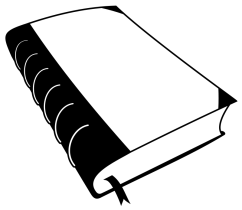

Aberystwyth University

Progress Report

Partridge: An Intelligent Literature Analysis and Recommendation Suite.

Status: Draft

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

Supervisor
Amanda Clare Maria Liakata

November 10, 2012

Contents

1 Project Summary **2**

1.1 Introduction 2

2 Current Progress **2**

2.1 Prior Art 2

3 Planning **3**

1 Project Summary

1.1 Introduction

Partridge is a web-based tool designed to assist in information processing and knowledge acquisition in 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[3]) across most scientific disciplines is providing researchers with an even larger volume of information to be processed. As the amount of available information grows, relevant material becomes increasingly difficult to find and the need for an automated information retrieval tool is ever 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 user 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.

To facilitate the above behaviour, Partridge will make use of Natural Language Processing (NLP) techniques. NLP enables the automated extraction of meaningful information from texts written in human languages such as English or French.

2 Current Progress

2.1 Prior Art

Within the field of scientific research and on the internet in general, there are already several systems for assisting users in retrieving useful and relevant information.

There are many existing search engines for finding information on the internet in general. Most people have heard of Google (`http://wwww.google.com`), and 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 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.

These search engines are often helpful in locating other pages within the World Wide Web. Unfortunately, the results they provide are usually too general to find scientific papers and journals. Search engines also 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 1).

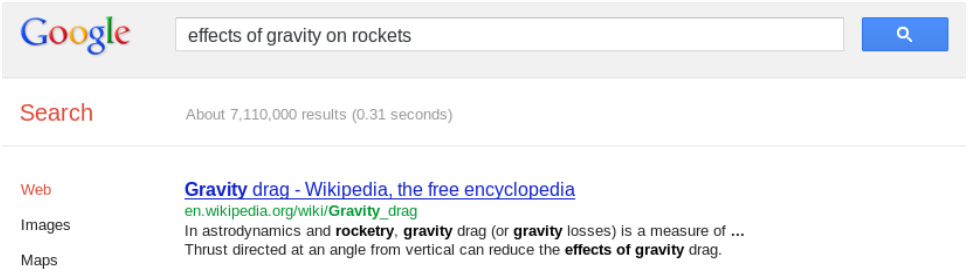

Figure 1: Google showing over 7M results for "effects of gravity on rockets"

There are also a number of search and indexing systems available for locating scientific papers by keyword or through the use of metadata to make recommendations. One of the most publicised paper search system is Google Scholar. This is an adaptation of Google's general search algorithm to specifically handle scientific papers.

3 Planning

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] 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.