УДК 681.3

И.А. Андреев, В.А. Башаев, В.В. Клейн, В.С. Мошкин, Н.Г. Ярушкина

СЕМАНТИЧЕСКАЯ МЕТРИКА ТЕРМИНОЛОГИЧНОСТИ НА ОСНОВЕ ОНТОЛОГИИ ПРЕДМЕТНОЙ ОБЛАСТИ

Андреев Илья Алексеевич, *студент факультета информационных систем и технологий Ульяновского государственного технического университета. Опубликовано несколько статей в области извлечения информации из текста. [e-mail: ares-ilya@yandex.ru].*

Башаев Виталий Александрович, *аспирант, окончил факультет лингвистики и международного сотрудничества Ульяновского государственного университета. Имеет статьи в области извлечения информации из текста. [e-mail: perevod73@yandex.ru].*

Клейн Виктор Викторович, *студент факультета информационных систем и технологий УлГТУ. Опубликовано несколько статей в области извлечения информации из текста. [e-mail: vikklein93@gmail.com].*

Мошкин Вадим Сергеевич, *аспирант кафедры «Информационные системы» УлГТУ, окончил факультет информационных систем и технологий УлГТУ. Имеет статьи в области интеллектуальных систем анализа данных. [e-mail: postforvadim@yandex.ru].*

Ярушкина Надежда Глебовна, *доктор технических наук, профессор, заведующая кафедрой «Информационные системы» УлГТУ. Имеет более 250 научных работ в области мягких вычислений, нечеткой логики, гибридных систем. [e-mail: jng@ulstu.ru].*

Аннотация

В данной статье описана семантическая метрика извлечения списка терминов из текстов конкретной проблемной области, основанная на анализе ее онтологии. Представлена формальная модель используемой OWL-онтологии, а также модели и алгоритмы оценки степени терминологичности слов или сочетаний слов текстовых массивов.

Помимо этого, приведены метрики оценки результатов работы представленных семантических алгоритмов, а также рассмотрена реализация формальных моделей представления знаний предметной области в онтологической форме и разработанных алгоритмов в программной системе извлечения терминологии из текста.

В заключении приведены результаты вычислительных экспериментов по извлечению терминов на основе онтологии эксплуатации токарно-фрезерного станка с числовым программным управлением (ЧПУ) из набора текстов соответствующей предметной области, а также подведены итоги проведенных исследований, выявлены наиболее эффективные алгоритмы оценки терминологичности слов/сочетаний слов и рассмотрена перспектива дальнейших научных изысканий в этой области.

Ключевые слова: извлечение терминов, семантическая метрика, онтология.

A SEMANTIC METRIC OF THE TERMHOOD BASED ON THE SUBJECT AREA ONTOLOGY

Ilya Alekseevich Andreev, *a student at the Faculty of Information Systems and Technologies at Ulyanovsk State Technical University; an author of several articles in the field of information extraction from text. e-mail: ares-ilya@yandex.ru.*

Vitaliy Aleksandrovich Bashaev, *a post-graduate student; graduated from the Faculty of Linguistics and International Cooperation of Ulyanovsk State University; an author of several articles in the field of information extraction from text. e-mail: perevod73@yandex.ru.*

Victor Victorovich Klein, *a student at the Faculty of Information Systems and Technologies of Ulyanovsk State Technical University; an author of several articles in the field of information extraction from text. e-mail: vikklein93@gmail.com.*

Vadim Sergeevich Moshkin, *a post-graduate student at the Department of Information Systems of Ulyanovsk State Technical University; graduated from the Faculty of Information Systems and Technologies at Ulyanovsk State Technical University; an author of articles in the field of data analysis intelligence systems. e-mail: postforvadim@yandex.ru.*

Nadezhda Glebovna Yarushkina, *Doctor of Engineering, Professor, Head of the Department of Information Systems at Ulyanovsk State Technical University; an author of more than 250 papers in the field of soft computing, fuzzy logic, and hybrid systems. e-mail: jng@ulstu.ru.*

Abstract

The article describes a semantic metric for a retrieval of a list of terms from the texts of a specific knowledge domain, based on the analysis of its ontology. A formal model of the used OWL-ontology, as well as models and algorithms for the degree evaluation of a word termhood or word combinations of text arrays are presented.

In addition, the evaluation metrics of the presented semantic algorithms performance are given. The implementation of the formal models of a domain knowledge representation in an ontological form and the algorithms developed in the software system for the terminology extraction from the text is considered.

In conclusion, the results of the computational experiments performed for the extraction of terms based on the ontology of the NC turning-milling machine operation from the texts of an appropriate knowledge domain are provided. A conducted research is summarized. The most efficient algorithms for the degree evaluation of a word termhood or word combinations are revealed, and a perspective for further scientific research in this area is examined.

Key words: term extraction, semantic metric, ontology.

Введение

Принцип работы существующих алгоритмов извлечения терминологии (term extraction) в лексикографии и терминоведении основан на статистических и лингвистических методах. В основе статистических методов лежит вычисление степени терминологичности на основании числовых закономерностей, присущих термину или нетермину. В основе лингвистических методов лежит отбор по определенным лексико-грамматическим шаблонам и другим лингвистическим признакам термина [1].

Главным недостатком использования статистических и лингвистических методов в процессе извлечения терминологии из текста является отсутствие возможности выделения из получившегося множества терминов только тех, которые относятся к рассматриваемой проблемной области [2].

На множестве информационных единиц в некоторых случаях полезно задавать отношение, характеризующее ситуационную близость информационных единиц, т. е. силу ассоциативной связи между информационными единицами. Его можно было бы назвать отношением релевантности для информационных единиц. Такое отношение дает возможность выделять в информационной базе некоторые типовые ситуации. Отношение релевантности при работе с информационными единицами позволяет находить термины, близкие уже найденным или заранее заданным в общей базе знаний [3].

При анализе больших массивов документации необходимо учитывать специфику ее предметной области, чтобы получить в качестве результата список терминов, характерных для конкретной предметной области. Именно для решения подобных задач используют семантические алгоритмы, базирующиеся на определенных семантических метриках.

В настоящее время одной из наиболее универсальных методик представления экспертных знаний с точки зрения полноты семантического описания информационной единицы предметной области является онтологический подход. Именно поэтому одним из важнейших направлений решения задачи извлечения терминологии из большого массива технической документации является разработка и использование семантических метрик на основе онтологических моделей [4].

1 Формальная модель онтологии предметной области «Эксплуатация токарно-фрезерного станка с ЧПУ»

Сущность онтологического подхода заключается в том, что предметная область представляется в виде организованной совокупности понятий, их свойств и связей.

Наиболее удобным форматом представления онтологии с точки зрения машинной обработки и наглядности описания особенностей предметной области является язык OWL.

Выделим обязательные требования к OWL-онтологии, используемой в рамках решения задачи извлечения терминологии:

• Онтология должна наиболее полно отражать особенности объектов предметной области. Задача решается посредством описания максимального количества объектов и классов рассматриваемой области во всем многообразии межклассовых отношений.

• Онтология не должна быть избыточной. Данная проблема решается разбиением отдельного класса на несколько с сохранением семантической целостности за счет определения вспомогательного предиката «имеет Отношение», отражающего некоторое взаимодействие объектов определяемых классов. Примером такого взаимодействия является триплет «Блокировка» – «имеет Отношение» – «Паллета».

• Онтология должна быть наглядной. Это требование в некоторой степени противоречит предыдущему, поэтому поиск компромиссных вариантов представления тех или иных объектов классов онтологии – одна из важнейших задач адекватного отражения особенностей рассматриваемой предметной области [5].

Онтологический подход хранения знаний предполагает представление их в следующем виде:

$$O=<T, R, F>. \quad (1)$$

Исходя из модели (1), онтология «Эксплуатация токарно-фрезерного станка с ЧПУ» имеет следующие составляющие:

1. T – термины прикладной области, которую описывает онтология. Например, объекты «Резцедержатель», «Станина», «Поплавковое реле».

2. R – отношения между терминами предметной области, при этом $R \subset \{R_{inc}, R_{add}, R_{term}, R_{lem}, R_{NC}\}$:

• R_{inc} – множество встроенных отношений объектов, таких как «sameAs» и «SubClassOf»;

R_{add} – множество отношений, позволяющих расширять набор объектов описываемой предметной области за счет сочетания лемм связанных объектов. Пример: свойства «имеет Отношение» и «является Частью»;

R_{term} – отношение «является Термином», имеющее логический тип значения. Это свойство является вспомогательным и определяется экспертом исходя из критерия, насколько данный объект онтологии является характерным конкретно для этой предметной области. Используется в процессе извлечения терминов согласно тезаурусному критерию терминологичности;

R_{lem} – отношение «имеет Лемму», имеющее строковое значение, полученное путем леммирования (приведения к начальной форме) наименования объекта с помощью программы Mystem компании Яндекс по соответствующим морфологическим признакам термина;

R_{NC} – множество отношений объектов, а также свойств типа данных, наиболее полно описывающих особенности взаимодействия объектов рассматриваемой предметной области. Пример: свойства «является Типом Смазки», «является Этапом», «состоит Из».

3. F – множество функций интерпретации (аксиоматизации), заданных на терминах и/или отношениях онтологии [6]. Примеры таких функций в разработанной онтологии представлены выражениями (2) и (3):

$$F_{СОЖ} : X_{ТипСверления} \rightarrow Y_{ТипПодачи}, \quad (2)$$

где $F_{СОЖ}$ – отношение «является Типом Подачи СОЖ»,

$X_{ТипСверления}$ – множество объектов класса «Тип Сверления»,

$Y_{ТипПодачи}$ – множество объектов класса «Тип Подачи СОЖ»;

$$F_{InEng} : X_{Контекст} \rightarrow Y_{Eng}, \quad (3)$$

где F_{InEng} – отношение «имеет Английский Эквивалент»,

$X_{Контекст}$ – множество объектов класса «Контекст»,

Y_{Eng} – множество объектов класса «Английский Аналог».

2 Семантическая метрика оценки терминологичности слов/сочетаний слов

Использование семантической метрики «термин/нетермин» на множестве слов конкретного текста с использованием заранее разработанной OWL-онтологии в процессе извлечения терминологии предполагает определение для каждого поступающего слова или сочетания слов степени близости к терминам рассматриваемой области. Применение такой метрики позволяет выделить из массива поступающих однословий/многословий только те термины и сочетания, которые относятся к данной предметной области.

Степень близости входных слов/сочетаний слов к терминам проблемной области (k_{Ont}) может иметь значение от 0 до 1: чем ближе полученное значение к 1, тем с большей долей вероятности данное одно-/многословие является термином [7].

В ходе решения поставленной задачи было разработано два критерия выделения терминов из предметной области посредством использования онтологии:

• тезаурусный критерий,

• критерий вложенных связей.

Результаты проведенных экспериментов должны показать, какой из данных семантических критериев является наиболее продуктивным и оптимальным применимо к имеющейся модели онтологии.

Тезаурусный критерий

Тезаурус представляет собой контролируемый словарь терминов на естественном языке, явно указывающий отношение между терминами и предназначенный для информационного поиска. Любая онтология является усложненной версией тезауруса [8].

Тезаурусный подход к извлечению терминологии предполагает непосредственный поиск вхождений лемм поступающих на вход слов и их сочетаний среди терминов, определенных в онтологии. Для этого в разработанной онтологии для каждого класса определено свойство «имеет Лемму», которое имеет строковое значение, полученное путем леммирования (приведения к начальной форме) имени объекта с помощью программы Mystem компании Яндекс по соответствующим морфологическим признакам термина.

Алгоритм определения степени близости слов/сочетания слов терминам проблемной области согласно тезаурусному критерию предполагает:

- Оценку степени близости поступающего на вход алгоритма слова/сочетания слов каждому объекту онтологии без учета онтологического критерия оценки;
- Определение опорного объекта онтологии, наиболее близко ассоциирующегося с входным одно-/многословием.

Опорный объект онтологии, используемый в дальнейшем анализе, имеет степень близости по отношению к входному слову/сочетанию слов, рассчитанную по следующей формуле:

$$k_t = \max_{i=1}^{m}\left(\frac{n_i}{p_i}\right), \qquad (4)$$

Рис. 1. Поиск опорного объекта онтологии

Рис. 2. Тезаурусный критерий

где m – количество всех объектов онтологии;

n_i – число слов из леммы входного многословия, найденных в лемме объекта онтологии;

p_i – общее число слов в лемме объекта онтологии.

Общая схема оценки степени близости слов/сочетания слов терминам проблемной области согласно тезаурусному критерию приведена на рисунке 1.

При этом порядок следования слов многословия в опорном объекте должен сохраняться.

Если несколько разных объектов онтологии имеют одинаковое значение коэффициента k_t, то опорным будет считаться тот объект, которому соответствует максимальное n_i. Если таких объектов несколько, то они все будут считаться опорными и анализ по онтологическому критерию будет проведен для каждого из этих объектов.

Структура онтологии рассматриваемой предметной области предполагает наличие у каждого из ее объектов свойства (DatatypeProperty) «является Термином», имеющее логический тип значения. Это свойство является вспомогательным и определяется экспертом исходя из критерия, насколько данный объект онтологии является характерным конкретно для этой предметной области.

Степень близости слова/сочетания слов терминам рассматриваемой предметной области в соответствии с тезаурусным критерием оценивается по следующей формуле:

$$k_{Ont} = \frac{k_t}{c+1}, \qquad (5)$$

где k_t – результат первого этапа анализа;

c – число отношений, связывающих опорный объект онтологии с ближайшим объектом, имеющим истинное значение свойства «является Термином».

В случае, если сам опорный объект имеет истинное значение данного свойства, то $c = 0$. Схема данного поиска приведена на рисунке 2.

Таким образом, процесс оценки степени близости одно/многословия к терминам проблемной области по метрике «термин/нетермин» в его онтологической составляющей представляет собой движение по графу, в узлах которого находятся объекты соответствующих классов онтологий.

Если опорный объект имеет ложное значение свойства «является Термином» и при этом не имеет никаких связей с другими объектами онтологии, либо все связанные объекты также имеют значение «false» этого свойства, то находится другой опорный объект для данного слова/словосочетания и оценка проводит-

ся заново. В аналогичной ситуации с другими опорными объектами либо в случае их отсутствия входное одно/многословие признается «нетермином» ($k_{Ont}=0$).

Критерий вложенных связей

Помимо оценки степени терминологичности отдельно взятого слова/сочетания слов, разработанная метрика позволяет извлечь термины из текста посредством их сопоставления с имеющимися объектами и сочетаниями лемм соответствующих объектов с помощью отношений R_{add}, определенных в онтологии.

Таким образом, при сопоставлении входных сочетаний и объектов предметной области, связанных между собой однонаправленными отношениями R_{add}, термином рассматриваемой предметной области будет считаться многословие, лемма которого полностью совпадает с объединением лемм соответствующих объектов онтологии.

Особенностью данного метода является необходимость представления объектов онтологии преимущественно в виде однословий с максимизацией числа отношений между объектами. Определяющими для использования этого метода являются отношения R_{add}, связь объектов посредством которых позволяет формировать словосочетания естественным образом. Пример формирования многословий с помощью свойства «имеет Отношение»:

1. Найденная цепочка объектов: «Вращение» + «имеет Отношение» + «Двигатель» + «имеет Отношение» + «Переменный ток»;

2. Объединение лемм объектов онтологии: «вращение двигатель переменный ток»;

3. Термин, извлекаемый из обрабатываемого текста: «вращение двигателя переменного тока».

Пример формирования многословий с помощью свойства «является Частью»:

1. Найденная цепочка объектов: «Подшипник» + «является Частью» + «Шпиндель»;

2. Объединение лемм объектов онтологии: «подшипник шпиндель»;

3. Термин, извлекаемый из обрабатываемого текста: «подшипник шпинделя».

При этом извлеченные термины, входящие в свою очередь в термины, состоящие из большего количества слов, не рассматриваются в качестве терминов с целью избегания избыточности.

Метрики оценки результатов

Рассмотрим метрики оценки, применимые к задаче классификации. Допустим, что мы знаем **правильные** категории для некоторого числа документов. Сгруппируем ответы нашего гипотетического анализатора следующим образом:

• Истинно-положительные (**true positives, tp**) – те категории, которые мы ожидали увидеть и получили на выходе;

• Ложно-положительные (**false positives, fp**) – категории, которых быть на выходе не должно, и анализатор их ошибочно вернул на выходе;

• Ложно-отрицательные (**false negatives, fn**) – категории, которые мы ожидали увидеть, но анализатор их не определил;

• Истинно-отрицательные (**true negatives, tn**) – категории, которых быть на выходе не должно, и на выходе анализатора они тоже совершенно правильно отсутствуют.

В этом случае мера точности (P, *precision*) определяется так:

$$P=\frac{\mathrm{tp}}{\mathrm{tp}+\mathrm{fp}}. \quad (6)$$

Мера точности характеризует, сколько полученных от классификатора положительных ответов являются правильными. Чем больше точность, тем меньше число ложных попаданий.

Мера точности, однако, не дает представление о том, все ли правильные ответы вернул классификатор. Для этого существует так называемая мера полноты (R, *recall*):

$$R=\frac{\mathrm{tp}}{\mathrm{tp}+\mathrm{fn}}. \quad (7)$$

Мера полноты характеризует способность классификатора «угадывать» как можно большее число положительных ответов из ожидаемых [9].

Помимо этого, удобно для характеристики классификатора, использующего разработанную семантическую метрику, использовать унифицированную метрику F_1:

$$F_1=2\times\frac{P\cdot R}{P+R}. \quad (8)$$

Фактически это просто среднее гармоническое величин P и R. Величина F_1 является одной из самых распространенных метрик для подобного рода систем. Именно F_1 используется, чтобы сформулировать пороговое качество разработанного семантического анализатора [10].

3 Структурно-функциональное решение системы извлечения терминологии

В рамках решаемой поставленной задачи были проведены следующие действия:

1. Экспертом в области эксплуатации токарно-фрезерного станка с числовым программным управлением построена OWL-онтология соответствующей проблемной области;

2. Была разработана онтологически-ориентированная система извлечения терминологии, применяющая описанные выше метрики для решения задачи определения терминологичности одно/многословий, извлекаемых из больших объемов технических текстов.

Онтология эксплуатации токарно-фрезерного станка с ЧПУ

Разработанная OWL-онтология имеет иерархическую организацию и включает в себя 261 экземпляр классов и

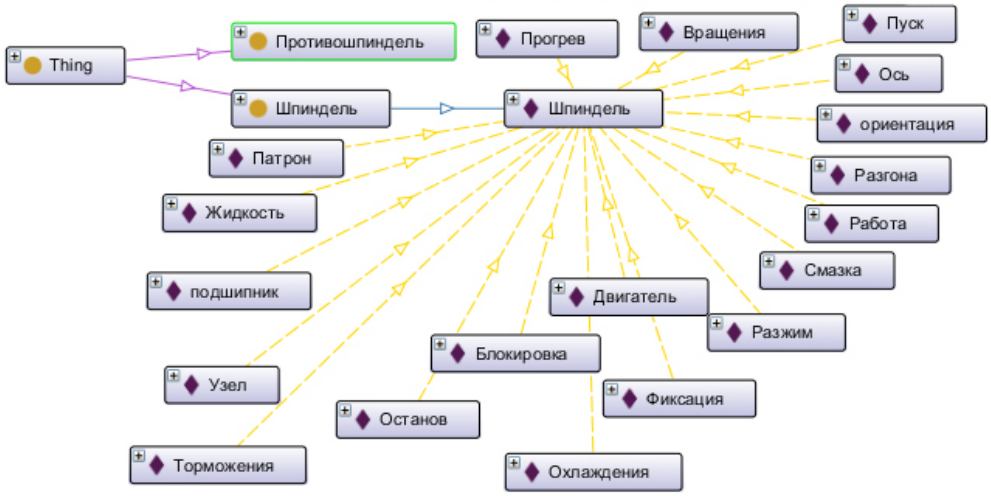

Рис. 3. Фрагмент онтологии

порядка 746 отношений объектов классов. На данный момент онтология имеет 4 уровня иерархии, что позволяет максимально конкретизировать термины предметной области, используемой при решении поставленной задачи. Фрагмент данной онтологии представлен на рисунке 3.

Онтология была разработана с помощью редактора Protégé 4.2, являющегося свободно распространяемым, открытым и имеющим легко расширяемую архитектуру за счет поддержки модулей расширения функциональности. Онтология хранится в файле с расширением .owl.

Пример объявления класса «Концевая фреза» разработанной OWL-онтологии:

```
<owl:Class rdf:ID=”Концевая_фреза “>
  <rdfs:label rdf:datatype=»http://www.w3.org/2001/
XMLSchema#string»
  >Концевая фреза</rdfs:label>
  <rdfs:subClassOf>
   <owl:Class rdf:about=“#Фреза»/>
  </rdfs:subClassOf>
 </owl:Class>
```

Онтологически-ориентированная система извлечения терминологии

Для реализации описанных алгоритмов была разработана программная система «Онтологически-ориентированная система извлечения терминологии» на языке C#, на платформе .NET 3.5, также была использована СУБД MS SQL Server 2008.

Концептуальная схема программной системы, реализующей рассматриваемую онтологическую метрику, включает в себя следующие компоненты:

- два программных модуля, каждый из которых взаимодействует со своей базой данных;
- модуль морфологического анализа Mystem от компании Яндекс;
- корпус текстов проблемной области;
- онтология проблемной области, составленная экспертом в редакторе Protégé (рис. 4).

Алгоритм работы разработанной системы извлечения терминологии предполагает следующую последовательность действий:

1. Обработка текста модулем морфологического анализа;
2. Подсчет лингвистических и статистических характеристик полученного текста, содержащего морфологическую разметку, основным модулем системы;
3. Подсчет семантических характеристик слов и словосочетаний обрабатываемого текста, базирующийся на представленных методиках с использованием разработанной OWL-онтологии.

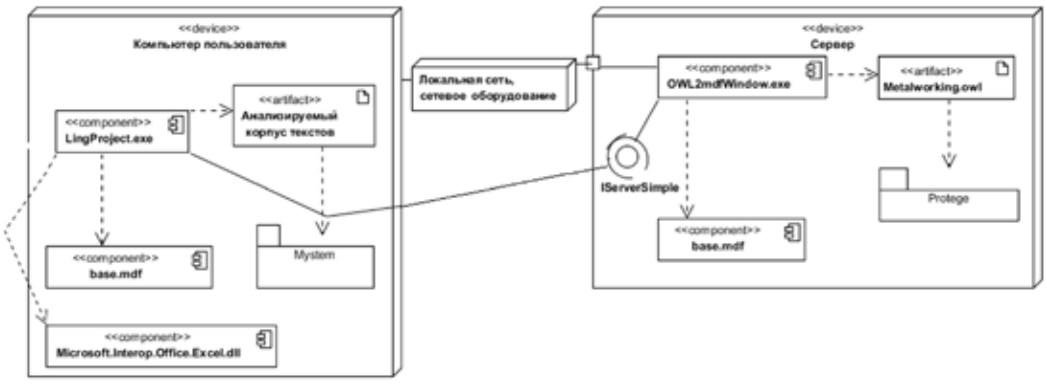

Рис. 4. Диаграмма развертывания онтологически-ориентированной системы извлечения терминологии

Помимо этого, у модуля подсчета семантических характеристик существует два режима функционирования: режим оконного приложения, обеспечивающий возможности обслуживания базы данных и быстрого тестирования работы разработанных алгоритмов, и режим сервера, запускаемого в фоновом режиме для обработки больших объемов данных.

4 Результаты вычислительных экспериментов извлечения терминов из текстов соответствующей предметной области

Приводимые результаты экспериментов имеют целью изучение эффективности разработанных показателей. Были рассмотрены результаты работы двух показателей: тезаурусного и критерия внутренних связей; четырех категорий словоупотреблений: одиночных слов, двух-, трех-, четырехсловных словосочетаний.

Для проведения эксперимента использовался текст объемом около 62000 слов из руководства по эксплуатации токарно-фрезерного станка с ЧПУ.

К особенностям текстов данной предметной области можно отнести высокую насыщенность терминами, влияние научного стиля на лексико-семантические, морфологические, синтаксические параметры и формализованность содержания, опирающегося на логико-понятийную схему предметной области.

Для оценки эффективности подсчета показателей рассмотрены меры $Precision$ (6), $Recall$ (7) и F_1-мера (8) для каждого показателя в каждой категории словоупотреблений.

Результаты экспериментов по извлечению терминов посредством применения тезаурусного критерия представлены в таблице 1, критерия вложенных связей – в таблице 2.

Так как в случае применения тезаурусного критерия оценивается терминологичность каждого слова/сочетания, поступающего на вход алгоритма, то для формального отделения терминов от нетерминов в результате его выполнения, необходимо ввести пороговое значение $k_{Ont} = 0{,}5$.

Анализ результатов выполнения разработанных методик необходимо рассматривать с учетом различий вариантов словоупотреблений:

- Одиночные слова.

Исходя из полученных выше результатов, следует отметить, что наилучшие показатели извлечения однословных терминов были получены при применении второго критерия. Почти все извлеченные алгоритмом одиночные слова являются терминами, в то время как всего было извлечено немногим более половины всех однословных терминов рассматриваемой проблемной области. $Recall$ у тезаурусного показателя для однословных терминов хоть и ненамного ниже, но значение $Recall$ позволяет судить о более низкой эффективности этого показателя. Таким образом, показатель вложенных связей оказался наиболее эффективным при вычислении однословных терминов, о чем свидетельствует и наивысшее значение F_1-меры среди показателей.

- Двухсловные словосочетания.

Исходя из результатов анализа, можно сделать вывод, что тезаурусный признак значительно уступает по полноте и точности второму критерию, имеющему лучшие значения $Precision$, $Recall$ и F_1-меры среди всех результатов. При достаточно высокой точности критерий вложенных связей извлекает более половины двухсловных терминов предметной области. Таким образом, для извлечения двухсловных терминов наиболее эффективным также является второй критерий.

Таблица 1

Результаты применения тезаурусного критерия

Категория	Выделено экспертом в категории	Кол-во слово-употреблений $k_{Ont} > 0{,}5$	Из них терминов	$Precision$	$Recall$	F_1-мера
1	294	120	88	0,73	0,29	0,42
2	631	305	133	0,43	0,21	0,28
3	361	379	214	0,56	0,59	0,57
4	107	196	120	0,61	1,12	0,79

Таблица 2

Результаты применения критерия вложенных связей

Категория	Выделено экспертом в категории	Кол-во слово-употреблений	Из них терминов	$Precision$	$Recall$	F_1-мера
1	294	168	154	0,91	0,52	0,66
2	631	431	372	0,86	0,58	0,69
3	361	370	327	0,88	0,9	0,89
4	107	159	129	0,81	1,2	0,97

• Трехсловные словосочетания.

Удовлетворительным можно считать результат работы тезаурусного признака по извлечению трехсловных терминов: извлекается более половины трехсловных терминов предметной области при среднем *Precision*. Результаты работы второго критерия можно назвать лучшими, о чем позволяют судить достаточно высокие значения *Precision* и *Recall*.

• Четырехсловные словосочетания.

Тезаурусный признак и критерий вложенных связей оказались сопоставимыми по эффективности. Значение *Recall*, превышающее 1 для обоих показателей, свидетельствует об извлечении ими терминов, ранее не выделенных в ходе экспертного анализа. Несмотря на схожие результаты, тезаурусный признак проигрывает второму признаку за счет более низкого значения *Precision*. Таким образом, признак вложенных связей оказался наиболее эффективным для извлечения и четырехсловных терминов.

Полученные результаты экспериментов по извлечению терминов из инструкции по эксплуатации токарно-фрезерного станка с ЧПУ с использованием разработанной онтологии соответствующей предметной области позволяют сделать вывод о высокой эффективности использования критерия вложенных связей для решения поставленной задачи, особенно в случаях анализа трех- и четырехсловий.

Заключение

Таким образом, предложенная в данной работе семантическая метрика «термин/нетермин» на основе онтологии проблемной области позволяет выделить из массива поступающих одно-/многословий только те термины и сочетания, которые относятся к данной предметной области, устанавливая для каждого из входных сочетаний слов численное значение степени их близости к терминам рассматриваемой предметной области.

Данная метрика может быть использована как в качестве самостоятельной, так и в сочетании с лингвистическими и статистическими метриками, используемыми в процессе извлечения терминологии с целью обеспечения всестороннего анализа поступающих данных.

СПИСОК ЛИТЕРАТУРЫ

1. Андреев И.А., Башаев В.А., Клейн В.В. Разработка программного средства для извлечения терминологии из текста на основании морфологических признаков, определяемых программой Mystem // Интегрированные модели и мягкие вычисления в искусственном интеллекте. – М. : Физматлит, 2013. – С. 1227–1236.

2. Yarushkina N. Soft computing and complex system analysis // International Journal of General Systems. 2001. Vol. 30, № 1. pp. 71–88.

3. Namestnikov A., Yarushkina N. Efficiency of Genetic algorithms for automated design problems // Известия Российской академии наук. Теория и системы управления. – 2002. – № 2. – С. 127–133.

4. Ярушкина Н.Г., Вельмисов А.П., Стецко А.А. Средства data minig для нечетких реляционных серверов данных // Информационные технологии. – 2007. – № 6. – С. 20–29.

5. Митрофанова О.А., Константинова Н.С. Онтологии как системы хранения знаний // Всероссийский конкурсный отбор обзорно-аналитических статей по приоритетному направлению «Информационно-телекоммуникационные системы». – 2008. – С. 54.

6. Добров Б.В., Лукашевич Н.В., Сыромятников С.В. Формирование базы терминологических словосочетаний по текстам предметной области // Электронные библиотеки: перспективные методы и технологии, электронные коллекции (RCDL-2003) : тр. 5-й Всерос. науч. конф. – СПб., 2003. – С. 201–210.

7. Афанасьева Т.В., Ярушкина Н.Г. Нечеткий динамический процесс с нечеткими тенденциями в анализе временных рядов // Вестник Ростовского государственного университета путей сообщения. – 2011. – Т. 3. – С. 7–16.

8. Кураленок И.Е., Некрестьянов И.С. Оценка систем текстового поиска // Программирование. – 2002. – 28(4). – С. 226–242.

9. Ярушкина Н.Г., Мошкин В.С Применение онтологического подхода к анализу состояния локальной вычислительной сети // Радиотехника. – 2014. – № 7. – С. 120–124.

10. Ярушкина Н.Г., Афанасьева Т.В. Нечеткие временные ряды как инструмент для оценки и измерения динамики процессов // Датчики и системы. – 2007. – № 12. – С. 46–50.

REFERENCES

1. Andreev I.A., Bashaev V.A., Klein V.V. Razrabotka programmnogo sredstva dlya izvlecheniya terminologii iz teksta na osnovanii morfologicheskikh priznakov, opredelyaemykh programmoy Mystem [Software Development for Morphological Terminology Extraction from the Text Specified by the Mystem-Program]. *Integrirovannye modeli i myagkie vychisleniya v iskusstvennom intellekte* [Integrated Models and Soft Computing in Artificial Intelligence]. Moscow, Fizmatlit Publ., 2013, pp. 1227–1236.

2. Yarushkina N. Soft Computing and Complex System Analysis. *International Journal of General Systems,* 2001, vol. 30, no. 1, pp. 71–88.

3. Namestnikov A.M., Yarushkina N.G. Efficiency of Genetic Algorithms for Automated Design Problems. *Izvestiya Rossiyskoy akademii nauk. Teoriya i sistemy upravleniya* [Theory and Management Systems. A Journal of the Russian Academy of Sciences], 2002, no. 2, pp. 127–133.

4. Yarushkina N.G., Velmisov A.P., Stetcko A.A. Sredstva data minig dlya nechetkikh relyatcionnykh serverov dannykh [Data Mining Facilities for Fuzzy Relational Data Servers]. *Informatsionnye tekhnologii* [Information Technologies], 2007, no. 6, pp. 20–29.

5. Mitrofanova O.A., Konstantinova N.S. Ontologii kak sistemy khraneniya znaniy [Ontologies as Knowledge

Storage Systems]. *Vserossiyskiy konkursnyy otbor obzorno-analiticheskikh statey po prioritetnomu napravleniyu 'Informatsionno-telekommunikatsionnyesistemy'*[All-Russian Competition to Find the best Review-analytical Article in the Advanced Field of Information-and-Telecommunication Systems], 2008, p. 54.

6. Dobrov B.V., Lukashevich N.V., Syromyatnikov S.V. Formirovanie bazy terminologicheskikh slovosochetaniy po tekstam predmetnoy oblasti [The Generation of Terminological Word-Combination Base at Knowledge-Domain Texts]. *Elektronnye biblioteki: perspektivnye metody i tekhnologii, elektronnye kollektsii (RCDL-2003): tr.5-y Vseros. nauch. konf.* [E-Libraries: Advanced Methods and Technologies, E-Collections (RCDL 2003): Proc. of the 5th All-Russian Scientific Conf.]. Sankt-Peterburg, 2003, pp. 201-210.

7. Afanasieva T.V. and Yarushkina N.G. Nechetkiy dinamicheskiy protsess s nechetkimi tendentsiyami v analize vremennykh ryadov [Fuzzy Time Series with Fuzzy Tendency]. *Vestnik Rostovskogo gosudarstvennogo universiteta putey soobshcheniya* [Bulletin of Rostov State Transport University], 2011, vol. 3, pp. 7–16.

8. Kuralenok I.E., Nekrestyanov I.S. Otsenka sistem tekstovogo poiska [Evaluation of Text Retrieval Systems]. *Programmirovanie* [Programming], 2002, no. 28(4), pp. 226–242.

9. Yarushkina N.G. and Moshkin V.S Primenenie ontologicheskogo podkhoda k analizu sostoyaniya lokalnoy vychislitelnoy seti [Applying Ontological Approach to the Analysis of the State of Local Area Network]. *Radiotekhnika* [Radioengineering], 2014, no. 7, pp. 120–124.

10. Yarushkina N.G., Afanasieva T.V. Nechetkie vremennye ryady kak instrument dlya otsenki i izmereniya dinamiki protsessov [Fuzzy Time Series as a Tool for Process Dynamics Evaluation and Measurements]. *Datchiki i sistemy* [Sensors and Systems], 2007, no. 12, pp. 46–50.