QUT ife Institute for Future Environments

QUT Science and Engineering

CRDC Sustainability Repository and Data Graph Builder Version 1.0 Final Report

Richi Nayak

Sachi Banduthilaka

Sangeetha Kutty

Samuel T. Smith

Levi Swann

Erin E. Peterson

CRDC Project QUT1705 October 2019

About us

The Institute for Future Environments (IFE) is a multidisciplinary research institute at Queensland University of Technology (QUT) in Brisbane, Australia. Hundreds of QUT researchers and students from across the fields of science, engineering, law, business, education and the creative industries collaborate at the IFE on large-scale research and development projects. Our mission is to generate knowledge, technology and practices that make our world more sustainable, secure and resilient.

Table of Contents

Table of Figures

1 Introduction

The CRDC Sustainability Repository and Data Graph Builder system has been designed and developed to ease the workload of CRDC and other relevant staff to collect and process the sustainability reporting data on Australian Cotton indicators and targets. The Australian research body on Cotton, CRDC (Cotton Research and Development Corporation) publishes data about 120 cotton indicators on a quadrennial sustainability report explaining their sustainable practices on Australian cotton farms. The standard practice is to utilize a manual approach to extract information from heterogeneous sources.

Industries like Agriculture, where the IT resources are scarce, may not possess a centralized repository with temporal and spatial information. This becomes more difficult when the data is scattered over diverse locations in diverse formats. The QUT-CRDC project (2016 – 2019) started with collecting information on various social, environmental and economic indicators and targets, and proposing new ones where applicable. The final component of the project focused on studying the feasibilities of using heterogeneous data sources to extract useful knowledge on cotton indicators and propose an autonomous system that allows to collect, extract, process and query the relevant sustainability indicators information. A data source can be a pdf document, a doc file, an excel sheet or a html page that may contain relevant information for cotton indicators.

In this project, a novel data mining based methodology has been developed to automate the data acquisition, processing and reporting of cotton sustainability indicators information that may be available on multiple heterogeneous data sources. The intuitive tool based on this methodology provides access to social, economic and environmental sustainability indicators, enabling users to generate information and graphics that communicate repository query results to stakeholders efficiently and effectively.

The project team consisted of experts on data mining, software engineering, visualization, environmental science, and design science.

This report gives a brief description of the prototype of the “Sustainability Repository and Data Graph Builder” system.

2 Purpose of the system

Reporting factual information about the agricultural cycle gives a better overview of economic, social and environmental aspects of an industry. In good practice, many agricultural industries release an annual report about their sustainable approach and results to illustrate sustainability performance.

The Australian Cotton Sustainability report publishes once every four years. It contains information on about 120 cotton indicators. It is a challenge to extract all relevant information manually with limited manpower. CRDC faced common problems such as the lack of the information available on sustainability aspects or the information is dispersed throughout the online repositories or intranet. Another common difficulty was the lack of continual values on these aspects so a temporal evolving pattern can be observed.

The automated repository and reporting system is designed to extract relevant knowledge for cotton indicators, store information in a centralized repository and present them in a visualized form as graphs. The final graphs are ready to download as image files and can be directly used in sustainability reports.

The proposed system is supposed to reduce the workload of CRDC staff in order to collect the relevant data that is a time-consuming and rigorous task. It allows the relevant data, that is distributed across multiple sources, to be stored in a centralized repository. It enables users to generate the graphs on an indicator of interest in an automatic fashion. This empowers the stakeholders to know the trends of an indicator and benchmark its performance.

The proposed design contains two main components: (1) a back end module which extracts data from heterogeneous sources and organizes the relevant information in a database; and (2) a front end module which allows users to interact with the data via a visualization interface where relevant graphs can be built with a minimal effort.

3 Concept

The back end module of the system is called as CRDC Information Repository System and the front end module is called CRDC Data Graph Builder. The system is developed as a web application. This section briefly explains the theories and concepts behind the design of the overall system.

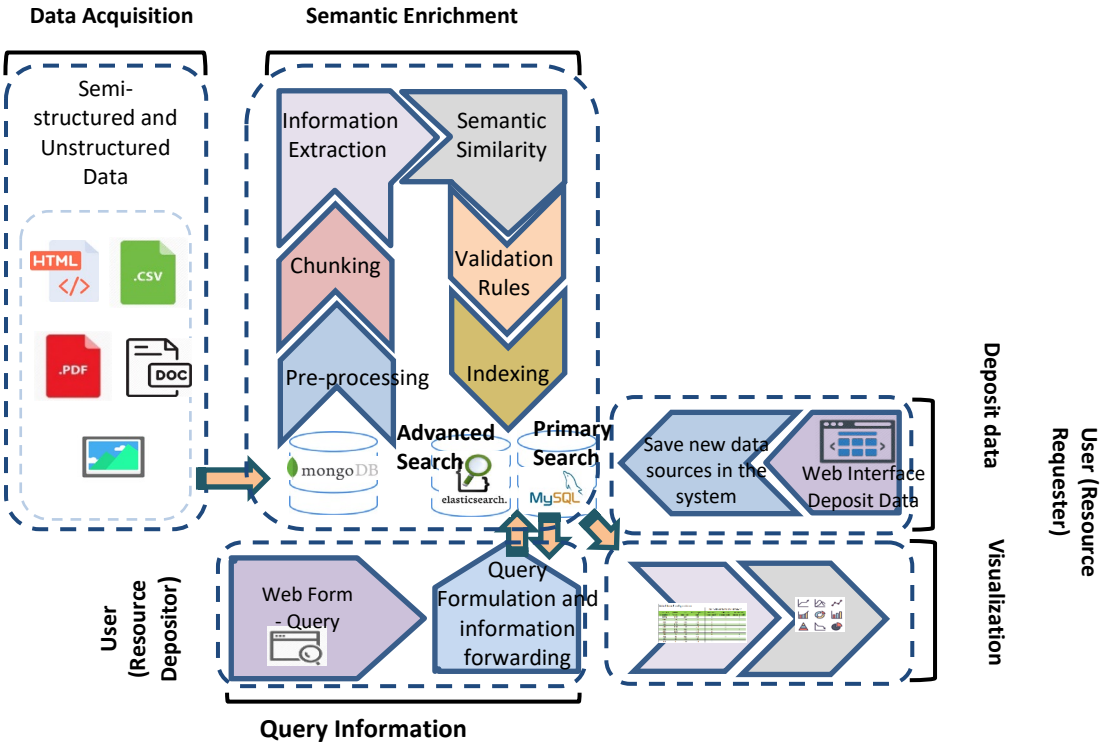

Figure 1: Design architecture - CRDC Repository

3.1 CRDC Information Repository System - Design and Development

The CRDC Information Repository System is an automated knowledge discovery system. It enables automatic extraction of useful information, especially on cotton indicators from heterogeneous data sources.

The design started with identifying the industry requirements, i.e., sustainable indicators which is used to measure sustainable performance and data sources they would usually depend. Some of these data sources were in structured format while some of them are in unstructured or semi-structured formats like pdf documents, excel sheets or html pages. As majority of data sources were

came from unstructured format, the team proposed to utilize text mining methodologies to extract meaningful information from sources. Figure 1 illustrates the brief overview of the use of various technologies in the proposed system.

3.1.1 Design

The proposed system consists of five modules as shown in design architecture (Refer Figure 1).

Firstly, text data is extracted from a data source. Different types of text scrappers were developed depending on the type of data source. The web scrapper goes through the given URL links and only filters out the web pages using “cotton” word. If any given URL or recursive URL has “cotton” word contained, it is marked as positive for text scrapping. Web pages follow HTML markup language to represent web documents and the web scrapper uses these markup notations to distinguish pure text component of a document. The PDF scrapper is responsible for scrapping text from pdf documents. Once scrapped, the text will be stored in a MongoDB collection. MongoDB is an open source, document oriented database. Unlike structured databases, MongoDB gives user friendly platform to manage unstructured data without difficulty.

In the Semantic Enrichment phase (Refer Figure 1), the stored text data in MongoDB is processed and indexed into MySQL and Elastic Search Engine. The first process in Semantic Enrichment is pre-processing text. Pre-processing in vital in Natural Language processing (NLP) to convert natural text into predictable and analyzable format. A simple data pre - processing method was carried out to clean the collected text data, e.g., removal of headers and footers in pdf text. Coming from unstructured data sources, these data may have substances that may lead the system to take inaccurate decisions on indicators.

Once, the text data is ready for information extraction, an additional process was done to identify the values for the indicators. First the text data will be chunked into sentences based on the assumption that a single sentence will discuss about only one indicator. Chunking is the process of parsing these sentences into meaningful text chunks with individual tags. This step is necessary in order to extract value and unit pairs for an indicator. The Semantic Enrichment module also tags and extracts spatial and temporal information from sentences.

To identify a sentence that contains information about an indicator, it is assumed that it will contain the same or semantically similar words to express that sustainable indicator. A word embedding approach has been applied to identify the semantically similar words in the sentence. Several sentences that included information about an indicator are selected to extract the value for the indicator. The entity detection with chunking and regular expressions has been applied to locate precise indicator values and their corresponding units in a sentence. Identifying units is important because the units are used to validate the results. For example, annual crop yield should be in kg (kilograms) or tonnes but not in km (kilometres).

Once the semantic similarity step is done, the processed data is ready to save in a centralized repository. During this process, the text data are stored in MongoDB collections in several formats such as text and sentences.

3.1.2 Implementation

The prototype system is implemented using open-source software. In the Semantic enrichment module, MongoDB is used to store scrapped text documents. Depending on the datasize, the processes in this can module can be resource consuming due to the need of processing a large text data such as single or multiple pdf files or html files.

Open source NLP Python libraries such as PDFminer, Beautifulsoup and Gensim are used to extract and process text and relevant information. After extracting cotton sustainability indicator related information, the data is stored in an open-source relational database, MySQL. The data in MySQL database can be accessed via Primary Search.

The system has an additional functionality which allows users to search processed documents via keyword search as advanced search. Advanced search is based on popular open source search engine Elastic Search. Elastic Search is a full-text search engine and it allows to store, search and analyze big volumes of data quickly.

The front end of the system, the CRDC Data Graph Builder is the web application of the system, developed on Python based web framework called Django. Django provides skeleton of a base web application where we can customize it to cater our requirements. The CRDC Data Graph Builder is linked to backend MySQL database and Elastic Search engine allowing users to access and query data. Users are allowed to customize their graphs using this CRDC Data Graph Builder.

3.2 CRDC Data Graph Builder – Design and Development

The design of the interface, and some of the features, was initially specified by the graphic design team. The approach to the user interface design was to organize information based on a logical sequential interaction pattern that a user would go through when using the graph builder. Development of the graphs and graph builder followed an iterative pattern of design, implementation and feedback from the other members of the group. Work completed on the graphs and application were presented at weekly meetings, and feedback was applied afterwards.

3.2.1 Graph design and development

The intent behind the visual design of the graphs was to develop visually engaging graphs that incorporated familiarity to ensure effective communication and comprehension of information. This approach can be broadly described as one of balancing objectively familiar elements (e.g. visual language and format from established graph types) with subjective visual elements (e.g. icons and symbolism) that would appeal to the audience and have suitability to the context of cotton sustainability.

In the initial design phase, a broad range of graph types were explored, including diverse permutations of radial, bar and line graphs. Each design variation was explored while applying subjective visual elements, including representations of cotton bolls, water, harvesters, cotton thread, among others. These initial designs were evaluated among the project team to identify and select which designs were effective and that should be developed further, and to identify any issues to be addressed in the next iterations. Many of the initial graph design, although looking "interesting", were evaluated as having readability issues. For example:

- A radial bar graph may cause confusion due to bars with similar values having vastly different lengths around the circumference of the circle.

- The representation of values as the size of circles in a bubble graph may make differences in value look greater in magnitude than they actually are, as reducing the radius/diameter of the circle based on the value would cause a larger decrease in the area.

In the next phase, the revision design phase, selected graph designs were developed further. The revised graphs would mostly be variants of bar and line graphs, to ensure the values in them could be easily interpreted by users. Some of the designs still included novel elements, such as a bar graph

that was themed to look like a harvester ploughing fields. The position, size and colour of other elements of the graphs, such as axes and their labels would also be based on these designs. A focus of this stage was to create a mix of standard graphs that could be broadly applied, and novel graph designs, which incorporated subjective visual elements and could be used for specific purposes (e.g. a bar graph incorporating water symbolism and iconography that would be suitable for explaining any indicators related to water use and efficiency).

3.2.2 Graph Builder design and development

The process of implementing the graph builder application started with determining the feasibility of the main features: drawing a graph and outputting it to a file. As this was to be a web application, the D3 library was selected to draw them as an SVG file. The "saveSvgAsPng" library was found later and used to export these graphs to both PNG and as an SVG.

The first graphs that were implemented in this testing phase were based upon the initial designs given by the graphic design team. At the time of implementation the initial designs were evaluated to determine how suitable they were for displaying the indicator data, and iterated upon as required. A regular bar graph, "bubble" graph with the size of each bubble indicating the magnitude of the value, a bar graph with rounded ends, and a "radial" bar graph were implemented in this prototype of the graph builder. As each graph type and its elements was implemented in the graph builder, they were shown at weekly meetings and changed based on the feedback of the other group members.

The design of the interface, and some of the features, was initially specified by the graphic design team. The approach to the user interface design was to organise information based on a logical sequential interaction pattern that a user would go through when using the graph builder. This primarily involved grouping options and function based on the applicability to creating a new graph; resuming and existing graph, and exporting the graph for use.

The graph builder user interface design underwent few changes throughout the development period, usually due to the addition of new features. A features would be suggested during weekly meetings; the difficulty and feasibility of these features would be determined afterwards, and implemented accordingly.

An additional team member, with no previous ties to the project, was introduced later on in the development life cycle as an outsider who could test the application and give feedback. The conjecture was that if the original project team miss certain issues because of their familiarity with the application, whereas this person would not. Their feedback would also work in a similar iterative process to what was already occurring in the regular group meetings. This new team member also assisted in the writing of the guide for using the graph builder.

4 Conclusion

The proposed system is a step forward in the field of sustainability reporting. It uses the advanced methods of data mining, text mining, information retrieval, visualization and web development to design an autonomous system to be used by CRDC. The proposed system has been customized to extract and report cotton sustainability indicators. However, it can also be applied to any other information as the proposed methodology is generic.

Contact

LOCATION

Institute for Future Environments

Level 6, P Block

QUT Gardens Point campus

2 George Street

Brisbane QLD 4000

ENQUIRIES

Email ife@qut.edu.au

Web www.qut.edu.au/ife

Phone +61 7 3138 9500

Mail GPO Box 2434, Brisbane QLD 4001

Fax +61 7 3138 4438