Automated research

Tom Rochette <tom.rochette@coreteks.org>

December 21, 2025 — 77e1b28a

0.1 Context

Scientific research is a method through which state-of-the-art (SotA) techniques, tools and methodologies are developed and shared with the scientific community. It consists of numerous steps through which an individual goes in order to familiarize themselves with the current SotA.

Currently, such process is generally defined but certainly not globally accepted. Furthermore, and this is the part we are interested in here, most of the process could be automated.

0.2 Learned in this study

0.3 Things to explore

- How is originality produced?
- Can Parsey McParseface be used to extract references?
- Grammar induction
- Reference consistency verification (are the reference years the same or lower than the paper's publication year?)

0.4 Questions

- How do you limit the depth of related papers?
  - Limit on the number of references per paper
  - Hard-coded/Defined limit
  - Limit on year range
- How should information be fed into the system?
  - Support only arxiv?

0.5 Observations

- If a reference is referred to often because it is large in size and can appeal to a wide variety of interest, it does not necessarily make it valuable for your current purpose

1 Overview

In this article we explore the idea of automating scientific research. We attempt to cover the various research phases such as papers retrieval, assessment of the domain, determination of the core papers and authors in the field, construction of a bibliography and more.

The purpose of the automation process is to reduce the effort required for an individual to get into a (new) research field. The true end goal would be to provide topics and receive new, genuine papers on the provided topic generated automatically by the computer through the agglomeration, analysis and synthesis of existing research.

2 Research procedure

2.1 General approach

- The user defines a topic of interest (ex. artificial general intelligence)
- Automatically
  - Find papers related to the topic and download them
    - Find relevant sources of papers in the domain (arxiv, nature, science, acm, springer, google scholar, etc.)
  - Create a graph of paper references (which paper references which paper)
    - This graph will be used to determine which papers are foundation and which papers are extensions
    - The more a paper is referenced, the more likely is it worth reading
      - Use some algorithms, maybe similar to Google Page Rank to determine the "quality" of the paper
    - References are extracted from the papers themselves (or by using a reference engine)
    - References are cross-referenced
  - Extract writing style (tf-idf, most frequent words, sentence/paragraph/section length) as well as various features (number of charts, tables, figures, etc.)
  - Extract paper format (abstract, number of figures, types, tables, graphs, etc.)
  - Compile a list of references based on the papers extracted above
  - Compile a list of the more prominent writers in the field/topic
  - Create summaries of the different presented ideas with links to the related articles

2.2 Alternative approach

- Provide one or many papers which have been of interest to you
  - Go through the general approach

3 Components

Here we attempt to list the necessary components that will be used during the research procedure.

3.1 Paper searcher

Use various search engines in order to retrieve relevant papers. This has the obvious shortcoming that the quality of the papers retrieved is directly related with the quality of the search engines used.

3.2 File downloader

Download papers of interest for further processing.

3.3 Text extractor

As PDF is not a structured format, we need to extract the text content and attempt to build a logical view of it. At this point we only have a set of strings, which we will need to structure.

3.4 Paper extractor

Using the text extracted previously, we run a program with various heuristics that will attempt to extract and structure the content of a paper.

If papers were to follow a structured format convention, the text extractor and paper extractor would most likely be unnecessary components, except to process old papers that did not respect such convention.

3.5 Paper evaluator

We attempt to evaluate the quality of a paper (a metric akin to a page rank). This is based on various metrics such as:

- Author "prestige" (number of referenced articles, number of references)
- Number of references in the paper
- Number of papers referencing this paper
- etc.

3.6 Text summarizer

In order to simplify the task of the human researcher, we attempt to build a summary of the paper (generally provided as the abstract).

3.7 Reference extractor

An important component of papers are their references. They provide us with a list of manually and humanly assessed papers that are relevant to the currently analyzed paper.

3.8 Indexer

As we process papers, we want to make sure not to reprocess already processed papers. This is the job of the indexer. Furthermore, it allows us to slowly build up an index/bibliography.

3.9 Reference list generator

With numerous papers processed, we can now construct a list of all references that were extracted.

3.10 Reference graph generator

One step further after having built a reference list is to build a reference graph. A reference graph is a representation of the papers referring to other papers.

3.11 Author list generator

By extracting authors from papers and their references, it is possible for us to build a list of researchers in a given domain, which can prove to be a useful tool if one wants to find resources that may be useful in their own research.

3.12 Author reference graph generator

As with the reference graph, the author reference graph is a representation that helps the researcher observe which author refers to which other authors. This can be an important tool to discover when there's only cross-referencing or very little external referencing in someone's work.

4 Features extraction

In order to extract meaningful features out of scientific articles, we need to determine the features we are interested in. To do so, we inspect a small amount (~10/50) of articles and extract the elements we want to construct a database with.

4.1 Paper structure

- Paper header
- Paper body
- Paper footer

4.1.1 Paper header

- Title
- Author(s)
- Email(s)
- Affiliation(s)
- Abstract

4.1.2 Paper body

- Content
  - Sections
    - Header
    - Body

4.1.3 Paper footer

- References
- Annexes

4.1.3.1 Reference structure

- Authors
  - Author
    - First name
    - Middle name
    - Last name
- Title
- Publication year
- Journal
- Pages
  - From
  - To
- Conference

4.2 Feature extraction procedures

4.2.1 Extraction of references from PDF documents

- Create a training set with labels
- Build a tree of the different combinations based on the training data
- Find a way to hierarchically express these combinations
- Run the regex on the whole reference for each type of segment to extract, then attempt to construct a reference by making sure that segments do not overlap (see https://github.com/tensorflow/models/tree/master/syntaxnet transition-based dependency parser for a similar idea)

4.2.1.1 Difficulties

- Many different reference format
- Format is unstructured
- Syntactic rules not necessarily respected

- Extraction appears to be easy to a human due to the use of existing knowledge to differentiate between names segments and title segments
  - The ability for a human to infer potential extraction rules such as "format A or format B or format C" simply by induction on multiple examples and using existing knowledge

5 Variables

5.1 Reference

- Position of subsegments (author, title, year)

5.2 Author

- Character sequence
- Length
- Structure class (First(Middle) Last|First (M.) Last|F. (M.) Last|Last, First(Middle)|Last, F.(M.))

6 Issues with test data

6.1 Problems

- Different tag names (year vs date)
- Different tagging (all authors under one author tag vs one author per tag)
- Tag formats (space padded tags)
- How should the precision/recall be computed on partial matches?

6.2 Solutions

- Reformat the test data according to our format
- Make test code generalize data

7 Known

> The class of languages will be considered learnable with respect to the specified method of information presentation if there is an algorithm that the learner can use to make his guesses, the algorithm having the following property: Given any language of the class, there is some finite time after which the guesses will all be the same and they will be correct.[1]

Given there is a maximum amount of different reference formats, smaller than the number of all valid permutations of the different token types supported, it is thus possible to inductively build a grammar that will allow us to extract the information within the reference correctly, every time.

8 Not classified

- Paper/Not paper classifier
- File hash (vs content hash)
- Paper clustering
- Display top X papers
- Author publications per year
- Author publication period
- Author references per year
- Paper keywords

[1] Gold, E. Mark. "Language identification in the limit." Information and control 10.5 (1967): 447-474.

- Author keywords
- Paper PR
- Author PR
- Recognize that two papers are the same but different iterations
- Automatically collect new papers by authors/topics/based on a pool of papers
- Rank papers by personal valuation
- Use the rank to determine the most interesting papers to share with the user
- Paper summarization
- Rank papers by citation count, papers that are the most cited should provide the most value/utility to the reader (they are more broadly useful, like money)

Growing the pool of papers

- Extract authors
- Extract keywords
- Extract period
- Extract establishment
- Extract email domain name
- Extract emails
- Record from what combination of information a new paper was found from, where it was found, when, etc.

Use case

- Drop a bunch of papers in a web page
- Papers are analyzed
- Topics are suggested or given by the user
- User ID asked to rank the papers
- Papers are grouped by research project

Agent keeps searching search engines regularly for updates

- Google
- Arxiv
- Reddit

What are the phases of a research?

- Topic selection
- Information collection
- Elicitation of questions
- Hypothesis
- Experimentation/prototyping/development
- Data collection
- Paper writing
- Formatting
- Peer review
- What we want from such a system is for it to answer to our questions. If we want to know how to do X, then it should know how to do it, as well as to list all the dependent knowledge you will need in order to understand/make use of the answer provided. As such, it will need to have a good model of your own knowledge

9 See also

10 References

- SCIgen - An Automatic CS Paper Generator
- http://www.cs.cornell.edu/cdlrg/reference%20linking/extraction.pdf
- https://github.com/CrossRef/pdfextract - Crashes on Windows
- https://github.com/metachris/pdfx - Does not extract text reference (only URL/DOI/arxiv)
- http://pythonhosted.org/refextract/ - Not compatible with Python 3.5 (unicode regex)
- http://www.dlib.org/dlib/september13/kern/09kern.html
- http://aye.comp.nus.edu.sg/parsCit/
- http://freecite.library.brown.edu/
- https://www.comp.nus.edu.sg/~kanmy/papers/lrec08b.pdf
- https://anystyle.io/
- http://www.lib.ncsu.edu/tools-citation
- http://pitt.libguides.com/citationhelp
- http://support.ebsco.com/knowledge_base/detail.php?id=5563
- http://www.scientificstyleandformat.org/Tools/SSF-Citation-Quick-Guide.html
- http://citeseerx.ist.psu.edu/index
- http://csxstatic.ist.psu.edu/about/scholarly-information-extraction
- http://www.lib.ncsu.edu/citationbuilder/#/article-journal/apa
- https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/
- https://academicgraph.blob.core.windows.net/graph/index.html
- http://academic.research.microsoft.com/VisualExplorer
- Councill, Isaac G., C. Lee Giles, and Min-Yen Kan. "ParsCit: an Open-source CRF Reference String Parsing Package." LREC. Vol. 8. 2008.
- Teufel, Simone, and Min-Yen Kan. Robust argumentative zoning for sensemaking in scholarly documents. Springer Berlin Heidelberg, 2011.
- http://allenai.org/semantic-scholar/citeomatic/?hootPostID=4cce060995ac4d34282766d010cb7f35
- Hamming, "You and Your Research" (June 6, 1995)
- Doing Your Research Project: A Guide For First-Time Researchers
- https://www.connectedpapers.com/
- https://www.citationgecko.com/

10.1 Labeled training data

- https://github.com/knmnyn/ParsCit/tree/master/doc
- https://github.com/knmnyn/ParsCit/blob/master/crfpp/taggeddata