Automating data extraction from documents with a linked data model

Details

Ensemble Team at Lancaster University and environmental scientists at LEC.

Collaborators

Bio

Dr Vatsala Nundloll has a Masters (with Distinction) in Advanced Computer Science and a PhD in Computer Science (Distributed Systems and Semantic Web Technologies). She has worked in various roles as analyst programmer, web developer, lecturer, researcher, and currently as a data engineer. Vatsala Nundloll has a passion for solving data challenges.

Further reading


For more information about previous work carried out at Lancaster University, and other areas of research, visit: www.youtube.com/watch?v=GcXmfq26Bj0.

Funding

This work was funded by the Engineering and Physical Sciences Research Council (EP/P002285/1).

Vatsala Nundloll
When someone mentions data, we are inclined to think of structured datasets made up of numbers in a spreadsheet. But data comes in many forms, including text, video, audio, and imagery. These unstructured data types don’t adhere to a column-row format, so accessing and analysing the information contained in these unstructured formats can be challenging, and valuable information can remain hidden.

Accessing and analysing information stored in unstructured formats can be challenging.

Querying unstructured data is difficult so valuable information remains hidden.

Dr Vatsala Nundloll combines machine learning, natural language processing, and semantic web techniques to bring together information from unstructured and structured sources into a linked data model.

This unified data model enables richer querying of information from multiple data sources.

Dr Vatsala Nundloll answers this question with a unified model that extracts data and information from unstructured textual documents. She demonstrates this approach in case studies focusing on ecology, conservation science, and flood risk management.

An intelligent method for automated data extraction

Machine learning (ML) and natural language processing (NLP) are currently receiving a lot of interest in the world of artificial intelligence (AI), particularly in the fields of data mining and information extraction. Essentially, machine learning involves teaching a machine to think like a human by using computer algorithms which train the machine to learn from experience. NLP entails teaching a machine to interpret human language. Combining computer science and linguistics, NLP examines interactions between a machine and human language, helping the machine to understand meaning and syntax.

To extract information from textual sources, a machine is taught to read or interpret the document and to identify the relevant pieces of information before extracting them. In this work, these emerging technologies are used to create an intelligent methodology that automates the extraction of data from a text, queries the information, and brings together additional relative information from other data sources.

Drawing additional data from other sources would enable richer queries and yield more extensive and insightful information.

When training the machine

Nundloll used Prodigy for AI (an annotation tool for creating training and test data) to train the model. As part of its requirement, training and test datasets need to be created from the digitalised journal. Training data is used to train the ML model to recognise the entities. The model is fed a list of plant species names so that it can distinguish how a plant species name is written. This is repeated for the other entities (pieces of information that we want to identify) namely location, topographic attributes, observer names, and abundance. The model’s accuracy was evaluated using the test set made up of a few sections derived from different parts of the journal. When the machine identified the required pieces of information, they were extracted from the journal and stored in a database.

Nundloll demonstrates this approach in a case study centred on the Journal of Botany, published in 1885. The journal holds a great deal of information about plant species, their locations in the Lake District (eg, on the slope of a hill or on a river bank), their abundance, and the people who observed and recorded them. This rich floristic information is of interest to historians and ecologists as it offers an understanding of the evolution of plant species in the Lake District, but the relevant pieces of information are ‘buried’ in the journal and not easily accessed.

Writing additional data from other sources would enable richer queries and yield more extensive and insightful information.
The resulting database can be queried, but it only contains information extracted from the *Journal of Botany*. Drawing additional data from other sources, however, would enable richer queries and yield more extensive and insightful information. For example, a group of environmental scientists wanted to use this data with a GIS application that could highlight the locations of plants named in the journal on a map. They already had geo-coordinate data for each of the locations in a separate dataset, so they needed some way to combine this data with the data extracted from the journal to create a map of species and locations. They also had another data set that included the taxonomy of plants and synonyms of species names that they wanted to incorporate. The challenge was to bring these data sets together and make richer queries.

### Semantic web technologies linking data

Search engines, such as Google, use semantic data embedded in web pages to provide its users with richer search results. Semantic Web technologies enable data to be defined so we can build data dictionaries. This makes it possible for data to be linked to other related data, creating a linked data model. Rather than just making simple queries about plants or their observers, we can make richer queries drawing on multiple sources, and therefore obtain more extensive information.

Nundloll brought together floristic information from the plant taxonomy, plant species synonyms, plant metadata (including a description of the plant), and geocoordinates of locations where the plants were found to build a linked data model. Integrating the data allows information extracted from this multisource data model to be used in different applications, including GIS. Other plant-related data sets can also be added to the model, if required.

### Data integration for flood risk management

Flood risk management usually depends on measurement data such as river levels, but there is an abundance of information in flood reports that isn’t currently being used. This information could help flood scientists understand unknown parameters and reduce the uncertainty in flood modelling. The unstructured nature of this data forms a barrier to its use, as it contains both numerical and textual data. To make best use of the data, we need to extract numerical values from the text and not lose their contextual information. So, how can we extract this information and integrate it with the measurement data?

Working with flood risk management experts, Nundloll shows how the linked data model uses Semantic Web and NLP technique to bring together data from structured and unstructured sources forming a unified model enabling flood scientists to make richer queries drawing on both information sources. This unified view of all flood risk management data offers better support for decision-making, informing policies in flood risk management.

The combination of ML, NLP, and Semantic Web techniques demonstrates how information from unstructured and structured sources can be brought together into a linked data model. This unified data model enables richer querying of information from multiple data sources. Moreover, it offers greater insight by bringing information that would otherwise have remained buried to the surface.

---

**Personal response**

*In addition to ecology and conservation science, what fields do you think could benefit most from your linked data model?*

A linked data model can be used in almost any other field. It is about bringing relevant pieces of information together. Semantic Web works at weaving through different fabrics of data that contain related information. In this regard, I would say that a linked data model can be used in the medical domain, government sector, media, etc.

*What was the biggest challenge you faced when it came to training the model?*

The challenge was to enable the machine to interpret complex words such as the plant species names. Often these names would have abbreviations as well. Moreover, given that these plant names consisted of two words, they ‘looked’ similar to the observer names (also two words most of the time). Whilst a human can easily differentiate between these two kinds of names, it is quite an arduous task when it comes to getting the machine to make this differentiation. This requires intensively training the machine to understand how to interpret both.

*What advice would you give a young researcher who’s interested in a career in data engineering?*

If someone is passionate about data and solving data challenges, then data engineering is one route to consider. While data engineering is about how to set up a data pipeline in order to ingest data for consumption, and also how to process the data for any transformation that is required, the art of interpreting data and analysing data is part of data science. This particular work falls under the latter category. However, they can be intertwined. Therefore, one should look into which aspect of the data challenges incites them the most, and take it from there.

---

*This data integration allows information extracted from this multisource data model to be used in different applications.*