Χαρακτηρίζοντας περιόδους χρηστών αφιερωμένες σε συγκεκριμένες δραστηριότητες ηλεκτρονικού ¨ξεφυλλίσματος¨ : Μια περιπτωσιολογική μελέτη

Characterizing e-print user sessions: A case study of arXiv.org users

Γεωργία Ροϊδούλη[1], Ηλίας Σάββας[2]
Κεντρική Βιβλιοθήκη, ΤΕΙ Λάρισας[1], Τμήμα Πληροφορικής & Τηλεπικοινωνιών, ΤΕΙ Λάρισας[2]
roidouli@teilar.gr, savvas@teilar.gr

Georgia Roidouli[1], Ilias Savvas[2]
Central Library, TEI of Larissa[1], Department of Computer Science & Telecommunications, TEI of Larissa[2]
New Buildings, Larissa, 41 110, Greece
roidouli@teilar.gr, savvas@teilar.gr

Περίληψη

Η επίδραση του Ιστού (Web) έχει ως αποτέλεσμα να επιτρέπει όλο και σε περισσότερους χρήστες του Διαδικτύου να έχουν τις ευκαιρίες απομακρυσμένης πρόσβασης, του ηλεκτρονικού ¨ξεφυλλίσματος¨, της αναζήτησης και της απευθείας διάθεσης τεκμηρίων από πληροφοριακές πηγές. Η εξέλιξη των e-prints ηλεκτρονικών αρχείων παρέχει νέες υπηρεσίες αναζήτησης για τους χρήστες σε αντίγραφα ερευνητικών εργασιών για να εντοπίζουν τις πληροφορίες ευκολότερα, γρηγορότερα και χωρίς κόστος.

Ένα καινοτόμο παράδειγμα e-print ηλεκτρονικού αρχείου είναι το arXiv.org αρχείο. Παραδοσιακά χρησιμοποιείται από τους φυσικούς που είναι πρωτοπόροι στον τρόπο ηλεκτρονικής επικοινωνίας. Για περισσότερα από 13 έτη ανταλλάσσουν τα ερευνητικά αποτελέσματά τους ηλεκτρονικά και μπορούν να ενημερώνονται για οτιδήποτε νέο κυκλοφορεί στον κλάδο τους άμεσα.

Σ΄ αυτή τη μελέτη παρουσιάζονται οι τρόποι ανάκτησης πληροφοριών των φυσικών και ειδικότερα εξετάζεται πώς οι χρήστες επιτυγχάνουν την πρόσβαση στα μεμονωμένα έγγραφα και στα σύνολα εγγράφων σε συγκεκριμένες περιόδους δραστηριοτήτων "ξεφυλλίσματος" και "αναζήτησης".

Σαν μεθοδολογικό εργαλείο έχει χρησιμοποιηθεί ένα πρότυπο διάγραμμα πλοήγησης κάθε δραστηριότητας ώστε να βοηθήσει να μελετηθεί η συνολική συμπεριφορά των χρηστών. Με τα αποτελέσματα αυτής της μελέτης εξασφαλίζουμε την εγκυρότητα της χρησιμότητας του συγκεκριμένου arXiv ηλεκτρονικού αρχείου και επιθυμούμε να παρακινήσουμε ερευνητές από διάφορα υπόβαθρα να χρησιμοποιήσουν τα ίδια πρότυπα

ανάκτησης πληροφοριών ώστε να ενισχυθεί η παροχή ανοικτής πρόσβασης στο ερευνητικό έργο τους.

Λέξεις Κλειδιά: *Ψηφιακή Βιβλιοθήκη, Ηλεκτρονικά αρχεία, Ανάκτηση πληροφοριών, Συμπεριφορά χρηστών*

Abstract

The effect of the Web has been to increase the opportunities for access, allowing copies of research papers to be downloaded across the Internet. The evolution of 'e-print archives' introduces new search services for users to locate information easier, faster and at no cost. Traditionally physicists have been innovators in methods of scholarly communication by using preprints for over 30 years. The introduction of arXiv.org e-print service has increasingly changed the conventional way of scientific communication without usurping the role of traditional publications. ArXiv users are an interesting group of people. For more than 13 years they exchange their research results electronically and they can find what is new on the archive just as soon as everyone does.

This paper identifies the physicists' information retrieval patterns and in particular examines how the users gain access to individual papers and sets of papers in their periods of "browsing" and "searching" activity. As a methodological tool, a model of a navigational diagram of each user activity has been applied to allow aggregate behaviour of many hundreds of user sessions to be visually compared.

Having the results of a study is more useful than merely guessing what arXiv.org users' current behaviour is in publicising their work in electronic form and how they search for academic papers in their community. Knowing that they make use of certain information retrieval patterns is also useful in this context because researchers from various backgrounds can be informed and get motivated to help institutions provide open access to their own research output on a large scale by setting up institutional digital archives of refereed research papers (e-prints).

Keywords: *Digital Library, arXiv, Preprints, E-prints, Information Seeking, Patterns*

Introduction

Initially, information technology has been used by libraries to automate the processing functions of the library, e.g. bibliographic services, inter-library lending and document delivery. Over the past years, there is a trend towards using information technology to enhance services to end-users. Issues involved here include: increased speed of access and delivery, access to resources not contained in the library, and distance access by the clients. For those particular reasons, a number of techniques and related standards has been used which allow documents to be distributed in electronic form.

There is a branch in the area of digital libraries (within the WWW) known as online archiving which evolved from the practice of authors emailing pre-prints of their

papers to peers for informal feedback. With the online archives, authors can deposit their pre-refereed work (pre-prints) and published work (post-prints) into an archive for all to see. ArXiv.org is a digital archive that according to Luce (2001), "acts as a repository for electronic versions of papers in physics and mathematics, providing a rapid and convenient way for scientists to rapidly share their results with colleagues".

Chronologically ordered, the archive started with the *hep-th* (High Energy Physics – Theory) as a postal service supported with an email interface. In 1992, the FTP interface was added and then *hep-ph* (High Energy Physics – Phenomenology) and *hep-lat* (High Energy Physics – Lattice) were added locally and *alg-geom* (Algebraic Geometry), *astro-ph* (Astrophysics) and *cond-mat* (Condensed Matter) were added remotely. By December 1993, a web interface was added and in November of 1994, data at some remote archives became mirrors. 1996 marked the growth of mirror networks and in June the web upload facility for author submissions was added. (Luce, 2001)

The arXiv and all the other e-print services "are seen as the catalyst for the scholarly and scientific literature from the cost barriers imposed by access-tolls" (Pinfield *et al.* 2001). Along with new value added services such as Arc (http://arc.cs.odu.edu/), OAIster (http://oaister.umdl.umich.edu/cgi/b/bib/bib-idx?c=oaister;page=simple), and Citebase (http://citebase.eprints.org/cgi-bin/search), they have impacted in researchers' reading behaviour.

The research of users' information seeking behaviour of this e-print service was chosen because of its high visibility, use and impact. The physicists' model of scholarly communication via arXiv is formal. Although papers are unrefereed the archives can be used by other services as the basis for refereeing papers (Taubes, 19996a), most obviously in the form of 'overlay' journals. Studies also show that the level of arXiv's usage is significant. Citations to electronic e-prints such as those accessible from the physics archive "have nearly doubled every year since 1992" (Youngen, 1998). A study of arXiv.org showed that in the case of highly cited papers there is a significant positive correlation between how often a paper is cited and how often is downloaded (Harnad *et al*. 2001), showing how access can enhance impact for the best papers.

Traditionally, when people retrieve information in the arXiv.org their activities are classified into two distinct patterns: searching and browsing (Pinfield, 2001). Searching implies that the user knows exactly what to look for, while browsing implies that the user can navigate among correlated searchable items to look for something new or interesting. The thesis examines physicists' use of arXiv.org, and in particular how they gain access to individual papers and sets of papers in their periods of "browsing" and "searching" activity.

The Study

As a methodological tool to help us gather detailed interaction data, we developed a model of a navigational diagram that allowed us to aggregate the current information seeking behaviour of 16 individual users. In summary the users came from the

University of Southampton in the Maths and Physics department respectively. They were identified from the web logs of the UK arXiv.org mirror.

The web logs were analysed and broken down by time and IP address into individual user sessions (http://data.archive.ecs.soton.ac.uk/export/raid/lac/georgia).
Each user session which is visualised by a navigational diagram is comprised of a sequence of downloads that describes the information process activity of each user.

Characterizing arXiv.org user sessions

The general rule to understand the meaning of the navigational diagrams is that time progresses depth-first down the diagram and across from left to right. According to the graph layout program, each page is categorized (browse, search, abstract, full-text document, bibliography) and labeled according to its category. For example: [SEARCH (find), BROWSE (recent, new), ABSTRACT (abs), DOCUMENT (ps, pdf, html, format), BIB (citation, reference), EXTERNAL SOURCE (SPIRES or ADS link, personal web-pages, universities web sites)].

BROWSING & SEARCHING SESSIONS

The Browsing and Searching sessions derive from the archives' interface.

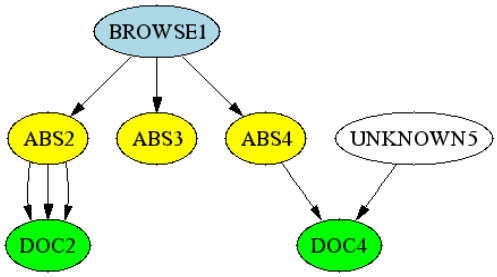


Figure 1: A browse/abs/doc session

Figure 1 shows a typical browse/abs/doc (BAD) session, where the user starts from a browse page [actually the list of HEP-TH (=high energy physics theory category in the archive) papers in June 1999, the session taking place on Dec 1st 2000]. From the browse page, the user clicks on a particular abstract and from there extracts the paper itself (making several attempts to get two formats, .ps and. tar.gz). The user then returns to the browse page and clicks on another abstract (this time the full paper is not requested). Finally the user returns to the browse page, chooses another abstract and this time fetches the full text of the article (the *unknown* box indicates a browser retry while waiting for the PostScript to be regenerated from the original TeX source and is an artifact of the web logs).

Another typical session of information seeking can be seen in figure 2. In a search /document (SD) session, four specific searches lead directly to the downloads of a number of full text documents respectively. Note that this user did not download the abstracts first.

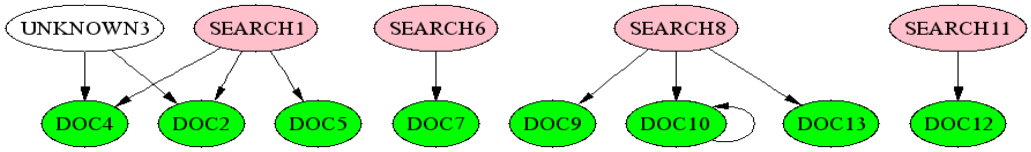


Figure 2: A search/doc session

Following on from this, it is easy to describe other session types:

1. SAD, where the search results cause the user to read the abstracts before downloading the documents.
2. BD, where a browse leads directly to documents, ignoring the abstracts.
3. BA and SA where only the abstracts are read.

A different kind of activity is seen in figure 3, where bibliography links are followed to determine the next most relevant article to read. In the specific case of this session, the 'bibliography' links are provided by SPIRES* and are to documents which have cited the current document. Therefore, we can tell that DOC4 (had it been downloaded) was cited by DOC5 (which was downloaded).

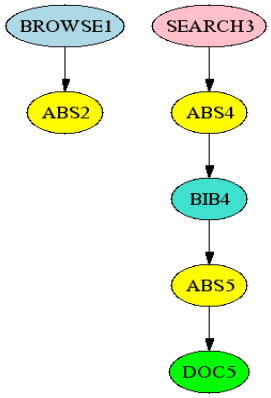


Figure 3: Search/Abstract/Bibliography/Abstract/Document session

* ***SPIRES*** is a service generated from Stanford Linear Accelerator Centre (SLAC), and is heavily used by theoretician physicists. SPIRES HEP literature database contains more than 500,000 high-energy-physics-related articles including journal papers, preprints, e-prints, technical reports, conference papers and theses http://www-library.desy.de/spires/hep/

COLLEAGUE RECOMMENDATION SESSIONS

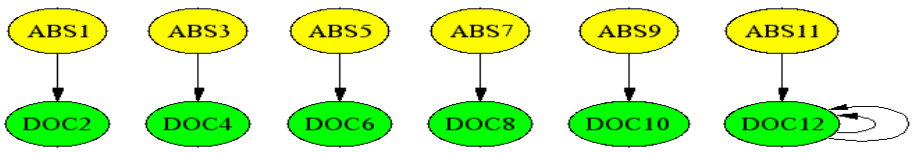


Figure 4: Abs/doc session

The retrieval pattern in Figure 4 derives from the users' knowledge of a specific number of an article that is entered in the arXiv.org. This usually happens for older

papers that exist in the archive and they are well known around the research community. The archive web interface allows users to cite references by archive number and take them first to abstracts and then to documents. Colleague recommendations or explicit citations result a lot in this specific retrieval pattern.

FIXED EXTERNAL SESSIONS

Figure 5 shows users who enter the archive from external web pages and services. The most common can be personal web pages, SPIRES, ADS* and various universities web sites (e.g. theoretical research groups of UK Universities at Imperial College, Southampton, and Cambridge etc.). Frequently these external web pages are external searches (e.g. at a SPIRES site).

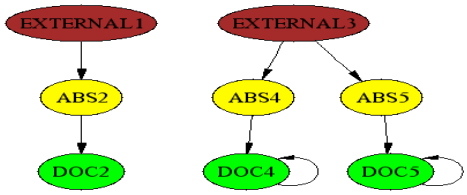
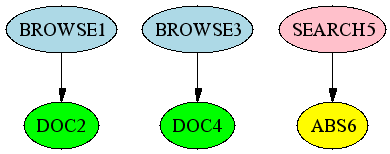


Figure 5: Fixed external session

***ADS** (Astrophysics Data System). It is a NASA funded project which maintains four bibliographic databases containing more than 3.2 million records: Astronomy, and Astrophysics, Instrumentation, Physics and Geophysics and pre-prints in Astronomy. http://adswww.harvard.edu/

MIXED SESSIONS



Figures 6: A Mixed session

Mixed sessions derive from the combination of two or more types of sessions described above. Figure 6 shows a mixed session: the user takes one document from each of two browse pages and then issues a search which results in a single abstract being downloaded. Mixed sessions can be multimodal, long and sometimes complex depending on the user's strategy of surveying the literature in the archive. The example given in Figure 7 illustrates such a session.

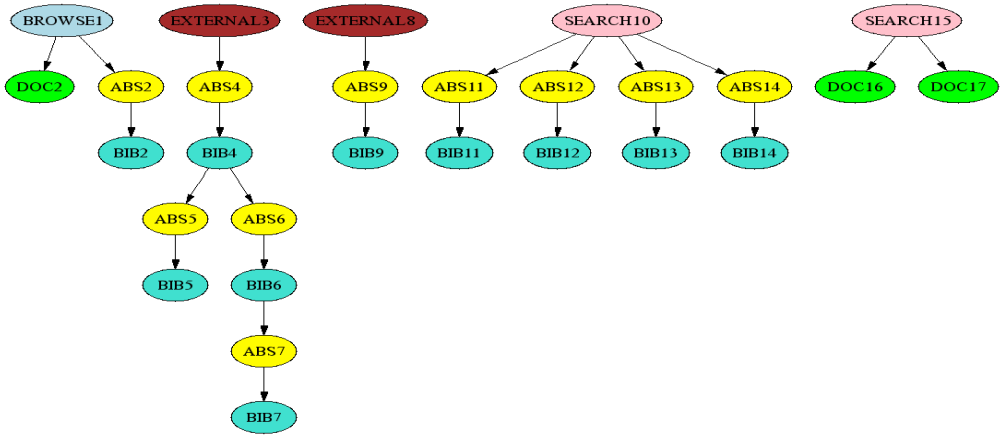


Figure 7: A multimodal long mixed session

Information Seeking Activities

The web logs observations show that article downloads in arXiv can happen in any of the following ways: searching, browsing, recommendations by people, from external links, SPIRES, and using a combination of all (mixed sessions).

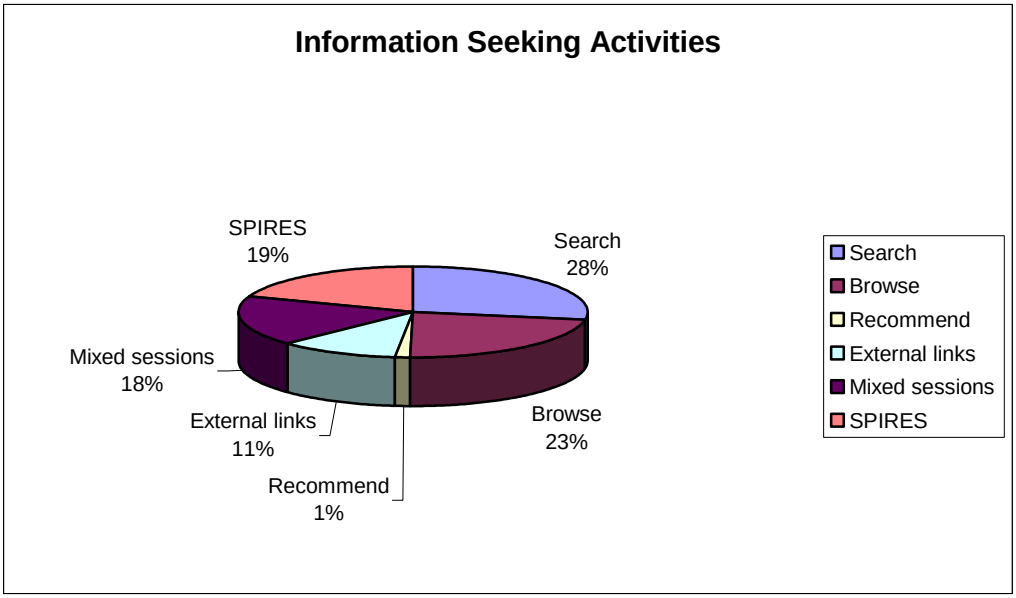


Table 1. Information Seeking Activities

An interesting finding that is apparent from Table 1 is that physicists use SPIRES as well as arXiv. This is very useful because we can figure out that not all publications of physicists go into arXiv. This piece of evidence gives the opportunity to new additional spaces (e.g. e-print institutional archives) to be created, which either by themselves or in conjunction can give a complete picture of research output from an individual, school or a university. But the question is why should the researchers be persuaded to publicise their work in electronic form, since they have a strong alternative system that works for them?

Having the results of an interview study is more useful than merely guessing what arXiv.org users' current behaviour is in publicising their work in electronic form and how they search for academic papers in their community.

Users' current behaviour

- Most researchers consult the archive to see what is new rather look up new articles in journals
- The archive constitutes a quick way of finding an article one wants to read, even if it is already published in paper form
- The archive is the main source of their work
- It provides faster awareness of other authors, much easier to track the work done by a researcher and others through the archive
- Immediate communication, ease of retrieval
- It attracts attention and makes it easy for people to access
- When they are working in a hot topic this is the best way to give to the others information about their work
- ArXiv preprints are read daily by everyone
- New results can be made public very quickly
- It helps making papers more known
- Instant distribution of work carried out, easy accessibility of past papers
- The article reaches the readers much faster than it would be in a journal.

Conclusions

Researchers who contribute to arXiv now consider it to have a central place in their work. They use it to disseminate both pre-prints and post-prints refereed articles. Interestingly, they still wish to have their work accepted by journals but do not see journals as the only means of distributing their work. In other words self – archiving is not seen as a substitute for publishing in peer-reviewed journals, but rather a useful supplement to the journal publishing process that makes research output widely available.

Looking at how arXiv is used was very helpful for us. It has helped us identify some of the things that are most important to researchers and to consider how these might be of practical significance in running similar e-print services.

The conclusion of this project is that ArXiv is a program that must be continued and become an exemplar for other disciplines to establish their own e-print services.

The fact that they use Spires as well as arXiv is very useful finding because it gives an incentive to researchers to provide services where they could deposit their research output and disseminate their results as quick as physicists do.

Also the piece of evidence proven by the web logs analyses and the interview study that the information seeking process is straight forward it could attract the attention by other researchers and make them change their culture in searching for academic papers in their community.

ArXiv is not just an archive that exists on the web. ArXiv preprints are read daily by everyone. Knowing that they use it persistently it is also important because it could persuade other scholars to might think of establishing an e-print service as well as having the traditional system of publishing their work.

Finally, the fact that users have immediate access through the archive's interface means that the archive constitutes an effective way of locating an article that someone wants to read. Researchers from other disciplines having this in mind they should start aiming of freeing up research output and thereby improving research communication by introducing similar e-print services. Within this context projects such as TARDIS (http://tardis.eprints.org) is exploring ways to influence the growth of institutional archives as the subject areas embracing electronic open archiving become broader.

References

Harnad, S. (2001). *"Research access, impact and assessment"*. Times Higher Education Supplement 1487: p16 longer version: http://www.cogsi.soton.ac.uk/~harnad/Tp/thes1.html

Luce, Richard (2001). "*E-prints intersect the Digital Library: Inside the Los Alamos arXiv*". Issues in Science and Technology Librarianship, Winter, 2001.

Pinfield, Stephen. (2001). "*How do physicists use an E-print Archive?"* D-Lib Magazine 7. no.12

Taubes, G. (1996a). "*Electronic Preprints Point the Way to author empowerment"*. Science, Vol.271, No. 5250, 9th February, 767

Youngen, G., (1997). *"Citation Patterns of the Physics Preprint Literature with Special Emphasis on the Preprints Available Electronically"*. University of Illinois at Urbana-Champaign Physics and Astronomy Library http://www.physics.uiuc.edu/library/preprint.html