

ОРГАНИЗАЦИЯ ХРАНЕНИЯ МЕТАДАННЫХ НАУЧНЫХ СТАТЕЙ

Лайком Д. Н.

Силич В. А.

Томский политехнический университет

wedun@tpu.ru

Введение

В современной науке большую роль играет документирование научных исследований. Научные статьи позволяют систематизировать результаты проведённого эксперимента. Опубликованные статьи позволяют эффективно делиться результатами проделанной работы с коллегами, работающими в этой или смежных научных областях. Также большую ценность представляют метаданные, по которым написаны статьи.

При больших объемах проанализированной информации необходимо организовать эффективный способ хранения метаданных статей. Метаданные могут быть использованы для развития научного направления, по которому была написана статья, либо для опровержения результатов проделанной работы.

В связи с этим учёные всё чаще сталкиваются с проблемой неэффективной работы с публикациями из-за высокой сложности поиска информации в публикациях, большого времени поиска, высоких затрат на анализ опубликованных статей. Дополнительной проблемой является комплексная система для комплексной обработки информации. Для успешного воспроизведения эксперимента ученым необходимо знать исходный набор данных, используемые методы исследования, промежуточные данные и результаты исследования. Все эти данные вместе с публикациями авторов требуют эффективной обработки. Помимо эффективного хранения требуется обеспечивать поиск по результатам экспериментов с целью упрощения доступа к результатам прошедших экспериментов. Доступ необходим для обеспечения повторяемости экспериментов (для подтверждения, опровержения или развития результатов прошедших экспериментов). Помимо упрощения работы над результатами прошедших экспериментов необходимо разработать метод предсказания наборов исходных данных, которые будут полезны в будущих экспериментах [1].

Описание метаданных научных статей

К метаданным статей можно отнести информацию об авторах, выпусках журналов, статьях и источниках, по которым была написана статья. Выделим список атрибутов для каждого из типов метаданных:

Метаданные для авторов:

- Имя

- Фамилия
- Id автора во внешних системах
- Место работы

Метаданные для статей:

- CDSID
- CDSInternal
- CDSReportNumber
- ArXivCode
- DOI
- Введение
- Дата создания
- Ключевые слова

Метаданные публикаций:

- Объем статьи
- Заголовок
- Год выпуска

Этим набором параметров можно однозначно описать опубликованные статьи и их авторов. Кроме этого вместе со статьями необходимо хранить наборы данных, которые были проанализированы при написании статей.

Цель исследования

Решить проблему хранения метаданных научных статей и последующего воспроизведения результатов экспериментов [2].

Задачи исследования:

1. Определить текущую модель хранения экспериментальных данных.
2. Определить схему хранения результатов экспериментов.
3. Подобрать и настроить программный комплекс для хранения и представления результатов экспериментов.
4. Перенести текущие результаты эксперимента в новую систему.
5. Настроить автоматическую работу каталога знаний.

Общие требования к системе:

- Качество анализа статей (извлечение семантической информации)
 - Скорость анализа статей
 - Полнота предоставления информации
 - Возможность анализа данных в базе (Агрегация информации из разных источников)
 - Стоимость разработки и внедрения
- Выбор СУБД

На этапе проектирования системы был сформирован ряд требований к СУБД:

1. Работать с неструктурированными данными.
2. Обработать сложные запросы без прерывания работы.

3. Иметь большую скорость чтения при работе в распределённой среде.
4. Иметь возможность атомарно хранить информацию об объектах системы.

Недостатком реляционных баз данных является высокая сложность написания запросов, а также сложностью гарантированно получить результат из-за высокой сложности операций, выполняемых в запросе [3].

Проанализировав требования было решено использовать СУБД Virtuoso. Кроме изложенных выше требований она имеет функциональный и разветвлённый API.

Способ хранения данных

Зная, что в базе будут собраны научные статьи из разных баз данных, было принято решение для каждого объекта в базе данных использовать уникальные идентификаторы (GUID). Объекты в базе данных Virtuoso описываются при помощи триплетов вида: «субъект-предикат-объект». Пример такоготриплета:

<document/1072d978-3ac1-4fa6-b3bc-854493a6a7c6><hasKeyword>'QCD'.
ЯзыкомзапросовкбазеданныхявляетсяязыкSPARQL. Он позволяет эффективно и выразительно описывать запросы к данным.

Способы решения задач

Исходя из поставленных задач сбыли сформированы способы решения каждой из них.

Таблица 1. Задачи системы и способы их решения.

Название задачи	Способ решения
Метод обработки научных статей.	Проанализировать тексты научных статей, выделить ключевые слова. Произвести лексический и синтаксический разбор приложений в тексте. Проанализировать рисунки и табличные данные.
Технология представления информации по запросу.	На основе обработанных данных статей, разработать систему, которая позволит ученым получать информация о публикациях по заданной теме.
Методика предсказание наборов данных для будущих исследований.	На основе проанализированных текстов научных статей в автоматическом режиме строить карту вероятности использования заданных наборов данных для будущих исследований.

Функции системы

К основным функциям разрабатываемой системы можно отнести:

Функции системы

- Метод обработка научных статей.
- Технология представления информации по запросу.
- Методика предсказание наборов данных для будущих исследований.

Вывод

Предложен новый метод организации метаданных научных публикаций. Он позволяет решить проблему эффективного хранения и обработки метаданных статей.

Современные исследования проводятся большим количеством научных коллективов. Для совместной работы необходим эффективный инструмент для обмена результатами исследований. Для обеспечения эффективной работы системы хранения документов необходимо учитывать особенности работы используемых СУБД.

Полученные результаты могут быть использованы структурными подразделениями государственных и образовательных учреждений в процессах упорядочивания баз данных научных работ, опубликованных их сотрудниками, а также другими научными коллективами при выполнении научно-исследовательских работ.

Список использованных источников

1. 1.Ахаев А.В. Формирование базы знаний экспертной системы на основе онтологии с использованием оригинального языка представления знаний/ А.В. Ахаев, И.А. Ходашинский // Электронные средства и системы управления. – 2013. - № 2. – С. 3-7
2. 2. Тузовский А.Ф. Разработка систем управления знаниями на основе единой онтологической базы знаний / А.Ф. Тузовский // Известия томского политехнического университета. - 2007 – № 2. – С. 182-185
3. 3. Park S.C. Dynamic rule refinement in knowledge-based data mining systems / S.C. Park, S. Piramuthu, M.J. Shaw // Dynamic rule refinement in knowledge-based data mining systems. - № 2. – С. 205-222
4. 4. Квятковская И. Ю. Использование онтологий для создания баз общих знаний при классификации информации о предметной области / И. Ю. Квятковская // Вестник Астраханского государственного технического университета. – 2007. - № 4. – С. 22
5. 5. Боровский А. А. Перспективы применения технологий машинного обучения к обработке больших массивов исторических данных / А. А. Боровский // Кибернетика и программирование. – 2015. - № 1. – С. 77-11