ИНЖИНИРИНГ ОНТОЛОГИЙ

УДК 004.83

DOI: 10.18287/2223-9537-2021-11-3-351-363

Подход к автоматическому построению лингвистической онтологии для определения интересов пользователей социальных сетей**А.М. Наместников, Н.Д. Пирогова, А.А. Филиппов***Ульяновский государственный технический университет, Ульяновск, Россия***Аннотация**

Социальные сети предоставляют исследователям возможности получения массива текстовых данных для дальнейшего анализа в рамках некоторой предметной области. Каждая предметная область имеет свой специфический профессиональный словарь и стиль написания текста. При определении предметной области текстового материала большую проблему представляет построение словарей, тезаурусов, онтологий. В данной статье под онтологией рассматривается лингвистическая онтология, направленная на определение предметной области текстового материала. Представлен алгоритм для автоматического построения онтологии на основе графа знаний *Wikidata*. Задача состоит в том, чтобы отобразить множество объектов графа знаний *Wikidata* на множество сущностей лингвистической онтологии. В статье предложен алгоритм определения степени принадлежности текстового материала предметной области. Эксперименты по оценке времени построения онтологии и применимости полученных лингвистических онтологий к задаче определения степени принадлежности текстовых материалов предметной области показали: время работы алгоритма и количество терминов в сформированной онтологии прямо пропорционально зависят от количества анализируемых свойств и объектов *Wikidata*; сформированная лингвистическая онтология применима к задаче определения степени принадлежности текста предметной области.

Ключевые слова: лингвистическая онтология, автоматизация, *Wikidata*, текстовый документ, предметная область.

Цитирование: Наместников, А.М. Подход к автоматическому построению лингвистической онтологии для определения интересов пользователей социальных сетей / А.М. Наместников, Н.Д. Пирогова, А.А. Филиппов // Онтология проектирования. – 2021. – Т.11, №3(41). – С.351-363. – DOI: 10.18287/2223-9537-2021-11-3-351-363.

Введение

На данный момент социальные сети предоставляют исследователям возможность получения массива текстовых данных для дальнейшего анализа в рамках некоторой предметной области (ПрО). Каждая ПрО содержит данные различного характера, имеет свой профессиональный словарь и стиль написания текста.

Например, при составлении личностного портрета пользователя социальной сети наибольший объём информации может быть получен из текстовой информации с его страницы. Смысловое содержание текстовых данных определяет тему данного материала, т.е. его ПрО. Определение ПрО текстового материала представляет собой трудоёмкую задачу и включает построение словарей, тезаурусов, онтологий. Эта работа по большей части выполняется вручную.

Известны два подхода к созданию и исследованию онтологий. Первый (формальный) основан на логике (предикатов первого порядка, дескриптивной, модальной и т.п.). Вторым

(лингвистический) основан на изучении естественного языка (в частности, семантики) и построении онтологий на больших текстовых массивах, т.н. корпусах.

В данной статье рассматривается применение лингвистической онтологии в задачах определения ПрО текстового материала. Существует несколько методов автоматического построения лингвистических онтологий.

- *Автоматическое построение онтологии по коллекции текстовых документов.* Данный подход описан Е.С. Мозжериной [1]. В статье обосновывается подход к автоматизации процесса построения онтологии по коллекции текстовых документов, относящихся к одной тематике, на основании статистических методов анализа текстов на естественном языке. Предполагается, что термины и некоторые базовые отношения между ними могут быть выделены автоматически из коллекции текстовых документов на основании статистических данных [2]. В данной статье рассматриваются первые два этапа построения онтологии: выделение классов и отношений между ними. Выделение классов из текстов на естественном языке сводится к определению терминов рассматриваемой ПрО.
- *Подход на основе лексико-синтаксических шаблонов.* Данный подход относится к группе методов автоматического построения онтологий, использующих лингвистические средства [3]. Для построения онтологий используются все уровни анализа естественного языка: морфология, синтаксис и семантика. Для автоматического построения онтологии используется один из методов семантического анализа текстов на естественном языке – лексико-синтаксические шаблоны, которые представляют собой характерные выражения и конструкции определённых элементов языка. Данная методика семантического анализа не является специализированной для определённой ПрО.
- *Автоматическое построение онтологий на основе машинного обучения* [4]. Для этого разрабатываются модели: генерации системы продукций (на основе применения генетического программирования); генерации преобразователей (на основе генетического и автоматного программирования); генерации систем логического вывода (также на основе генетического и автоматного программирования); аппарата активации продукций (на основе применения автоматного программирования).
- *Автоматическое построение онтологий на основе общедоступных тезаурусов и графов знаний.* В работе [5] представлен метод формирования лингвистической онтологии на основе тезауруса *WordNet*¹, в котором учитываются отношения между сущностями тезауруса. Основы открытых для пользователей графов знаний реализованы в 2007 г. в базе знаний *DBpedia*, созданной в результате семантической обработки статей *Wikipedia*. Со временем в *DBpedia* были добавлены подробные схемы данных (онтология), географические данные и связи с другими графами [6]. В настоящее время *DBpedia* считается одним из стандартов графов знаний и содержит более 6 млрд. связанных фактов. В 2008 г. был разработан граф *YAGO* [7]. Его отличительная особенность состоит в использовании семантического тезауруса *WordNet* и детальной иерархии классов сущностей. В настоящее время *YAGO* содержит около 120 млн. фактов. В 2010 г. была запущена система *Never-Ending Language Learner*, которая автоматически выделяет факты из текста веб-страниц. В настоящее время *Never-Ending Language Learner* содержит около 50 млн. фактов. Запущенный в 2007 г. граф знаний *Freebase* позволяет пользователям самим назначать категорию описываемой сущности. В настоящее время *Freebase* преобразован в *Google Knowledge Graph* [8]. Граф *Wikidata* предназначен для хранения знаний в *Wikipedia* на различных языках [9]. В большинстве публикуемых графов знаний используется модель *Wikidata* или связываются свои сущности с имеющимися в *Wikidata* [10].

¹ <https://wordnet.princeton.edu/>

При решении задачи анализа предпочтений пользователей социальных сетей нет необходимости формировать лингвистическую онтологию сложной структуры, т.к. достаточно сопоставить отдельные термины текстовых ресурсов с признаками, описывающими классы интересов пользователя.

В работе рассмотрен подход к автоматическому построению онтологии на основе графа знаний *Wikidata* для определения предпочтений пользователя социальной сети.

1 Алгоритм автоматического построения лингвистической онтологии на основе графа знаний *Wikidata*

Данные в *Wikidata* структурированы в виде множества сущностей, у каждой сущности есть страница. На данный момент в системе имеется два типа сущностей: объекты и свойства. В терминах онтологии объекты представляют индивидуальности и классы, а свойства *Wikidata* напоминают свойства *RDF* [11]. Например, доступ к странице объекта для представления английского писателя Дугласа Адамса можно получить по адресу <https://www.wikidata.org/wiki/Q42> (рисунок 1).

Рисунок 1 – Пример страницы с описанием объекта в графе знаний *Wikidata*

Идентификатор данной страницы – «Q42», поскольку *Wikidata* является многоязычным сайтом. Следовательно, объекты идентифицируются не меткой на определённом языке, а непрозрачным идентификатором, который назначается автоматически при создании объекта и не может быть изменён позже. Страница каждого объекта содержит следующие основные части:

- метка или наименование (например, «Дуглас Адамс»),
- краткое описание (например, «английский писатель и юморист»),
- список псевдонимов (например, «Дуглас Ноэль Адамс»),
- список утверждений,

- список ссылок на страницы с информацией об объекте в Википедии и на других сайтах.

Метка, описание и псевдонимы вместе определяют множество терминов. С объектом могут быть связаны термины на любом языке, поддерживаемом *Wikidata*, а также другие объекты графа знаний. Например, с помощью свойства «*instance of*» объект «Q42» (Дугласа Адамса) связан с объектом «*human*».

Таким образом, граф знаний *Wikidata* можно представить как:

$$W = \{\langle \hat{O}_i, P_1, O_1 \rangle, \dots, \langle \hat{O}_i, P_j, O_j \rangle, \dots, \langle \hat{O}_i, P_n, O_n \rangle\},$$

где $\hat{O}_i$ – объект, страница с описанием которого открыта (текущий объект);

$P_j \in P^S \cup P^C$ – j -е свойство текущего объекта $\hat{O}_i$. Данное свойство может принадлежать множеству предопределённых (P^S) или созданных (P^C) в процессе формирования *Wikidata* свойств;

O_j – j -й объект, с которым текущий объект $\hat{O}_i$ связан с помощью свойства P_j .

Лингвистическую онтологию можно представить следующим образом:

$$D = \langle C, T, R \rangle,$$

где C – множество классов лингвистической онтологии. Каждый класс определяет некоторую ПрО, к которой может быть отнесён текстовый материал, например, музыка, политика, спорт и т. д.;

T – множество терминов лингвистической онтологии. Термины представляют собой признаки, присутствие которых в текстовом материале позволяет отнести такой материал к некоторой ПрО;

R – множество отношений между элементами онтологии вида:

$$R = \{R^C, R^T\},$$

где R^C – родовидовые отношения между классами онтологии;

R^T – функциональные отношения ассоциации между классами и терминами онтологии.

Задача автоматического построения онтологии состоит в том, чтобы отобразить множество объектов графа знаний *Wikidata* на множество сущностей лингвистической онтологии.

Разработан алгоритм, который на основе структуры классов лингвистической онтологии и настраиваемых параметров формирует множество запросов на языке *SPARQL* к *Wikidata Query Service* для извлечения терминов для каждого класса:

$$F(W, Args, Stop, \ddot{D}) \rightarrow \tilde{D},$$

где F – разработанный алгоритм;

W – граф знаний *Wikidata*;

$\ddot{D}$ – лингвистическая онтология, содержащая иерархию классов;

$Args$ – параметры алгоритма:

- количество свойств $Args^P$ – максимальное количество извлекаемых свойств для анализируемого объекта на каждом этапе итерации;
- количество объектов $Args^O$ – максимальное количество извлекаемых объектов для каждого анализируемого свойства на каждом этапе итерации;
- количество итераций $Args^I$ – максимальная глубина анализа свойств и объектов;

$Stop$ – словарь «стоп-свойств» ($Stop \subset P^S \cup P^C$). Данные свойства будут пропускаться в процессе анализа графа знаний *Wikidata*;

$\tilde{D}$ – лингвистическая онтология, наполненная терминами.

Алгоритм автоматического построения лингвистической онтологии на основе анализа графа знаний *Wikidata* можно представить в виде следующих шагов:

- 1) Формируется очередь классов лингвистической онтологии:

$$C \in \ddot{D}, \text{ например, музыка} \in C.$$

- 2) Следующий класс онтологии устанавливается в качестве текущего объекта для анализа:

$$C_i \rightarrow \hat{O}, C_i \in C, \text{ например, } \hat{O} = \text{музыка.}$$

- 3) Для текущего объекта формируется запрос на языке *SPARQL* к *Wikidata Query Service*:

```
select ?prop ?propLabelen ?propLabelru with {
  select ?prop (COUNT(?item) AS ?count) where {
    item ?p wd:  $\hat{O}$ .
    ?prop a wikibase:Property;
    wikibase:directClaim ?p.
  }
  group by ?prop
  ORDER BY DESC(?count)
  LIMIT  $Args^P$ 
} as %result where {
  include %result.
  SERVICE wikibase:label {
    bd:serviceParam wikibase:language "en".
    ?prop rdfs:label ?propLabelen.
  }
  SERVICE wikibase:label {
    bd:serviceParam wikibase:language "ru".
    ?prop rdfs:label ?propLabelru.
  }
}
```

После выполнения запроса извлекается массив свойств P текущего объекта $\hat{O}_i$ для дальнейшего анализа.

- 4) Если отдельно взятое свойство не содержится в словаре «стоп-свойств», то объекты, связанные данным свойством с текущим объектом извлекаются с помощью запроса:

```
select ?item ?itemLabelen ?itemLabelru where {
  ?item wdt:%s wd:%s.
  SERVICE wikibase:label {
    bd:serviceParam wikibase:language "en".
    ?item rdfs:label ?itemLabelen.
  }
  SERVICE wikibase:label {
    bd:serviceParam wikibase:language "ru".
    ?item rdfs:label ?itemLabelru.
  }
} LIMIT  $Args$ ,
```

$$\hat{P} \rightarrow O, \hat{P} \subset P, P \cap Stop = \emptyset.$$

Извлечённые объекты добавляются в очередь, и устанавливается специальный «флаг» для определения начала новой итерации, например:

$$\begin{aligned} &\{\langle \text{музыка, практикуется, пианист} \rangle, \\ &\langle \text{музыка, обладает свойством, музыкальный жанр} \rangle, \dots \} \rightarrow O \\ &O = \{\text{пианист, музыкальный жанр}\}. \end{aligned}$$

- 5) Объекты из очереди записываются в онтологию в виде терминов и связываются функциональным отношением ассоциации с текущим классом лингвистической онтологии:

$$\begin{aligned} O_j &\rightarrow T_j, \hat{P}_j \rightarrow T_j, T_j R^T \hat{O}, \text{ например,} \\ \text{пианист} &\in T, \text{ музыкальный жанр} \in T, \\ \text{пианист} &R^T \text{ музыка, музыкальный жанр} R^T \text{ музыка.} \end{aligned}$$

- 6) Если «флаг» начала новой итерации установлен, то в качестве текущего объекта устанавливается следующий объект из очереди, затем происходит переход к шагу 3, например, $\hat{O} = \text{пианист}$.

- 7) Если очередь объектов пуста или текущее количество итераций равно максимальному количеству итераций $Args^I$, то происходит переход к шагу 2.

В результате работы алгоритма будет получена онтология в формате *OWL*, в которой с классами будут связаны термины, например:

```
<ClassAssertion>
  <Class IRI="#музыка"/>
  <NamedIndividual IRI="#музыка"/>
</ClassAssertion>
<ClassAssertion>
  <Class IRI="#музыка"/>
  <NamedIndividual IRI="#пианист"/>
</ClassAssertion>
<ClassAssertion>
  <Class IRI="#музыка"/>
  <NamedIndividual IRI="#музыкальный_жанр"/>
</ClassAssertion>
```

2 Алгоритм определения степени принадлежности текстового материала к ПрО

Алгоритм определения степени принадлежности отмечается простотой по сравнению с представленными в работах [12, 13]. Для определения предпочтений пользователей достаточно для каждой области интересов задать непересекающееся множество признаков, описанных терминами текста на естественном языке.

Задача определения предпочтений сводится к классификации множества текстовых материалов пользователя:

$$D = \{d_1, d_2, \dots, d_n\}.$$

Задачей классификации является нахождение наиболее вероятной категории из множества классов онтологии C для текстового материала d_i . Предложенный метод классификации текстовых материалов основан на предположении, что тексты, относящиеся к одной категории, содержат одинаковые признаки (слова или словосочетания) [14]. Наличие или отсутствие таких признаков в текстовом материале показывает его принадлежность или непринадлежность к той или иной теме.

Каждый текстовый материал рассматривается в разрезе терминов:

$$d_i = \{T_1^d, T_2^d, \dots, T_k^d\}.$$

Решение об отнесении текстового материала d_i к категории C_j принимается на основе пересечения терминов материала и категории (класса онтологии):

$$T^d \cap T.$$

Метрика для расчёта степени соответствия текстового входа (пост, комментарий) категории имеет вид:

$$v_j = \frac{|T^d \cap T^C|}{|T^C|}, v = [0..1], \quad (1)$$

где $|T^d \cap T^C|$ – количество совпавших терминов текстового материала d_i и класса онтологии C_j соответственно;

$|T^C|$ – количество терминов, связанных отношением ассоциации с классом онтологии C_j .

В результате для каждого текстового материала формируется множество степеней его соответствия классам онтологии:

$$E = \langle v_1, v_2, \dots, v_l \rangle.$$

Для вычисления итогового значения степени принадлежности текстового материала d_i к конкретной категории интересов пользователя используется следующее выражение:

$$j = \max(E).$$

На рисунке 2 представлен пример лингвистической онтологии, используемой для определения категории текстового материала.

Для определения степени принадлежности текстового материала «Депутаты Госдумы приняли в первом чтении законопроект об изоляции российского сегмента Интернета» к некоторой категории необходимо:

- 1) Выполнить разделение текста на слова с последующей лемматизацией каждого слова.
- 2) Используя выражение 1 определить степень соответствия текстового входа каждой категории:

- $v_1 = \frac{0}{5} = 0$;
- $v_2 = \frac{1}{5} = 0.2$ (Интернет);
- $v_3 = \frac{3}{5} = 0.6$ (депутат, Госдума, законопроект).

Таким образом, данный текстовый материал со степенью принадлежности 0.6 относится к категории «Политика» и степенью принадлежности 0.2 к категории «ИТ». Следовательно, данный текстовый материал скорее всего относится к категории «Политика».

Рисунок 2 – Пример лингвистической онтологии

3 Эксперименты

Были проведены эксперименты для определения зависимости времени работы алгоритма формирования лингвистической онтологии и количества терминов в ней от значений параметров алгоритма: количества свойств, объектов, итераций.

Начальными и максимальными значениями для экспериментов были выбраны соответственно: количество свойств – 5 и 70, количество объектов – 5 и 350, количество итераций – 2 и 350. Было установлено, что при больших значениях время работы алгоритма и количество терминов в сформированной онтологии не изменяются.

Результаты эксперимента представлены на рисунках 3-5 и в таблице 1.

Рисунок 3 – Зависимость времени работы алгоритма и количества терминов в сформированной онтологии от количества свойств

Таблица 1 – Результаты экспериментов для определения зависимости времени работы алгоритма формирования лингвистической онтологии и количества терминов в ней от значений параметров алгоритма

Количество свойств	Количество объектов	Количество итераций	Время (ч:мм:сс)	Количество терминов
Предметная область «Политика»				
5	5	2	0:01:06	547
20	5	2	0:05:46	2472
40	5	2	0:06:39	2825
50	5	2	0:06:42	2913
70	5	2	0:07:35	3179
5	20	2	0:04:27	2802
5	40	2	0:05:29	2635
5	50	2	0:07:51	5727
5	70	2	0:08:37	6046
5	90	2	0:12:49	8505
5	100	2	0:13:22	9755
5	125	2	0:14:55	10760
5	150	2	0:16:46	11804
5	175	2	0:21:15	13191
5	200	2	0:26:00	14068
5	250	2	0:29:01	15736
5	300	2	0:31:16	17493
5	350	2	0:34:24	18818
5	5	5	0:01:41	970
5	5	10	0:02:05	1063
5	5	15	0:03:04	1374
5	5	25	0:10:26	3563
5	5	50	0:14:03	3760
5	5	75	0:16:55	4848
5	5	100	0:32:25	6167
5	5	125	0:35:59	7818
5	5	150	0:40:36	8975
5	5	200	0:44:00	11445
5	5	250	0:58:24	14442
5	5	300	1:00:18	17662
5	5	350	1:09:00	20428
Предметная область «Музыка»				
5	5	2	0:00:57	514
20	5	2	0:08:31	3234
40	5	2	0:10:37	4084
50	5	2	0:10:07	4107
70	5	2	0:11:02	4177
5	20	2	0:03:19	2241
5	40	2	0:06:02	4066
5	50	2	0:06:29	4439
5	70	2	0:09:43	6418
5	90	2	0:11:05	8120
5	100	2	0:13:39	8854
5	125	2	0:16:21	10422
5	150	2	0:20:46	11068
5	175	2	0:22:05	13101
5	200	2	0:23:46	14506
5	250	2	0:27:34	15853
5	300	2	0:28:34	17634
5	350	2	0:33:03	20134
5	5	5	0:01:08	552
5	5	10	0:01:35	783
5	5	15	0:02:23	1224
5	5	25	0:04:32	2286
5	5	50	0:06:31	2803
5	5	75	0:09:25	3865
5	5	100	0:15:35	6256
5	5	125	0:18:53	8324
5	5	150	0:21:15	9030
5	5	200	0:29:06	12403
5	5	250	0:33:52	13714
5	5	300	0:36:46	14719
5	5	350	0:51:58	21422

Рисунок 4 – Зависимость времени работы алгоритма и количества терминов в сформированной онтологии от количества объектов

Рисунок 5 – Зависимость времени работы алгоритма и количества терминов в сформированной онтологии от количества итераций

При увеличении значений любого из трёх настраиваемых параметров время работы алгоритма и количество терминов в сформированной онтологии увеличивается.

Были проведены эксперименты по применимости полученных лингвистических онтологий к задаче по определению степени принадлежности текстовых материалов к ПрО. Было сформировано восемь онтологий по четыре ПрО: политика, музыка, медицина, спорт. По каждой ПрО было сформировано по две онтологии: первая онтология была сформирована на основе пустого словаря «стоп-свойств», вторая - на основе словаря «стоп-свойств» с 34 значениями. Количество терминов в каждой лингвистической онтологии от 2 000 до 5 000. Для экспериментов было выбрано четыре текста по одному на каждую исследуемую ПрО. Результаты экспериментов представлены в таблице 2.

На основе результатов эксперимента удалось определить, что сформированная с помощью предложенного алгоритма лингвистическая онтология применима к задаче определения степени принадлежности текста к ПрО, так как максимальные полученные значения по степени принадлежности ПрО соответствуют ПрО текста.

Заключение

Рассмотрен подход к автоматическому построению лингвистической онтологии на основе графа знаний *Wikidata* для определения предпочтений пользователя социальной сети. Получаемая лингвистическая онтология имеет простую структуру: иерархия понятий, заданная

вручную, и множество непересекающихся терминов, описывающих понятия, извлечённые из графа знаний *Wikidata*.

Приведено формальное описание алгоритмов формирования онтологии и определения ПрО текстового материала. Параметры алгоритма позволяют регулировать количество терминов в онтологии и время её формирования.

Таблица 2 – Результаты экспериментов по применимости полученных лингвистических онтологий к задаче по определению степени принадлежности текстовых материалов ПрО

Пустой словарь		Словарь со значениями	
Текст на тему «Политика»			
Политика	56%	Политика	59%
Спорт	35%	Спорт	26%
Музыка	9%	Музыка	15%
Медицина	0%	Медицина	0%
Текст на тему «Музыка»			
Музыка	86%	Музыка	91%
Политика	10%	Политика	5%
Спорт	4%	Спорт	4%
Медицина	0%	Медицина	0%
Текст на тему «Медицина»			
Медицина	82%	Медицина	73%
Музыка	18%	Политика	18%
Политика	0%	Музыка	9%
Спорт	0%	Спорт	0%
Текст на тему «Спорт»			
Спорт	79%	Спорт	86%
Политика	13%	Музыка	5%
Музыка	6%	Политика	5%
Медицина	3%	Медицина	3%

Эксперименты для определения зависимости времени работы алгоритма формирования лингвистической онтологии и количества терминов в ней от значений параметров алгоритма (количеств свойств, объектов, итераций) показали, что при увеличении значений любого из трёх настраиваемых параметров время работы алгоритма и количество терминов в сформированной онтологии увеличивается. При значении максимального количества свойств больше 70 время работы алгоритма и количество терминов в сформированной онтологии не изменяются.

Эксперименты по применимости полученных лингвистических онтологий к задаче определения степени принадлежности текстовых материалов ПрО показали, что сформированная лингвистическая онтология применима к задаче определения степени принадлежности текста ПрО.

Благодарности

Исследование выполнено в рамках государственного задания № 075-00233-20-05 по проекту «Исследование интеллектуального предиктивного мультимодального анализа больших данных и извлечения знаний из разных источников».

Исследование выполнено при финансовой поддержке РФФИ и Правительства Ульяновской области в рамках научных проектов № 19-47-730003, 19-47-730005.

Список источников

- [1] **Мозжерина, Е.С.** Автоматическое построение онтологии по коллекции текстовых документов / Е.С. Мозжерина // Электронные библиотеки: Перспективные методы и технологии, электронные коллекции (RCDL). 2011. С.293-298.
- [2] **Ермаков, А.Е.** Автоматизация онтологического инжиниринга в системах извлечения знаний из текста / А.Е. Ермаков // Материалы конференции «Диалог». 2008. С.4-8.
- [3] **Рабчевский, Е.А.** Автоматическое построение онтологий на основе лексико-синтаксических шаблонов для информационного поиска / Е.А. Рабчевский // Электронные библиотеки: перспективные методы и технологии, электронные коллекции (RCDL). 2009. С.69-77.
- [4] **Найханова, Л.В.** Основные аспекты технологии создания методов автоматического построения онтологий / Л.В. Найханова // Материалы конференции ЗНАНИЯ-ОНТОЛОГИИ-ТЕОРИИ (ЗОНТ). 2009.
- [5] **Лукашевич, Н.В.** Проектирование лингвистических онтологий для информационных систем в широких предметных областях / Н.В. Лукашевич, Б.В. Добров // Онтология проектирования. 2015. №. 1 (15). С.47-69.
- [6] DBpedia. - <https://wiki.dbpedia.org/>.
- [7] **Suchanek, F.M.** YAGO: A large ontology from wikipedia and wordnet / F.M. Suchanek, G. Kasneci, G. Weikum // Journal of Web Semantics. 2008. Vol. 6, 3. P.203-217.
- [8] **Муромцев, Д.** Индустриальные графы знаний-интеллектуальное ядро цифровой экономики / Д. Муромцев, Д. Волчек, А. Романов // Control Engineering Россия. 2019. №. 5. С.32-39.
- [9] **Shibaki, Y.** Constructing large-scale person ontology from Wikipedia / Y. Shibaki, M. Nagata, K. Yamamoto // Proceedings of the 2nd Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources. 2010. P.1-9.
- [10] **Erxleben, F.** Introducing Wikidata to the linked data web / F. Erxleben, M. Günther, M. Krötzsch, J. Mendez, D. Vrandečić // International semantic web conference. 2014. P.50-65.
- [11] **Samuel, J.** Towards understanding and improving multilingual collaborative ontology development in Wikidata / J. Samuel // Companion of the The Web Conference 2018 on The Web Conference. 2018. P.23-27.
- [12] **Рогущина, Ю.В.** Разработка онтологической модели информационной потребности пользователя при семантическом поиске / Ю.В. Рогущина // Онтология проектирования. 2014. №. 2 (12). С.7-31.
- [13] **Городецкий, В.И.** Онтологии и персонификация профиля пользователя в рекомендующих системах третьего поколения / В.И. Городецкий, О.Н. Тушканова // Онтология проектирования. 2014. №. 3 (13). С.60-82.
- [14] **Павлыгин, Э.Д.** Разработка программного комплекса для интеллектуального анализа социальных медиа / Э.Д. Павлыгин, А.Г. Подлобошников, Р.А. Савинов, Н.Г. Ярушкина, А.М. Наместников, А.А. Филиппов, А.А. Романов, В.С. Мошкин, Г.Ю. Гуськов, М.С. Григоричева // Автоматизация процессов управления. 2019. №. 2. С.23-36.

Сведения об авторах

Наместников Алексей Михайлович, 1974 г. рождения, доктор технических наук, доцент, окончил Ульяновский государственный технический университет (УлГТУ). Профессор кафедры «Информационные системы» УлГТУ. Имеет более 80 работ в области автоматизированного проектирования и интеллектуальных систем. Author ID (RSCI): 392690; Author ID (Scopus): 9277806100. nam@ulstu.ru.

Пирогова Наталья Дмитриевна, 1997 г. рождения. Окончила факультет информационных систем и технологий УлГТУ в 2021 г. Магистр по направлению «Прикладная информатика». Имеет 5 статей в области автоматизации бизнес-процессов и онтологического инжиниринга. natufochka73@gmail.com.

Филиппов Алексей Александрович, 1987 г. рождения, кандидат технических наук, окончил УлГТУ, доцент кафедры «Информационные системы» УлГТУ. Имеет около 80 статей в области онтологического инжиниринга и интеллектуального анализа данных. ORCID: 0000-0001-5275-7628; Author ID (RSCI): 708454; Author ID (Scopus): 57191472723. al.filippov@ulstu.ru.

Поступила в редакцию 27.07.2021, после рецензирования 23.09.2021. Принята к публикации 27.09.2021.

An approach to the automatic linguistic ontology construction to determine the interests of social networks users

A.M. Namestnikov, N.D. Pirogova, A.A. Filippov

Ulyanovsk State Technical University, Ulyanovsk, Russia

Abstract

Social networks provide researchers with the opportunity to obtain an array of text data for further analysis within a certain subject area. Each subject area has its own specific professional vocabulary and writing style. When defining the subject area of text material there is a big problem with building dictionaries, thesauri, and ontologies. In this article a linguistic ontology is considered under ontology and which is aimed to determine the subject area of text material. An algorithm for the automatic construction of an ontology based on the Wikidata knowledge graph is presented. The task is to map a set of objects of the Wikidata knowledge graph to a set of entities of a linguistic ontology. The article proposes an algorithm for determining the degree of belonging of the text material to the subject area. Experiments on assessing the time of building an ontology and the applicability of the obtained linguistic ontologies to the problem of determining the degree of belonging of text materials in the subject area have shown: the running time of the algorithm and the number of terms in the formed ontology are directly proportional to the number of analyzed properties and Wikidata objects; the formed linguistic ontology is applicable to the problem of determining the degree of belonging of a text to a subject area.

Key words: linguistic ontology, automation, Wikidata, text document, subject area.

Citation: Namestnikov AM, Pirogova ND, Filippov AA. An approach to the automatic linguistic ontology construction to determine the interests of social networks users [In Russian]. *Ontology of designing*. 2021; 11(3): 351-363. DOI: 10.18287/2223-9537-2021-11-3-351-363.

Acknowledgment: The authors acknowledge that the work was supported by the framework of the state task of the Ministry of Science and Higher Education of the Russian Federation No.075-00233-20-05 "Research of intelligent predictive multimodal analysis of big data, and the extraction of knowledge from different sources". The reported study was funded by RFBR and the government of Ulyanovsk region according to the research projects: 19-47-730003 and 19-47-730005.

List of figures and tables

Figure 1 - Example of a page describing an object in the Wikidata knowledge graph

Figure 2 - Example of linguistic ontology

Figure 3 - Dependence of the algorithm operation time and the number of terms in the formed ontology on the number of properties

Figure 4 - Dependence of the algorithm operation time and the number of terms in the formed ontology on the number of objects

Figure 5 - Dependence of the algorithm operation time and the number of terms in the formed ontology on the number of iterations

Table 1 - Experimental results for determining the dependence of the algorithm operation time and the number of terms in the formed linguistic ontology on the algorithm parameters values

Table 2 - Experimental results on the applicability of the obtained linguistic ontologies to the problem of determining the degree of belonging of text materials of the subject area

References

- [1] **Mozgerina EC**. Automatic construction of an ontology from a collection of text documents [In Russian]. Digital Libraries: Advanced Methods and Technologies, Digital Collections (RCDL). 2011. P.293-298.
- [2] **Ermakov AE**. Automation of ontological engineering in systems for extracting knowledge from text [In Russian]. Dialogue conference proceedings. 2008. P.4-8.

- [3] **Rabchevsky EA.** Automatic construction of ontologies based on lexical-syntactic templates for information retrieval [In Russian]. Digital Libraries: Advanced Methods and Technologies, Digital Collections (RCDL). 2009. P.69-77.
- [4] **Naikhanov LV.** The main aspects of the technology for creating methods for the automatic construction of ontologies [In Russian]. Knowledge-Ontology-Theory (KONT) conference proceedings, 2009.
- [5] **Lukashevich NV, Dobrov BV.** Designing linguistic ontologies for information systems in broad subject areas [In Russian]. *Ontology of Designing*. 2015; 5(1): 47-69.
- [6] DBpedia. - <https://wiki.dbpedia.org/>.
- [7] **Suchanek FM, Kasneci G, Weikum G.** YAGO: A large ontology from wikipedia and wordnet. *Journal of Web Semantics*. 2008; 6(3): 203-217.
- [8] **Muromtsev D, Volchek D, Romanov A.** Industrial knowledge graphs as the intellectual core of the digital economy [In Russian]. *Control Engineering Россия*. 2019; (5): 32-39.
- [9] **Shibaki Y, Nagata M, Yamamoto K.** Constructing large-scale person ontology from Wikipedia. Proceedings of the 2nd Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources. 2010. P.1-9.
- [10] **Erxleben F, Günther M, Krötzsch M, Mendez J, Vrandečić D.** Introducing Wikidata to the linked data web. International semantic web conference. 2014. P.50-65.
- [11] **Samuel J.** Towards understanding and improving multilingual collaborative ontology development in Wikidata. Companion of The Web Conference 2018 on The Web Conference. 2018. P.23-27.
- [12] **Rogushina UV.** Development of an ontological model of the user's information needs in semantic search [In Russian]. *Ontology of Designing*. 2014; 2(12): 7-31.
- [13] **Gorodeckiy VI, Tushkanova ON.** Ontologies and personification of the user profile in third generation recommender systems [In Russian]. *Ontology of Designing*. 2014. 3(13): 60-82.
- [14] **Pavligin ED, Podlaboshnikov AG, Savinov RA, Yarushkina NG, Namestnikov AM, Filippov AA, Romanov AA, Moshkin VS, Guskov GU, Grigoricheva MS.** Development of a software package for intellectual analysis of social media [In Russian]. *Automation of Control Processes*. 2019; 2: 23-36.

About the authors

Aleksey Mihaylovich Namestnikov (b. 1974), graduated from the Ulyanovsk State Technical University in 1996, D. Sc. Eng. (2018). He is a Professor at the Department of information systems of the Ulyanovsk State Technical University. He is a co-author of about 80 scientific articles and abstracts in the field of CAD and AI. Author ID (RSCI): 392690; Author ID (Scopus): 9277806100. nam@ulstu.ru.

Natalya Dmitrievna Pirogova (b. 1997) graduated from the Ulyanovsk State Technical University in 2021, Master of Sciences in Engineering (2021). She is a co-author of 5 scientific articles and abstracts. natufochka73@gmail.com.

Aleksey Aleksandrovich Filippov (b. 1987), graduated from the Ulyanovsk State Technical University in 2009, PhD (2013). He is an Associate Professor at the Department of information systems of the Ulyanovsk State Technical University. He is a co-author of about 80 scientific articles and abstracts in the field of ontology engineering and data mining. ORCID: 0000-0001-5275-7628; Author ID (RSCI): 708454; Author ID (Scopus): 57191472723. al.filippov@ulstu.ru.

Received July 27, 2021. Revised September 23, 2021. Accepted September 27, 2021.