УДК 004.4'414

МЕТОД АВТОМАТИЧЕСКОЙ СЕМАНТИЧЕСКОЙ РАЗМЕТКИ МАТЕМАТИЧЕСКИХ ОБРАЗОВАТЕЛЬНЫХ ТЕКСТОВ

Николаев К.С.[1], Невзорова О.А.[2]

[1,2] Казанский федеральный университет, Казань

[1] konnikolaeff@yandex.ru, [2] onevzoro@gmail.com

Аннотация

Автоматическая разметка семантического содержимого документа является важной задачей при создании образовательных и энциклопедических ресурсов. В данной работе мы предлагаем алгоритм и программное обеспечение для разметки математических текстов терминами из онтологии школьной математики OntoMathEdu, а также визуализации и ручного редактирования автоматической разметки. Представленные результаты будут использоваться при разработке курса по обучению планиметрии в системе дистанционного образования КФУ.

***Ключевые слова**: семантическая разметка, дистанционное образование, предметные онтологии.*

ВВЕДЕНИЕ

Онтологии и семантические технологии часто применяются для образовательных задач. Например, в [1] проводится анализ сложности вопросов в тестах самоконтроля и затем формируется персонализированная обратная связь. Технологии Semantic Web в данном случае применяются для хранения пользовательской информации и генерации обратной связи. В [2] приводится онтология, направленная на организацию процесса обучения компьютерным технологиям. В [3] приводится описание онтологии, связанной с образовательной деятельностью. В [4] онтологии и семантические технологии применяются для корректировки и персонализации назначения заданий студентам. Авторы [5] предлагают использование онтологий в качестве основы рекомендательной системы.

Применение онтологии к учебным текстам позврляет обогатить текст гиперссылками на дополнительные ресурсы, которые полезны в процессе изучения учебного материала. В Казанском федеральном университете ведется разработка онтологии школьной математики OntoMathEdu, которая содержит понятия и их отношения, изучаемые в курсе школьной планиметрии для 7–9 классов общеобразовательной школы. Онтология OntoMathEdu является концептуальным ядром цифровой экосистемы OntoMathEdu [6], структура которой включает следующие компоненты (схема приведена на рисунке 1):

1. Интеллектуальная цифровая образовательная платформа для школьной математики. Этот компонент занимает центральное место в экосистеме. Он используется в преподавании школьной математики с помощью онтологических и семантических технологий.

2. Коллекция вопросов. Эта коллекция представляет собой выходные данные компонента «Генератор тестов». Вопросы, хранящиеся в этой коллекции, используются при формировании наборов вопросов при мониторинге уровня знаний учащихся.

3. Хранилище формул. Этот компонент состоит из формул, извлеченных из школьных учебников геометрии, представленных в различных форматах (обычный текст, LaTeX, OpenMath).

4. Цифровые образовательные ресурсы. Этот компонент объединяет все вспомогательные источники данных, расположенные в Интернете.

5. Онтология OntoMathEdu. Онтология OntoMathEdu является отражением уровня знаний, соответствующего уровню школьной математики.

6. Сервис по обогащению онтологии. Этот компонент включает в себя набор методов, которые используются для уточнения взаимосвязей между понятиями в онтологии и улучшения горизонтальной связности онтологии.

7. Тестовый генератор. Этот компонент используется для автоматического создания новых тестовых заданий на основе анализа структуры и понятий существующих задач.

8. Семантический поиск формул. Это поисковой модуль, который выполняет семантический поиск математических текстов, присутствующих в экосистеме.

9. Рекомендационная система. Позволяет пользователям изучать понятия, связанные с их текущим образовательным процессом.

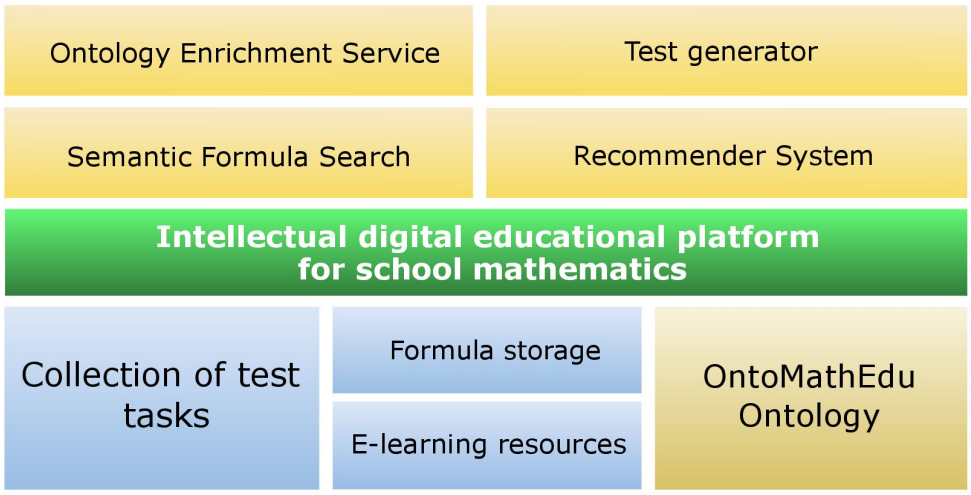

Рисунок 1. Состав экосистемы OntoMathEdu

Компонент «Интеллектуальная цифровая образовательная платформа для школьной математики» размещен на платформе цифрового образования Казанского федерального университета (edu.kpfu.ru). Одним из сервисов платформы является сервис по отображению подробной информации о математическом понятии (сервис «Карточка понятия»), который использует данные онтологии OntoMathEdu.

Для связи сервиса «Карточка понятия» и содержимого учебного курса было разработано программное обеспечение, которое предназначено для автоматической разметки текста страниц курса понятиями из онтологии с возможностью ручного редактирования результатов разметки.

Основной задачей программного обеспечения является замена участков текстов, содержащих математические понятия из онтологии OntoMathEdu, на гиперссылки, ведущие на сервис «Карточка понятия».

В следующем разделе статьи будет приведено описание алгоритма и технологий, использованых при разработке данного сервиса.

АЛГОРИТМ РАЗМЕТКИ МАТЕМАТИЧЕСКИХ ТЕКСТОВ

Разметка математических текстов понятиями из онтологии выполняется по следующей схеме. Вначале тексты проходят стандартный процесс предварительной подготовки, а именно: разбиение на предложения, разбиение каждого предложения на слова и приведение слов к нормальной форме. При этом слова и предложения, длина которых меньше 4-х символов игнорируются.

Далее, в пределах каждого предложения выбираются существительные и прилагательные в полной форме (так как именно эти части речи используются в названиях понятий). Из набора отобранных слов формируются цепочки слов с длинами от 1 до N, где N равно количеству значимых слов в предложении. Таким образом, максимальное количество цепочек для предложения с количеством значимых слов N, составляет:

$$\sum_{k=1}^{n} C_n^k = \sum_{k=1}^{n} \frac{n!}{(n-k)!\,k!}$$

На практике, количество значимых слов в предложении не превышает 8, и количество цепочек не превышает 255, что делает работу алгоритма умеренно быстрой.

Следующим этапом алгоритма является выполнение тех же операций на названиях понятий, содержащихся в онтологии OntoMathEdu и поиск схожих множеств слов с помощью классической меры Жаккара (1), где A – это множество слов из цепочки, полученной из текста, а B – множество слов из названия понятия онтологии:

$$\mathrm{J} = \frac{|A \cap B|}{|A \cup B|} \qquad (1)$$

В случае полного совпадения указанных множеств слов, каждому слову в цепочке присваивается указатель на понятие из онтологии. В случае, когда одно слово связано с двумя и более понятиями, выбирается понятие с наибольшим количеством значимых слов (например, в тексте «… в прямоугольном треугольнике» на слово «треугольник» будут претендовать понятия «Треугольник» и «Прямоугольный треугольник», и в результате понятие «Треугольник» будет игнорироваться).

УЧЕТ КОНТЕКСТА ДЛЯ РАСПОЗНАВАНИЯ СОКРАЩЕННЫХ НАЗВАНИЙ МАТЕМАТИЧЕСКИХ ПОНЯТИЙ

Дополнительной задачей при поиске математических понятий в тексте является распознавание названий понятий, которые имеют сокращенную форму по сравнению с названием, заданным в онтологии.

Решение данной задачи состояло в следующем. При написании учебных текстов было рекомендовано разделять тексты на абзацы, при этом в первом предложении абзаца указывать главное понятие, характеризующее весь абзац. Таким образом, из первого предложения выделяется главное понятие, которое добавляется к каждой цепочке слов, составленных из предложений в текущем абзаце.

Так, к примеру, в онтологии есть понятие «Биссектриса треугольника», но нет понятия «Биссектриса», которое обычно используется в тексте в контексте треугольника. На рисунке 2 приведен фрагмент учебного текста, отражающий такую ситуацию. В первом предложении будет успешно распознано понятие «Треугольник», и добавлено в качестве временного элемента к цепочкам второго предложения.

Такой учет контекста заметно повышает вероятность распознавания понятий в текстах.

Рассмотрим треугольник (Треугольник) ABM. AK - биссектриса (Биссектриса треугольника).

Рисунок 2. Пример учебного фрагмента

ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ ДЛЯ ВИЗУАЛИЗАЦИИ И РЕДАКТИРОВАНИЯ РАЗМЕТКИ

В дополнение к алгоритму разметки математических текстов был разработан программный продукт для ручного редактирования результатов разметки, основной функционал которого позволяет:

1. Загрузить на страницу курса разметку страницы из файла и из буфера обмена;
2. Запустить процесс автоматической разметки для текущего документа;

3. Визуализировать результаты разметки и выделять различные понятия разными цветами;
4. Отмечать последовательность слов в тексте и привязывать к ней некоторое понятие из онтологии;
5. Выполнять предпросмотр веб-версии текущего документа с размеченными понятиями;
6. Автоматически сохранять текущие результаты разметки в файл.

На рисунке 3 приведен пример визуализации автоматической разметки документа. Оттенками зеленого выделены понятия, однозначно определенные алгоритмом разметки. Оттенками синего цвета выделены понятия, размеченные пользователем вручную.

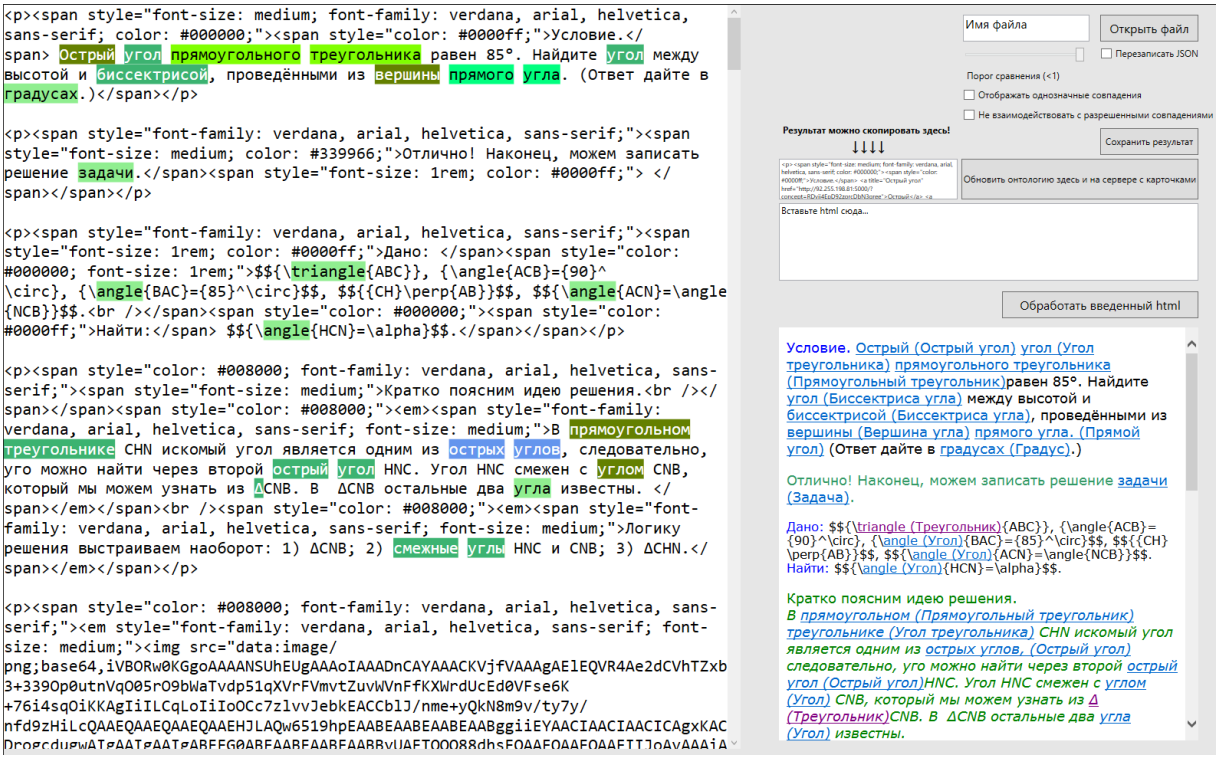

Рисунок 3. Интерфейс программы для визуализации и редактирования разметки

Помимо участков текста, явно отображающих слова из понятий онтологии, алгоритм разметки распознает символные и latex-обозначения математических терминов (например, символ Δ и тег \triangle распознаются алгоритмом раз-

метки как слово «Треугольник», а тег \angle – как слово «Угол»). Словарь соответствия символьных преставлений и математических терминов легко расширяется при необходимости. На рисунке 4 приведено сравнение версий документа до автоматической разметки и после.

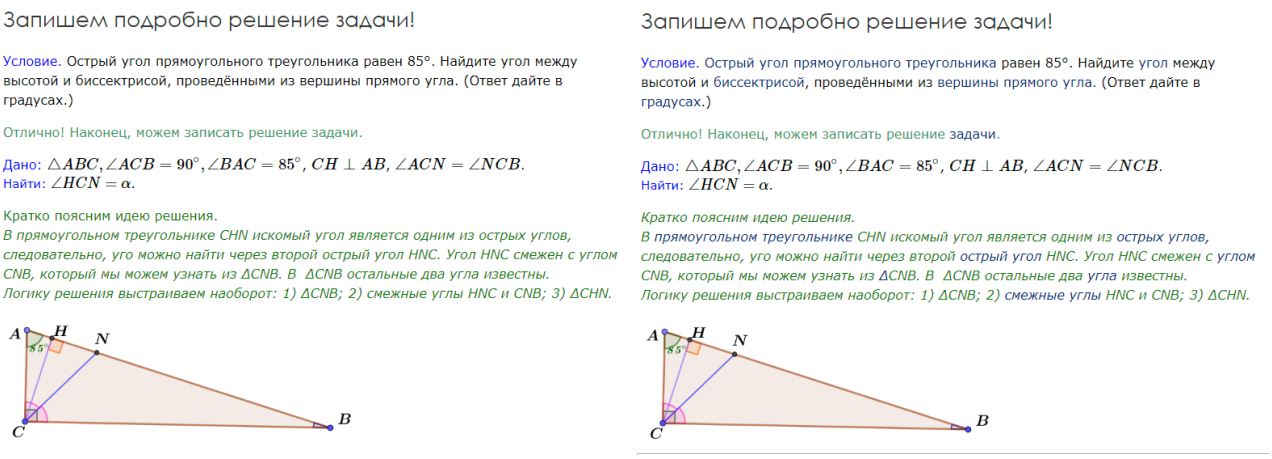

Рисунок 4. Сравнение исходного и размеченного документа

ЗАКЛЮЧЕНИЕ

В статье был описан алгоритм и программный продукт для автоматической разметки учебных текстов по планиметрии. Данный продукт будет использован при разработке учебных курсов в Казанском федеральном университете.

Программный продукт не требователен к формату входных данных и может быть использован для разметки других документов (не только математических), при наличии онтологии, отражающей соответствуюшую предметную область.

Благодарности

Исследование выполнено при поддержке Российского фонда фундаментальных исследований, грант № 19-29-14084.

СПИСОК ЛИТЕРАТУРЫ

1. *Cheniti Belcadhi L.* Personalized feedback for self-assessment in lifelong learning environments based on semantic web // Computers in Human Behavior. 2016. V. 55. P. 562–570.

2. *Da Nobrega G.M., de Araujo G.G., Cruz F.W.* Towards collaborative ontology construction for learning Computer Science in Education // 2021 International Conference on Advanced Learning Technologies (ICALT). 2021. P. 305–307.

3. *Tanwar S., Kumar Malik S.* Towards Blending Semantics with an Education Based Ontology Using Protege 5.2.0 A Revisit // 2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence). 2018. P. 733–738.

4. *Radovic M., Petrovic N., Tosic M.* An Ontology-Driven Learning Assessment Using the Script Concordance Test // Applied Sciences. 2022. V. 12. No. 3. P. 1472.

5. *Obeid C., Lahoud I., El Khoury H., Champin P.-A.* Ontology-based Recommender System in Higher Education // Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18. 2018. P. 1031–1034.

6. *Nikolaev K., Nevzorova O., Falileeva M.* Developing the OntoMathEdu Ecosystem for Educational Applications // DTTL 2021. 2021. P. 81–87.

THE METHOD OF AUTOMATIC SEMANTIC MARKUP OF MATHEMATICAL EDUCATIONAL TEXTS

Konstantin Nikolaev[1], Olga Nevzorova[2]

[1,2]Kazan (Volga Region) Federal University, Kazan

[1] konnikolaeff@yandex.ru, [2] onevzoro@gmail.com

Abstract

Automatic markup of the semantic content of a document is a popular task when creating educational and encyclopedic resources. In this paper, we propose an algorithm and software for marking mathematical texts with terms from the ontology of school mathematics OntoMathEdu, as well as visualization and manual editing of automatic markup. The presented results will be used in the development of a course on teaching planimetry in the digital education platform in KFU.

Keywords*: semantic markup, distance education, subject ontologies*

REFERENCES

1. *Cheniti Belcadhi L.* Personalized feedback for self-assessment in lifelong learning environments based on semantic web // Computers in Human Behavior. 2016. V. 55. P. 562–570.

2. *Da Nobrega G.M., de Araujo G.G., Cruz F.W.* Towards collaborative ontology construction for learning Computer Science in Education // 2021 International Conference on Advanced Learning Technologies (ICALT). 2021. P. 305–307.

3. *Tanwar S., Kumar Malik S.* Towards Blending Semantics with an Education Based Ontology Using Protege 5.2.0 A Revisit // 2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence). 2018. P. 733–738.

4. *Radovic M., Petrovic N., Tosic M.* An Ontology-Driven Learning Assessment Using the Script Concordance Test // Applied Sciences. 2022. V. 12. No. 3. P. 1472.

5. *Obeid C., Lahoud I., El Khoury H., Champin P.-A.* Ontology-based Recommender System in Higher Education // Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18. 2018. P. 1031–1034.

6. *Nikolaev K., Nevzorova O., Falileeva M.* Developing the OntoMathEdu Ecosystem for Educational Applications // DTTL 2021. 2021. P. 81–87.

СВЕДЕНИЯ ОБ АВТОРАХ

НЕВЗОРОВА Ольга Авенировна – доцент кафедры информационных систем Института вычислительной математики и информационных технологий Казанского федерального университета, к.т. н. Основные направления научных исследований: обработка естественного языка, искусственный интеллект.

Olga Avenirovna NEVZOROVA – Kazan Federal University, Institute of Computational Mathematics and Information Technologies, Associated Professor of the Department of Information System, PhD. Major fields of scientific research are Natural Language processing, artificial intelligence.

email: onevzoro@gmail.com
ORCID: 0000-0001-8116-9446

НИКОЛАЕВ Константин Сергеевич – ассистент кафедры системного анализа и информационных технологий Института Вычислительной математики и информационных технологий Казанского федерального университета. Основные направления научных исследований: обработка естественного языка, искусственный интеллект.

Konstantin Sergeevich NIKOLAEV – Assistant of the Department of System Analysis and Information Technologies of the Institute of Computational Mathematics and Information Technologies of Kazan Federal University. Major fields of scientific research are Natural Language processing, artificial intelligence.

email: konnikolaeff@yandex.ru
ORCID: 0000-0003-3204-238X

Материал поступил в редакцию 26 января 2022 года