УДК 33 **DOI: 10.34670/AR.2020.91.1.033**

Создание и обучение онтологий на основе анализа контекста и метаданных слабоструктурированного контента

Волчек Дмитрий Геннадьевич

Кандидат технических наук
Преподаватель
Национальный исследовательский университет ИТМО
197101, Российская Федерация, Санкт-Петербург, Кронверкский проспект, 49
e-mail: dvolchek@yandex.ru

Романов Алексей Андреевич

Кандидат технических наук
Преподаватель
Национальный исследовательский университет ИТМО
197101, Российская Федерация, Санкт-Петербург, Кронверкский проспект, 49
e-mail: romanov@itmo.ru

Аннотация

В статье описаны подходы к обучению онтологий на основе анализа метаданных и контекста слабоструктурированного контента. В качестве основных результатов работы можно выделить модель совместного представления контента и его метаданных системы управления контентом. Для извлечения терминов был использован ансамблевый метод, сочетающий в себе алгоритмы извлечения терминов как с использованием контрастного корпуса, так и без него, а также расширение метаданными признакового пространства кандидатов. Кроме того, описаны методы построения таксономических отношений на основе векторного представления слов и нетаксономических отношений посредством анализа универсальных зависимостей.

Поставленные задачи, связанные с разработкой онтологий, извлечением данных из Open EdX, и автоматическим извлечением концептов с помощью алгоритмов NLP и классификации, были выполнены. Помимо этого, был разработан метод построения таксономических отношений на основе векторного представления слов с последующей иерархической кластеризацией, а также метод извлечения нетаксономических отношений посредством анализа универсальных зависимостей.

В качестве возможного продолжения работы можно рассматривать задачу отбора признаков (из числа как показателей терминологичности, так и привлекаемых метаданных), более детальную проработку векторного представления слов и использования операций над векторами для представления терминов, состоящих из нескольких слов. Кроме того, возможно дополнительное обучение модели построения нетаксономических отношений, позволяющей извлекать отношения не только между двумя концептами в рамках одного предложения, но и, если присутствует более сложное отношение, связывающее 3 и более концептов.

Для цитирования в научных исследованиях

Волчек Д.Г., Романов А.А. Создание и обучение онтологий на основе анализа контекста и метаданных слабоструктурированного контента // Экономика: вчера, сегодня, завтра. 2020. Том 10. № 1А. С. 303-312. DOI: 10.34670/AR.2020.91.1.033

Ключевые слова

МООК, онлайн-курсы, Открытое Образование, семантические технологии, онтология, онтологическое моделирование, семантическая близость, обучение онтологий.

Введение

На протяжении достаточно длительного периода и по сегодняшний день наблюдается бурный рост объемов информации. Это с одной стороны хорошо, так как позволяет человечеству расширять горизонты и все больше познавать окружающий мир, использовать знания для своей собственной выгоды, с другой стороны порождает ряд существенных проблем. Чем больше объем накопленной информации и знаний, тем сложнее эти данные структурировать, интегрировать, актуализировать. Отдельно стоит упомянуть про поиск необходимой информации в огромном количестве разрозненных источников. Не зря такие технологичные компании как Google, Yandex и другие уделят столько внимания своим поисковым сервисам, создавая новые алгоритмы для повышения релевантности поисковой выдачи. Одним из последних новшеств является идея персональной выдачи результатов поискового запроса, когда двум различным пользователям может быть показана разная информация по результатам одного и того же запроса, опираясь на уже известные сведения и предпочтения самого пользователя. Другим направлением является выдача точных ответов на тот или иной вопрос пользователя, когда в поисковой выдаче формируется не просто список документов или интернет страниц, а непосредственный ответ на запрос. Такая возможность обеспечивается благодаря концепции связанных данных, когда помимо самой информации известно и ее семантическое значение. В этом случае работать информацией, представленной в таком виде, могут не только люди, но и средства автоматической обработки информации. Для реализации концепции связанных данных используются структуры под названием графы знаний – это большие хранилища информации представляющие собой направленные графы, где вершинами являются некоторые сущности, а ребрами – отношения между этими сущностями. Существенным отличием графов знаний от классических графов в математике является то, что ребра графа знаний могут характеризовать различные виды отношений между сущностями. Для построения и наполнения графа знаний необходимо сначала создать его структуру, которая называется онтологической моделью. На сегодняшний день таких моделей существует достаточно большое количество, а область их применения варьируется от образования и логистики, до проектирования умных производств, бизнес-процессов, поисковых сервисов и так далее. Таким образом крайне актуальной является задача создания или расширения существующих онтологических моделей для различных предметных областей.

Онтологический инжиниринг

Онтологическое моделирование – достаточно трудоемкий процесс, требующий серьезного подхода. Различные методы, сопутствующие проблемы и пути их решения в контексте

разработки онтологий – объекты множества исследований. В частности, в статье [Митрофанова, Константинова, 2015] описываются различные теоретические аспекты онтологического моделирования, в том числе классификация онтологий, принципы построения, возможности автоматического создания. Авторы работы [Гаврилова, Гулякина, 2008] выдвигают предположение о том, что специфика каждой отдельно взятой предметной области является превалирующей в процессе проектирования онтологий, что влечет разработку отдельных подходов к проектированию в каждом отдельно взятом случае. В работе [Gavrilova, Gladkova, 2014] авторы также отмечают необходимость использования различных подходов в процессе онтологического моделирования той или иной предметной области с целью учета всех особенностей последней и для создания онтологии, которая будет непосредственно учитывать все особенности моделируемого домена. В последнее десятилетие наблюдается существенный сдвиг в парадигме создания онтологий. Если в работах 15-20 летней давности [Гаврилова, Хорошевский, 2001; Частиков, Гаврилова, Белов, 2004] ключевым объектом исследований и основным действующим лицом выступал инженер-онтолог, то в более поздних работах [Asim, Wasim, Khan, Mahmood, Abbasi, 2018] основной фокус направлен на сами данные моделируемой предметной области.

Подход, когда онтологическая модель строится не экспертом предметной области, а непосредственно на основе имеющихся данных, получил название (по аналогии с машинным обучением) обучение онтологий. Основная идея заключается в том, чтобы строить онтологическую модель в полуавтоматическом режиме, а роль эксперта заключается в валидации полученных результатов, проектировании алгоритмов обучения, выбора гиперпараметров, разметки наборов данных и так далее. При этом те данные о моделируемой предметной области, которые представлены в структурированном виде, достаточно просто отображаются на онтологическую модель. Совсем иначе дело обстоит со слабоструктурированным контентом, однако таких данных о моделируемой предметной области обычно преобладающее количество. На практике обычно это текстовые документы, хранящие информацию о самой предметной области. Это могут быть регламенты, протоколы, техническая документация, проектная документация, описания сущностей или процессов и так далее. Поэтому для извлечения онтологической структуры из таких текстовых документов необходимо использовать методы обработки естественного языка (NLP). С точки зрения общего алгоритма обучения онтологий считается весьма устоявшимся следующая последовательность шагов, которая получила название «Слоеный пирог обучения онтологий», стоит отметить, что с небольшими вариациями такая схема уместна и для создания онтологической модели непосредственно экспертом предметной области.

Так как обучение производится на основе текстовых документов, то необходимо сначала выделить ключевые слова или термины, которые описывали бы сущности моделируемой предметной области. Следующим этапом считается поиск синонимов с целью исключения двойственности трактовки того или иного объекта. После установления всех семантических сходств термины предметной области становятся концептами этой области. Далее строятся таксономические отношения между извлеченными концептами по принципу Класс – Подкласс. Помимо иерархии концептов необходимо выделить и нетаксономические отношения, которые и вносят основную семантическую полноту создаваемой онтологической модели. После валидации полученных отношений экспертом предметной области формируются аксиомы, которые и образуют онтологическую модель. Отдельно стоит упомянуть, что на завершающем этапе возможно добавление правил, ограничений, логического вывода и так далее.

Извлечение терминов на основе анализа метаданных

Обучение онтологий происходит на основе данных о моделируемой предметной области, при этом зачастую эти данные хранятся в так называемых системах управления контентом (CMS). Это могут быть интернет порталы, сопровождения бизнес-процессов, платформы онлайн образования и проч. При этом такие системы хранят огромное количество метаданных (количество обращений к документу, время работы пользователя с документом, количество повторных обращений пользователя к тому или иному документу и многие другие), которые косвенно могут характеризовать важность терминов, содержащихся в этом документе с точки зрения моделирования той или иной предметной области. Существующие алгоритмы извлечения терминов никак не учитывают метаданные о документе, из которого и происходит экстракция терминов. С другой стороны, важным является и контекст, то есть если какой-то термин встречается не только в самом документе, на основе которого происходит обучение онтологии, но и в каких-то внешних источниках, то он с большей степенью вероятности является концептом рассматриваемой предметной области.

Существующие методы извлечения терминов из текста на естественном языке оперируют так называемым понятием «Терминологичности» того или иного кандидата в термины. Под терминологичностью понимается то, насколько тот или иной кандидат в термины описывает рассматриваемую предметную область. Глобально алгоритмы извлечения кандидатов в термины можно разделить на два класса:

– выполняющие экстракцию непосредственно из самого документа;
– использующие контрастный корпус, то есть набор документов не специфичных для моделируемой предметной области. Тогда если кандидат встречается как в рассматриваемом документе, так и в контрастном корпусе, его терминологичность снижается (предполагается, что это общий термин, не специфичных для рассматриваемой предметной области).

В настоящей работе предлагается подход, позволяющий учитывать метаданные для построения модели классификации (является ли кандидат в термины термином моделируемой предметной области). Формирование вектора атрибутов кандидата представлено на рисунке 1.

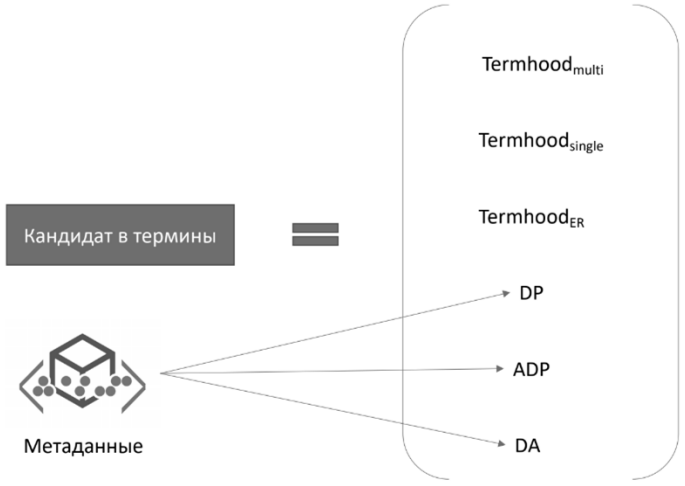

Рисунок 1 – Формирование вектора атрибутов кандидата в термины

Первые 3 атрибута характеризуют терминологичность кандидата с точки зрения алгоритмов извлечения:

$Termhood_{multi}$ – терминологичность кандидата в результате применения алгоритмов на основе использования контрастного корпуса;

$Termhood_{single}$ – терминологичность кандидата в результате применения алгоритмов, не использующих контрастный корпус;

$Termhood_{ER}$ – количество внешних источников, которые содержат рассматриваемого кандидата

DP – количество обращений к документу, из которого происходит извлечение;

ADP – среднее количество обращений;

DA – время с последнего обновления документа.

Для корректности работы алгоритма классификации необходимо произвести нормировку значений атрибутов:

$$\tilde{x}_{ik} = \frac{x_{ik} - x_{\min i}}{x_{\max i} - x_{\min i}}$$

Построение таксономических отношений

Существующие алгоритмы построения таксономических отношений Term subsumption и анализ формальных понятий (FCA) [Ganter, Wille, 2012]. Оба этих подхода основаны на статистических техниках, что дает существенное преимущество с точки зрения универсальности применения в различных предметных областях. Однако, не всегда удается достичь приемлемого уровня точности построения отношений. В настоящей работе предлагается использовать векторное представление слов и последующую иерархическую кластеризацию с целью извлечения таксономических отношений.

Для векторного представления слов предлагается использовать технологию Word2Vec [Mikolov, 2013]. Эта технология выгодно отличается от аналогичных, в том плане, что при выполнении операций над векторами возможно получение осмысленных конструкций. Например, сами авторы работы описывают ситуацию, когда при сложении векторов «Россия» и «Река» будет получен вектор крайне близкий к вектору «Волга».

Построение нетаксономических отношений

Для извлечения нетаксономических отношений предлагается использовать механизм анализа универсальных зависимостей. Для этого необходимо проанализировать каждое предложение и построить так называемое дерево зависимостей. Пример такого дерева представлен на рисунке 2.

В результате построения можно рассматривать полученное дерево зависимостей, как граф, связывающий отдельные слова в предложении. Далее для анализа используются извлеченные на первом этапе концепты, а именно рассматриваются всевозможные пары концептов и выявляется их совместное использование в предложении. Под рассмотрение попадают предложения, содержащие только пары концептов. Можно выделить несколько ситуаций для каждого такого предложения:

– отсутствие какого-либо отношения между концептами;

– связь имеет только один концепт;

– связь установлена между двумя концептами.

Если удалось установить связь между двумя концептами, то составляет тройка «Объект – Предикат – субъект», которая добавляется в список аксиом.

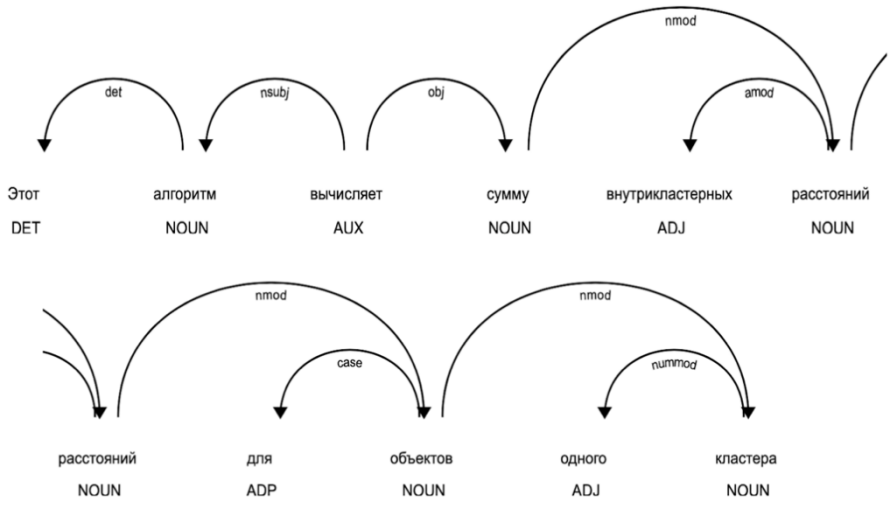

Рисунок 2 – Дерево универсальных зависимостей

Результаты

В качестве данных для эксперимента были использованы массовые открытые онлайн курсы Университета ИТМО. Курсы размещены на платформе Open EdX, что позволяет извлекать не только контент, но и метаданные для последующего использования в процессе обучения онтологий.

В качестве следующего этапа было произведено отображение данных CMS на разработанную базовую онтологию. Для этого необходимо извлекать и аннотировать данные из Open EdX в соответствии с базовой онтологией. В результате чего получен граф знаний, который хранит информацию не только о структуре платформы, курсов на ней расположенных, но также и метаданные для каждого отдельного элемента курса.

Для обучения модели классификации было отобрано 20 документов, содержащих 437 терминов. В качестве тестовой выборки использовались 5 документов, содержащих 143 кандидата в термины. Для решения задачи классификации были рассмотрены несколько алгоритмов. Результаты приведены на рисунке 3.

Метод	**Точность**	**Полнота**	f_1 **мера**
LogReg	0.56	0.52	0.54
Naive Bayes	0.52	0.45	0.48
Decision Tree	0.52	0.54	0.53
KNN	0.51	0.49	0.49
SVM	**0.63**	**0.55**	**0.59**

Рисунок 3 – Классификация терминов

Для построения таксономических отношений была обучена модель векторного представления слов Word2Vec skipgram (размерность пространства: 300, ширина окна 10), после чего была произведена иерархическая кластеризация. Результат представлен на рисунке 4.

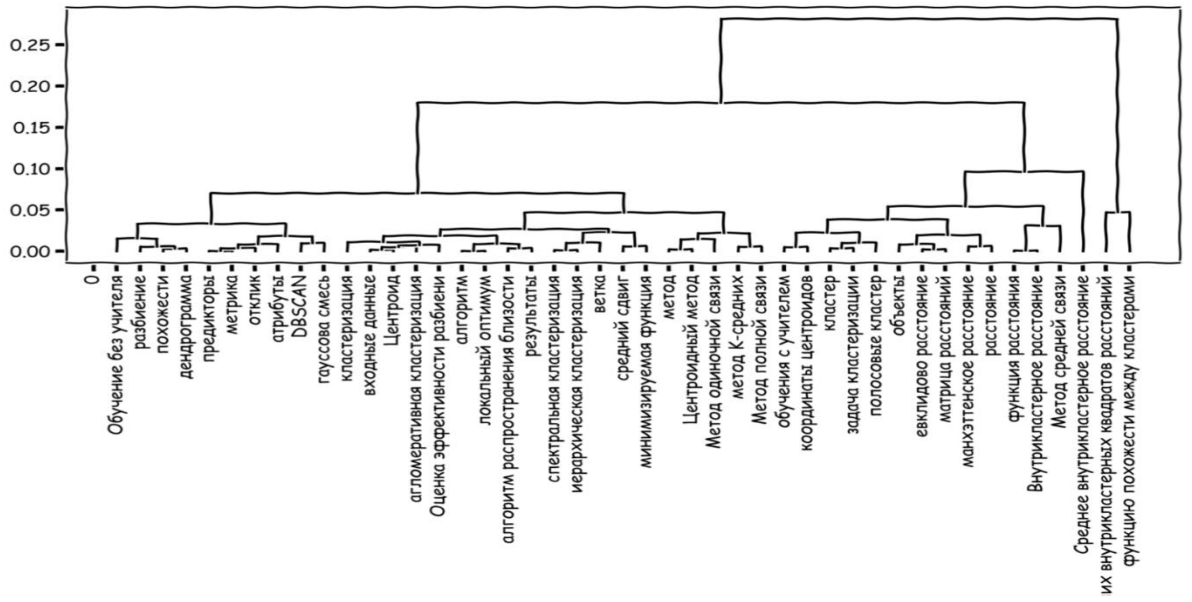

Рисунок 4 – Иерархическая кластеризация

Для построения нетаксономических отношений была обучена модель на основе анализа универсальных зависимостей, позволяющая устанавливать отношения между двумя концептами в рамках предложения. Пример представлен на рисунке 5.

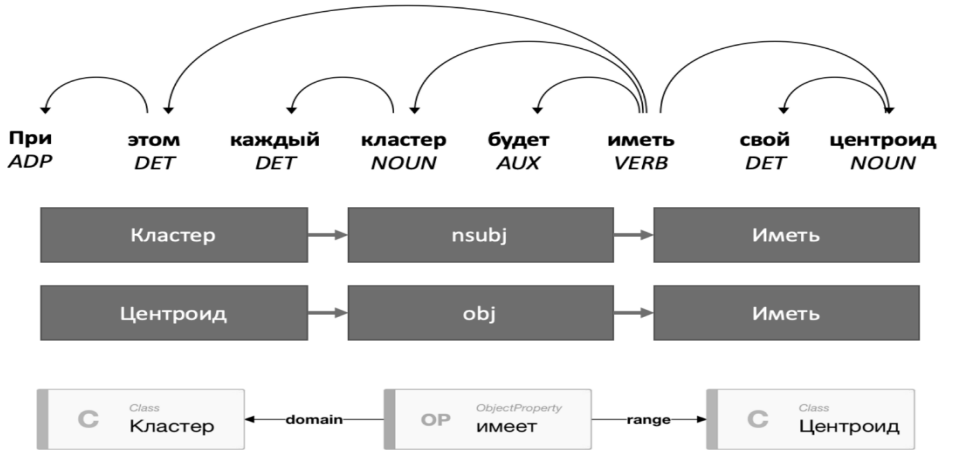

Рисунок 5 – Построение нетаксономических отношений

В результате можно сделать вывод, что описанные методы могут успешно использоваться в процессе обучения онтологий на основе контента, его метаданных и сопутствующего контекста.

Заключение

Поставленные задачи, связанные с разработкой онтологий, извлечением данных из Open EdX, и автоматическим извлечением концептов с помощью алгоритмов NLP и классификации, были выполнены. Помимо этого, был разработан метод построения таксономических отношений на основе векторного представления слов с последующей иерархической кластеризацией, а также метод извлечения нетаксономических отношений посредством анализа универсальных зависимостей.

В качестве возможного продолжения работы можно рассматривать задачу отбора признаков (из числа как показателей терминологичности, так и привлекаемых метаданных), более детальную проработку векторного представления слов и использования операций над векторами для представления терминов, состоящих из нескольких слов. Кроме того, возможно дополнительное обучение модели построения нетаксономических отношений, позволяющей извлекать отношения не только между двумя концептами в рамках одного предложения, но и, если присутствует более сложное отношение, связывающее 3 и более концептов.

Библиография

1. Митрофанова О. А., Константинова, Н. С. Онтологии как системы хранения знаний. // Санкт-Петербургский государственный университет, Факультет филологии и искусств, Кафедра математической лингвистики – 2015.
2. Гаврилова Т.А., Гулякина Н.А. Визуальные методы работы со знаниями: попытка обзора. // Искусственный интеллект и принятие решений — 2008. С.15-21.
3. Gavrilova T., Gladkova M. Big data structuring: the role of visual models and ontologies // Procedia Computer Science. – 2014. – Т. 31. – С. 336-343.
4. Гаврилова Т.А., Хорошевский В.Ф. Базы знаний интеллектуальных систем. // Учебник. - СПб, Изд-во «Питер», 2001.
5. Частиков, А. П., Гаврилова Т. А., Белов Д. Л. Разработка экспертных систем. Среда CLIPS. (2004)
6. Asim, M. N., Wasim, M., Khan, M. U. G., Mahmood, W., Abbasi, H. M. (2018). A survey of ontology learning techniques and applications. Database, 2018.
7. Buitelaar,P., Cimiano,P. and Magnini,B. (2005) Ontology learning from text: an overview. In: Ontology Learning from Text: Methods, Evaluation and Applications, Amsterdam, IOS Press, 123, 3–12
8. Ganter B., Wille R. Formal concept analysis: mathematical foundations. – Springer Science and Business Media, 2012
9. Mikolov T. et al. Distributed representations of words and phrases and their compositionality //Advances in neural information processing systems. – 2013. – С. 3111-3119.

Creation and training of ontologies based on the analysis of context and metadata of poorly structured content

Dmitrii G. Volchek

PhD in Technical Sciences
Lecturer
ITMO National Research University
197101, the city of St. Petersburg, Kronverksky Prospekt, 49
e-mail: dvolchek@yandex.ru

Aleksei A. Romanov

PhD in Technical Sciences
Lecturer
ITMO National Research University
197101, the city of St. Petersburg, Kronverksky Prospekt, 49
e-mail: romanov@itmo.ru

Abstract

The article describes approaches to training ontologies based on the analysis of metadata and the context of poorly structured content. As the main results of the work, we can single out a model for the joint presentation of content and its metadata in a content management system. To extract the terms, the ensemble method was used, combining the algorithms for extracting terms both with and without contrast housing, as well as expanding the metadata of the candidate attribute space. In addition, methods for constructing taxonomic relationships based on the vector representation of words and non-taxonomic relationships through analysis of universal dependencies are described.

The tasks associated with developing ontologies, extracting data from Open EdX, and automatically extracting concepts using NLP algorithms and classification were completed. In addition, a method was developed for constructing taxonomic relations based on a vector representation of words with subsequent hierarchical clustering, as well as a method for extracting non-taxonomic relations by analyzing universal dependencies.

As a possible continuation of the work, one can consider the task of selecting features (from both terminological indicators and metadata involved), a more detailed study of the vector representation of words and the use of vector operations to represent terms consisting of several words. In addition, additional training is possible for the model for constructing non-taxonomic relations, which allows one to extract relations not only between two concepts within the same sentence, but also if there is a more complex relation connecting 3 or more concepts.

For citation

Volchek D.G., Romanov A.A. (2020) Sozdanie i obuchenie ontologij na osnove analiza konteksta i metadannyh slabostrukturirovannogo kontenta [Creation and training of ontologies based on the analysis of context and metadata of poorly structured content]. *Ekonomika: vchera, segodnya, zavtra* [Economics: Yesterday, Today and Tomorrow], 10 (1A), pp. 303-312. DOI: 10.34670/AR.2020.91.1.033

Keywords

MOOCs, online courses, Open Education, semantic technologies, ontology, ontological modeling, semantic proximity, ontology training.

References

1. Mitrofanova O. A., Konstantinova, N. S. Ontology as a system for storing knowledge. // St. Petersburg State University, Faculty of Philology and Arts, Department of Mathematical Linguistics - 2015.
2. Gavrilova T.A., Gulyakina N.A. Visual methods of working with knowledge: an attempt to review. // Artificial Intelligence and Decision Making - 2008. P.15-21.
3. Gavrilova T., Gladkova M. Big data structuring: the role of visual models and ontologies // Procedia Computer Science. - 2014 .-- T. 31 .-- S. 336-343.

4. Gavrilova T.A., Khoroshevsky V.F. Knowledge bases of intelligent systems. // Textbook. - St. Petersburg, Publishing House "Peter", 2001.
5. Chastikov, A. P., Gavrilova T. A., Belov D. L. Development of expert systems. CLIPS environment. (2004)
6. Asim, M. N., Wasim, M., Khan, M. U. G., Mahmood, W., Abbasi, H. M. (2018). A survey of ontology learning techniques and applications. Database, 2018.
7. Buitelaar, P., Cimiano, P. and Magnini, B. (2005) Ontology learning from text: an overview. In: Ontology Learning from Text: Methods, Evaluation and Applications, Amsterdam, IOS Press, 123, 3–12
8. Ganter B., Wille R. Formal concept analysis: mathematical foundations. - Springer Science and Business Media, 2012
9. Mikolov, T. et al. Distributed representations of words and phrases and their compositionality // Advances in neural information processing systems. - 2013 S. 3111-3119.