K-Words Lab

Philippe P. Breucker, Audrey Baneyx

Abstract Poster Conference IFRIS, 12/2009

K-Words Lab:

studying the dynamics of a scientific field with keyword analysis

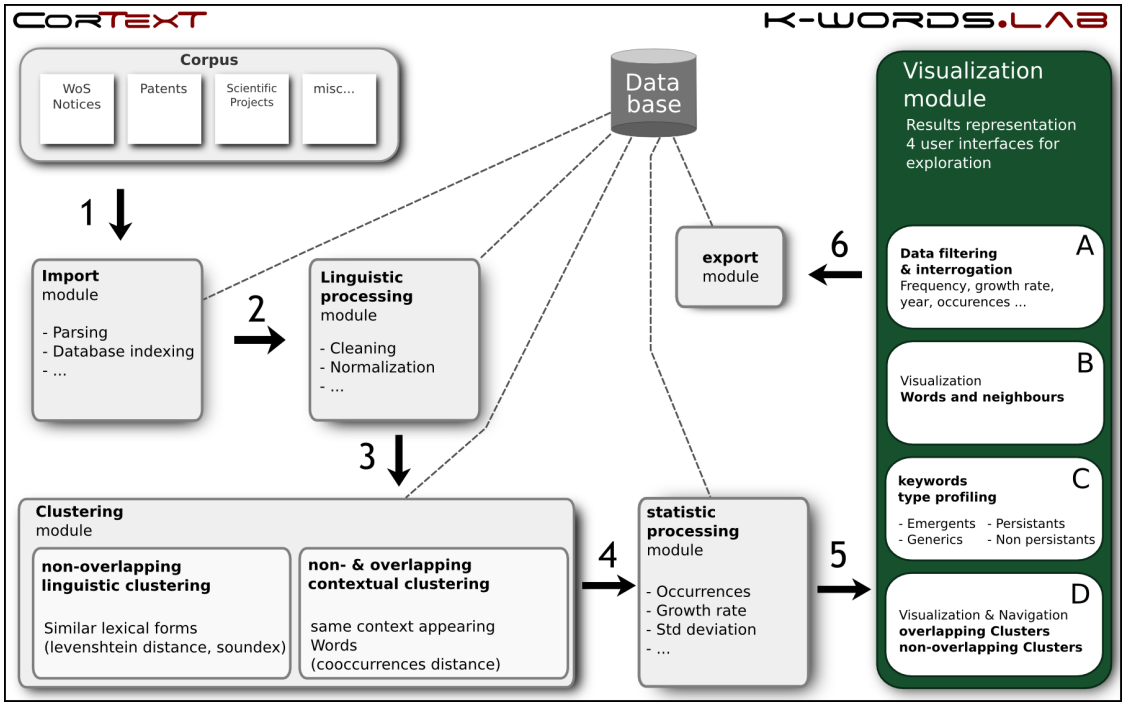
Audrey Baneyx[1], Philippe Breucker[2]

[1]Université Paris-Est, IFRIS, LATTS UMR CNRS 8134
[2]INRA Sens, Université Paris-Est, IFRIS

Research subject and hypothesis

We argue that it is possible to follow the dynamics of a scientific field as it emerges. Being able to do such allows, for example, industrial actors to position themselves regarding their competences and public policy makers to support the development of the field. For that matter, we follow traces of its emergence through publications analysis.
Traditional methods used in bibliometric analysis (journal analysis, co-authors analysis, co-word analysis...) are relevant when the field is structured and bonded. These methods cannot be used when the main authors of the field and its boundaries are unknown. The keywords we use are either determined by the author or given by the Web of Science, or are included in the publication abstract and title. The characterisation of that kind of sources requires programs that are robust enough to handle large amounts of data (over 500 000 publication abstracts for the nano sciences and technology corpus).

Material and methodology : the K-Words Lab 's chain of treatments



K-Words Lab is a tool based on php5/MySql5 technology, wich allows the robustness and large data handling we needed. The web based interface is written in php/HTML/CSS and some features in Flash (with flex) and JavaScript. Calculations have been made on Intel Xeon quad core processors with 4 GBytes of Ram memory. They required several hours to several days to complete, depending on the corpus processed.

The methodology implemented in K-Words Lab uses acquired knowledge brought by the automatic processing of language, artificial intelligence, statistics and science sociology. The software provides an automatic processing chain which calls upon several modules :

1. **import**: the import module is able to transform and index some semi-formated data (like biliographic notices, patents) into our database structure.
2. **linguistic**: one of the difficulty in processing keywords is the many different lexical forms of an expression : for example “carbon nanotube”, “carbon nano-tube” and “carbon nanotubes” should be considered as one expression. The linguistic module implements a distance calculation between all words, based on using, in the first place, the SoundEx algorithm (NARA, 2007), then an algorithm based on Levenshtein distance (Levenshtein, 1965) and then, the QTClust algorithm (Heyer, 1999) to reunite the different lexical forms of each keyword. It can process a large number of keywords (over 8 millions).
3. **clustering**: the clustering module is divided in two parts and is used in two purpose: non-overlapping and overlapping clustering for linguistic and contextual clustering.
   Linguistic clustering is used to find the different lexical forms of a word, where contextual clustering regroups words that are often co-cited with each other.
   We implement several clustering algorithms (like QTClust). A co-occurrence distance is

calculated in order to process the keywords by pair. Then the chosen clustering algorithm regroups them according to a distance threshold.

4. **statistic**: this module calculates several statistical data, like the number of occurrence by year, the average growth rate and its standard deviation. It also processes the data to make the viewing module faster.
5. **user interface**: It allows the user to filter the data from the database in terms of year, number of occurrences, average growth rate and so on.

   It also gives the user a visualization of his corpus by offering a cluster by cluster navigation, in a textual and graphic mode. A word by word visualization is also available witch displays a view of the statistical data of a single expression and lists its neighbours.

   But the most important part of the interface is its ability to rapidly show a set of words of a particular profile, like “emergent” or “persistent” keywords.
6. **export**: this module exports whatever list of keywords has been selected with the filters, in csv text format. Along with the keyword comes all its data (occurrences over the years, average growth rate,...)

Results

The results obtained by exploring the field of nano-sciences have proved that K-Words Lab can automatically analyse large sets of data. We have been able to process over 500 000 publications with 8 millions of keywords occurrences.
A two-level topology of the field is available :

- Keywords classification, with a differentiation between persistent and non-persistent, emergent and generic keywords.
- Exploration between the major concepts of the field and the visualization of their evolution over time.

In the future we will make K-Words Lab evolve to a visual application of the clusters in flex technology which will allow us to dynamically explore a network of the main concepts of a domain online.

Keywords : knowledge extraction, data mining, emergent phenomena detection, knowledge representation, classification, nanosciences