных, представляющая собой набор тезаурусов, хранящих ключевые слова текста, лексемы и предложения; A – множество абзацев текста; RA – пересечение множества абзацев с множеством аспектов.

Таким образом, выражение (4) представляет усовершенствованную логико-лингвистическую модель, содержащую множество аспектов и проекцию аспектов на множество абзацев исследуемого текста, что позволяет реализовать многоаспектный фактографический поиск.

Метод многоаспектного анализа для фактографического поиска. Метод включает в себя несколько этапов, каждый из которых, в свою очередь, содержит определенные подэтапы.

На предварительном этапе происходит формирование онтологии (1) и семантического пространства (2). Создаются таблицы отношений между аспектами. Каждая запись таблицы соответствует терму из тезауруса, а каждое поле – аспекту. Элементы таблицы – бинарные, сигнализирующие о наличии или отсутствии данного терма в данном аспекте.

Формируются эталонные матрицы частотных связей аспекта на основе анализа стопроцентно релевантных данному аспекту текстов. Эти тексты формируют или подбирают эксперты.

Этап 1. Обработка запроса.

1.1. Из запроса выбираются ключевые слова и лексемы.

1.2. Определяются ведущие аспекты запроса путем анализа таблицы отношений аспектов.

1.3. Уточнение у пользователя аспектов и их важности (пользователь может сам ранжировать аспекты и /или добавлять новые). Возврат к п.1.2. Если пользователь согласен, то переход к п. 1.4.

1.4. Группировка ключевых слов и лексем по аспектам. Генерация запросов по аспектам.

При анализе запроса каждому аспекту присваивается весовой коэффициент

$$\alpha_j = \sum_{i=1}^{n} a_{ij}, \qquad (5)$$

где a_{ij} – маркер участия i-го терма в j-м аспекте; n – количество термов в тезаурусе.

Этот коэффициент будет использован при формировании результатов фактографического поиска.

Этап 2. Поиск и группировка релевантных документов.

Вначале из исходного множества отбираются документы, для которых согласно векторному методу степень релевантности ПрО превышает заданный порог. Это классическая процедура, используемая во всех поисковых системах.

Далее производится сопоставление вектора ключевых слов запроса q_k и вектора ключевых слов документа x_k для первичного отбора документов в рабочие подмножества аспектов A_j.

$$r_j = \sum_{k=1}^{L_q} x_k q_k , \qquad (6)$$

где L_q – мощность множества термов в запросе с учетом расширения запроса по аспектам; k – номер терма.

Для каждого документа из рабочего подмножества рассчитывается нормированная частота совместной встречаемости ключевых слов аспекта и документа:

$$r_j = \frac{\sum_{k=1}^{t} x_k^j q_k}{\sum_{k=1}^{t} (x_k^j)^2 + \sum_{k=1}^{t} (q_k)^2 - \sum_{k=1}^{t} x_{k\,k}^{j} q_k} . \qquad (7)$$

Полученные значения частот подвергаются фаззификации с нечеткими оценками «низкая», «средняя», «высокая». Частоты представляются в нечеткой интерпретации и степени истинности указанных термов сравниваются со степенями истинности этих термов в документах-образцах из библиотеки СППР. Результат – сформированные подмножества релевантных документов по аспектам.

Этап 3. Формирование рабочих матриц частотных связей под конкретный документ. На данном этапе из исходной эталонной матрицы связей аспекта удаляются строки и столбцы тех терминов, которые не обнаружены в документе.

Этап 4. Обработка каждого найденного документа для фактографического поиска.

4.1. Предложения и абзацы в документе индексируются. Пары термов индексируются по порядку номеров.

4.2. Для i-той каждой пары термов, зафиксированной в тезаурусах ПрО по аспектам, рассчитываются относительные частоты fs_i появления этих пар в предложениях (sentenses) документа

$$fs_i = \frac{n(t_1, t_2)}{n(t_1) + n(t_2)} . \qquad (8)$$

Результат:

– квадратная матрица MLS_j связей термов в предложениях документа t по каждому аспекту с индексом j;

– таблица (матрица) индексов «предложение–частота встречаемости пары термов».

4.3. Рассчитываются относительные частоты fp_i появления пар термов в абзацах (paragraphs) документа для каждой пары термов, зафиксированной заранее в тезаурусах ПрО по аспектам. Результат:

– квадратная матрица связей MLP_j термов в абзацах документа;

– таблица (матрица) индексов «абзац–частота встречаемости пары термов».

4.4. Полученные рабочие матрицы MLS_j и MLP_j сравниваются с эталонными матрицами по аспектам. Для каждой рабочей матрицы вычисляется расстояние Кемени (KD) по формуле (3) относительно эталонной матрицы аспекта. С учетом того, размер матриц может быть разным, для последующего отбора лексем и абзацев следует вычислять нормированное расстояние

$$F = n^2 / KD , \qquad (9)$$

где n – фактический размер рабочей матрицы.

Используя пороговое преобразование, где T – заданный порог,

$$H = \begin{cases} 1 \; if \;\; F \ge T \\ 0 \;\; \text{otherwise} \end{cases} , \qquad (10)$$

определяем степень принадлежности каждого абзаца к каждому аспекту. Абзац (прецедент) может относиться к нескольким аспектам, поэтому результат фактографического поиска может оказаться избыточным. Но в данном случае избыточность предпочти-

тельнее потери информации при рассмотрении реальной задачи в разных аспектах.

Этап 5. Ранжировка аспектов по весовым коэффициентам, ранжировка найденных абзацев (прецедентов), фраз и значений параметров по степени релевантности определенным аспектам и выдача пользователю подмножества прецедентов, фраз и параметров по каждому аспекту.

В качестве примера практического использования покажем фрагментарно решение задачи «Поиск новых покупателей (каналов сбыта) для предприятия сельскохозяйственного машиностроения». В табл. 1 показаны ключевые слова и лексемы для аспектов данной задачи. Согласно предлагаемому методу после введения запроса и извлечения из него ключевых слов и лексем определяются ключевые аспекты, и система пытается определить задачу. Пользователю предлагается уточнить тип задачи. После уточнения типа задачи происходит формирование вторичных запросов, поиск в глобальной сети и отбор релевантных документов. Для каждого отобранного документа формируются рабочие эталонные матрицы частотных связей (этап 3 метода). Далее происходит формирование рабочих матриц документов, и рассчитываются частоты совместной встречаемости терминов и лексем в документах, абзацах и предложениях. Фрагмент матрицы частотных связей документа и аспекта «Функционал» показан в табл. 2.

Матрицы «абзац–частота встречаемости пары термов» и «предложение–частота встречаемости пары термов» не могут быть показаны из-за ограниченного объема статьи. Выбранные из документов фрагменты текста группируются согласно вычисленному значению расстояния (9).

Таблица 1 – Ключевые слова и лексемы для аспектов задачи «Поиск новых покупателей (каналов сбыта)»

Поиск новых покупателей (каналов сбыта)			
Аспекты			
Покупка–продажа	Продукция	Клиенты–покупатели–поставщики	Функционал
Ассортимент	Ассортимент	Адрес	Высокосортная мука
Дилер	Жатка	Ассортимент	Оборудование
Дистрибютор	Жатка зерновая	Валюта	Отбор отрубей
Заказ	Жатка навесная	Выставка	Крупа
Импорт	Жатка валковая	Госзаказ	Мука
Клиент	Жатка валковая прицепная	Дилер	Переработка
Куплю	Жатка зерновая навесная	Дистрибютор	Просеивание муки
Объем продаж	Жатка валковая прицепная	E-mail	Производство
Партнерство	Мельница	Форум	Помол зерновых
Покупка	Мельница “Фермер”	Объем	Отопление
Покупатель	Мельница агрегатная	Телефоны	Точного высева
Потребитель	Мельница малогабаритная		Оптовая торговля
Продажа	Навесное оборудование		Розничная торговля
Покупка	Сеялка		Сельхозпродукты
Сбыт	Тракторы		
Сервис	Котлы отопительные		
Скидки	Круполиния		
Тендер			
Цена			
Экспорт			

Таблица 2 – Матрица частотных связей документа и аспекта

Наименование документа и аспекта	Оборудование	Крупа	Мука	Переработка	Производство	Помол зерновых	Оптовая торговля
Оборудование	–	0,34	0,76	0,96	0,85	0,73	0,0
Крупа		–	0,54	0,43	0,32	0,62	0.03
Мука			–	0,21	0,57	0,78	0,11
Переработка				–	0,23	0,15	0,12
Производство					–	0,16	0,13
Помол зерновых						–	0,11
Оптовая торг-ля							–

Решение указанной задачи с использованием СППР, реализующей предлагаемый метод фактографического поиска, позволило отделу сбыта и маркетинга выявить потенциальных клиентов и увеличить прибыли предприятия.

ВЫВОДЫ. Для упрощения процесса многоаспектного поиска фактографических данных с применением концепции латентно-семантического анализа построена формальная модель многоаспектного семантического пространства, выявлены типы логических связей, которые могут быть использованы для многоаспектного фактографического поиска. Выбраны признаки и метрики для кластеризации контекстов.

Сформирована специализированная логико-лингвистическая модель для выявления фактографических данных. Предложен метод многоаспектного анализа текстов, выявления релевантных контекстов и формирования фактографического отчета для пользователя.

ЛИТЕРАТУРА

1. Башмаков А.И., Башмаков И.А. Интеллектуальные информационные технологии: учебное пособие. – М.: МГТУ им. Баумана, 2005. – 304 с.

2. Вавіленкова А.І., Ланде Д.В., Литивненко О.Є. Теоретичні основи аналізу електронних текстів: монографія. – К.: НАУ, 2015. – 258 с.

3. Foltz W., Gilliam S., Kendall S. Supporting content based feedback in online writing evaluation with LSA // Interactive Learning Environments. – 2000. – No. 8. – PP. 111–128.

4. Gries S.Th. Corpus-based methods and cognitive semantics: the many meanings of to run // Corpora in cognitive linguistics: corpus-based approaches to syntax and lexis, 2006. – PP. 57–99.

5. Evans V. Lexical concepts, cognitive models and meaning-construction // Journal of Cognitive semiotics. – 2006. – PP. 73–107.

6. Кобозева И.М. Лингвистическая семантика. – М.: Эдиториал УРСС, 2000. – 352 с.

7. Поспелов Д.А. Логико-лингвистические модели в системах управления. – М.: Энергоиздат, 1981. – 232 с.

8. Осуга С., Саэки Ю. Приобретение знаний. – М.: Мир, 1990. – 304 с.

9. Джарратано Д. Экспертные системы: принципы разработки и программирование – 4-е изд. / Пер. с англ. – М.: ООО «Вильямс», 2007. – 1152 с.

10. Корпусна лінгвістика / В.А. Широков, О.В. Булгаков, Т.О. Грязнухіна та ін. – К.: Довіра, 2005. – 471 с.

11. Design of a conceptual level programming environment based on task ontology / S. Kazuhisa, I. Mitsuru, K. Osamu, M. Riichiro // Proc. of Successes and failures of knowledge based systems in real world applications, 1996. – PP. 11–22.

12. Вавиленкова А.И. Извлечение смысла из предложений естественного языка // Программные продукты и системы. – 2012. – № 4 (100). – С. 87–90.

13. Палагин А.В., Крывый С.Л., Петренко Н.Г. Знание-ориентированные информационные системы с обработкой естественно-языковых объектов: основы методологии и архитектурно-структурная организация // Управляющие системы и машины. – 2009. – № 3. – С. 42–55.

14. Артамонов В.В., Тертышный В.А. Разработка модели информационного поиска с использованием связанных данных // Системи обробки інформації. – 2015. – № 10. – С. 69–75.

15. Тертышный В.А. Модель специализированной системы поиска сущностей на основе связанных данных // Вісник Кременчуцький національний університет імені Михайла Остроградського. – 2014. – Вип. 5/2014 (88). – С. 112–117.

16. Тоценко В.Г. Методы и системы поддержки принятия решений. Алгоритмический аспект. – К.: Наукова думка, 2002. – 381 с.

MODEL AND METHOD FOR MULTI-ASPECT SEARCH OF FACTOGRAPHIC FACTS FOR SUPPORT OF DECISION-MAKING

V. Tertishniy, I. Shevchenko

Kremenchuk Mykhailo Ostrohradskyi National University

vul. Pervomayskaya, 20, Kremenchuk, 39600, Ukraine. E-mail: mainhousepost@hotmail.com

Purpose. Formal model of aspect semantic space was built. Types of logical connections, which can be used for multi-aspect search, were identified. Features and metrics for clustering contexts were chosen. Specialized logical-linguistic model was formed for indication of factographic data. **Results.** Method of multi-aspect analysis indication of relevant contexts and formation of factographic reports was offered. During the analysis of request every aspect has got the weight coefficient. At the stage of relevant documents search an unclear frequent analysis is used. At the stage of factographic search matrixes of connections of terms in the documents and matrixes of terms by paragraphs have been built. **Practical value.** The Kemeny method has been applied in order to compare the aspects of these matrixes with those of the etalon matrixes. At the stage of defining rate of every paragraph belonging to each aspect threshold conversion is used. At the last stage the ranging of aspects and paragraphs which have been identified is conducted; the rate of their relevance to definite aspects is taken into account. References 16, tabl. 2, figure 1.

Key words: factographic search, logical-linguistic model; multi-aspect texts analysis.

REFERENCES

1. Bashmakov, A.I., Bashmakov, I.A. (2005), *Intellektualnyie informatsionnyie tehnologii: uchebnoe posobie* [Intelligent information technology: a training manual], MGTU im. Baumana, Moscow, Russia.

2. Vavilenkova, A.I., Lande, D.V., Litvinenko, O.E. (2015), *Teoretichni osnovi analizu elektronnih tekstiv: monografiya* [Basics of E-Teoretichni analizu tekstiv: monograph], NAU, Kyev, Ukraine.

3. Foltz, W., Gilliam, S., Kendall, S. (2000). "Supporting content based feedback in online writing evaluation with LSA", *Interactive Learning Environments*, no. 8, pp. 111–128.

4. Gries, S.Th. (2006). "Corpus-based methods and cognitive semantics: the many meanings of to run", *Corpora in cognitive linguistics: corpus-based approaches to syntax and lexis*, pp. 57–99.

5. Evans, V. (2006). "Lexical concepts, cognitive models and meaning-construction", *Journal of Cognitive semiotics*, pp. 73–107.

6. Kobozeva, I.M. (2000), *Lingvisticheskaya semantika* [Linguistic semantics], Editoreal URSS, Moscow, Russia.

7. Pospelov, D.A. (1981), *Logiko-lingvisticheskie modeli v sistemah upravleniya* [Logico-linguistic models in control systems], Energoizdat, Moscow, USSR.

8. Osuga, S., Saeki, Yu. (1990), Priobretenie znaniy [Acquisition of knowledge] Mir, Moscow, USSR.

9. Dzharratano, D. (2007), *Ekspertnyie sistemyi: printsipyi razrabotki i programmirovanie* [Expert Systems: Principles of design and programming], LLC «Vilyams», Moscow, Russia.

10. Shirokov, V.A., Bulgakov, O.V., Gryaznukhina, T.O. et al. *Korpusna lingvistika* [Corpus linguistics], Dovira, Kiev, Ukraine.

11. Kazuhisa, S., Mitsuru, I., Osamu, K., Riichiro, M. (1996). "Design of a conceptual level programming environment based on task ontology", *Proc. of Successes and failures of knowledge based systems in real world applications*, pp. 11–22.

12. Vavilenkova, A.I. (2012), "Removing the meaning of natural language sentences", *Software products and systems*, no. 4 (100), pp. 87–90.

13. Palagin, A.V., Kryivyiy, S.L., Petrenko, N.G. (2009), "Knowledge-based information systems with the processing of natural language include the basics of methodology and architectural and structural organizationtion", *Control systems and machines*, no. 3, pp. 42–55.

14. Artamonov, V.V., Tertyishnyiy, V.A. (2015), "Developing a model of information retrieval using linked data", *System on-timid Inḟormácie*, no. 10, pp. 69–75.

15. Tertyishnyiy, V.A. (2014), "Model entities specialized search engine on the basis of related data", *Transaction of Kremenchuk Mykhailo Ostrohradskyi National University*, iss. 5, no. 88, pp. 112–117.

16. Totsenko, V.G. (2002), *Metodyi i sistemyi podderzhki prinyatiya resheniy. Algoritmicheskiy aspekt* [Methods and decision support systems. Algorithmic aspect], Naukova dumka, Kiev, Ukraine.

Стаття надійшла 24.10.2016.