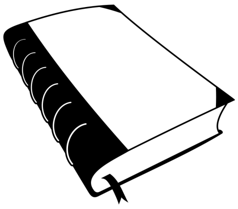


Aberystwyth University

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: 678f948
November 13, 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. B-C Björk (2009) estimates that approximately 1.4 Million papers were published in the year 2006 alone[3]. Moreover, the growing popularity of Open Access publishing (making papers available for free online[21]) across most scientific disciplines[3][8] 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. The problem is even more vital for General Practitioners. Goldacre (2008:97) points out that "there have been an estimated 15 million medical academic articles published so far, and 5000 journals published every month... picking out what's relevant is a garganutan task."[?]

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 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 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 use 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 cutting edge Artificial Intelligence (AI) techniques in order to analyse and process the papers in a more in depth way. AI is a very complex and field and implementing the above features will be incredibly challenging. To help with this, Partridge will build upon the system implemented by Liakata et al. for classifying papers on a sentence-by-sentence basis[12] and make use of Natural Language Toolkit for Python [?]

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

In science fiction literature and films, Artificial Intelligence (AI) and the ability of machines to automatically process and understand human language is almost always present. The current state of AI is a long way behind these fantastic visions. Dale(2000) comments that “Even the most ardent exponent of artificial intelligence research would have to admit that the likes of HAL in Kubrick’s 2001: A Space Odyssey remain firmly in the realms of science fiction[7]”.

Despite lacking behind the imagination of authors and script writers, over the last 60 years there has been a huge amount of progress in AI techniques. The phrase ‘Artificial Intelligence’ was coined in 1956[19] and can be used as an umbrella term, describing many subfields from “learning and perception to... diagnosing diseases*(Ibid).*”

Turing(1950) proposed a test for determining whether a machine could be considered intelligent or not[22]. Turing’s test is based upon whether a computer can communicate in a natural language proficiently enough to deceive a human into thinking that the machine is also human. A machine able to pass such a test would need to possess the ability to represent and learn from knowledge, to be able to reason about what it knows and to be able to process natural language[19]. Partridge will also need to be able to process natural language and learn from any observations it makes. It is reasonable to consider Partridge’s backend system to be an Artificial Intelligence problem.

2.1.1 Natural Language Processing

Natural Language Processing (NLP) is still a relatively unexplored discipline and as such is a very active area of research and development within the Artificial Intelligence community[13]. NLP enables the automated extraction of meaningful information from texts written in human languages such as English or French. NLP has already been applied to many text classification and data extraction problems. Studies have been carried out in the detection of emotions in suicide notes[11], the classification of a web page’s genre on the World Wide Web[15] and emotional polarity (is the sentence positive or negative) of a phrase or sentence[23]. Liddy(2001) 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 *(Ibid).*

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. Krug (2005), 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[9]. In order to help humans to find relevant research papers, Partridge will represent all papers in its repository in a logical hierarchical structure. This will facilitate better information retrieval for keyword searches and enable more advanced NLP techniques. Partridge's document storage format should therefore be standardised to provide a uniform way of processing each document.

2.1.2 CISP, CoreSC and SAPIENTA

Soldatova and Liakata(2007) proposed 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 articles in a logical ontology[20]. 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[10].

It is proposed that Partridge uses the CoreSC schema for document storage to provide a standard way to access and process the papers and thus eliminate dealing with multiple documents during its learning phase. Using CoreSC also offers the advantage of making Partridge compatible with the ART corpus. This is a set of physical chemistry and biochemistry research papers that have been pre-processed and already annotated using CISP concepts[14]. This would provide Partridge with a set of papers that could be used as a training set for classification tasks and could be used for information retrieval tests from early on in the project.

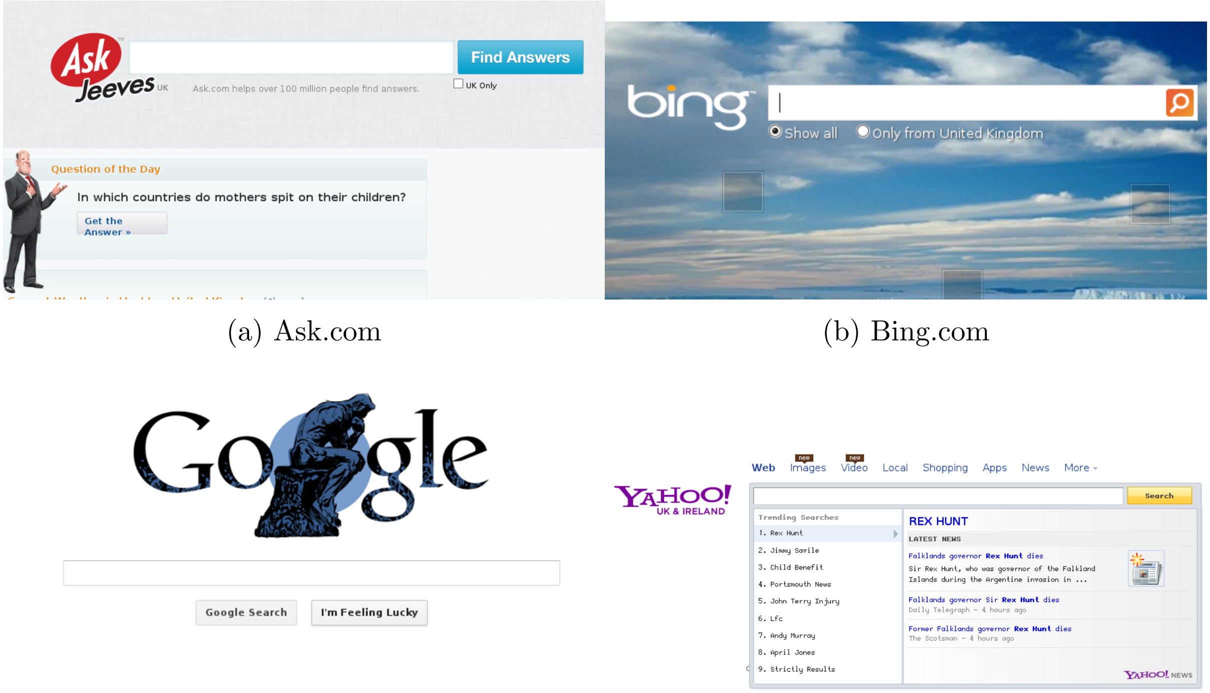
Liakata et al. (2012) describe a system for automatically processing and categorising sentences in a research paper according to their respective CoreSC element[12]. SAPIENTA (`http://www.sapientaproject.com`) was trained using the ART corpus and can achieve promising accuracies when categorising sentences (*Ibid*). With the authors' permissions, Partridge will make use of SAPIENTA when introducing new papers into its repository. It is hoped that preclassifying papers will make the development of Partridge more manageable within the timeframe available and will also allow more effort to be put into other classification problems rather than re-implementing a SAPIENTA clone. Any modifications that have to be made to SAPIENTA to integrate it into Partridge will be submitted back to the original authors to help them improve their works.

After pre-processing by SAPIENTA is complete, Partridge will need to do some classification work of its own to determine the area of study (i.e. Physics, Chemistry, Biology), type of paper (Literature review, Case Study, Experiment) and the polarity of the result (positive, negative, inconclusive). These tasks involve implementing existing AI techniques and applying them specifically to the task of text classification. To avoid the time consuming task of re-implementing these methods, it was decided that a library should be used instead. Different NLP libraries and their properies are discussed below in Section 2.3.3.

2.2 Related Works

2.2.1 Search Engines

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.



In which countries do mothers spit on their children?

Trending Searches

1. Rex Hunt
2. Jimmy Savile
3. Child Benefit
4. Portsmouth News
5. John Terry Injury
6. Lfc
7. Andy Murray
8. April Jones
9. Strictly Results

REX HUNT

LATEST NEWS

Falklands governor **Rex Hunt** dies
Sir Rex Hunt, who was governor of the Falkland Islands during the Argentine invasion in ...

Falklands governor Sir **Rex Hunt** dies
Daily Telegraph - 4 hours ago

Former Falklands governor **Rex Hunt** dies
The Scotsman - 4 hours ago

(a) Ask.com

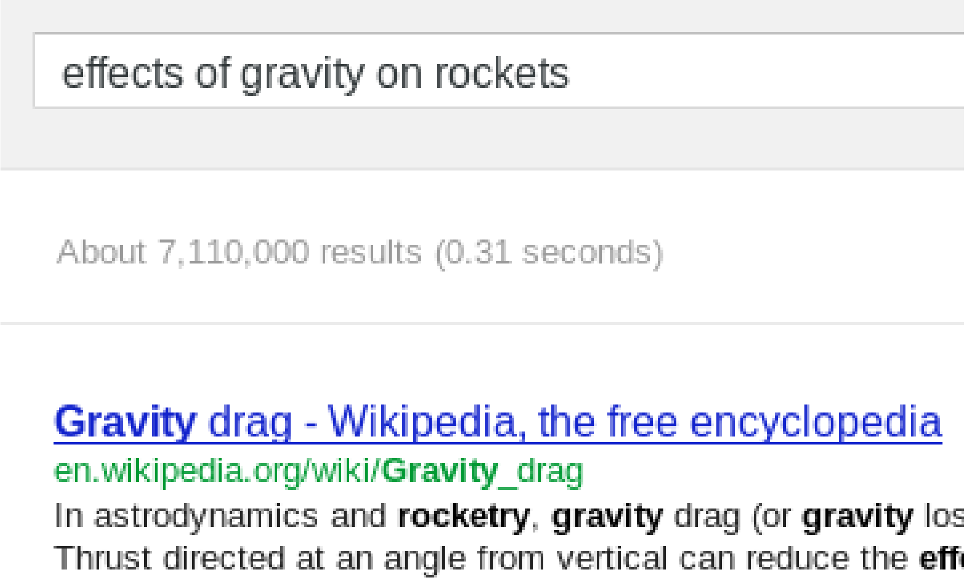
(b) Bing.com

(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).



effects of gravity on rockets

About 7,110,000 results (0.31 seconds)

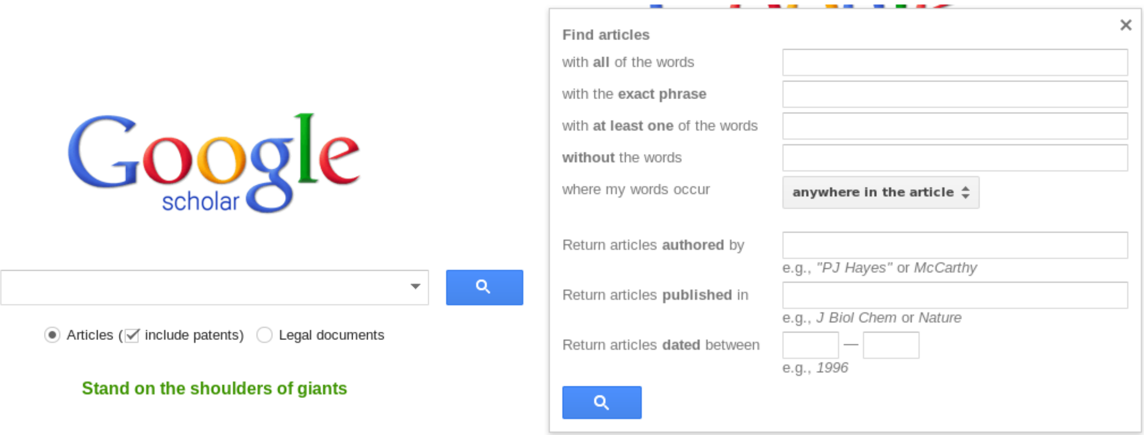
Gravity drag - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Gravity**_drag
In astrodynamics and **rocketry**, **gravity** drag (or **gravity** los
Thrust directed at an angle from vertical can reduce the **eff**

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

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.

2.2.2 Scientific Paper Search Engines

There are also a number of search and indexing systems that specifically look for scientific papers as opposed to web pages. One of the most publicised and well known paper search systems 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.



○ Articles (☑ include patents) ○ Legal documents

Stand on the shoulders of giants

Find articles

with **all** of the words

with the **exact phrase**

with **at least one** of the words

without the words

where my words occur: **anywhere in the article**

Return articles **authored** by

e.g., *"PJ Hayes"* or *McCarthy*

Return articles **published** in

e.g., *J Biol Chem* or *Nature*

Return articles **dated** between —

e.g., *1996*

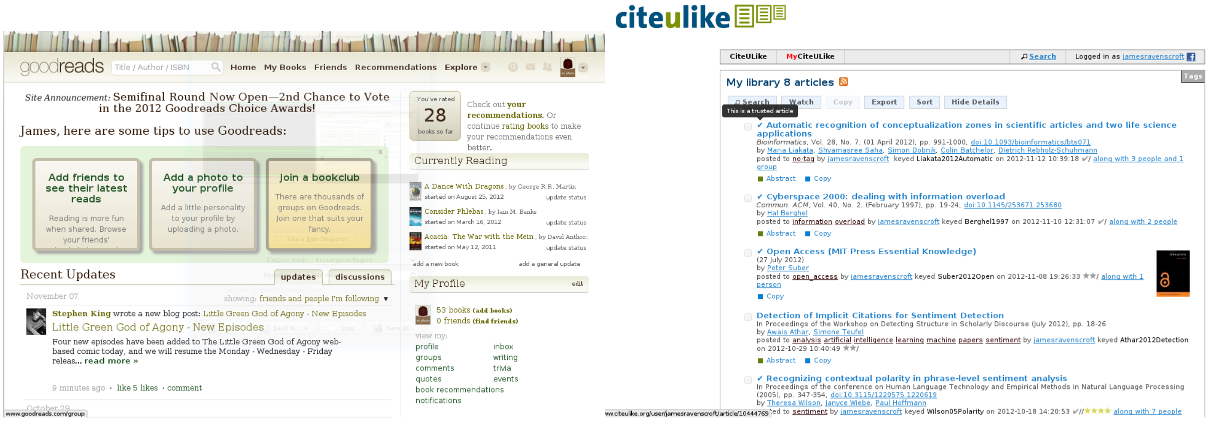
(a) Google Scholar's General front page (b) Advanced search features

Figure 3: Google Scholar's user interface

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 Liakata et al's SAPIENTA project for classifying each sentence of paper as it is added to the repository.

2.2.3 Social Citation and Recommendation Engines

Social citation and recommendation engines also provide a partial solution to the 'information overload' problem. Services like Goodreads (`http://www.goodreads.com/`) and CiteUlike (`http://www.citeulike.org/`) allow you to register your interest in specific authors and subjects. This allows the sites to build up a profile of the sorts of materials that you might be interested in and provide lists of recommendations as in Figure ??.



goodreads | Title / Author / ISBN | Home | My Books | Friends | Recommendations | Explore

Site Announcement: Semifinal Round Now Open—2nd Chance to Vote in the 2012 Goodreads Choice Awards!

James, here are some tips to use Goodreads:

Add friends to see their latest reads – Reading is more fun when shared. Browse your friends'

Add a photo to your profile – Add a little personality to your profile by uploading a photo.

Join a bookclub – There are thousands of groups on Goodreads. Join one that suits your fancy.

Recent Updates | updates | discussions

November 07 — showing: friends and people I'm following ▾

Stephen King wrote a new blog post: Little Green God of Agony - New Episodes

Little Green God of Agony - New Episodes

Four new episodes have been added to The Little Green God of Agony web-based comic today, and we will resume the Monday - Wednesday - Friday releas... **read more »**

9 minutes ago · like 5 likes · comment

October 29

www.goodreads.com/group

You've rated 28 books so far

Check out **your recommendations.** Or continue rating books to make your recommendations even better.

Currently Reading

A Dance With Dragons, by George R.R. Martin — started on August 25, 2012 — update status

Consider Phlebas, by Iain M. Banks — started on March 16, 2012 — update status

Acacia: The War with the Mein, by David Anthony — started on May 12, 2011 — update status

add a new book — add a general update

My Profile — edit

53 books **(add books)**

0 friends **(find friends)**

view my:

profile, inbox, groups, writing, comments, trivia, quotes, events, book recommendations, notifications

citeulike

CiteULike | MyCiteULike | Search | Logged in as jamesravenscroft

My library 8 articles | Tags

Search | Watch | Copy | Export | Sort | Hide Details

This is a trusted article

✓ Automatic recognition of conceptualization zones in scientific articles and two life science applications
Bioinformatics, Vol. 28, No. 7. (01 April 2012), pp. 991-1000, doi:10.1093/bioinformatics/bts071
by Maria Liakata, Shyamasree Saha, Simon Dobnik, Colin Batchelor, Dietrich Rebholz-Schuhmann
posted to no-tag by jamesravenscroft keyed Liakata2012Automatic on 2012-11-12 10:39:18 ✓/ along with 3 people and 1 group
Abstract | Copy

✓ Cyberspace 2000: dealing with information overload
Commun. ACM, Vol. 40, No. 2. (February 1997), pp. 19-24, doi:10.1145/253671.253680
by Hal Berghel
posted to information overload by jamesravenscroft keyed Berghel1997 on 2012-11-10 12:31:07 ✓/ along with 2 people
Abstract | Copy

✓ Open Access (MIT Press Essential Knowledge)
(27 July 2012)
by Peter Suber
posted to open_access by jamesravenscroft keyed Suber2012Open on 2012-11-08 19:26:33 ★★/ along with 1 person
Copy

Detection of Implicit Citations for Sentiment Detection
In Proceedings of the Workshop on Detecting Structure in Scholarly Discourse (July 2012), pp. 18-26
by Awais Athar, Simone Teufel
posted to analysis artificial intelligence learning machine papers sentiment by jamesravenscroft keyed Athar2012Detection on 2012-10-29 10:40:49 ★★/
Abstract | Copy

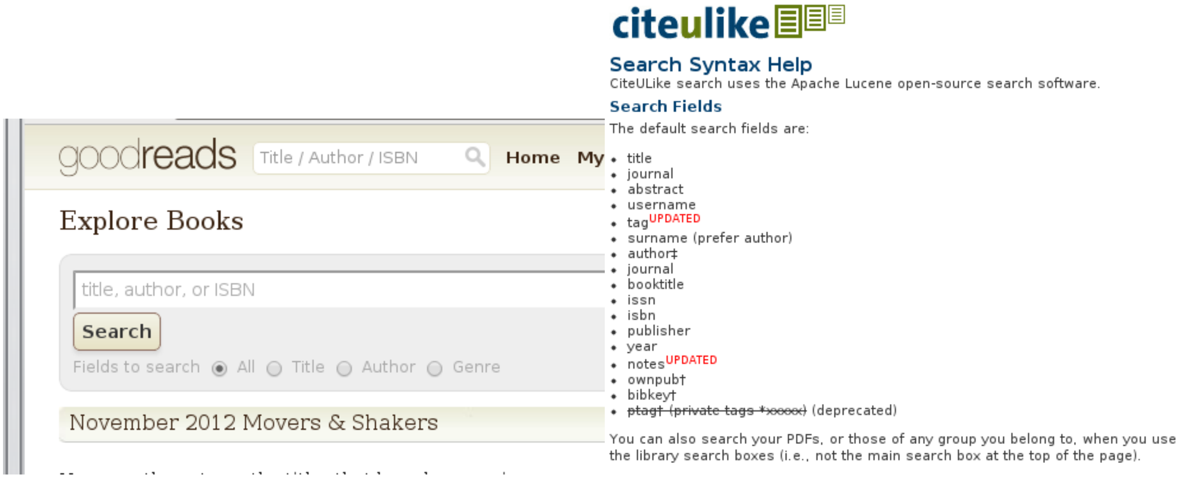
✓ Recognizing contextual polarity in phrase-level sentiment analysis
In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (2005), pp. 347-354, doi:10.3115/1220575.1220619
by Theresa Wilson, Janyce Wiebe, Paul Hoffmann
...ted to sentiment by jamesravenscroft keyed Wilson05Polarity on 2012-10-18 14:20:53 ✓//★★★★ along with 7 people

ww.citeulike.org/user/jamesravenscroft/article/10444769

(a) Goodreads user profile page (b) a CiteULike user profile page

Figure 4: Goodreads and citeulike social recommendation systems

These systems have the ability to make recommendations to the user without requiring specific keywords or search terms. They do this by learning the user's profile and taking into account the preferences of their 'friends' and their browsing history. However, the above-named systems do not take into account the content of the paper or book. They only deal with metadata as can be seen in Figure 5. This means that important discriminatory information that could be contained within the actual document content is overlooked completely. Partridge will analyse documents on a sentence-by-sentence and possibly word-by-word basis, thereby taking account of any embedded information that could be missed by these social metadata systems.



Search Syntax Help

CiteULike search uses the Apache Lucene open-source search software.

Search Fields

The default search fields are:

- title
- journal
- abstract
- username
- tagUPDATED
- surname (prefer author)
- author‡
- journal
- booktitle
- issn
- isbn
- publisher
- year
- notesUPDATED
- ownpub†
- bibkey†
- ~~ptag† (private tags *xxxxx)~~ (deprecated)

You can also search your PDFs, or those of any group you belong to, when you use the library search boxes (i.e., not the main search box at the top of the page).

(a) Goodreads advanced search page

(b) CiteULike advanced search options

Figure 5: Goodreads and citeulike search only deal with metadata.

2.3 Methodology

Partridge is based upon several fairly complex and new ideas in AI and NLP. For the project to be a success it is imperative that the right language, environment and toolkits are used to avoid unnecessary work on existing technologies and allow priority to be given to the novel aspects of the project.

The language that Partridge will be written in is one of the most fundamental and important choices to be made. The language choice effects not only the end performance of the application, but the speed at which the project is developed and the portability of the end code [4]. To avoid a learning overhead, it was decided that the programming language should be one that the author is familiar with. This reduced the choice of language down to C, Java and Python.

The C programming language was invented by Kernighan and Ritchie and published in 1978[18]. The language is small and optimised[17] and therefore compiled C programs tend to run extremely fast. Unfortunately, C is designed to be general and lack restriction [?] which is often an advantage for programmers who favour optimised code over excessive error checking. However, this means that debugging C programms can be quite long winded and challenging. Since the NLP aspects of Partridge present an amibitious challenge in themselves, having to debug applications without help from a managed programming environment is not desirable.

The Java programming language on the other hand, does provide memory management and error checking[5] at the cost of program performance. Java is a pseudo-compiled language that is translated into an intermediate bytecode which is then interpreted at runtime by an interpreter on the client computer called a Virtuam Machine(VM). Gosling (2000) describes java as as a "general-purpose... programming language [that allows] developers to write a program once and then be able to run it everywhere on the internet[?]." Java provides a solid and stable programming environment and stringent error checking, both of which would be advantageous in the development of Partridge. The language syntax is very verbose and a lot of 'boiler-plate' code must be written before a program can be executed. Therefore Java is not suitable as a rapid prototyping language. Since Partridge's development requires a steep learning curve for the author, a language that allows quick prototyping with minimal programming.

Python is a programming language with a famously shallow learning curve, readable syntax and dynamic inspection that facilitates rapid prototyping [2]. The language was invented by Guido Van Rossum as a multi-use, flexible, interpreted scripting language, extensible via compiled C modules[?]. Developing Partridge in python would provide a flexible prototyping environment and readable syntax allowing for investigative work to be carried out quickly and efficiently with minimal overheads for writing boiler-plate code and debugging. The ability to extend the language using native code would give Partridge an even greater advantage. Since NLP techniques can be quite processor intensive and Python is an interpreted language, and in its very nature, slower than a compiled language, inefficient sections of Python code could be re-written as C extensions and compiled into the application. For these reasons, Python was selected as the preferred programming language for Partridge development.

2.3.1 User Interface Style

2.3.2 Web Presentation Frameworks

2.3.3 Natural Language Libraries

. There are several existing libraries to facilitate Natural Language Processing. Many are written for Java [16][6] and are very complex or not well documented. The Natural Language Toolkit (NLTK) is a simple and intuitive library written for Python. Bird(2009) states that NLTK was designed "to provide an intuitive framework along with substantial building blocks, giving users a practical knowledge of NLP[2]". The project is relatively mature in comparison to the above named Java libraries. There is also a free book that accompanies the project available at `http://nltk.org/book/` which provides a huge amount of information on how to implement many popular NLP techniques using the library. For this reason NLTK was chosen as the primary accompanying library for the project.

2.4 Prototyping/Pilot Studies

2.5 Subsequent Changes to Methodology

3 Planning

3.1 Development Methodology

3.1.1 Agile

3.1.2 Waterfall Development

3.1.3 Partridge's methodology

3.2 Work Timeline

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] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*, ser. Oreilly Series. O'Reilly Media, Incorporated, 2009. [Online]. Available: http://books.google.co.uk/books?id=KGIbfiiP1i4C

This book provides an interesting preface on NLP and some reference for the NLTK python library

[3] 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).

[4] C. Britton, “Choosing a programming language,” http://msdn.microsoft.com/en-us/library/cc168615.aspx, 2008.

Brief paper discussing how to choose a programming language successfully

[5] N. Coffey, “Java for c programmers,” http://www.javamex.com/tutorials/how_to/java_for_c_programmers.shtml retrieved on 13/11/2012, 2008.

Provides some notes on the differences between C and Java

[6] H. Cunningham, D. Maynard, and K. Bontcheva, *Text processing with gate.* Gateway Press CA, 2011.

Another NLP toolkit for java. This one seems very complex compared to NLTK

[7] R. Dale, H. Moisl, and H. Somers, *Handbook of Natural Language Processing.* Marcel Dekker, 2000. [Online]. Available: http://books.google.co.uk/books?id=VoOLvxyX0BUC

Provided some good background information on NLP and an interesting preface on modern AI capabilities

[8] 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.

[9] 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.

[10] 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.

[11] 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

[12] 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

[13] 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.

[14] L. M. and S. L, "The art corpus," 2009.

[15] J. E. Mason, M. Shepherd, and J. Duffy, "An N-GramBased Approach to Automatically Identifying Web Page Genre," in *System Sciences, 2009. HICSS &#039;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.

[16] A. K. McCallum, "Mallet: A machine learning for language toolkit," http://mallet.cs.umass.edu, 2002.

An NLP toolkit written for Java, deemed to be overcomplicated and underdocumented

[17] P. Prinz and U. Prinz, *C pocket reference.* O'Reilly Media, Incorporated, 2002.

A pocket reference guide to the C programming language, introduction discusses why C is popular

[18] D. Ritchie, S. Johnson, M. Lesk, and B. Kernighan, "The c programming language," *Bell Sys. Tech. J*, vol. 57, pp. 1991–2019, 1978.

The original specification for the C Programming language describes the motivation behind C

[19] S. Russell and S. Norvig, *Artificial Intelligence: A Modern Approach*, ser. Prentice Hall Series in Artificial Intelligence. Prentice Hall, 2010. [Online]. Available: http://books.google.co.uk/books?id=8jZBksh-bUMC

Provided some general background in AI as well as a lot of information about machine learning techniques to be used as a backend for Partridge's NLP system.

[20] 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

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

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

[22] A. Turing, "Computing machinery and intelligence," *Mind*, vol. 59, no. 236, pp. 433–460, 1950.

This is one of the first papers to discuss artificial intelligence and provides a link between natural language processing and AI

[23] 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.