УДК 004.91

ПРИКЛАДНЫЕ ЗАДАЧИ СЕМАНТИЧЕСКОГО АНАЛИЗА ТЕКСТОВЫХ ДОКУМЕНТОВ

Надеждин Е.Н.

Государственный научно-исследовательский институт информационных технологий и телекоммуникаций, Москва, e-mail: en-hope@yandex.ru

Настоящая статья посвящена вопросам исследования семантики текстовых документов в процессе релевантного поиска информации в интернет. Рассмотрена общая задача автоматизированного изучения смыслового содержания текстовых документов, предоставляемых поисковыми системами Internet по запросу пользователя. На содержательном уровне описаны модели актуальных прикладных задач, которые связаны с изучением семантики текстов. Показаны специфические особенности формальной постановки таких задач и трудности выбора метода их решения. С опорой на современные достижения в области искусственного интеллекта и в компьютерной лингвистике сформулированы общие принципы семантического анализа текста, представленного на естественном языке. Предложена концептуальная схема семантического анализа текста, в основе которой лежит механизм идентификации каркасной модели с выделением отношений между ключевыми словами. Предложена классификация прикладных задач семантического анализа текстовых документов.

Ключевые слова: текстовый документ, семантика текста, задача семантического анализа, концептуальная схема семантического анализа

APPLIED PROBLEMS OF THE SEMANTIC ANALYSIS OF TEXT DOCUMENTS

Nadezhdin E.N.

State Institute of Information Technologies and Telecommunications, Moscow, e-mail: en-hope@yandex.ru

The present article is devoted to questions of a research of semantics of text documents in the course of relevant information search in Internet. Discusses the General problem of computer-aided study of the semantic content of text documents, search engines provide Internet upon user request. On the content level model describes the actual application tasks that are associated with the study of the semantics of texts. The specific features of formal performances such problems and difficulties of choosing the method of their solution. Based on modern achievements in the field of artificial intelligence and computational linguistics formulated the General principles of semantic analysis of the text presented in natural language. The proposed conceptual framework of semantic analysis, which is based on the identification mechanism wireframe model highlighting the relations between the keywords. The classification of applied problems of semantic analysis of text documents.

Keywords: text document, the semantics of the text, the problem of semantic analysis, conceptual diagram of semantic analysis

Стремительное развитие интернета и экспоненциальный рост количества актуальной информации в различных областях знаний привели к усложнению задач информационного поиска. Значительная часть полезных для пользователя знаний содержится в документальных базах знаний, которые являются базами текстовых документов. По данным статистики число таких документов приближается к уровню, критическому для традиционного (ручного) способа аналитического анализа [3, 5]. С учётом тенденции глобализации научно-образовательных процессов сегодня актуальными становятся вопросы опережающей разработки технологий и специальных инструментальных средств, поддерживающих процедуры автоматизированного релевантного поиска и извлечения знаний из текстовых документов.

Целью статьи является анализ содержательных моделей избранных задач извлечения элементов знаний из текстовых документов (ТД) и выявление особенностей их формализованного решения на основе единого концептуального подхода к семантическому анализу.

Центральным понятием лингвистики, как известно, является термин «текст». Понимание текста как «целостного речемыслительного процесса произведения» сложилось в русском языкознании в результате многолетней дискуссии и нашло отражение в работах известных учёных Н.Ф. Алефиренко, О.С. Ахмановой, Н.С. Болотновой, И.Р. Гальперина, Г.В. Колшанского, З.Я. Тураевой, В.В. Одинцова и др. В нашей работе примем за основу следующее обобщённое определение текста: «*Текст – это целостное коммуникативное образование, компоненты которого объединены в единую иерархически организованную семантическую структуру коммуникативной интенцией (замыслом) его автора*» [1, с. 303].

Сегодня лингвисты уделяют всё больше внимания коммуникативному и когнитивному аспектам изучения текста, неизбеж-

ности включения его в какой-либо (исторически реальный или условный) контекст. Например, З.Я. Тураева обоснованно отмечает, что «*текст не только отражает действительность, но и сообщает о ней… в тексте пересекаются коммуникативная, когнитивная (познавательная) и эмотивная функции*» [11].

В когнитивном аспекте текст предстаёт как «опредмеченное знание» или как «вербально кодированный фрагмент знаний, являющийся органической частью целостной системы знаний о мире», как «особым лингвистическим образом представленное знание». В качестве объекта научного анализа текст может параллельно рассматриваться с коммуникативных и с когнитивных (как трансформированное знание) позиций [2]. Для извлечения знаний принципиальное значение имеет целостный подход к изучению текста, и для этого следует рассматривать не только связи внутри сверхфразового единства между отдельными предложениями, но и связь самих этих единств в рамках текста (И.Р. Гальперин, 1981).

Текстовые документы и, соответственно, специальные знания, содержащиеся в них, обычно являются слабо структурированными. Это обстоятельство вызывает существенные трудности при формальном описании и автоматизации процесса извлечения и обработки знаний. С другой стороны, профессионально подготовленный текст является результатом напряжённого интеллектуального труда автора, и для извлечения знаний необходим интеллектуальный инструментарий, адекватный содержанию, профилю и уровню сложности текста. Поэтому проблема аналитического анализа семантики и извлечения знаний из текстовых документов является достаточно сложной и требует разработки и применения специализированных интеллектуальных информационных систем (ИИС) [2, 3]. На практике путём введения ряда допущений можно осуществить декомпозицию указанной проблемы семантического анализа ТД и выделить несколько групп относительно независимых типовых задач.

К типовым задачам исследования семантики одиночных ТД следует отнести:

а) выделение в тексте ключевых слов;

б) выделение в тексте цепочки ключевых слов;

в) выявление контекстных слов (контекста);

г) аннотирование текста.

Для случая обработки набора ТД приоритетными являются следующие задачи:

а) классификация текстов по совокупности признаков;

б) каталогизация текстов;

в) поиск заданных информационных фрагментов в потоке ТД;

г) поиск (подбор) текстов, обладающих заданными свойствами.

Рассмотрим содержательные модели некоторых прикладных задач, связанных с извлечением элементов знаний из текстовых документов.

1. Задача классификации текстов по нескольким признакам [14]. В общем случае указанную задачу можно сформулировать следующим образом.

Пусть имеется исходное множество ТД – объектов X, заданных своими признаковыми описаниями. Для отображения объектов используются n признаков, т.е. каждый объект исходного множества X представлен как n-мерный вектор в признаковом пространстве $P_n : X \subseteq P_n$. Дано множество классов Y ($Y \neq 0$), заданных символьными метками: $Y = \{-1; 1\}$. Допустим, что распределение объектов исходного множества X по классам Y априорно известно только на обучающей выборке $X^* \subseteq X$, т.е. для каждого объекта обучающей выборки $x_i^* \in X$ известен класс $y_i^* \subseteq Y$, к которому этот объект относится. Требуется найти отображение $f : X \to Y$ исходного множества объектов X во множество классов Y, если значения этого отображения известны только для объектов обучающей выборки X^*: $\{(x_1^*, y_1^*), ..., (x_N^*, y_N^*)\}$. Здесь принято: $N = |X^*|$.

Формально описанная задача имеет несколько приложений. Одним из них является предварительный отбор ТД из некоторого исходного множества для последующего детального семантического анализа по критериям соответствия конкретной предметной области (ПрО). Другим примером задачи 1 может служить бинарная классификация текстовых документов с использованием формальной модели тезауруса заданной ПрО.

2. Задача выделения цепочки ключевых слов [5, 8]. Предположим, что в ходе семантического анализа исходного ТД выделены и отобраны ключевые понятия (слова, термы) Z_j, $j = \overline{1, r}$. Для автоматизации процесса формирования сети допустимых переходов между ключевыми словами в заданной предметной области воспользуемся апробированной технологией автоматического построения семантической сети текста на основе корпуса текстов конкретной ПрО [12, с. 370]. Основным инструментом формализованного представления семантики предметной области выступает семантическая сеть предметной области (текста). Далее с учётом выявленных особенностей и дополнительной информации о тематике и структуре текста могут быть осуществле-

ны типизация ключевых слов и отбор их допустимых парных сочетаний $(Z_i - Z_j)$, $i, j = \overline{1, r}, j \neq i$. Для отображения парных сочетаний ключевых слов вводят матрицу инциденций $C = \{c_{i,j}, i, j = \overline{1,n}\}$, элементы которой определяются по правилу:

$$c_{i,j} = \begin{cases} 1, \text{если элементы } (Z_i, Z_j) \in D_k; \\ 0, \text{если } j = i; \\ 0, \text{если элементы } (Z_i, Z_j) \notin D_k. \end{cases}$$

Здесь $D_k \in D$ – подмножество парных сочетаний ключевых слов.

Обязательным этапом автоматического формирования сети переходов является обоснование механизма выявления важных для описания ПрО ключевых фраз и предложений. Поэтому на основании имеющихся данных сформулируем задачу синтеза семантической модели знаний, которая заключается в построении обобщённого функционального графа $G(P, I)$ по моделям P_v, $v = \overline{1,m}$, выявленных цепочек ключевых слов.

Пусть семантическая модель предметной области отображается неориентированным графом $G^*(P, I)$, где $P = \{P_v, v = \overline{1,m}\}$ – множество вершин укрупнённого графа G^*, представляющих выделенные пары слов, а $I = \{I_w, w = \overline{1,r}\}$ – множество дуг этого графа, представляющих информационные связи между концептами P_v, $v = \overline{1,m}$. **Требуется** построить граф $G^0(P, I)$, который обладает следующим свойством $G^0 = \bigcup_j G_j$ для определённого набора условий, отражающих требования представительности сетевой модели ПрО.

Для решения задачи 2 можно воспользоваться методом, изложенным в авторской статье [8]. После выбора ключевых слов и значимых предложений текста на основе этих предложений строится ассоциативная сеть, которая является основой для идентификации сети переходов между ключевыми словами для этой группы предложений. Сеть, построенная на всех предложениях корпуса текстов, описывающего предметную область с рангом выше порогового, является сетью переходов между ключевыми словами для всей ПрО.

3. Задача выявления контекста [14]. В качестве признакового описания документов, отражающего их тематику, используется набор содержащихся в них слов – *термов*, каждому из которых по определенным правилам присвоен числовой коэффициент – *вес*. При вычислении весов термов учитывается их частота встречаемости в тексте документа. Порядок термов, как правило, не учитывается. Наиболее распространенный общий подход к вычислению веса терма реализует формула *TF·IDF* (*TF – term frequency*, *IDF – inversed document frequency*), где *TF* – частота встречаемости терма в данном документе, *IDF* – величина, обратная частоте встречаемости терма в остальных документах. В размеченных текстах может также учитываться наличие терма в заголовке, выделение терма цветом и т.п. Затем проводится нормализация по ТД так, чтобы сумма квадратов всех весов была равна единице.

Количество слов, выделенных из ТД, обычно велико. Поэтому применяют различные способы уменьшения размерности пространства признаков [4]. Как неинформативные исключаются из рассмотрения слова с наибольшими и наименьшими частотами встречаемости. Все словоформы и некоторые однокоренные слова заменяются одним словом. С этой же целью используется словарь синонимов. В общем случае терм представляет собой не слово (термин), а класс слов, объединенных по общему признаку (корню, значению). Описанный способ извлечения информации из ТД широко используется при решении различных задач, требующих автоматической смысловой обработки текстов [12]. На сегодняшний день для множества предметных областей экспертами разработаны тематические словари, большинство которых состоит не из термов, а из их сочетаний, устойчивых для данной предметной области. При классификации текстов по тематике учитывается эта особенность. При этом обычно устойчивые группы слов рассматриваются как самостоятельные термы.

На практике широко применяется анализ контекста терма без привлечения экспертов.

Анализ контекста термов с последующей заменой элементарных термов характерными группами требует значительных временных затрат, которые можно сократить, если анализировать контекст только наиболее весомых термов и использовать результаты анализа для пересчета весов, отказавшись от «укрупнения» термов.

Пусть тексту документа d сопоставлен набор термов с их ненормированными весами $f(d) = ((t_1, w_1), (t_2, w_2), \ldots, (t_N, w_N))$, упорядоченный по убыванию весов. Выберем $K \leq N$ термов, имеющих наибольшие веса, и определим их новые веса с учетом контекста.

Если для каждой категории C_i экспертом сформирован список устойчивых сочетаний термов $S_i = (s_1, s_2, \ldots, s_{|S_i|})$, причем

каждому сочетанию присвоен коэффициент значимости по шкале $\phi(s_j) \in (0,1]$, то можно определить новые веса термов w_k' с учетом их вхождения в устойчивые сочетания из соответствующих списков.

Положим

$$w_k' = w_k + \sum_{j=1}^{|S_i|} \delta(t_k, s_j) \cdot \phi(s_j),$$

где

$$\delta(t_k, s_j) = \begin{cases} \gamma_j, \text{ если } t_k \text{ входит в } s_j \\ 0, \quad \text{иначе} \end{cases},$$

здесь γ_j – количество вхождений s_j в текст документа d, $k = 1,2,\ldots,K$.

После пересчета весов выполняется нормализация по документу. Таким образом, вес терма будет тем больше, чем чаще он входит в состав устойчивых сочетаний и чем чаще устойчивые сочетания употребляются в тексте. Без привлечения экспертов контекст наиболее весомых термов можно учесть следующим образом. Положим

$$w_k' = w_k + \sum_{\substack{j=1 \\ j \neq k}}^{K} x_{kj},$$

где x_{kj} определяет, сколько раз термы t_k и t_j встретились в одном контексте, $k = 1,2\ldots$, K. После пересчета весов выполняется нормализация по документу. Таким образом, вес терма будет тем больше, чем чаще он употребляется в одном контексте с другими термами.

Выбор числа K зависит от количества термов N и ресурсов автоматического анализатора. В случае, когда «близость» между термами трактуется как их совместное вхождение в предложение (или в иной фрагмент, если речь идет о размеченных текстах), изложенный подход можно применять не только к парам термов, но и к тройкам, четверкам и т.д.

4. Задача аннотирования текстовых документов [9]. Важной прикладной задачей извлечения информации (IE – Information Extraction) из ТД является их аннотирование. Аннотирование можно трактовать как составление метаданных анализируемых документов. Семантические метаописания создаются с использованием терминологии онтологии предметной области и могут быть разделены на контекстные и контентные метаданные, которые соответственно описывают контексты и контенты (содержания) объектов-документов. Известные подходы к аннотированию ТД произвольной тематики и узкой направленности существенно различаются. Наиболее сложны задачи автоматического извлечения информации из неструктурированных или слабо структурированных ТД, причем трудности реализации процедур IE возрастают по мере расширения тематики исследуемого корпуса документов. В существующих ИИС процедуры извлечения знаний из таких ТД основаны на выявлении в текстах паттернов (словосочетаний, предложений), содержащих определенное ключевое слово (обычно глагол) вместе с сопутствующими словами, выполняющими такие роли, как «субъект», «инструмент», «цель». Одним из условий применения известного подхода для автоматического аннотирования ТД произвольной тематики является их синтаксическая и семантическая корректность. На практике аннотирование выполняется для структурированных ТД конкретной тематики. В первой группе методов, ориентированных на такое аннотирование, используется поиск и выявление в документах часто встречающихся слов, характеризующих конкретные события, ситуации, факты. К таким словам относятся экземпляры концептов, такие как собственные имена (NE – named entities), названия организаций, географических пунктов, даты, адреса и т.п. Во второй группе методов извлечение информации заключается в поиске специфических выражений, характерных для определенных ПрО. Полуавтоматическое аннотирование документов и извлечение нужных данных при этом обычно происходит на основе предварительного обучения системы IE.

5. Задача распознавания ключевых слов в потоке слитной речи [12]. В этом случае трудоёмкий процесс ручного формирования сети допустимых переходов между ключевыми словами в заданной ПрО может быть автоматизирован с использованием технологии автоматического построения семантической сети текста на основе корпуса текстов этой предметной области. Семантическая сеть ПрО является здесь одним из способов представления в сетевом виде семантики предметной области (текста). Особенностью автоматического процесса формирования сети переходов является необходимость выявления ключевых слов в заданной ПрО, а также существенно важных для описания ПрО предложений. Предложения, не несущие информации о ПрО, в этом случае отбрасываются. Чем тщательнее будет произведен отбор, тем более корректно будет работать система распознавания речи. После выбора ключевых слов, а также значимых предложений текста на основе этих предложений строится однородная семантическая (ассоциативная) сеть. Далее, эта сеть используется для пре-

имущественного выбора гипотез ключевых слов, которые входят в эту семантическую сеть и находятся на наименьших расстояниях от предыдущего распознанного слова.

В настоящее время накоплен опыт успешного решения задач 1…5 на базе применения разнообразных методов и средств [3, 6, 10, 12, 14].

В задачах смыслового анализа текстовых документов хорошо себя показал латентный семантический анализ (*Latent semantic analysis*) (ЛСА) [15]. В основе метода ЛСА лежат принципы факторного анализа и, в частности, процедура выявления латентных связей изучаемых явлений или объектов. Данный метод положительно себя зарекомендовал при извлечении контекстно-зависимых значений лексических единиц за счёт статистической обработки больших корпусов текстов. Метод ЛСА опирается на линейный алгебраический подход и использует приведение матриц к каноническому виду. Здесь изучается прямоугольная матрица данных с числом столбцов n, равным числу разных слов, и со строками, которые представляют семантически обособленные фрагменты текста (концепции), представленные предложениями, фразами или синтагмами.

Число повторений слова в «концепциях» характеризует их статистическую значимость и интерпретируется как мера смысла. На столбцах и строках могут быть введены априорные целевые функции (функции интереса) и изучены условия диффузии интереса при движении по матрице. Далее применяется алгебраическая процедура, которая формирует сингулярное разложение прямоугольной матрицы (*Singular value decomposition*). В ходе разложения матрица разбивается оптимальным образом на сумму декартовых произведений векторов строк на векторы слов с весами, равными собственным значениям матрицы. Тем самым в неявной форме решается классическая задача кластеризации в пространстве слов и концепций, что позволяет определить формальные решения для целого ряда типовых задач смыслового анализа. Основные недостатки метода ЛСА заключаются в формально-математическом подходе, в сложности строгой интерпретации численных характеристик, в существенной трудоёмкости вычислений, кубически зависящей от объёма исходного текста. Выделение структурных элементов (СЭ) освобождает текст от случайных (шумовых) вкраплений, однако информация, которую несут СЭ, может быть неактуальной. Это имеет место в случаях, если конкретный СЭ используется в более широких контекстах или представляет субъективно авторское изложение или типовую фразу (штамп).

В рамках принятой нами информационной концепции смысл каждой фразы, каждого предложения и документа определяется на фоне предыдущего (или объемлющего) текста и измеряется количеством новой информации, которую этот фрагмент несет. Поэтому идеи метода ЛСА могут оказаться продуктивными в вопросах идентификации семантических моделей знаний ПрО.

В работах А.А. Харламова для автоматического смыслового анализа текста предложена технология TextAnalyst, которая позволяет выявить ключевые понятия в их взаимосвязях в тексте, а также ранжировать их по степени их смысловой значимости в данном тексте [12]. В результате исследований строится искомая однородная семантическая (ассоциативная) сеть N как совокупность несимметричных пар понятий $<c_i, c_j>$, где c_i и c_j – понятия, связанные между собой отношением ассоциативности (совместной встречаемости в некотором фрагменте текста). Ранжирование ключевых понятий, в свою очередь, позволяет ранжировать предложения ТД и выбирать наиболее существенные из них для множества текстов, описывающих ПрО. Такая сеть может быть исходной для построения сети переходов между ключевыми словами в задаче распознавания ключевых слов в потоке слитной речи. В перспективе для более точной идентификации сети переходов за счет разметки ассоциативных связей между ключевыми словами и типами их отношений в предикатных структурах соответствующих предложений может быть создана интегрированная методика анализа на основе статистического и лингвистического подходов к автоматическому смысловому анализу текстов.

В начале XXI века новый импульс к развитию получили методы лингвистического анализа ТД на естественном языке. Классический лингвистический подход к анализу текста предполагает существование нескольких относительно независимых уровней анализа: морфологического, синтаксического и семантического. В настоящее время успешно культивируются логико-лингвистические методы автоматического анализа текстов, которые основываются на эвристических правилах, разработанных экспертами-лингвистами [1, 2, 11]. Используемые при этом механизмы разработки лингвистических ресурсов весьма ресурсозатратны, поскольку для создания автоматических систем необходима разработка модели представления знаний части естественного языка, что требует согласо-

ванных усилий высококвалифицированных лингвистов, системных программистов и аналитиков.

В работах В.Я. Цветкова предложена общая схема семантического анализа ТД [13], заключающаяся в выделении в исследуемом тексте семантических информационных единиц (СИЕ): слово, предложение, фраза – и в последующем изучении их семантического окружения. Совокупности связанных информационных единиц дают возможность оценки морфологической и смысловой сложности языковых конструкций. В отличие от классического системного анализа данный подход допускает разные критерии делимости контента.

Одним из перспективных направлений развития семантического подхода к исследованию ТД, по мнению экспертов, является применение онтологий. Модели знаний в онтологиях выражаются в виде множеств понятий (концептов, сущностей) и отношений между ними [6, 14]. Представление семантики ПрО с использованием онтологического подхода во многих аспектах отвечает концепции создания ИИС извлечения знаний из ТД. Например, аннотирование текстов с применением онтологий позволяет составлять аннотацию из терминов концептов или значений (экземпляров) концептов, найденных в тексте документа. Здесь по-прежнему популярно аннотирование на основе использования NE. Аннотация представляет собой сформированные высказывания, содержащие NE и выраженные в формате RDF [6, 9]. Аннотации представляются в форме онтологий, что позволяет использовать средства онтологического анализа как для самих документов, так и для их аннотаций.

Принципиальное значение для получения положительного результата семантического анализа ТД имеет построение корректной семантической модели ПрО, в которой должны быть интегрированы опыт и специальные знания экспертов с учётом особенностей решаемой задачи. Ядро методики семантического анализа ТД должны составить: семантическая модель ПрО; алгоритмы и процедуры формального представления ТД и экстрагирования в его составе основных СИЕ; алгоритмы и процедуры идентификации модели ассоциативных связей СИЕ в составе каркасной модели текста; алгоритмы и процедуры нечёткой классификации.

Заключение

Проведенные исследования позволяют сделать вывод, что для решения задач семантического анализа ТД актуальным является обоснование универсального инструментария, опирающегося на достижения в области компьютерной лингвистики, теоретической информатики и искусственного интеллекта. Методологическим основанием разработки формализованного подхода к анализу семантики ТД могут служить современные результаты филологических исследований в области теории текста [1, 2, 11]. При этом наилучшую перспективу для создания математического обеспечения универсального семантического процессора имеют методы системно-структурной, когнитивной и функционально-прагматической лингвистики, методы нечёткой идентификации и классификации информационных объектов, многоагентная технология и онтологический подход [1, 6, 7, 10].

Список литературы

1. Алефиренко Н.Ф. Спорные проблемы семантики: монография. – Волгоград: Перемена, 1999. – 274 с.

2. Болотнова Н.С. Филологический анализ текста. Лингвистическая экспертиза. Проведение лингвистических исследований договоров, статей, научных трудов. – 3-е изд., испр. и доп. – М.: Флинта: Наука, 2007. – 520 с.

3. Ваграменко Я.А., Фанышев Р.Г. Технология интеллектуального анализа текстовой информации в базах знаний образовательной экспертной системы // Педагогическая информатика. – 2011. – № 1. – С. 57–62.

4. Васенин В., Афонин С., Козицын А. Автоматизированный анализ текстовой информации // Информационные технологии. – 2009. – № 7. – С. 56–57.

5. Ермаков А.Е. Извлечение знаний из текста и их обработка: состояние и перспективы // Информационные технологии. – 2009. – № 7. – С. 50–55.

6. Ланин В.В. Применение онтологического подхода и мультиагентной технологии для создания интеллектуальной системы управления документами // Тр. Междунар. науч.-техн. конф. «Открытые семантические технологии проектирования интеллектуальных систем» (OSTIS-2011). – Минск, 2011. – С. 435–442.

7. Надеждин Е.Н. Теоретические аспекты семантического анализа междисциплинарных знаний в интеллектуальных обучающих системах / Е.Н. Надеждин; ФГБОУ ВПО «Тульский государственный педагогический университет имени Л.Н. Толстого». – Тула, 2013. – 18 с.: 4 ил. – Библиогр.: 23 назв. – Русс. – Деп. в ВИНИТИ 17.12.2013 г.; № 374-В2013.

8. Надеждин Е.Н. Задача выявления цепочки ключевых слов и предложений при семантическом анализе текста // Научный альманах. – 2015. – № 9 (11). – С. 773–778.

9. Осминин П.Г. Современные подходы к автоматическому реферированию и аннотированию // Вестник Южно-Уральского государственного университета. Серия: Лингвистика. – 2012. – № 25. – С. 134–135.

10. Паутов К.Г., Попов Ф.А. Информационная система анализа и тематической классификации веб-страниц на основе методов машинного обучения // Современные проблемы науки и образования. – 2012. – № 6.; URL: https://www.science-education.ru/ru/article/view?id=7680 (дата обращения: 22.12.2016).

11. Тураева З.Я. Лингвистика текста (Текст: Структура и семантика) – Учеб. пособие для студентов пед. ин-тов. – М.: Просвещение, 1986. – 127 с.

12. Харламов А.А., Ермоленко Т.В. Семантическая сеть предметной области как основа для формирования сети переходов при автоматическом распознавании слитной речи //

Тр. Междунар. науч.-техн. конф. Открытые семантические технологии проектирования интеллектуальных систем. Open Semantic Technologies for Intelligent Systems (OSTIS-2013): материалы III Междунар. научн.-техн. конф. (Минск, 21–23 февраля 2013 г.). Минск: БГУИР, 2013. – С. 369–374.

13. Цветков В.Я. Семантика информации // Дистанционное и виртуальное обучение. – 2012. – № 10. – С. 4–7.

14. Шабанов В.И., Андреев А.М. Метод классификации текстовых документов, основанный на полнотекстовом поиске // Труды первого российского семинара по оценке методов информационного поиска. Под ред. И.С. Некрестьянова. – СПб.: НИИ Химии СПбГУ, 2003. – С. 52–71.

15. Thomas K., Landauer T., Harshman R. Latent semantic analysis. J. Amer. Soc. of Information Science, 1990. – 41(6).

References

1. Alefirenko N.F. Spornye problemy semantiki: monografija. Volgograd: Peremena, 1999. 274 p.

2. Bolotnova N.S. Filologicheskij analiz teksta. Lingvisticheskaja jekspertiza. Pro-vedenie lingvisticheskih issledovanij dogovorov, statej, nauchnyh trudov. 3-e izd., ispr. i dop. M.: Flinta: Nauka, 2007. 520 p.

3. Vagramenko Ja.A., Fanyshev R.G. Tehnologija intellektualnogo analiza tekstovoj infor-macii v bazah znanij obrazovatelnoj jekspertnoj sistemy // Pedagogicheskaja informatika. 2011. no. 1. pp. 57–62.

4. Vasenin V., Afonin S., Kozicyn A. Avtomatizirovannyj analiz tekstovoj informa-cii // Informacionnye tehnologii. 2009. no. 7. pp. 56–57.

5. Ermakov A.E. Izvlechenie znanij iz teksta i ih obpabotka: sostojanie i pepspektivy // In-formacionnye tehnologii. 2009. no. 7. pp. 50–55.

6. Lanin V.V. Primenenie ontologicheskogo podhoda i multiagentnoj tehnologii dlja sozdanija intellektualnoj sistemy upravlenija dokumentami // Tr. Mezhdunar. nauch.-tehn. konf. «Otkrytye semanticheskie tehnologii proektirovanija intellektualnyh sistem» (OSTIS-2011). Minsk, 2011. pp. 435–442.

7. Nadezhdin E.N. Teoreticheskie aspekty semanticheskogo analiza mezhdisciplinarnyh znanij v intellektualnyh obuchajushhih sistemah / E.N. Nadezhdin; FGBOU VPO «Tulskij gosu-darstvennyj pedagogicheskij universitet imeni L.N. Tolstogo». Tula, 2013. 18 p.: 4 il. Bibliogr.: 23 nazv. Russ. Dep. v VINITI 17.12.2013 g.; no. 374-V2013.

8. Nadezhdin E.N. Zadacha vyjavlenija cepochki kljuchevyh slov i predlozhenij pri semantiche-skom analize teksta // Nauchnyj almanah. 2015. no. 9 (11). pp. 773–778.

9. Osminin P.G. Sovremennye podhody k avtomaticheskomu referirovaniju i annotirovaniju // Vestnik Juzhno-Uralskogo gosudarstvennogo universiteta. Serija: Lingvistika. 2012. no. 25. pp. 134–135.

10. Pautov K.G., Popov F.A. Informacionnaja sistema analiza i tematicheskoj klassifikacii veb-stranic na osnove metodov mashinnogo obuchenija // Sovremennye problemy nauki i obrazovanija. 2012. no. 6.; URL: https://www.science-education.ru/ru/article/view?id=7680 (data obrashhenija: 22.12.2016).

11. Turaeva Z.Ja. Lingvistika teksta (Tekst: Struktura i semantika) Ucheb. posobie dlja stu-dentov ped. in-tov. M.: Prosveshhenie, 1986. 127 p.

12. Harlamov A.A., Ermolenko T.V. Semanticheskaja set predmetnoj oblasti kak osnova dlja formirovanija seti perehodov pri avtomaticheskom raspoznavanii slitnoj rechi // Tr. Mezhdunar. nauch.-tehn. konf. Otkrytye semanticheskie tehnologii proektirovanija intellektu-alnyh sistem. Open Semantic Technologies for Intelligent Systems (OSTIS-2013): materialy III Mezhdunar. nauchn.-tehn. konf. (Minsk, 21–23 fevralja 2013 g.). Minsk: BGUIR, 2013. pp. 369–374.

13. Cvetkov V.Ja. Semantika informacii // Distancionnoe i virtualnoe obuchenie. 2012. no. 10. pp. 4–7.

14. Shabanov V.I., Andreev A.M. Metod klassifikacii tekstovyh dokumentov, osnovannyj na polnotekstovom poiske // Trudy pervogo rossijskogo seminara po ocenke metodov informa-cionnogo poiska. Pod red. I.S. Nekrestjanova. SPb.: NII Himii SPbGU, 2003. pp. 52–71.

15. Thomas K., Landauer T., Harshman R. Latent semantic analysis. J. Amer. Soc. of Information Science, 1990. 41(6).