Technology Forecasting using Data Mining and Semantics

[1]Wei Lee Woon and [2]Stuart Madnick
[1]Masdar Institute of Science and Technology, Abu Dhabi. [2] Massachusetts Institute of Technology, USA

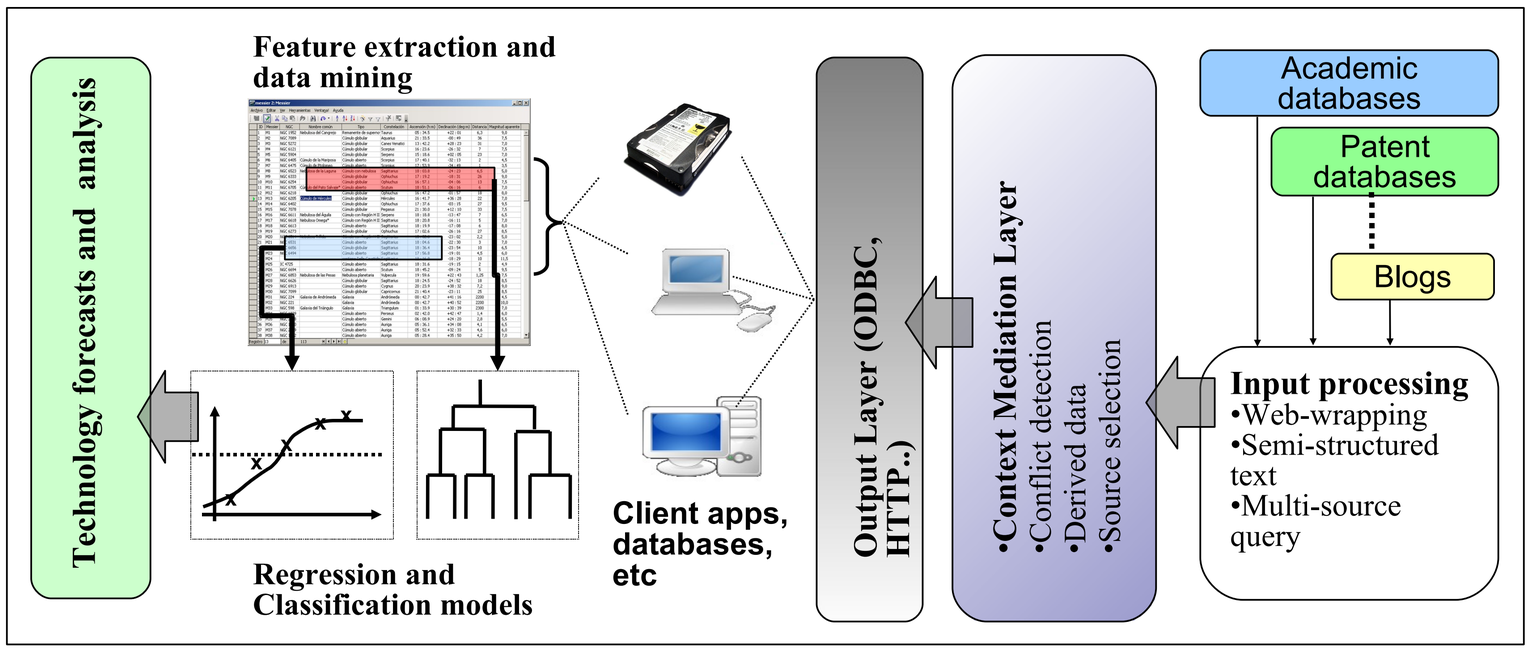

Figure 1: Envisaged system flow for mining technology databases

Motivation

The planning and management of research is increasingly ***data-intensive***; to keep updated on current developments, researchers and decision-makers need to sift through a variety of information sources including academic publications, technical reports and patents.

Research managers often rely on intuition and personal domain knowledge to arrive at management decisions. For example, funding decisions for NSF and NIH grants schemes are arrived at via peer review.

While the role of *expert opinion* in decision making has been an acceptable option for a long time, it still has a number of problems; in particular, expert decisions are subjective, and it can be difficult to record the reasons and contexts in which these decisions were made. Also, finding suitably qualified experts can be difficult and expensive.

This project focuses on the use of a novel combination of methods drawn from data mining and context mediation, as a means of incorporating empirical information into the R&D management process.

Objectives

- Apply *technology-mining* methods to study, visualize and predict the evolution of technological growth.
- Use a novel combination of data-mining and context-mediation techniques for improved performance
- Conduct a detailed case study in *renewable energy technologies*.

Renewable Energies: A case study

To focus research efforts, a case study in renewable energy and sustainability will be conducted. Energy is an important determinant of economic growth and is also a critical element in the general and continued well-being of society.

At a time when climate change, energy security and sustainable development are becoming increasingly critical, the development of technologies for renewable and sustainable energy is of unprecedented importance. Through this case study, we hope to contribute in some small way to the solution of this hugely pressing problem.

Key technical challenges

- Investigating the link between bibliometric measures and technological potential
- Normalization of data from heterogeneous sources
- Identifying suitable and reliable optimization routines
- Techniques for fusing multiple information sources
- Reconciliation of contextual differences
- Assignment of measures of confidence

Project overview

Figure 1 provides a conceptual view of the projected chain of research activities:

1. Information is extracted from sources of information which may be either *unstructured, heterogeneous or both.*

2. Extracted information must be pre-processed to resolve semantic and contextual inconsistencies.

3. Suitably prepared data can be directly analysed using spreadsheets or statistical packages, or passed to further stages for feature extraction and pattern recognition.

4. The ultimate goal is to extract meaningful information from the data, which can be in the form of predictions and/or intuitive visualisations.

Context mediation

One unique feature of the project is the emphasis on exploiting multiple, heterogeneous and unstructured sources of data. The current focus is on academic publications and blogs but we plan eventually cover other sources such as patents, the mainstream press and industrial reports. To deal with such a diversity of data, two issues must be addressed:

1. The *automatic* and *efficient* extraction of information from heterogeneous and unstructured websites.

2. *Reconciliation* of differences in terminology, semantics, time frames and numerical conventions.

Through previous research activities, members of the project team have already developed two technologies for addressing these issues:

- **Cameleon:** A highly scalable and flexible system for data extraction and "wrapping" from semi-structured sources of data.
- **COIN:** A framework for representing, processing, and reconciling heterogeneous data semantics.

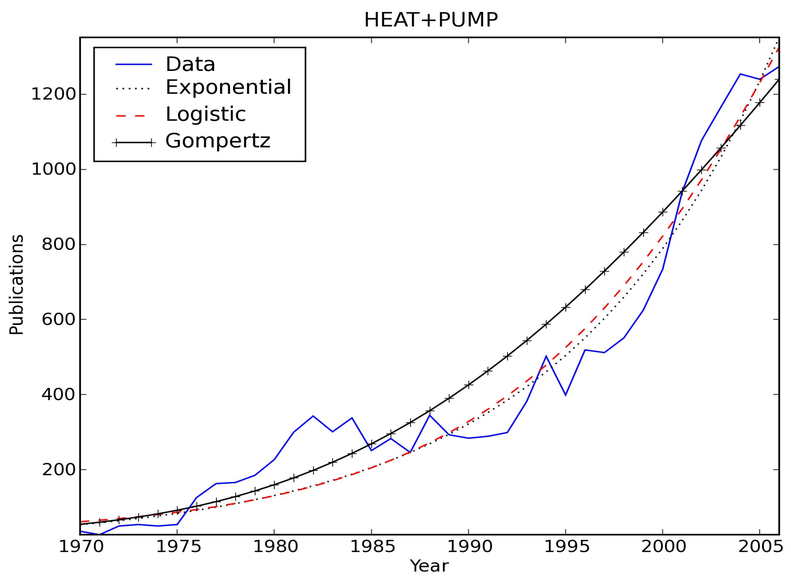

Figure 2: Fitting growth curves to publications data

Data Mining

Once the information has been collected and pre-processed, it is analyzed using data mining tools and techniques.

Currently, our research is centred around the analysis of the *term frequency* of technology-related keywords. This refers to the frequency of occurrence of the keyword in the academic literature, and is an indication of interest and activity within the scientific community. While this cannot be observed directly, we use *publication counts* obtained from academic search engines, as an approximation of this quantity.

One important activity that our team is currently working on fitting growth models to trends in the publication counts corresponding to a variety of energy technologies. For example, Figure 2 shows the growth curve in the number of publications related to "heat+pump"; three different growth models used to characterize this curve have also been superimposed in this figure. The growth potential of the technology can then be inferred from the estimated curve parameters.

A further research direction which we are emphasizing is the creation of techniques for intuitively representing the technological landscape. This has several potential applications. As a visualization tool, it can help give researchers and managers a broader view of the inter-relationships between research disciplines. It can also form part of the data mining process, by detecting clusters of related keywords which can be used as components in the bibliometric features.

Figure 3 shows a hierarchical tree representation built using the joint term frequency of the keywords displayed in the figure. Terms that are closely related in this "ancestry tree" tend to appear together in publications – this is helpful for designing improved bibliometric indicators for our research.

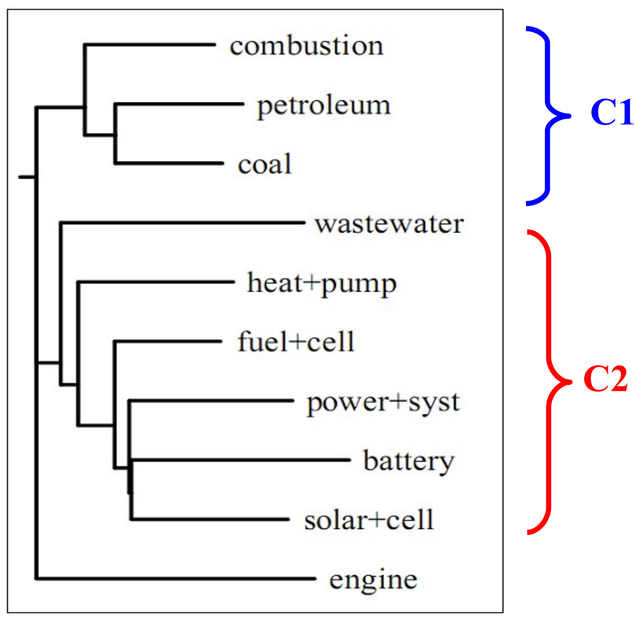

Figure 3: Visualization and clustering of research topics

Conclusions

In summary, while this research is still in a relatively early stage, we already see some instances where potentially useful information can be extracted from unstructured sources of information.

However, there is still much to investigate. In particular, it is important to develop a better understanding of the link between these curves and technological development. One question is: are there instances of key technological trends which could have been predicted using this approach?

It is also important that the accuracy and relevance of the data used to create these curves is ensured even while more diverse sources of information are incorporated. The incorporation of MIT's context mediation (COIN) technology will hopefully help to address this need.