УДК 303.732.4:332.1

ПОИСК СВЯЗЕЙ МЕЖДУ ИНДИКАТОРАМИ РАЗВИТИЯ СТРАН В БАЗЕ ДАННЫХ ВСЕМИРНОГО БАНКА

Носов А.С., Аверин Г.В.
Донецкий национальный технический университет
кафедра компьютерных систем мониторинга
E-mail: aleksandr.nosov.1991@mail.ru

Аннотация

Носов А.С., Аверин Г.В. Поиск связей между индикаторами развития стран в базе данных Всемирного банка. *Рассмотрена структура базы данных индикаторов развития стран мира предоставляемая Всемирным банком. Разработана структура системы интеллектуального анализа данных использующая базу данных Всемирного банка. Рассмотрены основные индикаторы и группы индикаторов развития стран мира. Рассмотрены основные направления и задачи Data Mining.*

Общая постановка проблемы

В настоящее время анализ социально-экономического развития стран мира осуществляется преимущественно путем применения индикативного метода. Суть метода заключается в выборе наиболее важных социальных, экономических и экологических показателей развития стран, которые отражают приоритетные стороны развития общества. В качестве индикаторов обычно используют:

- количество ВВП на душу населения;
- заболеваемость туберкулёзом;
- количество социальных выплат на душу населения;
- уровень образованности населения;
- импорт товаров и услуг;
- экспорт товаров и услуг;
- индекс человеческого развития;
- показатель экологического следа.

Общее количество индикаторов может достигать более полутысячи показателей, отражающих наиболее важные стороны общества. Задача поиска связей между индикаторами развития стран мира является актуальной, однако мало изученной. Отыскание связей между различными процессами, протекающими в странах мира, позволяет осуществить качественный прогноз в развитии изучаемого процесса и, возможно, впоследствии обеспечить принятие оптимального управленческого решения с целью защиты национальных интересов государства или общества.

Сегодня задача установления связей между индикаторами развития стран мира считается крайне актуальной среди экспертов международных организаций. Это связано с тем, что сегодня накоплены большие объёмы социально-экономической информации, которые позволяют применить методы интеллектуального анализа данных. В литературе имеются работы, которые связаны с применением методов классификации, кластеризации, ассоциации, деревьев решений, аналогий, многомерной визуализации при установлении связей между индикаторами социально-экономического развития или при ранжировании стран по уровню их развития. Подобный анализ невозможно сделать без современных баз данных и методов анализа больших объёмов информации.

Сложность данной задачи также связана с необходимостью использования различных форматов данных, наличием пробелов во временных рядах данных, а также сложностью используемых методов анализа.

Структура базы данных Всемирного банка

Существует много различных баз данных социально-экономического развития. Наиболее известная их них – это база данных индикаторов развития стран мира Всемирного банка с 2010 года стала доступной в открытом доступе сети Internet по адресу http://databank.worldbank.org/databank/_download/WDIandGDF_excel.zip. Она предоставляется в виде Microsoft Excel документа и имеет объём 61,8 Mb. Внутренняя структура таблицы имеет следующий вид:

- 1-й столбец код индикатора развития;
- 2-й столбец название индикатора;
- 3-й столбец код страны;
- 4-й столбец название страны;
- 5-55 столбцы – числовые значения индикаторов по годам наблюдений с 1960 по 2010 гг.

2-й и 4-й столбцы таблицы имеют специфический формат. Значения 4-го столбца повторяются циклически, в то время как значения 2-го столбца повторяются, пока не закончится цикл 4–го столбца и далее следует название следующего индикатора.

В значениях индикаторов(5-55 столбцы) присутствуют пробелы, причём количество пробелов увеличивается по мере убывания значения года в заголовке столбца.

Данная база данных включает в себя 18 различных категорий индикаторов:

- сельское хозяйство и развитие сельских районов 23 индикатора;
- повышению эффективности внешней помощи 25 индикаторов;
- изменение климата 42 индикатора;
- экономической политики и внешнего долга 38 индикаторов;
- образование 34 индикатора;
- энергия и горная промышленность 10 индикаторов;
- окружающая среда 27 индикаторов;
- финансовый сектор 27 индикаторов;
- пол 54 индикатора;
- здоровье 36 индикаторов;
- инфраструктура 29 индикаторов;
- труда и социальной защиты 24 индикатора;
- бедность 17 индикаторов;
- частный сектор 30 индикаторов;
- государственный сектор 16 индикаторов;
- наука и технологии 12 индикаторов;
- социального развития 26 индикаторов;
- градостроительство 17 индикаторов;

Целью данной работы является преобразование базы данных Всемирного банка в удобный для пользователя вид, подключить к этой базе данных открытые библиотеки интеллектуального анализа данных и разработать программный модуль, представляющий Wiki систему, которая ориентирована на пользователей, не являющихся специалистами в области информационных технологий. Решение данной задачи позволяет обеспечить доступ большого количества специалистов к анализу данных, которые заинтересованных в изучении социально-экономического развития стран и регионов мира.

Функциональные возможности библиотек статистического анализа

Открытые библиотеки статистического анализа позволяют выполнить следующие задачи анализа данных:

- получение описательных статистики;
- анализ многомерных таблиц;
- подгонка распределений;
- построение многомерной регрессии;
- построение нелинейной регрессии;
- построение логит и пробит регрессии;
- дискриминантный анализ;
- анализ соответствий;
- кластерный анализ;
- факторный анализ;
- многомерное шкалирование;
- анализ выживаемости;
- построение структурных моделей;
- построение деревьев классификации;
- прогнозирование временных рядов;
- анализ канонической корреляции;
- получение непараметрической статистики;
- анализ компонент дисперсии;
- логлинейный анализ таблиц частот;
- анализ надежности предпочтений;
- дисперсионный анализ;
- ковариационный анализ;
- анализ Монте-Карло;

Определение и задачи интеллектуального анализа данных

Интеллектуальный анализ данных или Data Mining - это процесс обнаружения в сырых данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

Интеллектуальный анализ данных предназначен для поиска в больших объемах данных неочевидных, объективных и полезных на практике закономерностей.

Задачи интеллектуального анализа данных:

- классификация - формализованная задача, в которой имеется множество объектов (ситуаций), разделённых некоторым образом на классы;
- кластеризация - задача разбиения заданной выборки объектов (ситуаций) на подмножества, называемые кластерами, так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались;
- ассоциация – задача поиска ассоциативных правил;
- визуализация – визуализация исходных данных и результатов применения Data Mining;
- прогнозирование;
- анализ и обнаружение отклонений;
- оценивание;
- анализ связей.

Не все эти методы могут быть применены к данной задаче. Однако многие из них могут быть использованы в связи с тем, что существуют разработанные и открытые алгоритмы.

Структура проектируемой системы анализа индикаторов развития стран мира

Структурная схема проектируемой системы доступа и анализа к данным изображена на рисунке 1.

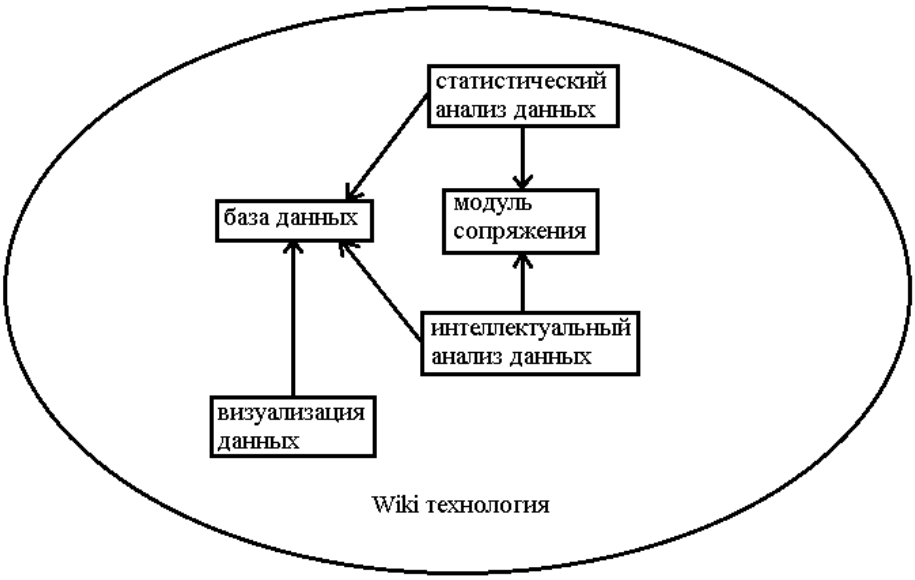

Рисунок 1 – Структурная схема проектируемой системы

Модуль базы данных может быть разработан с помощью технологии MySQL. Модуль статистического анализа данных разрабатывается с помощью языка программирования Visual C++. Модуль интеллектуального анализа данных планируется взять и использовать из открытого доступа сети Internet. Модуль визуализации данных удобнее всего выполнить с помощью одного из Web ориентированных языков программирования: PHP, ASP.NET, Perl, Ruby.

Выводы

Данная система создается, и она будет иметь спрос среди людей, занимающихся исследованиями в области социально-экономического развития. В перспективе планируется

Список литературы

1. База данных всемирного банка / Интернет ресурс. – Режим доступа: www/ URL: http://data.worldbank.org.
2. Википедия / Интернет ресурс. – Режим доступа: www/ URL: http://ru.wikipedia.org/wiki/Data_mining
3. Описание методов Data Mining / Интернет ресурс. – Режим доступа: www/ URL: http://www.basegroup.ru/library/methodology/ data_mining/
4. Описание методов Data Mining / Интернет ресурс. – Режим доступа: www/ URL: http://www.inftech.webservis.ru/it/database/datamining/ar2.html
5. Описание Wiki технологии / Интернет ресурс. – Режим доступа: www/ URL: http://ru.wikipedia.org/wiki/Вики
6. Описание задачи классификации / Интернет ресурс. – Режим доступа: www/ URL: http://ru.wikipedia.org/wiki/Задача_классификации