УДК 681.3

А.А. МАРЧЕНКО

МЕТОД АВТОМАТИЧЕСКОГО ПОСТРОЕНИЯ ОНТОЛОГИЧЕСКИХ БАЗ ЗНАНИЙ. III. АВТОМАТИЧЕСКАЯ ГЕНЕРАЦИЯ ТАКСОНОМИИ КАК ОСНОВЫ ОНТОЛОГИИ[1]

Аннотация. Разработан метод автоматического построения онтологических баз знаний. Создан алгоритм выделения явных семантических отношений между концептами онтологии из векторов их семантико-синтаксической валентности. Векторы семантико-синтаксических валентностей также использованы в качестве контекстных векторов в алгоритме формального концептуального анализа, что позволило разработать метод автоматической генерации таксономий высокого качества. В результате создан базовый алгоритм автоматического построения онтологических баз знаний на основе разработанной тензорной семантико-синтаксической модели естественного языка.

Ключевые слова: автоматическое извлечение знаний, корпусная лингвистика, онтология, неотрицательная факторизация тензоров.

ВВЕДЕНИЕ

Алгоритм обогащения онтологических сетей новыми отношениями между концептуальными узлами представлен в [1]. Для его эффективной работы, кроме текстовых корпусов электронной энциклопедии, необходимо наличие качественной таксономии. При разработке и тестировании описанного алгоритма использовалась таксономия лексико-семантической базы WordNet. В рамках предложенного подхода таксономия — это иерархическая основа онтологии, на которую алгоритм наращивает горизонтальные семантические связи. Таксономия является критически важным ресурсом, качество которого имеет определяющее влияние на точность и надежность выявления семантических отношений между концептами-узлами. Именно поэтому следующий этап разработки и развития описанной модели — создание методов автоматического построения таксономии на основе обработки текстовых корпусов. Данные методы рассматриваются в качестве начального этапа автоматического построения онтологической базы знаний.

АЛГОРИТМ АВТОМАТИЧЕСКОГО ПОСТРОЕНИЯ ТАКСОНОМИИ

Среди разработанных методов автоматического построения таксономии на основе обработки текстовых корпусов можно выделить два основных класса: методы кластеризации с применением мер семантической близости и теоретико-множественные. Данные методы работают с моделью векторного пространства, в котором слова, или термы, представлены в виде соответствующих им векторов признаков, полученных при обработке и анализе текстового корпуса.

Для методов кластеризации характерно использование некоторой меры семантической близости (например, мера косинуса угла между векторами слов) для поиска расстояния между векторами слов в целях определения, насколько они семантически подобны и должны ли объединяться в один кластер. Методы кластеризации, в свою очередь, подразделяются на агломеративные (кластеризация снизу вверх) и разделяющие (кластеризация сверху вниз). Наиболее эффективные методы данных классов описаны в [2–6].

С помощью теоретико-множественных методов осуществляют построение графа таксономии установлением частичного порядка на множестве слов-понятий, используя отношение включения между их множествами признаков. Одним из лучших методов этого класса является формальный концептуальный анализ (Formal Conceptual Analysis, FCA) [7].

[1] Начало см. в № 1, 2, 2016.

Тестирование и практическое использование методов свидетельствуют о более высокой точности таксономий, построенных методом FCA [8]. При этом точность достигает оценки $F = 68.23\%$, тогда как при агломеративной кластеризации $F = 62.92\%$, а при методе Bi-Section-KMeans — разделяющей иерархической кластеризации, $F = 62.80\%$. В данном контексте под точностью подразумевается оценка соответствия структуры построенной иерархии некоторому эталонному графу таксономии, которая вычисляется по методике, описанной в [8].

Отметим, что по сравнению со структурами, генерируемыми методами-конкурентами, именно таксономии, построенные методом FCA, лучше интерпретируются разработчиками-инженерами. Недостатком FCA является тот факт, что данный метод NP-полный и для его реализации в приложениях реального времени необходимо применять различные эффективные эвристики [9].

В качестве векторов-наборов признаков терминов в алгоритмах обычно используют лингвистический контекст, например, векторы инцидентности терминов с базисным набором глаголов, с которыми употребляются данные слова в позициях подлежащего или дополнения [8].

Использование векторов семантико-синтаксической валентности слов в качестве таких векторов-наборов признаков терминов оказалось удачным решением для реализации алгоритмов построения таксономии с помощью описанных подходов. Эксперименты показали, что в этом случае наблюдается стабильное увеличение точности построенных графов таксономии в среднем на 6–8 % практически для всех основных методов по сравнению с использованием обычных контекстных векторов инцидентности.

Рассмотрим подробнее оригинальный алгоритм построения таксономии.

Вход: текстовый корпус англоязычной Википедии — English Wikipedia.

Выход: таксономия слов-понятий.

Начальным этапом алгоритма является построение метрики, на основе которой выполняется процесс первичной кластеризации слов-понятий.

Шаг 1. Проводится частотный анализ статей Википедии со сборкой матрицы TD (Term $\times$ Document) по технологии латентного семантического анализа, при этом учитываются слова-термы, являющиеся словосочетаниями, а цепочка слов — название некоторой статьи Википедии, считается термом.

Шаг 2. Выполняется факторизация TD $(N \times M)$ методом Ли и Сунга, генерируются матрицы T $(N \times k)$ и D $(k \times M)$, $k << N, M$, такие что $TD = T \times D$.

Шаг 3. Для определения семантической близости между словами a и b выбираются векторы $T[a]$ и $T[b]$ и вычисляется скалярное произведение $(T[a], T[b])$, из которого находится косинус угла между данными векторами.

Таким образом, получается метрика $\mu(a, b) = \cos(\alpha(T[a], T[b]))$.

Основное свойство метрики в пространстве состоит в выполнении правила треугольника $\mu(a, c) \le \mu(a, b) + \mu(b, c)$. Это правило в данном случае может нарушаться ввиду существования слов, имеющих несколько значений. Например, bank — финансовое учреждение (Bank of America) и берег реки (Northern bank of the River Thames). Возможна ситуация, когда $\mu(\textbf{\textit{money}}, \textbf{\textit{water}}) \ge \mu(\textbf{\textit{money}}, \textbf{\textit{bank}}) + \mu(\textbf{\textit{bank}}, \textbf{\textit{water}})$, в этом случае необходимо выполнить операцию расщепления неоднозначного слова ***bank*** на ***bank* 1** и ***bank* 2**, т.е. расщепить его вектор T следующим образом:

— в векторе для ***bank* 1** оставить неизмененными значения в тех его позициях, которые коммутировали в скалярном произведении с векторами слов ***money***, ***finance***, ***credit*** и т.д., а другие обнулить;

— в векторе для ***bank* 2** оставить неизмененными значения в тех его позициях, которые коммутировали в скалярном произведении с векторами слов ***river***, ***water***, ***boat*** и т.д., а другие обнулить.

В качестве алгоритма начального этапа построения таксономии рассмотрим метод иерархической агломеративной кластеризации.

Шаг 1. Все слова образуют собственный кластер;

Шаг 2. While not End do

Begin

Найти два наиболее близких кластера: A и B, и объединить их;

Вычислить новый центроид-кластероид полученного кластера;

End.

Наиболее близкую пару кластеров можно найти следующим образом:

— вначале использовать $\mu(a, b)$;

— далее применить $\mu(a_c, b_c)$, где a_c, b_c — кластероиды, т.е. центры соответствующих кластеров;

— после объединения в кластер более двух слов для дальнейшего вычисления расстояний между кластерами нужно выбрать кластероид, т.е. найти слово, самое близкое к остальным словам кластера (с минимальной суммой расстояний от данного слова ко всем другим словам кластера). В качестве кластероида целесообразно выбирать однозначное слово, для которого всегда выполняется правило треугольника.

Алгоритм продолжает работу, пока не образует единого кластера, объединяющего все слова множества. Таким образом, процесс кластеризации генерирует некоторую иерархию.

Возможен следующий вариант алгоритма: для определения ближайшей пары кластеров расстояние вычисляется по формуле

$$M(P, Q) = \frac{1}{|P||Q|} \sum_{p \in P,\, q \in Q} \mu(p, q),$$

позволяющей эффективно находить слова, близко расположенные к различным кластерам, в состав которых они могут входить одновременно. И именно при обнаружении таких слов — нарушителей правила треугольника, можно эффективно расщеплять многозначные слова и их векторы. Если кластеры слов, к которым тяготеет найденное слово, достаточно заполнены, то процесс расщепления векторов будет весьма точным и надежным.

В процессе работы алгоритм генерирует иерархическую сеть, где узлы являются словами с некоторым зафиксированным значением (как ***bank* 1** или ***bank* 2**) в случае их изначальной неоднозначности. Внутри каждого кластера можно улучшить качество структуры использованием более сложных и точных методов сборки таксономии, например алгоритмом FCA, описанным в [9].

Рассмотрим процесс такой перестройки некоторой подсети таксономии.

Шаг 1. Для лексем из подмножества данного кластера получаем векторы семантико-синтаксической валентности слов из матриц W, H, X, Y, Z, как это описано в [10].

Шаг 2. Выполняем расщепление полученных векторов семантико-синтаксической валентности слов на составляющие векторы валентности значений-концептов, как это описано в [11], с привязкой расщепленных векторов к узлам-концептам в обрабатываемой подсети иерархии.

Шаг 3. Используем расщепленные векторы семантико-синтаксических валентностей концептов в качестве векторов лингвистического контекста для построения данной иерархии заново методом FCA.

Шаг 4. На более качественной иерархии, полученной на предыдущем шаге, снова проводим расщепление векторов семантико-синтаксической валентности слов данного подмножества на составляющие векторы валентностей концептов с привязкой расщепленных векторов к узлам-концептам вновь образованной иерархии.

Шаг 5. Используем заново полученные расщепленные векторы семантико-синтаксических валентностей концептов в качестве векторов лингвистического контекста для следующего построения данной иерархии методом FCA.

Шаг 6. Повторяем шаги 4 и 5 до тех пор, пока структура таксономии и векторы семантико-синтаксических валентностей концептов-узлов изменяют свою форму и значения.

Достигнутое стабильное состояние будет соответствовать неподвижной точке, где качество иерархии и качество векторов валентности уже взаимно не улучшаются и получено максимально достижимое качество иерархии.

ЭКСПЕРИМЕНТЫ ПО РАСЧЕТУ ОЦЕНКИ ЭФФЕКТИВНОСТИ АЛГОРИТМА

Для анализа эффективности предложенного алгоритма проведены эксперименты автоматического построения таксономий. В качестве эталонных таксономий использована лексико-семантическая база WordNet, из которой выделено множество L_N первых слов из всех синсетов, расположенных ниже некоторого узла N (включая первое слово синсета N).

Обработаны все статьи из English Wikipedia, названия которых содержат слова из L_N. Выполнен синтаксический анализ предложений этих текстов, построены их управляющие пространства синтаксических структур, после чего данные из полученных управляющих пространств перенесены в большие тензоры лингвистической модели, где также сохраняются данные обработки значительной части (больше одного миллиона) статей English Wikipedia. Далее осуществлялась факторизация тензоров для получения матриц векторов семантико-синтаксической валентности слов (матрицы W, H, X, Y, Z), которые гарантированно содержат векторы слов из L_N. Затем вычислялась метрика $\mu(a, b) = \cos(\alpha(T[a], T[b]))$ согласно описанной ранее методике: генерировалась матрица TD (Term × Document) на большом множестве статей English Wikipedia, включая все статьи, названия которых содержат слова из L_N. После неотрицательной факторизации матрицы TD получена мера близости $\mu(a, b) = \cos(\alpha(T[a], T[b]))$.

Предварительный нулевой этап тестирования — генерация таксономии T_0 слов из множества L_N с использованием агломеративного алгоритма кластеризации с мерой близости $\mu(a, b)$, после чего вычисляется оценка F соответствия сгенерированной таксономии T_0 эталонной иерархии (подсети WordNet с корнем N) согласно методике, описанной в [8].

Далее выполняется процедура расщепления векторов W, H, X, Y, Z для слов из L_N с привязкой расщепленных векторов к соответствующим узлам таксономии T_0. Точность работы процедуры расщепления и привязки оценивается согласно методике, описанной в [11].

После этого проводится процедура FCA для генерации таксономии T_1 для слов из L_N с использованием их расщепленных векторов W', H', X', Y', Z'.

Далее вычисляется значение оценки F соответствия сгенерированной таксономии T_1 эталонной подсети WordNet, после чего циклически повторяются этапы:

— расщепление векторов W, H, X, Y, Z для слов из L_N с привязкой к соответствующим узлам таксономии T_i;

— оценка точности расщепления и привязки заново сформированных векторов W', H', X', Y', Z' семантико-синтаксической валентности концептов к узлам таксономии T_i;

— выполнение процедуры FCA, т.е. генерации таксономии T_{i+1} для слов из L_N с использованием соответствующих им заново расщепленных векторов W', H', X', Y', Z';

— вычисление оценки F соответствия построенной таксономии T_{i+1} эталонной подсети WordNet.

Выход из цикла происходит, когда T_{i+1} полностью совпадет с T_i.

Эксперименты выполнялись для тестовых наборов N_0 = «еда» и N_0 = «транспорт», в результате которых наблюдалась сходимость оценки соответствия генерируемых таксономий эталонной иерархии WordNet. Для тестового набора N_0 = «еда» получены следующие оценки:

$$F(T_0, T_{\text{эталон}}) = 59.82\ \%;\ F(T_1, T_{\text{эталон}}) = 64.45\ \%;$$

$$F(T_2, T_{\text{эталон}}) = 65.03\ \%;\ F(T_3, T_{\text{эталон}}) = 66.19\ \%;$$

$$F(T_4, T_{\text{эталон}}) = 67.67\ \%;\ F(T_5, T_{\text{эталон}}) = 68.99\ \%;$$
$$F(T_6, T_{\text{эталон}}) = 71.15\ \%;\ F(T_7, T_{\text{эталон}}) = 73.64\ \%;$$
$$F(T_8, T_{\text{эталон}}) = 75.64\ \%;\ F(T_9, T_{\text{эталон}}) = 75.64\ \%.$$

Для тестового набора N_0 = «транспорт» получены следующие оценки:

$$F(T_0, T_{\text{эталон}}) = 61.61\ \%;\ F(T_1, T_{\text{эталон}}) = 65.18\ \%;\ F(T_2, T_{\text{эталон}}) = 67.98\ \%;$$
$$F(T_3, T_{\text{эталон}}) = 70.29\ \%;\ F(T_4, T_{\text{эталон}}) = 73.94\ \%;\ F(T_5, T_{\text{эталон}}) = 75.71\ \%;$$
$$F(T_6, T_{\text{эталон}}) = 77.04\ \%;\ F(T_7, T_{\text{эталон}}) = 79.04\ \%;\ F(T_8, T_{\text{эталон}}) = 79.04\ \%.$$

Эксперименты для тестовых наборов N_0 = «еда» и N_0 = «транспорт» проводились в целях оценки качества автоматической генерации таксономий концептов-существительных. Поэтому для построения иерархических графов процедура FCA использовала векторы X (существительное-подлежащее), Z (существительное-дополнение) и W (объект, определяемый в синтагматической кольцевой связи). Для концептов-существительных также можно использовать векторы H (объект, определяющий в синтагматической кольцевой связи), однако на практике наполнение этих векторов для существительных незначительно. Поэтому векторы H, а также векторы Y, описывающие в основном коммутационные свойства глаголов, для построения иерархий существительных не применялись.

Как видно из полученных данных, с переходом на алгоритм FCA и с использованием векторов семантико-синтаксической валентности концептов значительно растет оценка качества структуры таксономий. После существенного улучшения качества структур таксономий на следующем этапе заметно растет показатель точности расщепления векторов семантико-синтаксической валентности слов и привязки их к узлам заново построенной таксономии (табл. 1 и 2).

Далее можно наблюдать равномерный рост качества генерируемых таксономий и оценок точности расщепления и привязки векторов семантико-синтаксической валентности к узлам таксономий. Улучшение качества структуры таксономии T повышает точность расщепления и привязки векторов семантико-синтаксической валентности, что, в свою очередь, приводит к улучшению качества данной структуры на следующей итерации. Этот процесс взаимного рекурсивного улучшения продолжается до момента получения максимально достижимого уровня качества T (в данной конфигурации вычислительного процесса {алго-

Таблица 1. Результаты эксперимента для тестового набора N_0 = «еда»

Векторы семантико-синтаксической валентности	Оценки точности расщепления и привязки векторов семантико-синтаксической валентности к узлам таксономии T_i (%)								
	Этап 0	Этап 1	Этап 2	Этап 3	Этап 4	Этап 5	Этап 6	Этап 7	Этап 8
X	72.87	77.13	78.73	80.11	82.40	84.71	85.49	87.11	87.11
Z	71.39	75.83	76.49	77.32	79.14	80.23	81.84	83.59	83.59
W	66.18	72.95	74.07	76.81	78.21	81.04	83.78	86.43	86.43

Таблица 2. Результаты эксперимента для тестового набора N_0 = «транспорт»

Векторы семантико-синтаксической валентности	Оценки точности расщепления и привязки векторов семантико-синтаксической валентности к узлам таксономии T_i (%)							
	Этап 0	Этап 1	Этап 2	Этап 3	Этап 4	Этап 5	Этап 6	Этап 7
X	73.32	78.93	80.53	82.19	84.73	85.06	86.79	86.79
Z	69.04	75.81	78.09	80.54	84.29	84.82	85.11	85.11
W	68.81	73.98	75.22	77.37	78.96	81.59	83.27	83.27

ритм, данные}), когда структура таксономии окончательно фиксируется $T_i = T_{i-1}$. Будем называть такую таксономию T_i неподвижной точкой алгоритма построения таксономии.

Работа алгоритма завершается, когда он «попадает» в неподвижную точку, которой соответствует некоторый набор векторов семантико-синтаксической валентности XZW_i, не изменяющий структуры T ($T_i = T_{i-1}$). В таком случае алгоритм расщепления векторов семантико-синтаксической валентности слов и привязки полученных векторов семантико-синтаксической валентности концептов к узлам таксономии T_i, структурно идентичной T_{i-1}, не приведет к изменению данного набора векторов, т.е. $XZW_{i+1} = XZW_i$. Алгоритм FCA на тех же данных строит аналогичную структуру $T_{i+1} = T_i$ и попадает в состояние неподвижной точки, которая соответствует наилучшей достижимой структуре T_{best}.

Качество T_{best} ограничено полнотой данных в массивах базы системы. Алгоритм достигает того максимума, который обусловлен этой полнотой и качеством данных.

После генерации высококачественной таксономии получается также точная и надежная привязка векторов семантико-синтаксической валентности к узлам-концептам иерархической сети. Затем по этим векторам семантико-синтаксической валентности концептов можно переходить к описанию явных семантических связей-отношений между понятийными узлами генерируемой онтологии, используя алгоритм, описанный в [1].

Таким образом, получено описание полностью автономной алгоритмической модели системы автоматического построения онтологической базы знаний универсального типа. При этом единственным необходимым ресурсом входных данных для функционирования и саморазвития системы является текстовый корпус электронной энциклопедии English Wikipedia. Система читает Википедию и транслирует тексты ее статей на естественном языке во внутреннее представление в форме записей онтологической базы знаний. Чем лучше алгоритм понимает смысл предложений текстов, тем точнее и адекватнее семантические структуры создаваемой онтологии как понятийной системы.

ЗАКЛЮЧЕНИЕ

В статье описана модель естественного языка, реализованная с помощью факторизации лингвистических тензоров. На основе построенной модели разработан алгоритм пополнения онтологических баз знаний новыми семантическими отношениями между узлами-концептами. Также описан подход к автоматическому построению иерархической основы онтологий — таксономий, с применением структур данных и алгоритмов представленной модели естественного языка. Все данные компоненты, интегрированные вместе, представляют собой полностью автономную алгоритмическую модель системы автоматического построения онтологической базы знаний универсального типа. Система анализирует тексты на естественном языке, генерирует их семантико-синтаксические структуры, записывает их в специальные массивы данных, с помощью процедур факторизации преобразует эти данные во внутренний формат, интерпретирует их и добавляет полученные знания в онтологическую базу. Результаты эксперимента и тестирования свидетельствуют о корректности и надежности работы системы и ее отдельных компонентов, что доказывает адекватность и эффективность представленной модели, а также перспективность ее полномасштабной реализации и использования систем данного типа для создания, развития, пополнения и обогащения онтологий на практике.

СПИСОК ЛИТЕРАТУРЫ

1. Марченко А.А. Метод автоматического построения онтологических баз знаний. II. Автоматическое определение семантических отношений в онтологической сети // Кибернетика и системный анализ. — 2016. — **52**, № 2. — С. 30–36.

2. Caraballo S.A. Automatic construction of a hypernym-labeled noun hierarchy from text // Proceedings of the 37th Annual Meeting of the ACL-1999. — 1999. — P. 120–126.
3. Hindle D. Noun classification from predicate-argument structures // Proceedings of the Annual Meeting of the ACL-1990. — 1990. — P. 268–275.
4. Faure D., Nedellec C. A corpus-based conceptual clustering method for verb frames and ontology acquisition // LREC Workshop on Adapting Lexical and Corpus Resources to Sublanguages and Applications. — 1998. — P. 1–30.
5. Pereira F., Tishby N., Lee L. Distributional clustering of English words // Proceedings of the 31-st Annual Meeting of the ACL-1993. — 1993. — P. 183–190.
6. Bisson G., Nedellec C., Cañamero D. Designing clustering methods for ontology building // Proceedings of the ECAI Ontology Learning Workshop. — 2000. — P. 13–19.
7. Ganter B., Wille R. Formal concept analysis — mathematical foundations. — Berlin; Heilderberg: Springer-Verlag, 1999. — 284 p.
8. Cimiano P., Hotho A., Staab S. Comparing conceptual, divisive and agglomerative clustering for learning taxonomies from text // Proceedings of the European Conference on Artificial Intelligence (ECAI). — 2004. — P. 435–439.
9. Vychodil V. A new algorithm for computing formal concepts // Cybernetics and Systems 2008. Proceedings of the 19th European Meeting on Cybernetics and Systems Research. — 2008. — P. 15–21.
10. Марченко А.А. Метод автоматического построения онтологических баз знаний. I. Разработка семантико-синтаксической модели естественного языка // Кибернетика и системный анализ. — 2016. — **52**, № 1. — С. 23–33.
11. Анисимов А.В., Марченко А.А., Вознюк Т.Г. Определение семантических валентностей концептов онтологий с помощью неотрицательной факторизации тензоров больших текстовых корпусов // Кибернетика и системный анализ. — 2014. — **50**, № 3. — С. 3–16.

Надійшла до редакції 16.07.2015

О.О. Марченко

МЕТОД АВТОМАТИЧНОЇ ПОБУДОВИ ОНТОЛОГІЧНИХ БАЗ ЗНАНЬ. III. АВТОМАТИЧНА ГЕНЕРАЦІЯ ТАКСОНОМІЇ ЯК ОСНОВИ ОНТОЛОГІЇ

Анотація. Розроблено метод автоматичної побудови онтологічних баз знань. Створено алгоритм виділення явних семантичних відношень між концептами онтології з векторів їхньої семантико-синтаксичної валентності. Вектори семантико-синтаксичних валентностей також використано як контекстні вектори в алгоритмі формального концептуального аналізу, що дозволило створити метод автоматичної генерації таксономій високої якості. В результаті створено базовий алгоритм автоматичної побудови онтологічних баз знань на основі розробленої тензорної семантико-синтаксичної моделі природної мови.

Ключові слова: автоматичне добування знань, корпусна лінгвістика, онтології, невід'ємна факторизація тензорів.

O.O. Marchenko

A METHOD FOR AUTOMATIC CONSTRUCTION OF ONTOLOGICAL KNOWLEDGE BASES. III. AUTOMATIC GENERATION OF TAXONOMY AS THE FOUNDATION OF ONTOLOGY

Abstract. The author develops a method for automatic generation of ontological knowledge bases. An algorithm for extraction of explicit semantic relationships between concepts of ontology on the basis of their semantic-syntactic valence vectors has been developed. Vectors of semantic-syntactic valences of words have been also used as context vectors for formal concept analysis algorithm, which has allowed us to develop the method of automatic generation of high-quality taxonomies. A basic algorithm for automatic construction of ontological knowledge bases has been developed on the basis of the tensor semantic-syntactic model of natural language.

Keywords: automatic extraction of knowledge, corpus linguistics, ontologies, non-negative tensor factorization.

Марченко Александр Александрович,
доктор физ.-мат. наук, доцент Киевского национального университета имени Тараса Шевченко, e-mail: omarchenko@univ.kiev.ua.