Text Mining and Challenges

Author, Ajit Singh

A Data Science Foundation White Paper

April 2019

---

www.datascience.foundation

Data Science Foundation
Data Science Foundation, Atlantic Business Centre, Atlantic Street, Altrincham, WA14 5NQ
Tel: 0161 926 3641 Email: contact@datascience.foundation Web: www.datascience.foundation
Registered in England and Wales 4th June 2015, Registered Number 9624670

A potentially useful intellectual tool for researchers is the ability to make connections between seemingly unrelated facts, and as a consequence create inspired new ideas, approaches or hypotheses for their current work. This can be achieved through a process known as text mining (or data mining if it focuses on non-bibliographic datasets).

Text/data mining currently involves analysing a large collection of often unrelated digital items in a systematic way and to discover previously unknown facts, which might take the form of relationships or patterns that are buried deep in an extensive collection.

How Text Mining works

Text mining involves the application of techniques from areas such as information retrieval, natural language processing, information extraction and data mining. These various stages can be combined together into a single workflow.

Information Retrieval **(IR)** systems identify the documents in a collection which match a user□s query. The most well-known IR systems are search engines such as Google, which allows identification of a set of documents that relate to a set of key words. As text mining involves applying very computationally-intensive algorithms to large document collections, IR can speed up the discovery cycle considerably by reducing the number of documents found for analysis. For example, if a researcher is interested in mining information only about protein interactions, he/she might restrict their analysis to documents that contain the name of a protein, or some form of the verb □to interact□, or one of its synonyms. Already, through application of IR, the vast accumulation of scientific research information can be reduced to a smaller subset of relevant items.

Natural Language Processing **(NLP)** is the analysis of human language so that computers can understand research terms in the same way as humans do. Although this goal is still some way off, NLP can perform some types of analysis with a high degree of success. For example:

Part-of-speech tagging classifies words into categories such as nouns, verbs or adjectives

Word sense disambiguation identifies the meaning of a word, given its usage, from among the multiple meanings that the word may have

Parsing performs a grammatical analysis of a sentence. Shallow parsers identify only the main grammatical elements in a sentence, such as noun phrases and verb phrases, whereas deep parsers generate a complete representation of the grammatical structure of a sentence

The role of NLP is to provide the systems in the information extraction phase (see below) with linguistic data that the computer needs to perform its □mining□ task.

Information Extraction **(IE)** is the process of automatically obtaining structured data from an unstructured natural language document. Often this involves defining the general form of the information

that the researcher is interested in as one or more templates, which are then used to guide the extraction process. IE systems rely heavily on the data generated by NLP systems. Tasks that IE systems can perform include:

Term analysis, which identifies the terms in a document, where a term may consist of one or more words. This is especially useful for documents that contain many complex multi-word terms, such as scientific research papers

Named-entity recognition, which identifies the names in a document, such as the names of people or organisations. Some systems are also able to recognise dates and expressions of time, quantities and associated units, percentages, and so on

Fact extraction, which identifies and extracts complex facts from documents. Such facts could be relationships between entities or events

A very simplified example of the form of a template and how it might be filled from a sentence is shown in Figure 1. Here, the IE system must be able to identify that 'bind' is a kind of interaction, and that 'myosin' and 'actin' are the names of proteins. This kind of information might be stored in a dictionary or an ontology, which defines the terms in a particular field and their relationship to each other. The data generated during IE are normally stored in a database ready for analysis by the final stage, that of data mining.

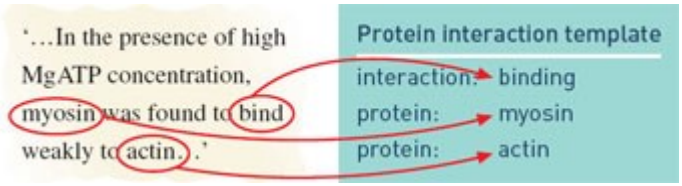

Fig 1: template-based information extraction

Data Mining **(DM)** (often known as knowledge discovery) is the process of identifying patterns in large sets of data. When used in text mining, DM is applied to the facts generated by the information extraction phase. Continuing with the protein interaction example, the researcher may have extracted a large number of protein interactions from a document collection and stored these interactions as facts in a separate database. By applying DM to this separate database, the researcher may be able to identify patterns in the facts. This may lead to new discoveries about the types of interactions that can or cannot occur, or the relationship between types of interactions and particular diseases, and so on.

The results of the DM process are put into another database that can be queried by the end-user via a suitable graphical interface. The data generated by such queries can also be represented visually, for example, as a network of protein interactions.

Text mining is not just confined to proteins, or even biomedicine though this is an area where there has been much experimentation using text/data mining techniques. Its concepts are being extended into

many other research disciplines. Increasing interest is being paid to multilingual data mining: the ability to gain information across languages and cluster similar items from different linguistic sources according to their meaning.

Examples of Text Mining

Research and development departments of major companies, including IBM and Microsoft, are researching text mining techniques and developing programmes to further automate the mining and analysis processes. Text mining software is also being researched by different companies working in the area of search and indexing in general as a way to improve their results. There are also a large number of companies that provide commercial computer programmes.

- AeroText - provides a suite of text mining applications for content analysis. Content used can be in multiple languages
- (AlchemyAPI - SaaS-based text mining platform that supports 6+ languages. Includes named entity extraction, keyword extraction, document categorization, etc.
- Autonomy - suite of text mining, clustering and categorization solutions for a variety of industries
- Endeca Technologies - provides software to analyze and cluster unstructured text.
- Expert System S.p.A. - suite of semantic technologies and products for developers and knowledge managers.
- Fair Isaac - leading provider of decision management solutions powered by advanced analytics (includes text analytics).
- Inxight - provider of text analytics, search, and unstructured visualisation technologies. (Inxight was bought by (Business Objects that was bought by SAP AG in 2008)

Challenges

1. Intellectual Property Rights

As it stands, each publisher maintain their own □digital silos□ of information, and cross searching among these separate silos is undertaken more in the breach than the observance. Yet it is only through the dismantlement of the legal protections around such silos that effective text and data mining can take place. The greater the common document source being mined the more effective the results achieved.

Such a cross-silo approach could be achieved in a number of ways. Either through agreements reached with the existing publishers to allow cross searching of text files among publisher silos on a licence basis. Or through the adoption by the industry at large of open access as the standard business model.

Most databases which include a sweat of the brow activity may only be accessible if the customer has paid a subscription or licence fee. Even if this hurdle is overcome, the terms of the subscription and licence may be such that the owner of the database will still not allow reformulation of the material in any way. Several commercial journal publishers have raised concerns that the creation of □derivative works□ could undermine the commercial opportunity facing their primary journals.

Nevertheless, a number of stm publishers have recently reached an agreement with the Wellcome Trust to allow text mining to take place on works which Wellcome has funded (through payment of author fees) but only within the terms of the licences agreed with each publisher. This still remains restrictive as far as text mining is concerned. Licences would need to be changed to open up the database to unrestricted mining activity, even if they lead to derivative works being created. This is what the user community wants, this is what Science needs, this is what the traditional publishing industry wants to avoid.

But we are seeing further instances of the licences slowly being adapted to meet this user demand.

2. Technical Issues

A key technical issue is whether text/data mining is undertaken from a single large accumulated database held centrally, or else whether a federated search system is adopted with knowbots being launched to pull in results from remote and privately held databases. A centralised database also raises issues of resources. Not only in terms of the infrastructure to support a large central file but also in the support services necessary to run it. Computation can take place in a more controlled environment on a single aggregated database, though this may not always be possible for a variety of technical and IPR reasons.

A distributed model raises issues around data normalisation, of performance levels, of other standardisation issues. A distributed or federated system requires conformity by all involved to common metadata standards to allow effective cross reference and indexing. If the need is to rely on a federated approach the issue of trust arises trust that the remote database of text and data will always be there, curated and consistent in its approach to metadata creation and full-text production.

In support of a federated approach to text and data mining one can see the emergence of the cloud as a mechanism for processing large amounts of data using the existing powerful computer resources made available by organizations such as Amazon, Yahoo, Microsoft, HP, etc. A federated powerful processing infrastructure is in place in the cloud.

Implications of Text and Data Mining

Providing a text/data mining facility for Science requires a new means of collaboration between existing and future stakeholders to accept data and text mining as being effective and acceptable processes. In particular, that such mining does not eliminate any significant role currently being performed by stakeholders, that it does not raise challenges and barriers to text/data mining applications, that it does not threaten publishers and librarians and their existence.

There is the rub. The battle will be whether the advantages which text and data mining confer are sufficiently powerful and attractive to the research community to enable it to sweep objections aside. At present all we can hypothesise is that data and text mining will happen is happening in select areas and will be another driver for change in the march towards full electronic publishing over the next few years. But how soon depends on a number of factors. Intellectual property rights and their protection will be at the forefront of these.

Text and data mining creates a new way of using information. It opens the horizons of researchers. But to fully appreciate the scope of the technology it requires some training for the researcher and the inclusion within their research process of text/data mining techniques.

Besides that it needs access to a large document database. As has been mentioned, this creates problems with regard to licensing. But text miners need text, and they need it in a form which is useful for the text mining systems.

Open Access

A review of text and data mining is not complete if one ignores other underlying trends in scientific communication. One of these is the changing business models which have come about in the past 6-8 years (in effect since the Budapest Initiative in 2002, the Bethesda Statement and the Berlin Declaration in 2003).

Text mining is believed to have a considerable commercial value. This is particularly true in scientific disciplines, in which highly relevant (and therefore monetarisable) information is often contained within written text. In recent years publishers have been effecting improvements to their publication systems without opening the doors to text and data mining. Some of the general initiatives taken, such as (Nature's proposal for an Open Text Mining Interface (OTMI) and NIH's common Journal Publishing Document Type Definition (DTD) which has been adopted by many of the larger publishers, do provide semantic cues to machines to answer specific queries contained within text, but without going as far as removing publisher barriers to public access.

Applications:

1. Relevance to Researchers

The burgeoning growth of published text means that even the most committed and avid reader cannot hope to keep up with all the published output in any one subject field, let alone relevant adjacent fields. There is a consistent expansion in research publications of between 3.5% and 4% per annum, driven largely by the competitive needs of inpidual researchers to gain recognition and esteem for the quality of their work. This will probably never change in our lifetimes. The consequence is that nuggets of insight or new knowledge are at risk of languishing undiscovered within the sheer burgeoning mass of published literature if they are not identified or mined in some structured way.

Text mining offers the scope for helping the researcher make serendipitous connections through the use of automatic systems. These automated systems are unaffected and undeterred by the ongoing expansion in the output of published scientific, technical and medical text. It is a process which is truly scaleable, in line with scientific output.

But will researchers, already faced by a vast array of sophisticated research tools in their own areas, and emerging search and discovery tools covering the whole of Science will they want to learn about another sophisticated tool, one which offers no guarantee that it will produce any meaningful results? The

pressure to adopt text and data mining may well come from intermediaries and gatekeepers acting on behalf inpidual and groups of scientists. It opens up a role for librarians.

2. Impact on Libraries

It is an oft-claimed requirement that for librarians to have a □future□ they must get closer to the faculty and the research staff of their institution. Librarians need to monitor what the faculty really need and how they are building up their knowledge resources. This will provide evidence with which to negotiate future licensing and subscription rights but also whether some new services □ such as text and data mining □ have any relevance for their clientele.

Standards setting and monitoring will become important for interoperability and advancing the art of text and data mining. Assisting in this standards setting process could become a responsibility of the library profession. Helping with the creation of ontologies and appropriate mark-up languages could become their future role.

Libraries may find a valuable role in support of text and data mining by proactivlely working with the faculty and their patrons to get the local institutional repository up and running, and full, and ensuring that the content is accessible and has appropriate metadata and other standards embedded.

3. Impact on Publishers

It is less easy to be sanguine about how text and data mining will impact on publishers.

With several thousand scholarly publishers worldwide, each with their own silo of digital data, it would take a substantial change in the industry mindset to create a large, consistent database, sufficient to make text mining an effective service. Cooperatives of publishers are few and far between, and the industry record on cooperation has (with a few exceptions) been poor. But the only way effective new e-Science services can be introduced, such as text/data mining, is if there is a large collection of digital material available. Whilst the stm industry remains so fragmented and unwilling and legally unable to cooperate to create such a collection is unlikely to arise on their backs.

About the Data Science Foundation

The Data Science Foundation is a professional body representing the interests of the Data Science Industry. Its membership consists of suppliers who offer a range of big data analytical and technical services and companies and individuals with an interest in the commercial advantages that can be gained from big data. The organisation aims to raise the profile of this developing industry, to educate people about the benefits of knowledge based decision making and to encourage firms to start using big data techniques.

Contact Data Science Foundation

Email:contact@datascience.foundation
Telephone: 0161 926 3641
Atlantic Business Centre
Atlantic Street
Altrincham
WA14 5NQ
web: www.datascience.foundation