УДК 004.048

РЕПОЗИТОРИЙ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

Т.Б. Шатовская[1], В.Б. Репка[2], И.В. Каменева[3], М.М. Марченко[4]

1ХНУРЭ, г. Харьков, Украина, shatovska@gmail.com,
2ХНУРЭ, г. Харьков, Украина, vic-repka@yandex.ru,
3ХНУРЭ, г. Харьков, Украина, irina.kamenieva@gmail.com,
4ХНУРЭ, г. Харьков, Украина, shami13@gmail.com;

В данной статье представлен анализ известных репозиториев статистических наборов даннях, а также на основе проведенного анализа была разработана модель интеллектуального научного репозитория статистических наборов даннях.

РЕПОЗИТОРИЙ, ИНТЕЛЛЕКТУАЛЬНЫЕ АГЕНТЫ, JADEX, ВЕБ-СЕРВИС, ORACLE, RDF, FTP

Вступление

Для проведения исследований в области интеллектуального анализа данных и машинного обучения необходимым условием является наличие различных входных наборов данных. В настоящее время исследователи создают свои базы данных таких наборов. Примером такой системы может служить: The UCI Machine Learning Repository [1], который является наиболее популярным среди исследователей благодаря своей классификации наборов данных, а также содержанием множества наиболее часто встречающихся выборок, Data Envelopment Analysis Dataset Repository [2], XMLData Repository [3], Frequent Itemset Mining Dataset Repository [4]. Наряду с вышеуказанными статистическими репозиториями целая плеяда от простых хранилищ файлов до специализированных хранилищ данных может быть использована исследователями при решении прикладных задач, проведении исследований собственных алгоритмов и научных проблем. Казалось бы, единственной сложностью для пользователя будет являться поиск и непосредственное понимание структуры столь разрозненных хранилищ информации. Однако детальное исследование таких хранилищ данных приводит нас к осознанию наличия более глубоких проблем в использовании данных. В частности наблюдается полное несогласование и жесткость структуры файлов с данными с SDMX – Statistical Data and Metadata Exchange – стандартом и структурой, используемой многими европейскими организациями, невозможность предварительной подготовки данных к конкретной прикладной задаче, отсутствие истории использования данных для тех или иных научных и прикладных задач.

1. Постановка задачи

Целью данной работы является концептуально новый подход к созданию интеллектуального хранилища статистических данных, основанного на взаимодействии онтологических моделей методов Data mining, пользователей системы и агентном подходе хранения и обработки файлов репозитория данных.

2. Анализ статистических репозиториев

The UCI Machine Learning Repository. Все данные о наборах данных хранятся в виде файлов как ftp-хранилище. Описание каждого набора данных также хранится в виде файла, что замедляет скорость поиска информации.

Преимущества репозитория: хорошо отсортированные данные, полнотекстовый поиск.

Недостатки репозитория: только текстовый формат данных, неудобен в использовании и для преобразований. Отсутствует система персонализации, пользователь может добавить выборку только по жесткому шаблону, что требует дополнительных затрат на администрирование системы.

Data Envelopment Analysis Dataset Repository (DEA Dataset Repository) позволяет пользователям загружать и искать наборы данных.

Преимущества репозитория: поиск по любому из критериев.

Недостатки репозитория: формат файлов, интерфейс. Поиск осуществляется только по одному из множества критериев, то есть нельзя совместить поиск по нескольким условиям. Неопытным пользователям сложно осуществлять поиск по данному репозиторию, нет просмотра всех наборов данных репозитория.

XML Data Repository. В XML Data Repository собраны данные только в виде XML, а также статистические данные для использования в научно-исследовательских экспериментах.

Недостатки данной системы: неудобство поиска, нет понимания, для каких задач можно использовать данную выборку, недостаточный объем информации по данным. Нет дополнительной информации о наборах данных, некоторые наборы данных очень большие по объему.

Преимущества: универсальный формат данных XML, прост для преобразования в любой другой формат и для программного использования. Отсутствие регистрации и получение данных по протоколу http позволяет использовать репозиторий агентами для поиска данных.

Frequent Itemset Mining Dataset Repository содержит небольшое количество наборов данных, которые не содержат классификации.

Недостатки данной системы: неудобство поиска, нет отображения дополнительной информации о наборах данных, некоторые наборы данных недоступны для скачивания.

Преимущества: для каждого набора данных существует описание экспериментального использования.

В конечном итоге были выявлены следующие недостатки существующих репозиториев: наличие только текстового формата данных, отсутствие систем персонализации, неудобство пользовательского интерфейса, неудобство поиска, недостаточный объем информации о данных.

3. Репозиторий интеллектуального анализа научных наборов данных

Система репозитория интеллектуального анализа научных наборов данных предназначена для использования исследователями, которые проводят эксперименты в области интеллектуального анализа данных и машинного обучения. Основной целью данной системы является хранение наборов данных, их автоматическая классификация, поиск, предобработка, персонализация работы с системой. Система представляет собой веб-приложение, которое использует агентную платформу Jadex[5], [6]. Общая схема работы системы изображена на рис. 1.

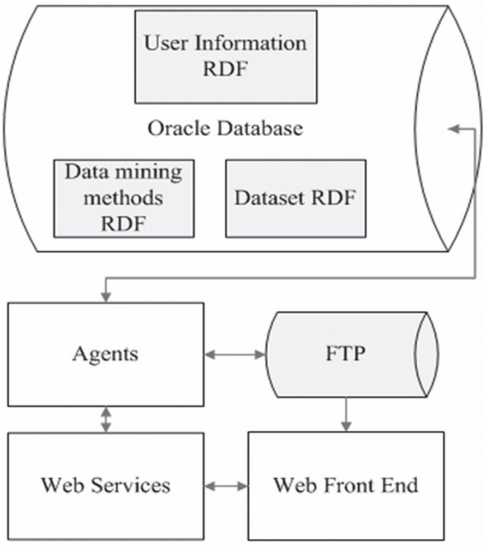

Рис. 1. Общая схема работы системы

Система состоит из следующих блоков: сервер баз данных, FTP сервер, веб-сервисы, агентная платформа Jadex, веб-интерфейс.

Сервер баз данных Oracle используется для хранения информации о пользователях системы, о методах исследований и хранит информацию о мета данных для наборов данных [7], [8].

FTP сервер используется для хранения файлов наборов данных.

Агентная платформа Jadex используется для проведения бизнес-логики системы [9].

Для реализации поиска в репозитории был разработан поисковый агент, основанный на цели, который может учитывать текущее состояние системы и различные критерии поиска, для того чтобы принять ту или иную стратегию поиска. При этом интеллектуальный агент действует не просто рефлексивно, но и принимает решение: какие именно действия необходимо выполнить, для того чтобы достичь поставленной цели в условиях текущего состояния среды.

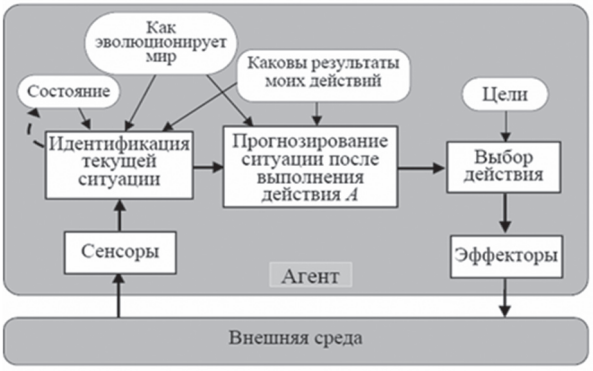

Рис. 2. Архитектура агента основанного на целях

Для реализации поискового агента в Jadex используются такие понятия как знания (beliefs), факты (facts), планы (plans), события (events), а также цели (goals). На основании имеющихся знаний и событий окружающей среды поисковый агент выбирает тот или иной план для достижения поставленной цели – поиска наиболее релевантных данных по запросу [9], [10].

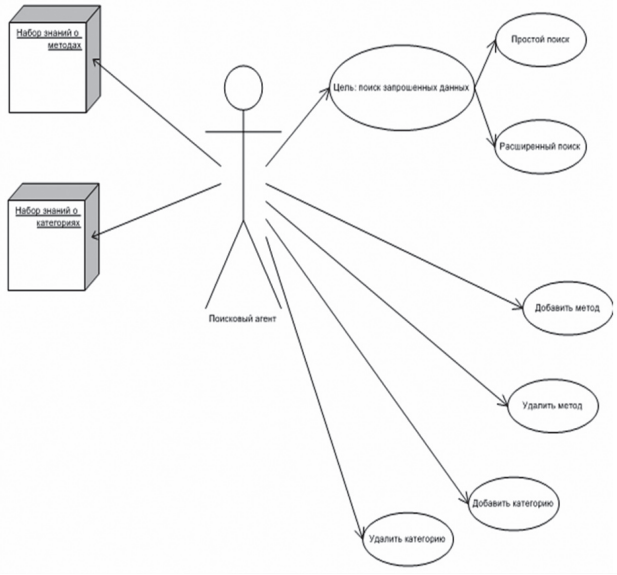

Рис. 3. Структурная схема поискового агента

Структурно-поисковый агент состоит из двух основных планов (SimpleSearchPlan и ExtendedSearchPlan) и нескольких вспомогательных (AddMethodPlan, RemoveMethodPlan, AddCategoryPlan, RemoveCategoryPlan), также двух наборов знаний.

В поисковом агенте все планы имеют пассивную стратегию работы. То есть это подразумевает, что для обработки каждого сообщения создается новый объект плана, и после обработки сообщения этот объект удаляется. Такая стратегия позволяет обрабатывать несколько сообщений одновременно.

В качестве входных данных для агента применяются сообщения с различным содержанием: строка с запросом поиска; набор пар: ключ и значение, а также события добавления или удаления сведений об имеющихся в хранилище методах и категориях. Результатом работы агента является набор ссылок на найденные данные.

Главной целью агента является поиск методов в хранилище. Для достижения этой цели агент использует SimpleSearchPlan и ExtendedSearchPlan планы. Для SimpleSearchPlan плана входными данными является строка, содержащая запрос поиска. При этом в строке запроса могут использоваться булевы выражения, например, “и”, “или”, “нет” и их комбинации. А ExtendedSearchPlan план используется для расширенного поиска. Входными данными является набор, содержащий пары: ключ и значение. На основании полученных данных поисковый агент строит запрос к репозиторию данных [11], [12].

Для того чтобы интегрировать Jadex платформу с веб, был разработан внешний интерфейс – Jadex Web Service eXtension (JWSX). Суть этого интерфейса заключается в создании внешнего доступа к функциям агента через веб сервисы [13], [14].

Веб-сервисы используются для взаимодействия пользователя веб-интерфейса с агентной платформой.

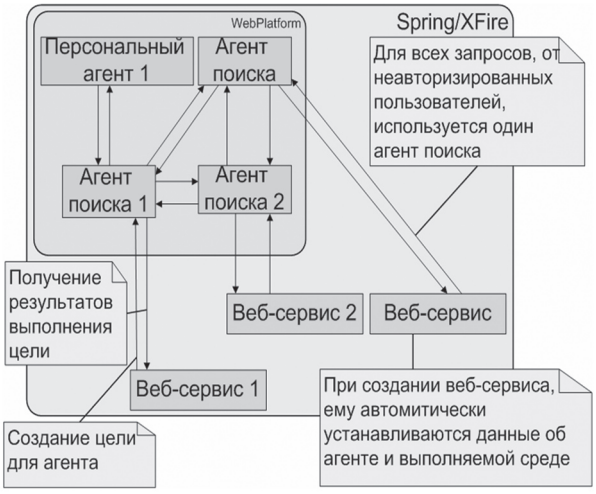

Рис. 4. Интеграция агента поиска с веб-сервисами

Варианты использования данной системы изображены на рис. 5.

В свою очередь каждый их представленных вариантов использования можно уточнить, используя несколько вариантов использования.

Рис. 5. Варианты использования системы

Вариант использования «Регистрация» включает в себя следующие варианты использования:

- регистрация начинающего пользователя;
- регистрация продвинутого пользователя.

Вариант использования «Просмотр наборов данных» включает в себя следующие варианты использования:

- просмотр всех наборов данных, находящихся в системе;
- просмотр конкретного набора данных в расширенном виде: все метаданные, комментарии, оценки.

Вариант использования «Фильтрация наборов данных» включает в себя следующие варианты использования:

- фильтрация наборов данных по методу исследования;
- фильтрация наборов данных по параметрам метаданных.

Вариант использования «Менеджмент комментариев» включает в себя следующие варианты использования:

- редактирование комментария;
- удаление комментария.

Вариант использования «Менеджмент пользователей» включает в себя следующие варианты использования:

- редактирование данных о пользователе;
- удаление пользователей;
- изменение статуса пользователя;
- назначение пользователя администратором.

Вариант использования «Менеджмент наборов данных» включает в себя следующие варианты использования:

• редактирование метаданных о наборе данных;
• редактирование местоположения файла набора данных;
• удаление набора данных;
• добавление методов исследований для использования набором данных.

Вариант использования «Менеджмент методов исследований» включает в себя следующие варианты использования:
• добавление метода исследования;
• редактирование данных о методе исследования;
• удаление метода исследования;
• изменение в иерархической структуре методов исследования.

Все варианты использования системы распределяются между агентами.

Данный репозиторий будет иметь следующие функциональные возможности:

а) регистрация пользователей;

б) аутентификация пользователей;

в) менеджмент научных наборов данных: добавление новых выборок в систему, редактирование выборок, удаление выборок;

г) менеджмент методов исследования: добавление новых методов, редактирование методов, удаление методов;

д) поиск по онтологии наборов данных по таким критериям: тип задачи; предметная область; имя набора данных; ключевые слова; вид (шкала) входной/выходной характеристики; количество элементов (объем); дата (помещения в репозиторий, модификации, создания); автор задачи/набора данных/экспериментов; методы, какими решалась задача; эффективность работы набора данных.

е) добавление результатов работы с выборками;

ж) определение степени валидности набора данных.

Выводы

Проведен сравнительный анализ современных поисковых алгоритмов в репозиториях научных наборов данных и поисковых сервисов, выявлены их преимущества и недостатки; разработан поисковый агент, исследована и реализована возможность интеграции интеллектуальных агентов и веб-сервисов.

Данная система позволит исследователям иметь общее рабочее пространство для подготовки к экспериментам с алгоритмами интеллектуального анализа данных и машинного обучения, используя основные функции системы, такие как: загрузка наборов данных, поиск наборов данных, поиск исследователей, интересующихся теми же методами анализа данных.

В дальнейшем данную систему можно улучшить развитием входящих в нее онтологий и увеличением числа агентов. Также можно добавить подсистему работы со статьями и результатами научных экспериментов исследователей, которые проводились с использованием наборов данных из репозитория.

Список литературы: 1. The UCI Machine Learning Repository [Электронный ресурс]. – Режим доступа: www/ URL: http://archive.ics.uci.edu/ml/ -09.09.2009 г. – Назв. с экрана. 2. Data Envelopment Analysis Dataset Repository [Электронный ресурс]. – Режим доступа: www/ URL: http://www.etm.pdx.edu/DEA/Dataset/default.htm/ – 09.09.2009 г. – Назв. з екрану.3. XMLData Repository [Электронный ресурс]. – Режим доступа: www /URL: http://www.cs.washington.edu/research/xmldatasets/ – 09.09.2009 г. – Назв. с экрана.4. Frequent Itemset Mining Dataset Repository [Электронный ресурс]. – Режим доступа: www/URL: http://fimi.cs.helsinki.fi/data/ – 09.09.2009 г. – Назв. с экрана. 5. Jadex BDI Agent System [Электронный ресурс]. – Режим достпа: www/URL:http://jadex.informatik.unihaburg.de/bin/view/About/Overview/ – 09.09.2009 г. – Назв. с экрана. 6. Jadex User Guide [Электронный ресурс]. – Режим доступа: www /URL: http://jadex.informatik.uni-hamburg.de /docs /jadex-0.96x /userguide /index.single.html/ – 09.09.2009 г- Назв. с экрана.7. Oracle Database 10g [Электронный ресурс]. – Режим доступа: www/URL:http://www.oracle.com/technology/products /database /oracle10g/index.html/ – 09.09.2009 г. – Назв. с экрана. 8. Oracle Database 10g Release 2 Spatial [Электронный ресурс]. – Режим доступа: www /URL: http://www.oracle.com/technology/products/spatial/10gr2_tech_info.html/ – 09.09.2009 г. – Назв. с экрана. 9. Jadex Tutorial [Электронный ресурс]. – Режим доступа: www/URL: http://jadex.informatik.uni-hamburg.de/docs /jadex0.96x/tutorial/index.single.html – 09.09.2009 г. – Назв. с экрана. 10. The Foundation of Physical Intelligent Agents [Электронный ресурс]. – Режим доступа: www/ URL: http://www.fipa.org/ – 09.09.2009 г. – Назв. с экрана. 11. Агентно-ориентированные системы: от формальных моделей к промышленным приложениям [Электронный ресурс]. – Режим доступа: www/ URL: http://window.edu.ru/window_catalog/pdf2txt?p_id=27142&p_page=7/ – 09.09.2009 г. – Назв. с экрана. 12. Гаврилова, Т. А. Базы знаний интеллектуальных систем [Текст]: учеб. / Т.А. Гаврилова, В.Ф. Хорошевский. – СПб.: Питер, 2000. – 384 с. 13. Hendler J. Agents and the Semantic Web, IEEE Intelligent Systems [Текст] / Hendler J. – John Wiley and Sons. – Apr. 2001, pp. 30–37. 14. Codehaus XFire – Home [Электронный ресурс] – Режим доступа: www/ URL: http://xfire.codehaus.org/ – 09.09.2009 г. – Назв. с экрана.

Поступила в редколлегию 28.10.2009

УДК 004.048

Репозитарій інтелектуального аналізу даних / Т.Б. Шатовська, В.Б. Репка, І.В. Каменева, М.М. Марченко // Біоніка інтелекту: наук.-техн. журнал. – 2009. – № 2 (71). – С. 75-78.

Проводиться порівняльний аналіз сучасних статистичних репозитаріїв. Представлена загальна схема роботи репозитарія інтелектуального аналізу даних, а також варіанти використання репозитарія інтелектуального набору даних.

Іл. 2. Бібліогр.: 14 назв.

УДК 004.048

Data Mining Repository / T.B.Shatovska, V.B. Repka, I.V. Kamenieva, M.M. Marchenko // Bionics of Intelligence: Sci. Mag. – 2009. – № 2 (71). – P. 75-78.

In the article dissects modern statistical repositories. A general scheme of data mining repository is present and variants of usage of Data mining repository also.

Fig. 5. Ref.: 14 items.