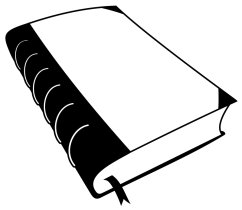


Aberystwyth University

Progress Report

# Partridge: An Intelligent Literature Analysis and Recommendation Suite.

*Author:*
James Ravenscroft
jrr9@aber.ac.uk
090407039

*Supervisor*
Amanda Clare Maria Liakata

Status: Draft
Revision: 7dff174
November 12, 2012

# Contents

# 1 Project Summary

Partridge is a web-based tool designed to assist in information processing and knowledge acquisition within the domain of scientific research.

Since the advent of the 'Digital Age' and the ability of computers to copy and reproduce information for a negligible cost, the amount of information available to researchers has been increasing drastically. A 2009 study by B-C Björk suggests that approximately 1.4 Million papers were published in the year 2006 alone[2]. Moreover, the growing popularity of Open Access publishing (making papers available for free online[11]) across most scientific disciplines[2][3] is providing researchers with an even larger volume of information to be processed. As available information increases, relevant material becomes progressively more difficult to find and the need for an automated information retrieval tool more apparent.

Partridge aims to autonomously process as many scientific papers as possible to facilitate researchers who would otherwise be required to manually read each paper themselves. This should reduce the amount of information that the reader is required to process themselves, thereby speeding up the research process. Partridge will achieve this through the use of several existing techniques in the field of Natural Language Processing which are discussed below.

From the point of view of it's users, Partridge will assist researchers in two ways. The system will provide filtering of papers based upon their specific domain (i.e. is the paper primarily concerned with methodology within an experiment in chemistry or is it about Ethics in Psychological studies?) and their result, whether the paper yielded positive, negative or inconclusive evidence for a hypothesis. Depending upon the time constraints of the project, it is hoped that Partridge will also offer a user 'profiling' system that provides recommendations for researchers based on their reading history. This feature should help users find relevant papers more quickly or find research that they may have otherwise overlooked.

There are already several tools that help researchers manage the vast library of internet journals available on the internet. Search engines such as Google (`http://www.google.com/`), and social citation management tools such as CiteULike (`http://www.citeulike.org`), do offer some assistance in tracking down relevant information. However, these tools are often too general or rely upon the user knowing exactly what keywords to search for before carrying out the search. These drawbacks are further discussed in Section 2.2 below.

To overcome the drawbacks of these existing systems, Partridge will make use of several modern Natural Language Processing (NLP) techniques. NLP enables the automated extraction of meaningful information from texts written in human languages such as English or French. There have already been several papers on using NLP for detecting emotions in suicide notes[6], the genre of a web page on the World Wide Web[9] and emotional polarity of a phrase or sentence[12]. Liakata et al. have also used NLP techniques to classify sentences within scientific papers to determine what scientific concept they relate to (i.e. does this sentence cover background information or is it a part of the

hypothesis?)[7]. Partridge will build upon and make use of these existing applications of NLP to filter and retrieve data from scientific papers in a novel way.

# 2 Current Progress

The Partridge project has been underway since the beginning of October. The following section looks at some literature on the subjects of Natural Language Processing and information retrieval within the domain of scientific papers. Some related works are investigated and compared to Partridge and details of prototyping work that has been carried out are given.

## 2.1 Literature Review

Natural Language Processing is still a relatively unexplored discipline and as such is still a very active area of research and development within the Artificial Intelligence community. E. Liddy defines Natural Language Processing as:

> A theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications. [8]

In the case of Partridge, scientific papers, constituting the naturally occuring texts, are processed at sentence-by-sentence and word-by-word levels of linguistics and represented in the form of Extended Markup Language (XML) documents, a format that is both human and machine readable. This information is then used for the purpose of classifying and searching papers in a human-like way.

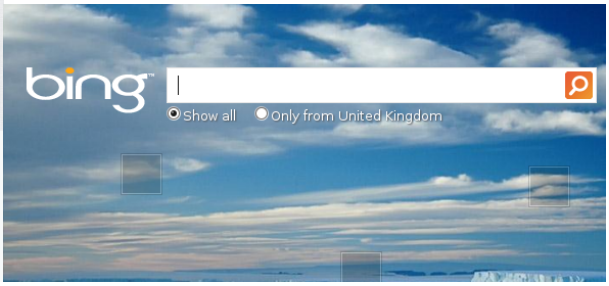
It is therefore necessary to define what a 'human-like way' of processing scientific papers on the internet consists of. Steven Krug is an expert in the field of human-computer interaction on the internet. In 'Don't Make Me Think', Krug suggests that when browsing the internet, humans find it much easier to locate specific information within a labelled and logically structured document than one that is provided as a single text entity.[4]. In order to help humans to find relevant research papers, Partridge will represent all research papers in its repository in a logical hierarchy that can be queried directly by the human users and also processed using other NLP techniques to extract further information.

Partridge's document storage format should therefore be formalised to provide a standard way of processing each document. In their 2007 paper, Soldatova and Liakata propose a methodology for storing the Core Information about Scientific Papers (CISP) as a way to formally represent scientific concepts that should be present in the papers in a logical ontology[10]. They then proceed to define a schema for their CISP ontology that defines the Core Scientific Concepts (CoreSC) as part of the XML document itself[5]. Using this format for Partridge would allow the system to

## 2.2 Related Works

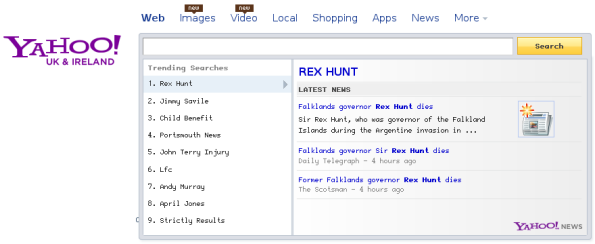
There are already many existing systems for finding and filtering information on the World Wide Web. Search engines are very useful for information retrieval in this very large and generalised search domain. Most people have heard of Google (`http://wwww.google.com`), Yahoo (`http://www.yahoo.co.uk`), Bing (`http://www.bing.com`) and Ask (`http://www.ask.com`). There are many more similar systems available for free general use across the internet. They all present very similar user interfaces (as shown in Figure 1) in which users are asked to supply keywords that might be linked to relevant documents and the search engine returns a list of Uniform Resource Locators (URLs) that they consider to match the user's query.


(a) Ask.com




(b) Bing.com


(c) Google.com




(d) Yahoo.com

Figure 1: 4 popular search engine interfaces

Search engines are helpful in locating pages and websites within the World Wide Web. Unfortunately, the problem space they deal with is usually too big for them to find scientific papers and journals given a set of keywords. Internet search engines index a huge proportion of irrelevant information compared to useful information[1], and as a result, even relatively specific queries such as effects of gravity on rockets" yield millions of results (as shown in Figure 2).

Partridge offers an advantage over these mechanisms as it will specifically index research papers rather than attempting to index the whole Internet. This means that there should be a higher proportion of useful information as output compared to the output of an Internet Search Engine.

There are also a number of search and indexing systems that specifically look for scientific

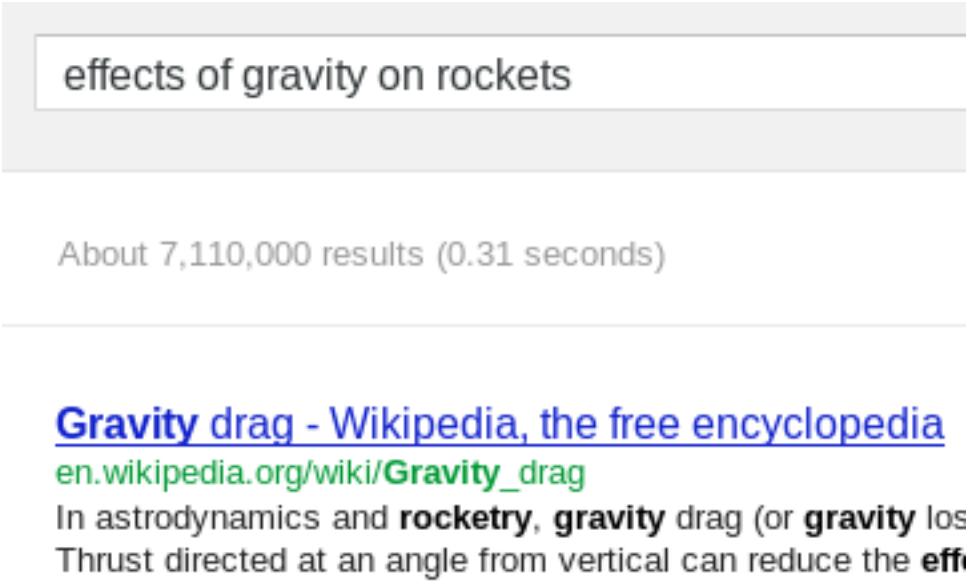
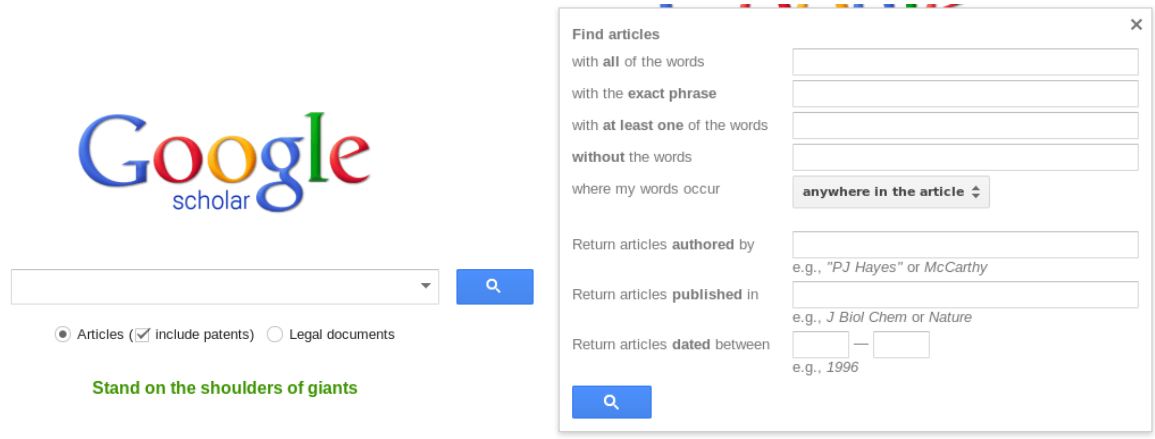



Figure 2: Google showing over 7M results for "effects of gravity on rockets"

papers as opposed to web pages. One of the most publicised and well known paper search system is Google Scholar (`http://scholar.google.com`). As can be seen in Figure 3a, This is an adaptation of Google's general search engine (discussed above) to specifically index and search scientific papers. Google also offers advanced query options specific to Scholar that allow searching by author, year and for words that occur only in the document title as shown in figure 3b. Whilst this does deal with the problem of 'information overload' and provides even more fine control over the information returned from searches, the user is still required to have a very good idea of what they are looking for in terms of keywords and/or specific authors. It is possible that a user would not know what they are looking for until they've seen it. Even if the user has a set of keywords to search for, they can only search the title of the paper or the content as a whole. This means that users who want to find a particular phrase within a CoreSC part of the paper (e.g. only look for this phrase in the 'Result' section of the paper) are unable to get results at their desired level of detail.

Partridge will provide the option to filter papers by subject and it is hoped that the system will also provide user-specific recommendations by profiling them through their reading history. This will make it easier for users to find relevant papers without knowing exactly which keywords they need to search for. Partridge will also offer facilitate searching for keywords within a specific CoreSC section by making use of Liaketa et al's SAPIENTA project for classifying each sentence of paper.

# 3 Planning




(a) Google Scholar's General front page (b) Advanced search features

Figure 3: Google Scholar's user interface

# References

[1] H. Berghel, "Cyberspace 2000: Dealing With Information Overload," *Commun. ACM*, vol. 40, no. 2, pp. 19–24, February 1997. [Online]. Available: http://dx.doi.org/10.1145/253671.253680

Paper presented in the ACM explaining 'information overload' and a summary of the shortfalls of modern search engines in information retrieval.

[2] B.-C. Björk, A. Roos, and M. Lauri, "Scientific journal publishing: yearly volume and open access availability," http://InformationR.net/ir/14-1/paper391.html], 2009.

This paper provided some insight into the growing area of online paper publishing and provided some figures on how many papers are published annually (or were in 2006).

[3] S. Harnad and T. Brody, "Comparing the impact of open access (oa) vs. non-oa articles in the same journals," *D-lib Magazine*, vol. 10, no. 6, 2004.

This paper observed that the popularity of Open Access articles is growing year by year - and so is awareness and visibility of OA.

[4] S. Krug, *Don't Make Me Think: A Common Sense Approach to the Web (2nd Edition)*. Thousand Oaks, CA, USA: New Riders Publishing, 2005.

This book provided some insights into how humans use the internet and some inspiration as to how Partridge could emulate this behaviour.

[5] M. Liakata and L. Soldatova, "Guidelines for the annotation of general scientific concepts," *Aberystwyth University, JISC Project Report http://ie-repository. jisc. ac. uk/88*, 2008.

This paper provides a set of guidelines on how to use CoreSC and CISP to annotate scientific documents.

[6] M. Liakata, J.-H. H. Kim, S. Saha, J. Hastings, and D. Rebholz-Schuhmann, "Three Hybrid Classifiers for the Detection of Emotions in Suicide Notes." *Biomedical informatics insights*, vol. 5, no. Suppl. 1, pp. 175–184, 2012. [Online]. Available: http://dx.doi.org/10.4137/BII.S8967

[7] M. Liakata, S. Saha, S. Dobnik, C. Batchelor, and D. Rebholz-Schuhmann, "Automatic recognition of conceptualization zones in scientific articles and two life science applications," *Bioinformatics*, vol. 28, no. 7, pp. 991–1000, Apr. 2012. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/bts071

This is Maria's key paper on SAPIENTA. It discusses some approaches her and her team took to annotating CoreSC in papers and how the system works

[8] E. Liddy, "Natural language processing," 2001.

In her encyclopedia entry, Liddy defines natural language processing and the components that make it up. This was used as a basis for some of Partridge's literature review.

[9] J. E. Mason, M. Shepherd, and J. Duffy, "An N-GramBased Approach to Automatically Identifying Web Page Genre," in *System Sciences, 2009. HICSS '09. 42nd Hawaii International Conference on.* IEEE, Jan. 2009, pp. 1–10. [Online]. Available: http://dx.doi.org/10.1109/HICSS.2009.68

This paper discusses classification of web page content and has a lot of useful information on how to use n-grams in an NLP application.

[10] L. Soldatova and M. Liakata, "An ontology methodology and cisp-the proposed core information about scientific papers," *JISC Project Report*, 2007.

In this paper, the CISP ontology is formalised and suggested as a way of providing better metadata for papers

[11] P. Suber, "Open Access Overview," http://www.earlham.edu/~peters/fos/overview.htm, October 2012.

This article gives a brief overview of Open Access publishing, what its about and how it works.

[12] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, ser. HLT '05. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005, pp. 347–354. [Online]. Available: http://acl.ldc.upenn.edu/H/H05/H05-1044.pdf

This paper discusses the best features for classifying the polarity of a phrase within the context of a machine learning NLP system.