УДК 681.3:658.56

А.Н. ГЛИБОВЕЦ, И.В. РЕШЕТНЁВ

МЕТОД ИТЕРАТИВНОГО ПОСТРОЕНИЯ ТЕРМИНОЛОГИИ В КОЛЛЕКЦИЯХ НАУЧНЫХ ТЕКСТОВ НА УКРАИНСКОМ ЯЗЫКЕ

Аннотация. Описан метод итеративного построения терминологий в коллекциях научных текстов на украинском языке. Освещена проблематика автоматизированного построения тезаурусов по составлению научной терминологии. Значительное внимание уделено анализу лексикографических особенностей характеристических фрагментов текста документов. Учтена специфика украиноязычных документов. Основное внимание уделяется решению прикладной задачи построения терминологии с описанием связей в формате RDF из входящих текстов в широкоупотребляемом формате pdf.

Ключевые слова: статистические методы, лексикографические методы, тезаурус, термин, связь «общее–частное», гипонимия.

ВВЕДЕНИЕ

Создание и актуализация специализированных словарей не успевают за прогрессом в исследованиях в силу объективных причин: сложности изучаемых сфер и изменчивости понятий со временем [1]. Вместе с тем для исследователей остается острой необходимость взаимопонимания на понятийном уровне, что требует как унифицированной и доступной терминологической базы, так и качественной поисковой системы научных документов.

Одним из эффективных способов улучшения релевантности поисковой выдачи таких систем — использование тезауруса [2]. Среди методов построения тезаурусов автоматизированный метод лучше всего подходит для сферы научных исследований в силу высоких темпов обновляемости информации и связанной с этим высокой себестоимостью участия экспертов в такой работе. В рамках ряда исследований, проведенных на кафедре информатики Национального университета «Киево-Могилянская Академия» (НаУКМА) по созданию поисковой системы научных документов, разработка компонента автоматизированного построения тезауруса улучшает ее качество.

Основные цели данной публикации — описание и реализация метода извлечения терминологии из входящих научных текстов, положенных в основу такой системы. В работе проанализированы существующие подходы к построению тезаурусов и описан разработанный метод автоматизированного определения важных украиноязычных терминов и терминологических связей между ними, который реализован в виде веб-сервиса. Анализ эффективности метода проведен на реальных данных научной украиноязычной периодики. Разработанный компонент стал естественной составляющей поисковой системы украиноязычных научных документов.

При разработке метода учитывалась ограниченность выпущенных документарных коллекций на украинском языке, что потребовало учета возможности итеративного добавления научных документов в терминологические базы с последующим обновлением содержания тезауруса. Акцентируется внимание на описании решения прикладной задачи построения терминологии с описанием связей в формате RDF из входящих текстов в широко употребляемом формате pdf.

1. ОБЗОР СУЩЕСТВУЮЩИХ ПОДХОДОВ

1.1. Роль тезауруса в информационном поиске. Тезаурусом называют управляемый словарь, содержащий семантические связи между терминами и улучшающий процесс поиска связанных терминов [3].

Обычно информационная потребность пользователя поисковой системы не соответствует терминам, которые встречаются в документах, или пользователь

неправильно понимает терминологию области знаний, в которой он осуществляет поиск. При таких условиях одним из методов улучшения поисковой выдачи является использование тезаурусов терминов предметных областей [4]. Тезаурусы представляют собой таблицы терминов и связей между ними с указанием типа связи (NT, BT, USE, RT) [3]. Информационные системы могут использовать тезаурусы на этапе индексации документов для более правильной классификации документов по категориям или во время поиска, расширяя поисковый запрос пользователя связанными терминами.

Главная проблема составления тезаурусов заключается в том, что для большинства коммерческих баз данных, которые распространяют научную информацию, они составляются экспертами из областей знаний, а также специалистами по составлению тезаурусов. В новейших областях знаний, где соответствующая терминология только формируется и выпускается большое количество новых публикаций, таких как биоинформатика или компьютерная инженерия, терминологические словари очень быстро устаревают, и их нужно обновлять чаще, снова привлекая экспертов. В противоположность такому подходу существуют методы автоматизированного построения тезаурусов, которые в качестве корпуса принимают все новейшие публикации по теме и строят на их основе взаимосвязи между терминами. С помощью такой системы обновлять терминологические связи значительно проще и дешевле. В [2] рассмотрены основные методы автоматизированного построения тезаурусов, которые имеют как различную эффективность и временную оценку сложности, так и принципы: статистический и лексикографический. Описанию разработки нового метода, использующего комбинацию идей, встречающихся в этих подходах, и посвящена данная работа.

В [5] дается следующее обозначение тезауруса. Тезаурус — это лексико-семантическая модель концептуальной реальности или ее представителя, которая выражена в форме системы терминов и их взаимосвязей, предлагает доступ с помощью многих аспектов и используется как система обработки и поиска внутри модуля информационной поисковой системы. Отметим, что автор акцентирует внимание на принципиальной неразрывности теоретической модели тезауруса и практического применения программных модулей с такой функциональностью.

1.2. RDF как формат представления тезаурусов. Формат RDF — один из самых распространенных способов представления данных и метаданных для технологий семантического веба. Несколько упрощая, можно сказать, что в основе данного формата лежит идея представления информации в виде триплетов «субъект – предикат – объект». Такая общая и простая, на первый взгляд, модель может удачно удовлетворить потребностям тезауруса для описания его содержания. Следующая важная особенность формата — широкая международная поддержка на уровне реализации прикладных систем. Как отмечалось ранее, роль тезауруса определяется не только точностью и объемом представленных терминологических связей, но и практической применимостью программного модуля, простотой доступа и пригодностью к машинной обработке. Именно благодаря возможности публикации данных, обработанных с помощью программной части тезауруса, непосредственно в Интернет в общепринятом формате, авторы предложили формат RDF и обеспечивающую систему веб-сервисов с программным интерфейсом в качестве конечного формата доступа к тезаурусу.

Среди конкретных спецификаций RDF формат JSON-LD, представленный в стандарте ISO-25964 [6], по мнению авторов настоящей статьи, наилучшим образом соответствует поставленной задаче публикации ресурсов тезауруса в виде веб-сервиса. К базовым концепциям формата относятся [7]: IRI — интернациональные идентификаторы ресурсов; контекст, который служит, в основном, для задания сокращений к IRI; идентификаторы узлов и типизированные значения.

Предложенных базовых элементов формата достаточно, чтобы данные тезауруса минимально удовлетворяли стандарту.

1.3. Автоматизированные методы построения тезаурусов. Методы автоматизированного составления тезаурусов можно разделить на два принципиальных класса: статистические, интенсивно использующие частотные и позиционные характеристики терминов в документах в качестве основы для различных моделей выявления связей между терминами, и лексикографические, использующие сведения из сферы обработки человеческой речи для осуществления синтаксического, морфологического и других видов анализа текста для установления семантических связей на основе информации, полученной исключительно из текста. В лексикографических методах обычно используются собранные экспертами корпусы языков, которые содержат правила общего употребления слов, словоформы и синонимические ряды. Реализации множества методов способствуют программные пакеты для проведения первоначального анализа свободного текста. В свою очередь, для статистических методов таким инструментом являются утилиты индексирования и ранжирования терминов.

Основой для многих статистических методов поиска зависимостей между терминами служит создание индекса терминов, описывающих содержание документов наилучшим образом, что обычно требует ранжирования терминов по степени важности. Наиболее применяемыми техниками взвешивания, разработанными для алгоритмов поисковых систем, является использование частоты термина (TF), обратной документарной частоты (IDF), а также их комбинаций.

Метод совместного употребления терминов — один из подходов в информационном поиске к формированию многословных терминов [8]. Основные элементы для вычислений в методе — частота вхождения термина в определенные разные по размеру контекстные рамки, такие как целый документ, главы документа, параграфы и другие элементы. При этом, чем ближе слова встречаются в контексте выбранной рамки, тем большей назначается мера совместного употребления. Некоторые авторы сомневаются в качестве найденных терминологических связей с помощью этого метода. Например, в [9] говорится о неэффективности составленного по данному методу тезауруса применительно к задачам поиска. Автор [9] предлагает свой подход, он вводит понятие концептуального пространства как сети терминов и взвешенных ассоциаций между ними, которые способны отобразить концепты и связи между ними в соответствующем информационном пространстве, представленном в виде коллекции документов в базе данных. Модель ассоциативного поиска, включенная в данный метод, приближена к ментальным способам представления информационных потребностей пользователей поисковой системы в виде сети терминов и связей между ними, которые, как правило, нечеткие.

Лексикографические методы поиска связей между терминами базируются на принципе прямого указания связи между словами с помощью языковых средств, причем характер связи можно определить, исходя из синтаксического и лексического строения высказываний. В разрезе лексикографических методов поиска связей в терминологии интересен способ составления терминологических словосочетаний, как один из самых продуктивных в словообразовании. Таким образом, использование лексикографических методов для наполнения тезауруса терминологическими отношениями представляется наиболее подходящим по своим принципам. В [9, 10] отмечалась одна из самых распространенных проблем всех статистических методов — проблематика индексирования фразовых терминов или, в нашем случае, терминологических словосочетаний. В частности, авторы отмечали необходимость построения качественных решений на основе лингвистических особенностей текстов, в том числе с использованием техники тегирования по частям речи, как одну из главных задач улучшения статистических методов в информационном поиске.

В следующем лексикографическом методе ключевую роль играет понятие гипонимии. Гипонимия — это отношение вида к роду в лексико-семантической системе. Родовые слова называют гиперонимами, а видовые — гипонимами. По-

нятно, что явление гипонимии непосредственно указывает на связь типа «общее–конкретное» между терминами и является неотъемлемой составляющей тезаурусов. Развивая эту идею, М. Хеарст [11] создал автоматизированный лексикографический метод выделения гипонимов из текста.

Две главные проблемы, решаемые с помощью данного подхода, — элиминация необходимости в предварительно составленных базах знаний по предметной области и возможность применения метода на разнообразных текстовых коллекциях. В [11] составлено множество лексико-синтаксических шаблонов, непосредственно указывающих на искомые лексические зависимости, которые легко распознать в тексте как программными средствами, так и самостоятельно. Гипотеза метода подтверждает наличие большого количества полезной информации о предметной области в самом тексте, которая может быть обнаружена как человеком, так и алгоритмом, не прибегая к слишком конкретным деталям определенных явлений и вещей, не требуя от системы глубокого лексикографического или семантического анализа.

Данную технику поиска таксономических связей предложил Альшави [12]. Он использовал иерархию шаблонов для интерпретации определений, состоявших преимущественно из индикаторов частей речи и символов-масок. Основным недостатком данного подхода авторы считают проблему подбора такого множества шаблонов, которые с одинаковой точностью указывали бы на направленность связи в текстах различных стилей.

Резюмируя основные достижения метода, можно указать на сравнительную дешевизну его применения для автоматизированного сбора семантических связей в документах. Метод позиционируется как альтернатива статистическим методам и по сравнению с ними имеет преимущество в точности работы на редких связях между терминами, которые встречаются в тексте единично и не могут удачно обрабатываться статистическими методами. Представленные в исследовании шаблоны и стратегии отсечения модификаторов существительных не претендуют на полноту и оставляют определенную свободу для будущих дополнений.

2. ИТЕРАТИВНЫЙ КОМБИНИРОВАННЫЙ МЕТОД ПОСТРОЕНИЯ ТЕРМИНОЛОГИИ

В этом разделе описаны основные этапы итеративного метода построения терминологии с помощью комбинации лексикографических и статистических методов.

2.1. Структурная схема алгоритма. Процесс построения терминологии на основе коллекции текстов можно разделить на два принципиальных шага: 1) выделение множества слов, встречающихся в текстах документов, отвечающих терминам в области знаний соответствующих документов; 2) установка на множестве данных терминов отношений, используемых в тезаурусе.

Задача выделения терминов из множества всех слов документа смысловым образом подобна обычной операции индексирования текстов поисковыми системами, что и было использовано в нашем методе для получения упорядоченного списка уникальных слов коллекции, с применением техники взвешивания TFIDF. При этом в начале такой последовательности содержатся слова, наилучшим образом характеризующие содержание документов, а следовательно, являющиеся кандидатами в термины.

Для ограничения такого списка слов можно ввести оператор, который предоставил бы возможность определить граничный элемент списка, после которого идут общеупотребительные слова, не являющиеся терминами.

Данный оператор может иметь следующие вариации для нашего метода. «Стоп-список» — оператор, отсекающий заданное параметром количество слов в хвосте последовательности. Такой подход использует один из популярных методов удаления стоп-слов в поисковых системах, однако остается чувствительным к размеру коллекции текстов. Пропорциональный оператор подобен оператору «стоп-список», с ограничением в качестве параметра определенного процента слов в хвосте последовательности, основанный на статистическом распределении числа терминов в коллекциях научных текстов.

В результате вычислительных экспериментов было решено остановиться на пропорциональном подходе к ограничению входного списка терминов, исходя из его преимуществ при обработке неподготовленных текстовых коллекций.

Проведено оценку терминов по метрике документарной частоты эталонной коллекции. Понятно, что способы ограничения списка слов будут работать только при условии применения надежной схемы взвешивания, что, в свою очередь, в нашем случае будет зависеть от способа подсчета составляющей документарной частоты терминов, чувствительной к составу и размеру коллекций.

В данной работе проблему малых коллекций текстов для надежного взвешивания предложено решать путем наполнения и использования справочной системы документарных частот терминов. Справочная система базируется на построении и индексации большой и разнообразной учебной коллекции текстов научной тематики с последующим хранением полученных документарных частот как эталонных. В качестве документарной основы для такой коллекции предложено полное собрание статей журнала «Научные записки НаУКМА».

После получения первоочередного списка терминов для составления тезауруса необходимо определить характер и направленность связей между терминами.

Введем понятие характеристического фрагмента текста, который является непосредственным вхождением термина в документ в определенном контексте. Из множества методов рассмотрения контекста употребления слов, например частей окружающих словосочетаний и оборотов, предложений, окон с фиксированным размером количества слов, мы выбрали именно предложения в качестве основы для наших исследований, исходя из имеющихся инструментов, которые позволяли бы применить методику тегирования по частям речи в качестве основы для лексикографических методов.

Следующий шаг — нахождение характеристических фрагментов текста всех терминов из списка. Данный поиск можно осуществить линейно, однако, предусматривая возможность масштабирования разработанного метода, предложено использовать одну из поисковых систем с открытым кодом, которая возвращала бы все документы из нашей коллекции, содержащие определенный термин, таким образом ограничивая пространство линейного поиска. Далее, среди найденных документов осуществляется линейный поиск характеристических фрагментов.

На следующей стадии работы метода анализируются все найденные характеристические фрагменты с применением различных методик для определения типа связи. Применение простого метода совместного употребления терминов внутри одного характерного фрагмента позволяет установить связь терминов (RT), если они входят в характерные фрагменты текста вместе с начальным термином; применения множества определенных лексикографических шаблонов позволяет найти связи типов BT, NT и RT.

Проведем расширение тезауруса с помощью терминологических словосочетаний. Применение лексикографических шаблонов базируется на методе нахождения терминологических словосочетаний, который, в свою очередь, в случае совпадения шаблона с текстом позволяет выделить не только однословные термины, но и состоящие из нескольких слов. Естественно, терминов второго типа намного больше. Таким образом, побочным продуктом применения лексикографических шаблонов является расширение первоочередного списка терминов терминологическими словосочетаниями. Этого нельзя было достичь на первом этапе с помощью индексирования в рамках использованных инструментов.

Для применения определенных нами лексикографических шаблонов введем такую формальную нотацию. Лексикографический шаблон (Lexicographic Pattern — LP) — упорядоченный список операторов сопоставления. Оператор сопоставления — команда, которая требует применения операции поиска совпадения типа существительного словосочетания (Noun Phrase — NP) или конкретного слова, или символа из синонимического ряда (Exact Word — EW).

NP — оператор сопоставления, выполняющий поиск существительного

словосочетания за счет применения указанных для каждого такого оператора списка правил совпадения по частям речи. Возвращает в качестве результата все найденные во фразе существительные словосочетания в порядке заданных правил совпадения, а также позиции найденных существительных словосочетаний в фразе. К операторам сопоставления данного типа в качестве параметра можно задать их роль (индексы 1 и 0).

Роль оператора *NP* — индекс 1 или 0, который указывает на главную или второстепенную роль данного оператора в шаблоне (записывается как NP_1 или NP_0).

EW — оператор сопоставления, осуществляющий поиск вхождения конкретного символа или слова в фразу из списка возможных альтернатив, возвращает позиции вхождений таких слов.

W — оператор окна, который указывает минимальные и максимальные рамки окна, играет роль маски совпадения с любыми последовательностями слов в предложении.

IT — оператор итерации, который обозначает повторяющуюся последовательность операторов в шаблоне.

Правило сходимости (MR) — заданная последовательность тегов частей речи, которой должна соответствовать подпоследовательность слов в предложении.

Теги частей речи (N, A, P) — параметры конфигурации правил сопоставления для выделения терминологических словосочетаний, обозначающих существительное (N), прилагательное (A) и предлог (P) соответственно.

Удовлетворение шаблона — нахождение множества удовлетворяющих операторам сопоставления подпоследовательностей слов, где каждая позиция такой подпоследовательности отвечает как порядку вхождения в фразу, так и порядку оператора, определенного в шаблоне. Все возможные совпадения по отдельным операторам должны быть объединены в результирующее множество путем ограничения по правилам.

Например, чтобы зафиксировать в нашей формальной нотации лексикографический шаблон, отвечающий за прямые определения с использованием тире, нужно записать следующее:

$$LP = (NP_0\,(MR < A,\ N >), EW(\text{“—”}, \text{”-”}), NP_1\,(MR < N,\ N >)).$$

Такому шаблону удовлетворяет фраза: «Социологическое исследование — система процедур для получения научных знаний о социальных явлениях и процессах». При этом первому оператору сопоставления будет отвечать терминологическое словосочетание «социологическое исследование», оператору сопоставления по слову было предоставлено две альтернативы — собственно символ «тире», а также дефис для обработки случаев замены данного символа в исходном тексте, последнему оператору соответствует словосочетание «система процедур».

Таким образом, операторы сопоставления типа *EW* в шаблоне играют роль фиксированных точек шаблона, в то время как операторы *NP* — роль наполняемых переменных, извлекающих словосочетания из фраз во время удовлетворения шаблона.

Проведем интерпретацию связей по совпадениям текста с шаблоном. При составлении шаблона параметрам *NP* дополнительно указывается параметр главной или второстепенной роли в шаблоне, которые интерпретируют связи между полученными совпадениями по *NP* следующим образом: между представителями NP_0 и NP_1 устанавливается связь BT; между представителями NP_1 и NP_0 устанавливается связь *NT*; между представителями одинаковых ролей — связь *RT*.

Основой для такой интерпретации является то, что в большинстве шаблонов на соответствующих местах терминологических словосочетаний по частям предложения бывают или однородные определения, или приложения, или обобщающие слова, или, например, в случае сопоставления с шаблоном прямых определений в тексте — соответственно термин и его родовая принадлежность. Таким образом, в тексте в случае совпадения с шаблоном направленность связи четко определена.

Таблица 1. Перечень разработанных лексикографических шаблонов в формальной нотации

Название	Формальная запись правил шаблона
MR1-9	$MR < NPNN >,\ MR < ANNN >,\ MR < ANAN >,\ MR < ANN >,$ $MR < NAN >,\ MR < NN >,\ MR < AN >,\ MR < N >$
LP1	$NP1, EW <$ '– '|'–'$>, EW <$ 'це' |'є'|'означає' |'вважається', $NP0$
LP2	$EW <$'такий'$>, NP1, EW <$'як'$>, \{ITNP0, EW <$','$>\}, EW <$''і|'або'|'й'|'та'$>, NP0$
LP3	$NP0, ITEW <$','$>, NP0, EW <$'і'|'або'|'й'|'та'$>, EW <$'інший'$>, NP1$
LP4	$NP1, EW <$','$>, EW <$'включаючи'|'а саме'|'зокрема'|'особливо'$>,$ $ITNP0, EW <$','$>, EW <$'і'|'або'|'й'|'та'$>, NP0$
LP5	$NP0, W < 0,3 >, EW <$'бути частиною'|'входити в'$>, W < 0,3 >, NP1$
LP6	$NP1, W < 0,3 >, EW <$'складатися з'|'підрозділятися на'$>, W < 0,3 >,$ $ITNP0, EW <$','$>, EW <$'і'|'або'|'й'|'та'$>, NP0$

Для реализации алгоритма поиска гипонимов по Хеарсту сначала необходимо научить систему распознавать фразовые словосочетания. Предлагаемый подход — фиксация существительных в предложении с последующим подбором окружающих слов по правилам.

Учитывая сходство научного стиля при подаче определений на многих языках, кажется удачной мысль о локализации разработанных Хеарстом шаблонов для украинского языка с добавлением новых.

Для того чтобы сузить рамки исследования и достичь определенного результата для специфических, и вместе с тем наиболее употребляемых способов создания терминологии, были привлечены только термины-существительные и существительные словосочетания.

Из шаблонов, отвечающих за связи между терминами в предложении, выбраны следующие категории:

- прямые определения и дефиниции с использованием характерных для украинского языка знаков пунктуации и слов-связей;
- шаблоны по Хеарсту;
- шаблон на обозначение связей часть–целое.

Все представленные шаблоны расширяются синонимичными и похожими в употреблении словами в формулах шаблона. При сопоставлении предложений с шаблоном все слова приводятся к нормальной форме, что позволяет уменьшить необходимое количество вариаций шаблона. Детально разработанные шаблоны представлены в табл 1.

При применении правил шаблона учитывается их очередность, таким образом, в первую очередь отыскиваются существительные в качестве элементов совпадения словосочетания, с большим количеством слов, а значит, более редкие в употреблении.

2.2. Математическая модель и алгоритмическая формализация метода. Введем следующие обозначения: D — множество текстовых документов, LP — множество лексикографических шаблонов, T — множество терминов тезауруса, T_F — отсортированный по метрике TFIDF и ограниченный функцией $limit(T)$ список важных однословных терминов коллекции, T_E — множество многословных терминологических словосочетаний, R — множество связей тезауруса, $R_i \in \{(T_1, T_2, Rl)\}$, где $Rl \in \{RT, BT, NT\}$ и $T_1, T_2 \in T_t$ — множество характеристических фрагментов текста для термина t, S_C — множество предложений характерного фрагмента C, Lem_S — множество лематизированных слов предложения S, M_{lp} — последовательность совпавших с лексикографическим шаблоном терминологических словосочетаний.

Также введем следующие функции:

$lm(T) : \{t \mid t \in T_S\} \to \{t' \mid t' \in T_F\}$ — функция ограничения отсортированного списка терминов;

1. $\forall d \in D$:
   $T_i := extract(d)$
   $(t)_{i=1}^{|T_i|} := sort(T_i, d)$
   $T_F := T_F \bigcup lm((t)_{i=1}^{|T_i|})$
2. $T_F := lm(sort(T_F))$
3. $\forall t \in T_F$:
   $C_t := findCF(t)$
   $\forall c \in C_t$:
   $S_C := split(c)$
   $\forall s \in S_C$:
   a. $\forall lem \in lem(s)$, если $lem \in T_F \Rightarrow R := R \bigcup (t, lem, RT) \bigcup (lem, t, RT)$
   b. $\forall lp \in LP$:
   1. $M_{lp} := match(lp, s)$
   2. $R_{lp} := inrs(M_{lp}); R := R \bigcup R_{lp}$
   3. $T_E := T_E \bigcup terms(M_{lp})$
4. $T := T_F \bigcup T_E$
5. $Thesaurus := (T, R)$

Рис 1. Алгоритм построения терминологии

$extract(d): D \to \{t \mid t \in T\}$ — функция извлечения терминов из документа; $sort(T, d): \{T\} \times D \to (t')_1^{|T|}, \forall t_i, t_j, i \geq j \Leftrightarrow tf(t_i, d) \cdot idf(t_i) \geq f(t_j, d) \cdot idf(t_j)$ — функция, которая строит последовательность отсортированных терминов документа по убыванию метрики TFIDF;

$tf(t, d): T \times D \to R$ — функция вычисления частоты термина в документе;

$idf(t): T \to R$ — функция, ставящая каждому термину в соответствие его инвертированную документарную частоту с эталонной коллекции;

$findCF(t): T \to \{c \in C_t\}$ — функция поиска характеристических фрагментов термина;

$split(c): C_t \to \{s \mid s \in S_C\}$ — функция разбиения характеристического фрагмента текста на предложения;

$lem(s): S_C \to (lem \mid lem \in Lem_S)$ — функция извлечения последовательности лем из предложения;

$match(lp, s): LP \times S_C \to \{m \mid m \in M_{lp}\}$ — функция удовлетворения шаблона, которая возвращает множество последовательностей совпавших терминологических словосочетаний в порядке следования позиций шаблона;

$inrs(M_{lp}): \{m \mid m \in M_{lp}\} \to R$ — функция установления связей на множестве последовательностей совпадений с шаблоном.

Разработанный метод построения тезауруса можно представить алгоритмом, показанным на рис. 1. Используется также следующая формализация правил совпадения с лексикографическим шаблоном:

$LP = \{(pe)_1^l \mid pe \in PE\}$ — множество лексикографических шаблонов, заданное как множество элементов шаблона. $PE = \{NP_0, NP_1, EW, W, IT\}$ — элементы шаблона;

$NP_0 = \{((mr)_1^m, 0) \mid mr \in MR\}$; $NP_1 = \{((mr)_1^m, 1) \mid mr \in MR\}$ — множества команд поиска терминологических словосочетаний с указанием главной (1) или второстепенной (0) роли словосочетания в шаблоне;

$MR = \{(tag)_1^k \mid tag \in \{'N', 'A', 'P'\}\}$ — множество правил совпадения, заданное последовательностями тегов частей речи;

$EW = \{(ew)_1^n \mid ew \in Lem\}$ — множество команд поиска прямого совпадения по слову, которое задано на последовательностях альтернатив лем;

$W = \{(\min, \max) \mid \min, \max \in N\}$ — множество команд поиска окон, которое задано парами минимальной и максимальной длины окна в предложении;

$IT = \{(it)_1^t) \mid it \in PE\}$ — множество команд писка итераций, которое задано на подпоследовательностях элементов шаблона;

$P_M = \{((l)_1^v, p) \mid l \in Lem,\ p \in N\}$ — множество фразовых совпадений, заданных парами последовательностей лем и позиций первой лемы;

$M_{lp} = (p)_1^v \mid \exists\, lp = (pe)_1^l, \forall s \in S, \forall p_i \in apply(pe_j, s),\ pe_j \in \{NP_0,\ NP_1\}$ — последовательность фразовых совпадений по операторам NP_0, NP_1 шаблона;

$apply(pe, s) : PE \times S \to \{p \mid p \in P_M\}$ — функция сопоставления элемента шаблона с фразой, которая ставит в соответствие множество фразовых совпадений;

$match(lp, s) : LP \times S_C \to \{m' \mid m' \in M_{lp}\}$, когда $\forall lp = (pe)_1^l,\ \exists\, (m)_1^l \mid m \in apply(pe_i, s)$ такая, что $\forall m_i, m_j, i < j \Leftrightarrow m_i = ((l)_1^{vi}, p_i), m_j = ((l)_1^{vj}, p_j), p_i \le p_j, \{m'_{\overline{1,n}}\} \subseteq \{m_{\overline{1,l}}\}$

$$inrs(M_{lp}) : \{m \mid m \in M_{lp}\} \to$$

$$\to \left\{ r \mid \left\langle \begin{array}{l} r = (T_1, T_2, BT) \mid \exists lp = (pe)_1^l, \exists s_1, \exists pm_1 \in apply(pe_i, s_1),\ pm_1 = (T_1, p_1), \\ \exists s_2, \exists pm_2 \in apply(pe_j, s_2),\ \ pm_2 = (T_2, p_2),\ pe_i \in NP_1,\ pe_j \in NP_0, \\ r = (T_1, T_2, NT) \mid pe_i \in NP_0,\ \ pe_j \in NP_1 \\ r = (T_1, T_2, RT) \mid (pe_i, pe_j \in NP_0) \cup (pe_i, pe_j \in NP_1) \end{array} \right\rangle \right\}$$

ЗАКЛЮЧЕНИЕ

В настоящей работе описано решение задачи итеративного построения терминологии в коллекциях научных текстов на украинском языке. На основе предложенного метода и разработанного алгоритма создан программный модуль в виде веб-сервиса с возможностями построения тезаурусов в формате RDF из исходных текстов формата pdf. Формат тезауруса JSON-LD выбран с учетом возможности публикации полученных терминологических связей в стандартизированном виде сетевого доступа к ресурсам и с позиций понимания тезауруса как полноценного программного модуля поисковой системы научных материалов. Из типов связей между терминами для поиска предпочтение отдано связям «общее–частичное», которые определялись с помощью лексикографического анализа предложений текстов на предмет содержания гипонимических связей между терминами.

В основу разработанного модуля построения тезаурусов положен описанный в данной работе метод поиска важных терминов и связей в тексте. Первый этап работы данного метода, который связан с поиском важных терминов в коллекциях документов, решен с помощью предложенного метода взвешивания, сортировки и фильтрации терминов документов с помощью метрики документарной частоты эталонной коллекции. В качестве такой коллекции использовался архив украиноязычной периодики «Научные записки НаУКМА», на основе которого построен справочный индекс документарных частот терминов.

Второй этап разработанного метода связан с применением лексикографических шаблонов для поиска гипонимических связей в исходных текстах. Для поиска успешной реализации использовалось открытое программное обеспечение, направленное на решение утилитарных задач лемматизации терминов и тегирование слов предложений по частям речи, а также адаптированы к украиноязычным правилам словоупотребления лексикографические шаблоны, предложенные в исследовании Хеарста [11]. Авторы настоящей публикации разработали расши-

ряемый программный пакет с функциональностью управления применением лексикографических шаблонов.

Тестирование реализации предложенного метода на тематических коллекциях научных текстов продемонстрировало эффективность первого этапа алгоритма, а также достаточную точность второго этапа в рамках разработанных шаблонов. Ограничение лексикографического метода поиска гипонимии не позволяют достичь полноты поиска связей в тексте из-за однозначности употребляемых в шаблонах контекстов терминологических связей и низкой статистической частотой их появления в тексте. Проблему можно устранить увеличением количества шаблонов, расширением синонимических рядов, определяющих шаблон слов, что требует привлечения экспертов по лексикографии, а также улучшением метода тегирование по частям речи с помощью стохастических методов устранения неоднозначности в определении частей речи отдельных слов.

Полученный программный модуль продемонстрировал прикладную применимость на тестовых коллекциях данных и может использоваться как составляющая поисковой системы научных материалов.

СПИСОК ЛИТЕРАТУРЫ

1. Л е н д а у С . І . Словники: мистецтво та ремесло лексикографії. — Київ: К.І.С., 2012. — 480 с.
2. L a s s i M . Automatic thesaurus construction // University Collage of Boras, Sweden. — 2002. 10 p. — http://www.academia.edu/506142/Automatic_thesaurus_construction.
3. Т и п ы связей в тезаурусе. — Веб. 10.05.2014 — http://publish.uwo.ca/~craven/677/thesaur/main06.htm.
4. C h e n H . , T o b u n D . N g , M a r t i n e z J . , S c h a t z B . A concept space approach to addressing the vocabulary problem in scientific information retrieval: an experiment on the worm community system // J. of the Amer. Soc. for Inform. Sci. — 1997. — http://arizona.openrepository.com/arizona/bitstream/10150/105991/1/chen21.pdf.
5. M i l l e r U . Thesaurus construction: problems and their roots // Inform. Proc. & Management. — 1997. — **33**, N 4. — P. 481–493.
6. “I S O 25964 — the International Standard for Thesauri and Interoperability with Other Vocabularies.” ISO 25964 Thesaurus Schemas. Web. 08 April 2014. — http://www.niso.org/schemas/iso25964/.
7. J S O N - L D 1.0. Web. 08 June 2014. — http://www.w3.org/TR/json-ld/.
8. C h e n H . , T a k Y i m , F y e D . , S c h a t z B . Automatic thesaurus generation for an electronic community system // J. of the Amer. Soc. for Inform. Sci. — 1995. — **46**, N 3. — P. 175–193.
9. C h e n H . , L y n c h K . , B a s u , K . , N g T . D . Generating, integrating, and activating thesauri for concept-based document retrieval // IEEE Expert. — 1993. — **8**, N 2. — P. 25–34.
10. G r e f e n s t e t t e G . Automatic thesaurus generation from raw text using knowledge-poor techniques. — Rank Xerox Research Centre, 1993. — http://www.academia.edu/4186829/AUTOMATIC_THESAURUS_GENERATION_FROM_RAW_TEXT_ USING_KNOWLEDGE-POOR_TECHNIQUES.
11. H e a r s t M . A . Automatic acquisition of hyponyms from large text corpora // Proc. of the 14th Conf. on Comput. Ling. Assoc. for Comput. Ling. — 1992. — **2**. — P. 539–545.
12. A l s h a w i H . Processing dictionary definitions with phrasal pattern hierarchies // Comput. Ling. — 1987. — **13**, N 3–4. — P. 195–202.

Поступила 03.07.2014