УДК 004

ОНТОЛОГИЧЕСКИЙ ПОДХОД К ОЦЕНКЕ ТЕМАТИКИ НАУЧНОГО ТЕКСТА

О.П. Кузнецов[1], В.С. Суховеров[2]

Институт проблем управления Российской академии наук, Москва, Россия
[1]olkuznes@ipu.ru, [2]suhoverv@ipu.ru

Аннотация

В работе предлагается подход к определению тематики научного текста, использующий онтологию предметной области. Излагается оригинальный принцип построения онтологий прикладных наук, при котором дерево онтологии содержит три обязательных ветви: «Фундаментальные теории», «Прикладные теории», «Области приложений», классами онтологии являются темы, а экземплярами классов - термины соответствующих тем. Описывается строение онтологии наук об управлении. Приводятся фрагменты онтологии управления и её словаря. Предполагается, что научный текст релевантен теме, если он содержит термины этой темы. Предлагается метод оценки степени релевантности научного текста различным темам, основанный на подсчёте числа вхождений в документ терминов этих тем. Результатом работы этого метода является «профиль документа» - вектор релевантностей документа темам онтологии. Описывается автоматическая система анализа тематики научных текстов из области теории и практики управления, разработанная на основании предложенного подхода. Обсуждаются лингвистические проблемы поиска терминов. Приведены некоторые статистические данные обработки тематических разделов журнала «Автоматика и телемеханика» и пример построения профиля для конкретной статьи журнала. Отмечаются возможные направления улучшения оценок релевантности.

Ключевые слова: *онтология, компетентность, релевантность, тема, термин, профиль документа, теория управления.*

Цитирование: *Кузнецов,О.П.* Онтологический подход к оценке тематики научного текста / О.П. Кузнецов, В.С. Суховеров // Онтология проектирования. – 2016. – Т.6, №1(19). – С. 55-66. – DOI: 10.18287/2223-9537-2016-6-1-55-66.

Введение

Определение тематики документа обычно является первым и необходимым этапом для принятия решения относительно дальнейших действий с содержимым документа и, возможно, с его автором. Поскольку алгоритмические подходы к определению тематики – как известные [1], так и предлагаемый – не анализируют качество содержания документа (корректность, новизну, актуальность, стиль и т.д.), то дальнейшие действия в значительной степени зависят от того, получил ли документ экспертную оценку (рецензию, экспертное заключение и др.) или нет. Наличие квалифицированной оценки качества документа (типичный случай – публикация в рецензируемых изданиях) означает, что он может служить свидетельством степени компетентности автора. Определение тематики таких документов помогает решать различные задачи управления персоналом в научных организациях – приём на работу, поручение наукоёмких заданий, подбор команды для работы над проектом и т.д. Неотрецензированный документ, как правило, требует формального или неформального рецензирования, экспертного заключения и т.д. Здесь типичный случай – рецензирование статей, поданных в научный журнал или докладов, присланных на конференцию. Определение тематики таких документов позволяет решать задачу назначения рецензентов.

Заметим, что сама задача определения тематики может решаться независимо от наличия или отсутствия оценки качества документа; такая оценка играет роль для возможных приложений. В настоящей работе задача определения тематики документа решается для научных документов: статей, докладов и др.

1 Автоматическая оценка тематики текста. Компетенция и компетентность

Прикладные задачи, для решения которых нужна автоматическая оценка тематики научного текста, можно разбить на два больших класса: 1) информационные и 2) управленческие задачи.

Под информационными задачами мы понимаем задачи классификации[1] и рубрикации. Решение такой задачи заключается в определении раздела информационного хранилища, куда следует поместить рассматриваемый документ. Об отклонении документа речь не идёт.

Управленческие задачи связаны с принятием решений, влияющих на дальнейшие действия с документом и взаимодействие с его автором. Здесь следует выделить два класса задач:

1) задачи, решаемые редколлегиями научных журналов, программными комитетами конференций, диссертационными советами: определение соответствия тематике журнала или конференции, выбор рецензентов и оппонентов;
2) задачи управления персоналом в научных организациях: приём на работу, поручение наукоёмких заданий.

В задаче 1 речь идёт о принятии решения по документу: принять или отклонить. При этом предполагается, что тексты не рецензированы, научное и литературное качество текста неизвестно; единственная задача автоматической оценки – отправить текст компетентному рецензенту. Решение о выборе рецензентов на основании автоматической оценки тематики принимает человек (ответственный секретарь редакции, председатель программного комитета и т.д.), которому известны компетенции рецензентов.

В задаче 2 речь идёт о принятии решения относительно научного работника по результатам оценки его компетентности на основании научных публикаций. Поскольку качество текста автоматически не оценивается, то рассматриваются тексты, качество которых уже было оценено, т.е. публикации в рецензируемых журналах. На основании установленной компетенции могут приниматься различные решения: о приёме на работу; о назначении экспертом по материалам, присланным на экспертизу; о выборе оппонентов и т.д. Если объединить обе задачи (оценку тематики текста и оценку компетентности сотрудника на основании его публикаций), то в дополнение к оценке тематики текста мы получаем оценку компетентности возможных рецензентов этого текста по набору их публикаций, что позволяет автоматизировать выбор рецензентов.

Уточним два понятия, которые в этой области постоянно используются: *компетенция* и *компетентность*. Компетенция – это понятие, в общем случае не связанное с конкретным лицом. В работе [2] компетенция описывается как совокупность следующих характеристик:

1) наименование компетенции;
2) категория или область, к которой принадлежит компетенция;

[1] Читателей просим обратить внимание также на опубликованную в этом номере журнала статью Микони С.В., посвященную классификации. Классификация (и как результат, и как процесс) важна для построения онтологий в различных предметных областях, а полученный опыт представляет особый интерес для онтологии проектирования. Поэтому результаты исследований в этой области для редакции журнала будут всегда приоритетными. *Прим. ред.*

3) описание компетенции, объясняющее, что может знать, уметь или делать обладатель компетенции;
4) свидетельства компетенции, по наличию которых можно определить, обладает ли сотрудник данной компетенцией.

В нашем случае п.2 (категория компетенции) – это область научного знания; в дальнейшем области знания будем называть темами. Как правило, это сформировавшиеся области, и п.3 можно опустить: описание компетенции предполагается известным. В качестве свидетельств компетенции рассматриваются публикации сотрудника.

Компетентность – это отношение человек-компетенция, означающее, что человек владеет данной компетенцией.

Отношение документ-компетенция будем называть релевантностью: документ релевантен некоторой теме, если в нём говорится об этой теме.

Такое понимание различий между компетенцией и компетентностью вполне традиционно (см., например, [3]).

Является ли документ, релевантный некоторой компетенции, свидетельством компетенции автора? Если документ – статья, опубликованная в рецензируемом журнале, то да, является. Если документ не отрецензирован (статья, поступившая в редакцию журнала, или доклад, присланный на конференцию), - нет, не является.

Очевидно, что отмеченные два отношения – компетентность и релевантность – не являются бинарными, поскольку важно оценить не только их наличие, но и уровень (степень) компетентности или релевантности в некоторой непрерывной шкале. Соответственно, и различные свидетельства компетенции могут обладать различной степенью убедительности.

2 Оценка тематики текста и онтологии прикладных наук

Задача определения тематики документа представляет собой позиционирование его тематики в некотором тематическом пространстве. Представление тематического пространства в виде онтологий рассматривалось в ряде работ, посвящённых управлению персоналом [4, 5]. Предлагаемый подход к автоматизированному решению этой задачи впервые был изложен в работе [6]. Он разбивается на три подзадачи.

- Описание тематического пространства как онтологии области научного знания с учётом специфики онтологий прикладных наук.
- Разработка метода позиционирования научного текста в заданной онтологии на основе поиска в нём терминов, характерных для определённых фрагментов онтологии.
- Разработка автоматизированной системы, реализующей предлагаемый метод для онтологии наук об управлении.

Прежде, чем заняться конкретным решением первой подзадачи – несколько слов о предлагаемом принципе построения онтологий прикладных наук.

Онтологии научных областей знаний обычно строятся по таксономическому принципу, т.е. традиционному принципу классификации. Классификатор – это ориентированное дерево, где корень соответствует области знания в целом, две смежные невисячие вершины находятся в отношении «класс-подкласс», а документ – это лист дерева (висячая вершина), который с единственной смежной с ней вершиной находится в отношении «экземпляр-класс».

Важным принципом таксономической классификации является принцип однозначности:

Каждый объект классификации (статья, книга, заявка на грант и т.д.) должен находиться ровно в одной вершине дерева. Чем дальше от корня находится объект, тем точнее (подробнее) он охарактеризован.

Так устроены универсальные классификаторы (УДК, классификаторы РНФ, РФФИ и др.). Таксономическая традиция возникла в естественных науках, где принцип однозначного отнесения природных объектов к какому-либо классу существует несколько столетий и, в общем, себя оправдывает. С классификацией текстов, о чём идет речь в данной работе, дело обстоит сложнее, поскольку в одном документе могут рассматриваться несколько тем. В ряде случаев однозначность классификации вынуждена: принятая статья может быть отнесена только к одной рубрике журнала, а принятый доклад включается в программу только одной секции конференции. Однако для принятия однозначного решения необходима информация обо всех возможных альтернативах.

В прикладных науках неоднозначность классификации научных текстов становится принципиальной. Статья, присланная на конференцию и посвящённая анализу данных, содержит определённый математический аппарат (например, методы математической статистики) и при этом говорит об использовании описанных результатов в различных прикладных областях: финансы, экология и т. д. Это означает, что для оценки данной статьи могут потребоваться три компетенции: в определённом разделе фундаментальной науки (математики, физики и т. д.), проблемной области (анализе данных) и в области приложений. Потребуются ли они в действительности – зависит от степени релевантности статьи этим компетенциям. Соответственно, задача автоматизированного выбора рецензентов должна заключаться в поиске наилучшего совпадения набора релевантностей документа с набором компетентностей рецензента.

С учётом этих соображений предлагается следующий подход к принципам построения онтологий прикладных наук.

- Строение онтологии в виде дерева в основном сохраняется (возможные локальные нарушения древовидности связаны с многозначностью терминов – см. ниже).
- С корнем дерева онтологии всегда смежны три вершины. Их имена всегда одинаковы: «фундаментальная наука», «прикладная теория» или «проблемная область», «область приложений». Эти вершины в дальнейшем будем называть главными вершинами, а поддеревья, корнями которых они являются, - главными поддеревьями.
- Все вершины дерева онтологии разбиваются на два вида: вершины-темы и вершины-термины. Вершины-темы образуют основной каркас дерева и связаны между собой отношением тема-подтема, имеющим все таксономические свойства отношений типа класс-подкласс. Вершина-термин связана с темой отношением термин-тема, которое является отношением экземпляр-класс. Все нижележащие вершины наследуют этот термин. Висячими вершинами являются все вершины-термины и только они.
- Экземплярами классов (висячими вершинами, листьями) являются термины, относящиеся к данному классу (тематическому разделу). Термин может принадлежать нескольким темам в разных ветвях дерева. В этом случае древовидность онтологии формально нарушается.

Сам документ в онтологии не хранится и экземпляром не является. При этом принцип однозначности отсутствует: в тематическом пространстве документ может относиться к разным ветвям дерева, т.е. принадлежать классам, находящимся на разных ветвях в разных поддеревьях, с разным весом, который отражает степень релевантности документа данному классу.

На рисунке 1 приведён малый фрагмент онтологии теории игр, где термины изображены пунктирными прямоугольниками, отношения тема-подтема – сплошными стрелками, а

отношения тема-термин – пунктирными стрелками. Отметим, что термин в общем случае может принадлежать нескольким темам. Здесь возможны два семантически различных случая. В первом случае термин принадлежит темам из разных главных поддеревьев. Например, термин «Передача данных» принадлежит как фундаментальной теории «Теория информации», так и прикладной проблемной области «Передача и обработка сигналов». Во втором случае термин в разных контекстах имеет разный смысл. Таков, например, термин «устойчивость», который используется в разных прикладных теориях и фактически представляет собой множество разных терминов, названных одним и тем же словом. В обоих случаях нарушается однозначность принадлежности термина теме и, соответственно, – древовидность онтологии.

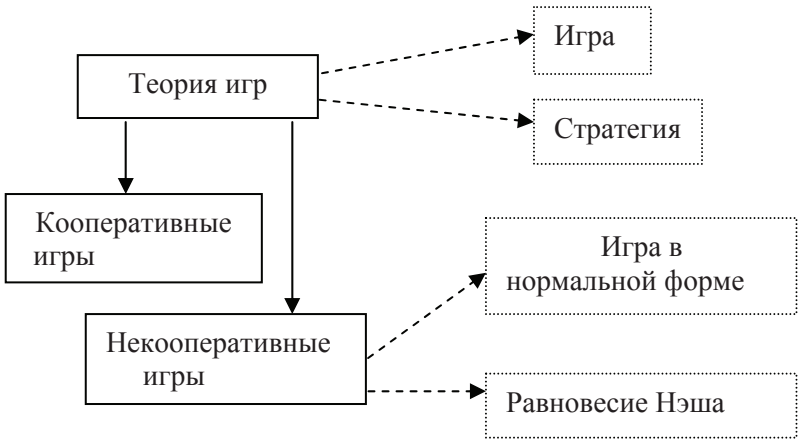

Рисунок 1 – Фрагмент онтологии теории игр

3 Онтология наук об управлении

Структуру создаваемой онтологии определяет круг задач, которые предполагается решать с её помощью. Онтологии, ориентированные на детальное описание предметной области (ПрО), имеют значительную глубину: длины путей от корня к листьям довольно велики. Разработанная в данной работе онтология наук об управлении предназначена для оценки тематики научных статей об управлении, присылаемых в журнал «Автоматика и телемеханика», с целью определения наиболее подходящих по компетентности кураторов этих статей – членов редколлегии, которые назначают рецензентов и принимают решения о судьбе статьи. Куратор не обязан иметь узкую компетенцию, совпадающую с тематикой статьи; однако он способен определить, кто из известных ему рецензентов этой компетенцией обладает, оценить аргументацию рецензентов и по итогам рецензирования принять решение о приёме или отклонении статьи. Для решения этой задачи большая глубина онтологии не нужна.

Построению онтологии предшествовало изучение архива статей журнала «Автоматика и телемеханика» за несколько последних лет. Сначала путём «ручного» извлечения терминов из статей журнала был создан словарь терминов, представляющий собой таблицу, столбцы которой соответствуют трём главным деревьям, описанным выше. Отнесение терминов к конкретным темам производилось экспертным образом. Фрагмент словаря приведён в таблице 1.

При формировании словаря важно было отбирать термины, специфические для конкретной темы, поэтому такие «безликие» термины, как «функция», «система», «уравнение», встречающиеся практически в каждой статье, бессмысленно включать в

словарь в качестве однословных терминов. Эти слова должны входить в состав многословных терминов: «функция Ляпунова», «уравнение Лагранжа первого рода». Более тонкие ситуации, возникающие при поиске терминов, - это наличие слов, которые в одних контекстах являются терминами, а в других – нет. Примеры: «игра», «множество». Такие коллизии решаются, как правило, экспертным путём.

Таблица 1 – Фрагмент словаря терминов с привязкой их к теориям и сферам применения

Термин	**Фундаментальные теории**	**Прикладные теории**	**Области применения**
летательный аппарат		Управление движением	Авиация
линейные матричные неравенства	Линейная алгебра	Теория автоматического управления	
орграф	Теория графов		
отношения предпочтения	Теория множеств и отношений	Теория выбора и принятия решений	
передача данных	Теория информации	Передача и обработка сигналов	Коммуникационные системы и сети

В результате анализа словаря был получен список тем, который затем был преобразован в дерево онтологии. На рисунке 2 показан фрагмент верхней части дерева онтологии. Уже на третьем уровне иерархии нет возможности показать все вершины. Под «Прикладными теориями», кроме показанных на рисунке, располагаются вершины «Автоматизация проектирования», «Анализ данных», «Искусственный интеллект», «Передача и обработка сигналов» и др., под «Областями приложений» – вершины «Социально-экономические системы», «Бизнес и финансы», «Технологические процессы», «Робототехника», и др.

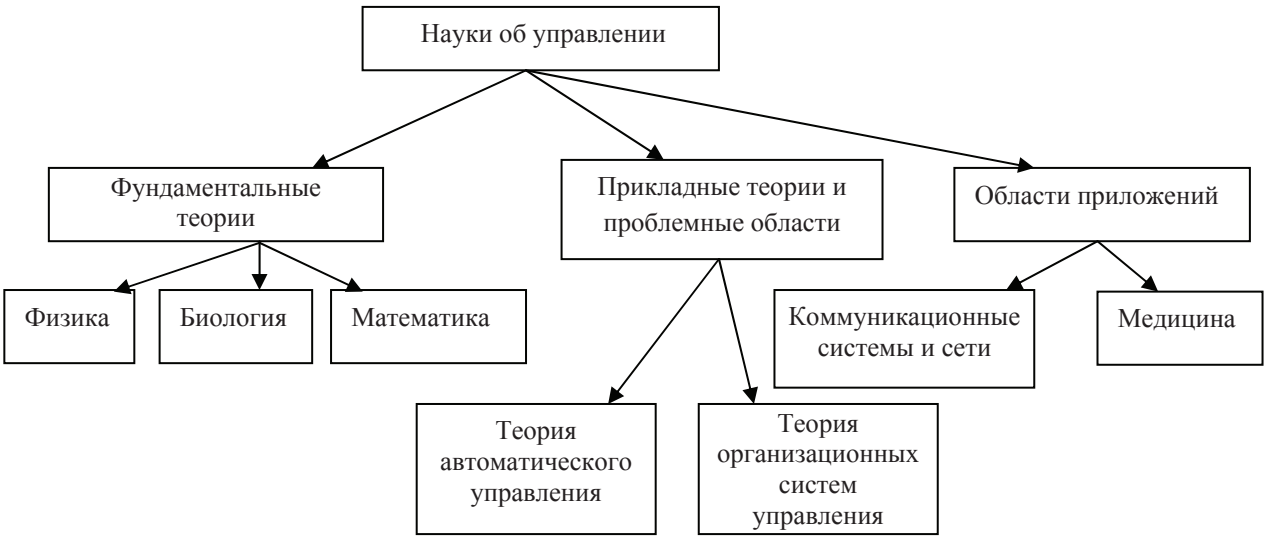

Рисунок 2 – Фрагмент верхней части дерева онтологии наук об управлении

В текущей версии нельзя говорить о полноте словаря и тематики: их расширение будет постоянно происходить в ходе опытной эксплуатации и по мере привлечения специалистов в отдельных областях. Кроме того, постоянно появляются новые теории, области приложений и соответствующая терминология. Поэтому как словарь, так и онтология должны в принципе быть постоянно открытыми к расширению и модификации.

4 Метод анализа тематики текста

Предлагаемый метод анализа тематики научного текста на основе онтологии, организованной указанным образом, состоит из двух этапов.

Первый этап заключается в последовательном поиске всех терминов онтологии в анализируемом документе, подсчёте числа вхождений терминов с учётом их принадлежности различным темам онтологии. При этом приходится решать лексические проблемы, традиционные для систем обработки текстов естественного языка. Одна из таких проблем – это проблема морфологии, которая состоит в том, что в стандартных словарях слова представлены в лемматизированной, т.е. «нормальной» форме (существительные, прилагательные, причастия – именительный падеж, единственное число; глагол – инфинитив), тогда как в текстах эти слова могут встречаться в различных словоформах. Эта проблема усложняется для многословных терминов типа «целочисленное линейное программирование».

В текущей версии системы эта проблема решается с помощью регулярных выражений — формального языка для записи, поиска и манипуляций с подстроками текста [7]. Для каждого термина (однословного или многословного) формируется и записывается в словарь регулярное выражение, содержащее все его возможные словоформы.

Пример: трёхсловному термину «целочисленное линейное программирование» соответствует регулярное выражение, содержащее все его возможные словоформы:
(целочисленн((ое)|(ого)|(ому)|(ым)|(ом))(\s|(\s\s))
линейн((ое)|(ого)|(ому)|(ым)|(ом))(\s|(\s\s))
программировани((е)|(я)|(ю)|(ем)|(и))).

На втором этапе на основе найденных терминов и их принадлежности темам формируется вектор релевантностей документа всем темам онтологии. Этот вектор называется *профилем документа*. Поскольку тем в онтологии много, а ненулевых релевантностей намного меньше, в качестве профиля документа можно рассматривать вектор ненулевых релевантностей. Профиль является результатом многомерной классификации документа.

Профиль сотрудника можно определить как профиль множества его публикаций. Вопрос о том, как конкретно формировать этот профиль из профилей публикаций, здесь не рассматривается.

Для описания алгоритма формирования профиля $P(d)$ документа d введём ряд понятий. Будем считать, что онтология содержит N тем, которые имеют фиксированные номера; в документе найдено K_d вхождений терминов; все они пронумерованы.

Значимость s^i_m *i-го термина* в m-й теме определим следующим образом:

$s^i_m = 0$, если i-й термин не принадлежит m-й теме; иначе

$s^i_m = 1/c_i$, где c_i – число тем, которым принадлежит i-й термин.

Таким образом, принадлежность термина нескольким темам уменьшает его вероятность принадлежности m-й теме и вносит неопределённость в его классификацию. Эта неопределённость будет разрешаться за счёт других терминов, возможно, экспертным путём.

Значимость всех найденных терминов для m-й темы определяется как $S_m = \sum^{Kd}_{i=1} s^i_m$.

Суммарная значимость S_d всех найденных терминов документа d для всех тем онтологии равна $S_d = \sum^{N}_{j=1} S_j$.

Тогда *релевантность* $R_m(d)$ документа d m-й теме определяется как сумма значимостей всех терминов m-й теме, нормированная по S_d :

$$R_m(d) = S_m/S_d .$$

Профиль $P(d) = (R_1(d), R_2(d), \ldots, R_N(d))$.

Проиллюстрируем приведённый алгоритм примером.

Пусть в документе *d* найдено 15 терминов. Термины {1, 2, …, 10} принадлежат теме 1, термины {9,…, 15} принадлежат теме 2. Термины {6, 7, …, 10} повторились 4 раза; остальные – 2 раза.

Тогда $s^1{}_1 = \ldots = s^8{}_1 = 1$; $s^9{}_1 = s^{10}{}_1 = 1/2$; $s^{11}{}_1 = \ldots = s^{15}{}_1 = 0$;

$s^1{}_2 = \ldots s^7{}_2 = s^8{}_2 = 0$; $s^9{}_2 = s^{10}{}_2 = 1/2$; $s^{11}{}_2 = \ldots = s^{15}{}_2 = 1$.

$S_1 = (s^1{}_1 + \ldots + s^5{}_1)\cdot 2 + (s^6{}_1 + \ldots + s^{10}{}_1)\cdot 4 = 5\cdot 2 + 3\cdot 4 + (0{,}5 + 0{,}5)\cdot 4 = 26.$

$S_2 = (s^9{}_1 + s^{10}{}_1)\cdot 4 + (s^{11}{}_1 + \ldots + s^{15}{}_1)\cdot 2 = (0{,}5 + 0{,}5)\cdot 4 + 5\cdot 2 = 14.$

$R_1(d) = 26/(26+14) = 0{,}65$; $R_2(d) = 14/(26+14) = 0{,}35$; $P(d) = (0{,}65;\ 0{,}35;\ 0;\ldots;\ 0)$.

При таком нормировании сумма релевантностей документа всегда равна 1. Однако полученные релевантности не нужно интерпретировать как вероятности принадлежности теме – напомним, что наша классификация многомерна, и предполагается, что документ может принадлежать разным темам. Величины релевантностей отражают предпочтительность компетенций предполагаемых рецензентов или экспертов.

5 Программная реализация

При разработке онтологии использовались программные средства Protégé [8]. Для представления онтологии в виде, удобном для совместной работы с программной системой, был выбран формат книги Excel, которая содержит таблицу-рубрикатор тем онтологии и таблицу-словарь терминов ПрО.

Рисунок 3 – Блок-схема программной системы определения профиля документа

Программный комплекс определения профиля документа содержит программу на языке Python, выполняющую поиск терминов в документе и вычисление количественной оценки тематики текста; экранный интерфейс пользователя и книгу Excel (рисунок 3).

Как уже отмечалось выше, словарь ПрО содержит лемматизированные термины на естественном языке, причём каждому из них соответствует единственное регулярное выражение [7], которое, в свою очередь, содержит все возможные словоформы термина (однословного или многословного).

Система тестировалась на текстовых вариантах файлов статей различных рубрик из архива журнала «Автоматика и телемеханика».

Одна из тестируемых рубрик – «Системы массового обслуживания». На рисунке 4 показано окно рабочей программы с результатом вычисления релевантностей (в процентах) текста статьи темам онтологии «Науки об управлении».

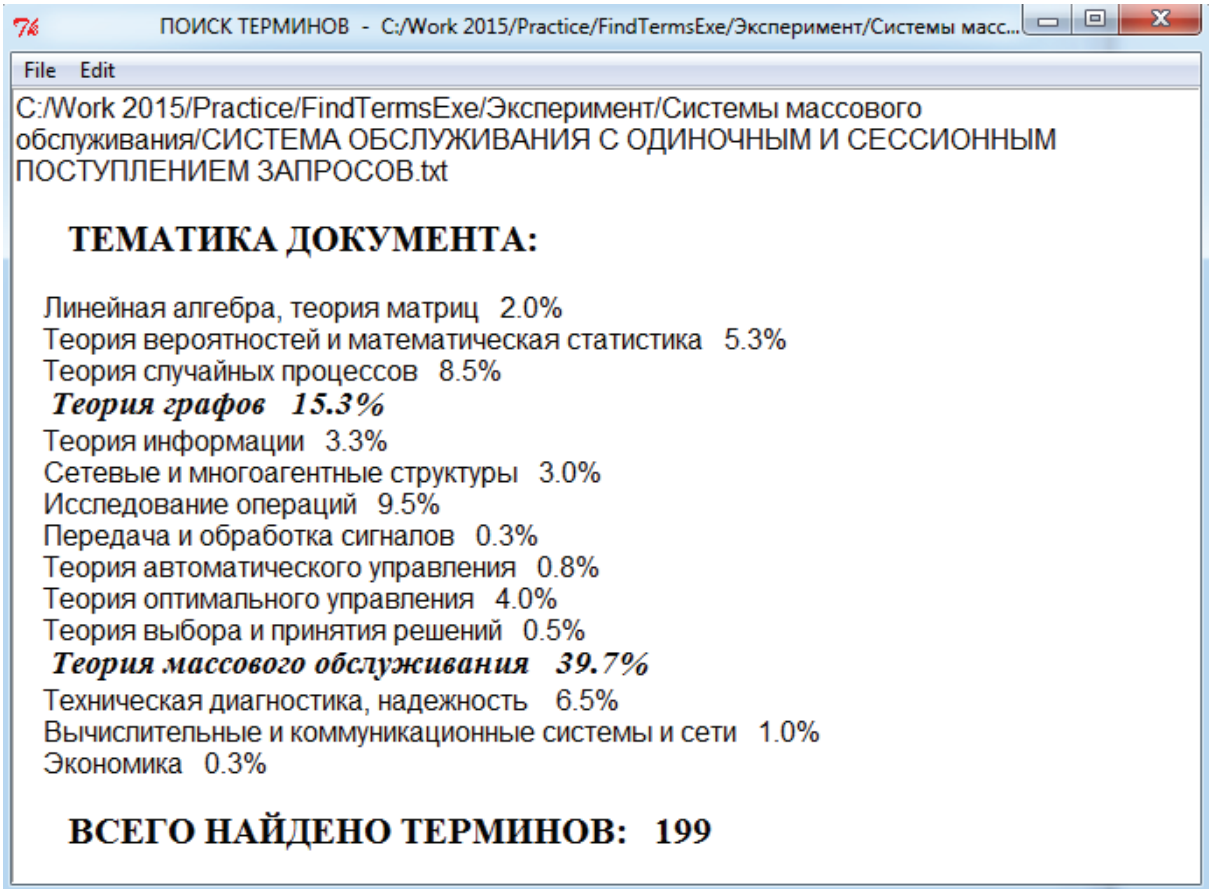

Рисунок 4 – Окно рабочей программы с результатом вычисления релевантностей текста статьи заданным темам

Жирным курсивом на рисунке выделены тематические разделы, для которых релевантность текста превышает 10%.

Другая тестируемая рубрика – «Теория линейных систем». Она была выбрана не случайно: ей во многом соответствует терминология тематики онтологии «Теория автоматического управления», содержащая более 100 терминов. В тестовых примерах объёмом 16-62 kB количество найденных терминов обычно превышало 50 единиц (рисунок 5), а в некоторых случаях превосходило 100 при объёме словаря менее 700 терминов. Но были и исключения, когда документ содержал менее 50 терминов; при этом в нём превалировали математические выкладки и рисунки-схемы.

В среднем прослеживается тенденция роста числа найденных терминов с увеличением объёма текста, хотя количество найденных терминов от документа к документу сильно изменяется (рисунок 5). Это обусловлено рядом причин: наполнение тем терминами неравномерно – разные темы содержат различное число терминов; объёмы статей также могут сильно различаться; статьи различаются и соотношением объёмов текста и математических выкладок; кроме того, текущая версия словаря пока не гарантирует полноты терминологии для конкретных тем онтологии.

При обработке данных тестирования статей рубрики «Теория линейных систем» были выделены кортежи тем с наиболее высокими значениями релевантности. Для них были вычислены суммарные значения по всем обработанным статьям (рисунок 6). Оказалось, что основной вклад – 73,5% - в профили статей этой рубрики вносят термины трёх тем онтологии: «Теория автоматического управления», «Теория устойчивости управляемых систем» и «Теория оптимальных систем». На графе онтологии ПрО каждая из вершин,

представляющих две последние тематики, является смежной с вершиной первой тематики «Теория автоматического управления».

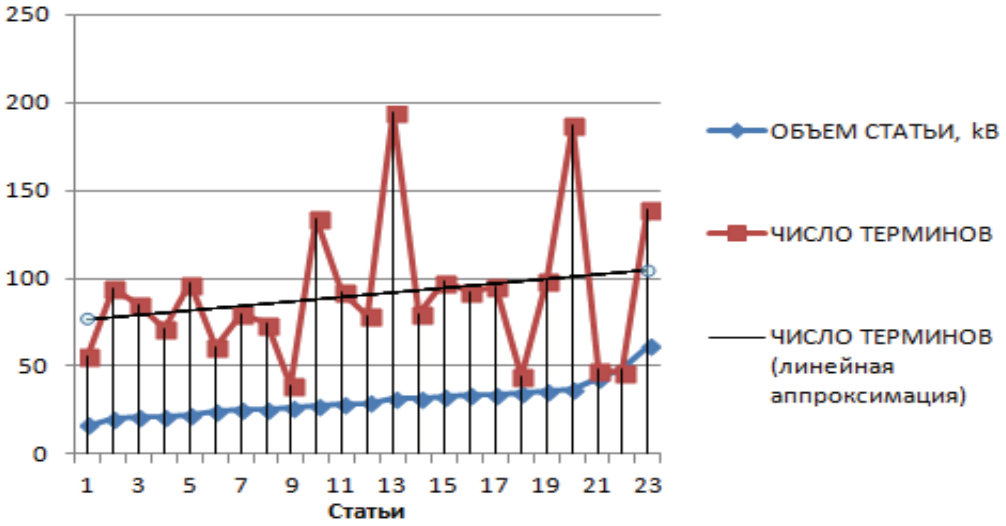

Рисунок 5 – Результаты тестирования рубрики «Теория линейных систем»

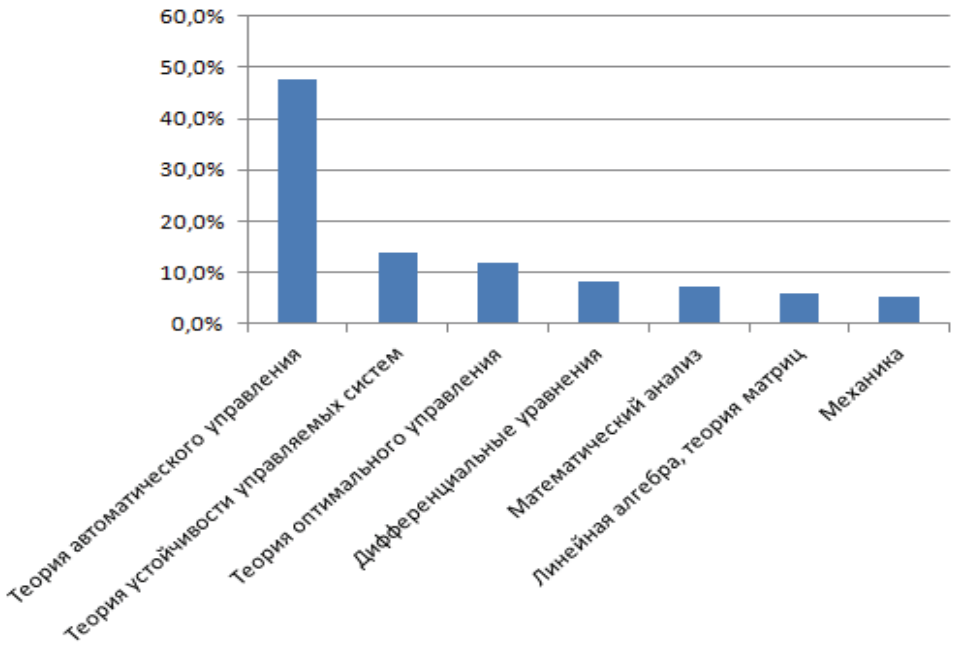

Рисунок 6 – Соотношение суммарных показателей релевантности статей из рубрики «Теория линейных систем» тематикам онтологии ПрО

Заключение

Качество работы автоматизированной системы существенно зависит от корректности и полноты онтологии, полноты словаря и равномерности распределения терминов словаря по темам. Для текущей версии полнота словаря особенно актуальна. Для её обеспечения необходима серьёзная работа со специалистами в соответствующих областях.

Возможно развитие более сложных оценок профиля документа, учитывающих гипонимические отношения между тематиками в онтологии и весовые коэффициенты значимости терминов, установленные экспертами. Точности оценки профиля документа может способствовать привязка релевантностей документа к структурным разделам содержательной части текста и получение раздельных оценок по каждому из них.

Реально создание дополнительных программ для автоматизации преобразования терминов на естественном языке в регулярные выражения при составлении и пополнении словаря. В перспективе представляется интересным увеличение словаря за счёт регулярных выражений, представляющих собой лексические обороты, семантически соответствующие определённой теме. Отдельный интерес представляет накопление в ходе эксплуатации системы статистики встречаемости терминов и её использование для уточнения профиля документа.

Список источников

[1] ***Ильвовский, Д.*** Системы автоматической обработки текстов / Д. Ильвовский, Е. Черняк // Открытые системы. 2014, №1. - С. 51-53.

[2] ***Draganidis F., Mentzas G.*** Competency based management: a review of systems and approaches/ Information Management & Computer Security, 2006, Vol. 14, No. 1, P. 51-64.

[3] ***Занина, Л.В., Меньшикова Н.П.*** Основы педагогического мастерства. – Ростов-на Дону: Феникс, 2003. – 288 с.

[4] ***Euzenat J., Shvaiko P.*** Ontology matching. – Springer-Verlag Berlin Heidelberg, 2007. - 332 p.

[5] ***Rogushina J., Gladun A.*** Ontology-based competency analyses in new research domains / Journal of Computing and Information Technology. V.20, N. 4, 2012. – P.277-293.

[6] ***Крюков, К.В.*** О понятии формальной компетентности научных сотрудников / К.В. Крюков, О.П. Кузнецов, В.С. Суховеров // Материалы III-й международной научно-технической конференции (OSTIS-2013, Минск). Минск, УО "Белорусский государственный университет информатики и радиоэлектроники" (БГУИР), 2013. - С.159-162.

[7] ***Фридл Дж.*** Регулярные выражения, 3-е издание. – Пер. с англ. – СПб.: Символ_Плюс, 2008. – 608 с.

[8] ***Smith M.K., Welty C., McGuinness D.L.*** OWL Web Ontology Language Guide, 2004. - https://www.w3.org/TR/2004/REC-owl-guide-20040210/

A ONTOLOGICAL APPROACH TO DETERMINING THE SUBJECT MATTER OF SCIENTIFIC TEXT

O.P. Kuznetsov[1], V.S. Sukhoverov[2]

Institute of Control Sciences of Russian Academy of Sciences, Moscow, Russia

[1]olpkuz@yandex.ru, [2]suhoverv@ipu.ru

Abstract

The paper proposes an approach to determination the subject matter of scientific text, using the domain ontology. It presents the original principle of building applied sciences ontologies, in which the ontology tree contains three required branches: the "Fundamental theory", "Applied theory", "Applications". Topics are classes of the ontology and terms of the matching topics are instances of classes. We describe the structure of the control theory ontology and present

fragments of this ontology and its dictionary. It is assumed that the scientific text relevant to the topic, if it contains terms of this topic. A method is proposed for evaluating of the degree of scientific text relevance to different topics based on counting the number of the terms occurrences of these topics in the document. The result of this method is the "the document profile" – vector of document relevance to ontology topics. We describe the automatic system for analysis the subjects of scientific texts from the field of the control theory and practice developed on the basis of the proposed approach. The linguistic problems of the terms search are considered. We present some statistics after processing of thematic sections of the journal "Automation and Remote Control" and an example of building a profile for selected article. Possible directions to improve the estimates of relevance are noted.

Key words: *ontology, competence, relevance, topic, term, document profile, control theory.*

Citation: Kuznetsov O.P., Sukhoverov V.S. A ontological approach to determining the subject matter of scientific text. *Ontology of designing*. 2016; 6(1): 55-66. DOI: 10.18287/2223-9537-2016-6-1-55-66.

References

[1] ***Il'vovskij D., Chernyak E.*** Sistemy avtomaticheskoi obrabotki tekstov [Systems of automatic processing of texts] //Otkrytye sistemy. 2014, №1: 51-53. (In Russian).

[2] ***Draganidis F., Mentzas G.*** Competency based management: a review of systems and approaches/ Information Management & Computer Security, 2006, Vol. 14, No.1: 51-64.

[3] ***Zanina LV., Men'shikova NP.*** Osnovy' pedagogicheskogo masterstva. [Fundamentals of pedagogical skills] – Rostov-na Donu: Feniks, 2003.

[4] ***Euzenat J., Shvaiko P.*** Ontology matching. – Springer-Verlag Berlin Heidelberg, 2007. 332 p.

[5] ***Rogushina J., Gladun A.*** Ontology-based competency analyses in new research domains / Journal of Computing and Information Technology. V.20, N. 4, 2012: 277-293.

[6] ***Kryukov KV., Kuznetsov OP., Sukhoverov VS.*** O ponyatii formal'noj kompetentnosti nauchnykh sotrudnikov [On the notion of formal competence of research staff] / Materialy III-j mezhdunarodnoj nauchno-tekhnicheskoj konferentsii (OSTIS-2013, Minsk). Minsk, UO "Belorusskij gosudarstvennyj universitet informatiki i radioelektroniki" (BGUIR), 2013: 159-162. (In Russian).

[7] ***Friedl Jeffrey EF.*** Mastering Regular Expressions, 3rd Edition – O'Reilly Media, 2006.

[8] ***Smith MK., Welty C., McGuinness DL.*** OWL Web Ontology Language Guide, 2004. - https://www.w3.org/TR/2004/REC-owl-guide-20040210/

Сведения об авторах

Кузнецов Олег Петрович. 1936 г. рождения. Окончил МГУ им. М.В.Ломоносова (1958 – философский факультет, 1966 - механико-математический факультет). К.т.н. – 1965, д.т.н. -1983, проф. 1998. Заведующий лабораторией Института проблем управления им. В.А.Трапезникова РАН. Председатель Научного Совета Российской Ассоциации искусственного интеллекта. Член редколлегий журналов «Автоматика и телемеханика», «Проблемы управления», «Искусственный интеллект и принятие решений». Автор 143 статей и 3 монографий.

Kuznetsov Oleg Petrovich (b.1936) graduated from Moscow State University in 1958 (philosophy department) and 1966 (mathematics and mechanics department). Head of department at Institute of Control Sciences V.A.Trapeznikov Academy of Sciences. Chairman of Scientific Council of Russian Association of Artificial Intelligence. He is author 143 scientific articles and 3 monographies.

Суховеров Виктор Степанович. 1945 г. рождения. Окончил МЛТИ (1970 – факультет электроники и счетно-решающей техники). К.т.н. – 2004. Старший научный сотрудник лаборатории методов интеллектуализации дискретных процессов и систем управления Института проблем управления им. В.А.Трапезникова РАН. Автор более 50 статей.

Sukhoverov Victor Stepanovich (b. 1945) graduated from the Moscow State Forest University Faculty of Computer Sciences in 1970, PhD (2004). Senior researcher of Laboratory of Intellectualization Methods of Discrete Processes and Control Systems – V. A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences. He is author and co-author of more than 50 publications.