# CONTENT MINING

Mark MacGillivray

Content mining is the activity of automating searches through content, usually large bodies of text such as research articles or books (but possibly also research data itself), seeking new or alternative knowledge and/or applications within the content. This can have valuable results because the process of creating content in the first place is one of focusing efforts on a particular problem, and presenting outcomes to a target audience; it is therefore possible that other unexpected uses can be found within such corpora. Content mining may also be referred to as text mining, data mining, text-and-data mining, TDM, data science, and so on. It can also be considered to include specialist activities in natural language processing (NLP), but in general the approaches taken include extracting text from source documents into database and/or indexing software, then searching for keywords relevant to a particular discipline, generating taxonomies, categorisation of content, frequency distribution measurements, pattern recognition, entity recognition and relational modeling, and many more. However, the technical possibilities are further complicated by legal issues.

## DEFINITION

A traditional approach to knowledge dissemination starts off sensibly by seeking to distribute materials of interest to relevant parties. This evolves into subscription by interested parties to sources that provide material tailored to their needs. This practice bundles knowledge up into content packages such as research articles, making them suitable for delivery in what is conceptually a production line or pipeline model of knowledge dissemination. However, the advancement of digital publishing and the web allows for these pipelines to themselves be treated as raw resource, suitable for exploration in the search for new meaning and understanding. With the ability to easily apply large amounts of computational power to digital resources, content mining has grown into a powerful tool for performing new research. Content mining can range from relatively straightforward techniques to cutting-edge research in its own right.

### CRAWLING AND SCRAPING

Content mining starts with finding useful raw resources for processing. The fundamental principle is that the material of other content resources, such as the text in a journal article, can be processed in ways that may yield new discoveries or that may expose otherwise unrecognised links between works. This is made possible by ingesting the raw material from its source into a computational environment, hence digital versions of such raw materials are highly suitable for content mining. With so much material now available in digital form on the web, this has become much easier. However there are still obstacles to overcome, such as the rate limits applied to certain websites, or the extraction of useful raw data from unsuitable formats such as PDF (which is designed for presentation, not for preservation of data). In the case of mining academic journal articles, different publishers can place different limitations on how much data can be extracted, and how quickly (Mounce, 2014). Some sources provide application programming interfaces (APIs) from which machine readable text and data can be directly extracted, but this is not always the case. For these sorts of reasons, content mining often begins with CRAWLING the web or other large source archives, searching for relevant material, and in some cases where no alternative is present, SCRAPING potentially useful text and data out of presentation layers such as human readable web pages or PDFs.

### NORMALISATION

With some suitable source materials retrieved, the next stage can involve some form of normalisation of the raw materials into a more uniform structure. For academic research, this can begin as identifying common sections in a paper - the methods section, for example, is often a very useful part of a scientific article. Processing out of formats such as PDF and into formats more suited for machine processing, such as HTML or XML, is also common. More complex processes can even extract data from images in figures within published articles (see Kuhn et al 2012 and Murray-Rust 2015). The normalisation process commonly ends with extraction of all text and data points into a relational database or indexing system, so that it is ready for processing by customised software packages.

### PROCESSING

The processes to perform depend on the results desired, and the raw resources being used. Typically, processing takes the form of compilation of a customised software package to search the normalised raw materials for generating taxonomies, categorisation of content, frequency distribution measurements, pattern recognition, entity recognition and relational modeling, and other techniques that allow for analysis that may yield previously unseen patterns to emerge. Data visualisation becomes a useful tool at this point, as it can be far easier to identify such patterns in a network diagram visualisation than in a spreadsheet of data points, for example. A specific case, for reference, is described by Ross Mounce of the ContentMine project, in which the literature "haystack" is searched for species names (Mounce 2015).

### FACTS AS OUTPUT

In the case of the aforementioned ContentMine project, the results of such content mining activities are presented as "facts", with the slogan "the ContentMine uses machines to liberate 100,000,000 facts from the scientific literature". The purpose of this form of presentation is to differentiate the resulting facts from the source material, and the key point of differentiation is a legal issue - facts are not subject to copyright whereas creative works are.

## LEGAL ISSUES

"Could it be true that laws designed more than three centuries ago with the express purpose of creating economic incentives for innovation by protecting creators' rights are today obstructing innovation and economic growth? The short answer is: yes."

(Hargreaves, 2011)

Although technical solutions exist for finding and using content as raw resource for mining, it is not necessarily the case that it can be used in such a way - even where that content can be legally accessed. This is because a great deal of content is submitted to aggregators that apply copyright licenses to the material, and provide contractual arrangements to allow access back to the content - and those arrangements often specify that the material can only be accessed by human beings for the purpose of reading (and not for text mining). However, as with all legal issues, there are differences across jurisdictions which make the situation somewhat more complex and a lot less simple when trying to determine what is legally possible. For example the United States has perhaps a wider application of "fair use", whereas Europe has more restrictive "database rights" that come into force when effort has gone into aggregating a set of data. There is a vast amount of detail in these legal issues that is beyond the scope of this article, however of particular note is the recent UK copyright exceptions.

The Hargreaves report, quoted above, led to the introduction of an exception to copyright law within the UK legal jurisdiction (IPO, 2014) to allow for text and data mining (content mining) for non-commercial research. This exception "allows researchers to make copies of any copyright material for the purpose of computational analysis if they already have the right to read the work (that is, they have 'lawful access' to the work)". The aforementioned ContentMine project makes use of this exception to perform such research within the UK, which allows the project researchers and their partners to process any article that they have the right to access, regardless of where the source material is published; the exception applies to activity in the UK, even if the source material originated elsewhere. This has greatly simplified the activity of content mining within the UK at least, and has opened the door to a number of useful applications that are currently only possible in the UK and three other jurisdictions in the world that provide similar exceptions (South Korea, Israel, and Japan).

## USEFUL APPLICATIONS

With the technical capability and the legal right to do so, content mining can be used to provide high quality search into vast corpora, even beyond that which can be accessed by the person performing the search; this is possible because the facts that can be extracted as part of the non-commercial research of the UK project are themselves not subject to copyright, so they can be exposed as search terms to enable others to find the works that are of most use to them before having to pay for access to those materials. This does still raise contentious issues in that some publishers wish to claim exclusive rights to do this, but on the other hand, provision of such services by others actually increases exposure of the very items the publishers with to sell.

Assuming access to materials is not an issue, content mining is ideally suited to any discipline where a systematic review must be carried out. By crawling a large corpus of material for review and normalising it, it can then be processed with packages designed to find articles that meet the requirements of the systematic review in question. This can greatly reduce the effort involved in for example performing double blind systematic reviews for the purposes of human or animal clinical trials, and it is already believed that such an approach "has the potential to transform the way we work" (Steel 2015).

Further applications include performance of literature reviews to identify gaps in current research. Whilst traditional reviews may help to form a picture of what has already been researched, and some inference may be possible in what remains outstanding, this can be achieved more easily by content mining the relevant published works and then visualising the relations between such works. This quickly allows for visualisation of "holes" in the relationship between published research in a given field, and can allow for targeted searches into those areas lacking coverage.

## FUTURES

There are numerous interesting futures involving content mining. Firstly there is the ongoing debate between publishers and lawmakers as to whether publishers should retain the right to perform content mining exclusively, or whether researchers themselves are better positioned and skilled to do such research directly. Discussions are currently ongoing in the European Parliament (Committee on Research, Industry and Energy, 2014) to consider the applicability of the UK copyright exception across the European Union, and if such an exception is already considered advantageous to the UK then it is likely that widening the exception to the EU will result in even greater potential for new discoveries to be made.

Whilst advances in content mining have perhaps been greater to date in the sciences, there is also considerable applicability of content mining techniques to the digital humanities, where large texts that are already undergoing digitisation could further be processed in ways that may provide new insights for humanities research. This could range from something as simple as digital annotation highlighting relations between texts to "digital remixing" of source materials via algorithmic processes that produce new hybrid works.

## REFERENCES

I. Hargreaves (2011)

Digital Opportunity: A review of intellectual property and growth

<http://www.ipo.gov.uk/ipreview-finalreport.pdf>

T. Kuhn et al (2014)

Mining images in biomedical publications: Detection and analysis of gel diagrams

<http://www.jbiomedsem.com/content/pdf/2041-1480-5-10.pdf>

R. Mounce (2014)

The right to read is the right to mine: Text and data mining copyright exceptions introduced in the UK

<http://blogs.lse.ac.uk/impactofsocialsciences/2014/06/04/the-right-to-read-is-the-right-to-mine-tdm/>

R. Mounce (2015)

Finding species in the literature haystack

<http://contentmine.org/2015/07/finding-species-in-the-literature-haystack/>

P. Murray-Rust (2015)

The importance of figures and captions

<http://contentmine.org/2015/06/the-importance-of-figures-and-captions/>

G. Steel (2015)

"We think ContentMine has the potential to transform the way we work!"

<http://contentmine.org/2015/06/we-think-it-contentmine-has-the-potential-to-transform-the-way-we-work/>

UK Government Intellectual Property Office (2014)

Exceptions to Copyright

<https://www.gov.uk/exceptions-to-copyright>

European Parliament Committee on Research, Industry and Energy (2014)

OPINION of the Committee on Industry, Research and Energy for the Committee on Legal Affairs on the implementation of Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society (2014/2256(INI))

<https://edri.org/files/Reda/Reda_ITRE_Opinion.pdf>