Citation-assigned Parsing as Supplements in Concept Space

Daisey Jacobs  
School of Information Science, Human Sciences Building  
Pretoria University  
Pretoria 0002  
South Africa  
jacobsd@postino.ac.za

S.Chandrappa  
Karnataka Legislature Assembly  
Bangalore 560 001  
India  
chandrappa@email.com

P.Pichappan  
Department of Library & Information Science  
Annamalai University  
Annamalainagar 608002.TN  
India  
ppich@vsnl.net

Abstract

The act of referencing to earlier scholarly papers is a norm in scientific literature. The references have been used as a workable parameter to establish semantic links in literature. With the explosive growth of networked information systems available, it is becoming increasingly significant to evolve methods to automatically extract the semantic links of large collections effectively. The online pages have offered the hyperlink mechanism to other online and print resources. When the references and hyperlinks are integrated, the semantic linking would become more meaningful. Through this paper, a prototype architecture is outlined with the results of a pilot study.

1. Introduction

Through this paper, we propose a prototype citation parsing architecture for universal linking of scholarly resources distributed in Web; the architecture aims to be a step towards the universal reference linking analysis proposed by Campbell [1].

A perfect information world is one that provides to all users, the interconnected corpus of knowledge and allows concept navigation across the spectrum of world knowledge. The information community is working in this direction aiming to integrate the information all-inclusive, which would serve the world community.

In a hybrid information world, where we have access to both print and electronic, people may prefer to electronic form over print for the reasons that electronic form has the added advantages than the corresponding print editions. The data available on online usage indicates that online information usage is increasing dramatically and the digital born resources are proliferating at an unexpected rate. Hence interlinking electronic documents is essential in the era, where online scholarly resources tend to rule the information world.

Semantic link analysis is carried out through several measures; for print documents the one, which is used is citation analysis. If a document is available in both print and online, citation analysis of print alone is enough. However the increasing of digital born resources with no analog antecedent of any kind force us to establish a linking mechanism. The hyperlinks available currently in online documents is a beginning in this direction and forms a component in linkage.

2. Background

Web holds a number of scholarly documents that need to be linked and the linking mechanism goes beyond indexing and hyperlinks. In a rich link environment all links are not just equal and the quality links speaks more. Links are a means to offer access to selected works, which is the function of the journals. One scenario in this analysis is that the published 'journal' will become a set of links.[2] Citing an early paper is a norm in scholarly medium and this norm has not changed in online medium of publication. Online documents provide links to print as well as online documents, i.e., to URLs and other files. The online and digital born resources tend to contain more URL links than the print documents.

The World Wide Web is the fabric woven on hypertext connections between documents, but for the scholarly resources, the link par excellence is formal citation of one paper by another. This is the way researchers have naturally been interconnecting their writings all
along, but until now it has only been possible to follow those connections off-line, piece-wise, mediated by a great deal of real footwork in between.[3]

The present web pages links are determined by crawlers, which do not understand the cognitive scholarly link. Citations contained in scholarly publications establish a semantic and cognitive link with other papers and the link is determined by the authors of the content rather than by a third party.

3. Related and ongoing studies

Citation parsing approach is being tested in many testbeds in different directions. In a major JISC-funded Open Journal and CogPrints Projects[4], Wendy Hall, Stevan Harnad, and Les Carr have successfully used citation linking to interconnect a small but interdisciplinary "seed" database of full texts in the Cognitive Sciences with a much larger 10-year set of abstracts and their reference lists from a subset of the ISI (Institute for Scientific Information). However it is limited to abstract linking only. Cornell is one of several places engaged in exploring methods for reference linking the online literature, partly through its involvement in the Open Citation project[5]. The Citeseer is a Autonomous Citation Indexing project aiming to have open archiving of citation linking to all class of literature.[6] The OAI Protocol for Metadata Harvesting [7](Van de Sompel and Lagoze) mandates simple, Dublin Core based metadata to achieve interoperability between archives. The advantage of the Open Archive Initiative is the application of the simple metadata harvesting; however it does not provide mechanisms to expose and harvest full content.[8] (Warner). The directions towards reference linking may yield fruits in the long run and one cannot expect the comprehensive benefits immediately as the citation linking of online resources is complex.

4. Issues in citation parsing

It is an established understandable assumption that the journal literature is intrinsically 'hyperlinked'. To translate the hyperlink structure to reality the online medium is a more natural environment for this feature.[9] (Hitchcock).

Citations allow gauging the information impact of publications and structured in a dynamic way and free from noise. Relevance identification through citations is independent of language, words, terms etc. Citation parsing has advantages over crawlers-fixed relevance. Tracking the online citations is beneficial and it taps the digital born resources, otherwise remain unearthed.

Citation parsing in online is not equal to print citation indexing as processing and retrieval of online requires different approach. Basically elements in online resource vary with the corresponding print versions. Digital-born online journals exhibit a different citation character. Very few electronic journal articles cited other online articles, and that about half of those references were inaccessible later in the same year[10].Preserving hyperlinks from one to another online resource is not in the control of the creator of an online document. Moreover digital information is considered by many to be more fragile than traditional technologies. A perfect online referencing linking has to consider many issues before initiating such an exercise.

5. Aim of the work

The paper will examine and test two possible linking mechanisms; one is citation link as manifested in scientific papers by authors and second one is the hyperlink approaches the current feature of online resources. By combining these two features, we believe that links can be organised in units at a higher level of granularity. This scaling approach would enhance the function of online scholarly resources in a better way. Thus, we intend to index, based on the referencing determined by the authors of web pages and papers which would serve as a comprehensive web description and pave way to web mapping.

The architecture is designed in a way to integrate concepts free from terminologies and navigate them seamlessly. We also decided to differentiate the parsing for citations at the end and hyperlinks at the body of web pages.

6. Architecture

A hyperlink from one to another online paper or URL is interesting and advantage for the user who can travel instantly to the world of interconnections.

We convince that citation-based navigation is really significant when linking arises from full text not limited to citations available at the end of papers. To say explicitly that hyperlinks available in text combined with the references
could benefit much. It would be interesting to perceive that both hypertext in web pages and references are the links determined not by crawlers but by humans.

Citation parsing is a component of establishing navigational structure of online resources, and it has influenced linking mechanisms greatly. We believe that the citation parsing if extended to the hyperlinks offered by authors and creators of online resources, it would firmly lead to the concept navigation across the web world. When the efforts of all groups working on establishing the online linkages are integrated, it would take the users to a universal bibliographic and citation database that would link every scholarly work ever written.

The simple system we intend to follow is given below.

1. Seed Paper (Citing/Cited)
2. Hyperlinks/URLs
3. Citing Seed Paper (3A & 3B Coupled)
4. Offline/Print (4A & 4B Co-cited)

The figure portrays the citation parsing architecture we propose. The seed paper {1}, is one where references and links emerge. {2A & 2B} are the hyperlinks originate from seed to either internal or external links and may have URLs or not. The {2A & 2B} are given in layers as the URLs specified in the seed paper that takes to the concerned URLs and it may be a component of the {2} in certain cases. They constitute web pages, email messages, newsgroup postings etc. {3A & 3B} are the online resources that refer the seed paper subsequently, which is not available in the seed paper as it is forward. The links taking the users to forward prediction is impossible. The {1} functions as both cited and citing paper. When the seed paper is analysed, it is the citing paper. When the papers that cite the seed paper, the {1} is the cited paper. Thus, the architecture ensures that seed paper tracking alone is not sufficient in building linkages. {4A & 4B} are the references offered by the creators of online resources to the off line and print sources. Arrows denote where links emerge and how they proceed. For instance all links emerge from {1}, except the ones from {3}. Since {3A & 3B} cite the seed paper together they are coupled together. {4A & 4B} are cited together by the seed paper. The bond between 4A & B may increase over time as the linkage identification mechanism is dynamic. The architecture ensures scalability in the extent that it takes beyond online linkages.

Mapping of all the links lead to the building of networks and online users travel to comprehend the displayed result list follow the right trail leading to a destination. The choices of routes for users are narrowed down only to the ones that will lead them to relevant information that they are looking for.

7. Candidate/Seed paper for indexing

The heuristics required to parse documents are difficult, as the web pages do not have structured format. All hyperlinks and references identified could be placed under one linking mechanism, which then placed at a two-tier hierarchy. One lists the citations contained in the web papers and another speaks of the hyperlinks in the pages that are also specified by authors.

For locating new papers, one can proceed to the search engines and heuristics. For identifying a seed paper as well as for retrieval of related papers on a concept, we suggest to use either the crawler or a known paper currently published in online. Online journals once fixed for indexing and linking would ensure a routine kind of parsing. Should the parsing is viewed as a regular mechanism, we suggest to begin with the current online journals. For identifying the papers using crawlers, one can use terms in a selected field followed by file extensions such as “.pdf”, “.ps”, “.ps.Z”, or “.ps.gz” etc. To confirm the selected paper as a research document, the programme developed check at the end of the paper, whether it contains references, the character of scientific papers.

8. Linkage Building
The seed paper is used to extract the basic data required for parsing such as the URL of the paper, author, title, format, key terms, abstract, hyperlinks in the text, references and links established by the creator to other URLs irrespective of their availability in text or reference. In the above elements, hyperlinks, references and links to URLs are the determinants of the relevance between one online source and other resources. Thus, link determination is by the creator of the text who only knows the information content absolutely than by a programme. In a similar project by Citeseer, besides the citations in the text, the citation context is also extracted from the text. This extraction is a powerful linking, we agree, however we would wait till the outcome of the project is completely documented.

We understand the limitations in doing so, since the data mentioned above may not be completely extracted due to the variations in format. We are hopeful in minimising the difficulties in element identification when we improve our algorithm. Tags are assigned for the identified fields that would serve as elements input for file creation.

The selected seed paper is scanned for a ‘three levels’ link identification. First, the hyperlinks presented in the text of the seed paper are extracted automatically and indexed in one file. This would ensure that when the elements of the seed papers are the queries, the hyperlinks would take them to the other resources. If the hyperlink directs to an online full text (provided access is permissible), the user can travel to the online in proper path.

The second level linking is for the papers that refer the seed paper. The concept of citation indexing rests firmly on identifying the links between citing and cited papers. In the printed journals, citation linking the source population is fixed and recording of all journals data is a regular exercise, which helps in establishing links. However for online journals we do not have currently any fixed population and indexing is not carried out. Moreover, fixed population undermines the basic principle of universal indexing and thus we have resorted for extracting the links for the second stage through the following way. The online resources that cite a seed paper is infinite and no mechanism is available to detect the linkages, we have resorted to the following method. We searched the online resources using search engines by initiating the search with the terms such as author, title, heading and the years following the publication of the seed paper to narrow down the search. Fixing of tags for the field, ‘references’, identifies the references in the retrieved resources. The references are stored in separate files. For the third type of linkages fixing, as specified in the first stage, the identified URLs are separated and kept in one file. We have proposed the link fixing for the off line sources cited by the seed paper. All the three files now have the heading for the seed paper with key terms and other basic description. Any one searching the terms in the retrieval package would retrieve the seed paper as well as the linkages. The searching may translate words into concepts in searching and concepts into words in retrieval. The resulting structure we have retrieving in the pilot testing is given below.

9. RETRIEVAL RESULTS FROM LINKAGES

<table>
<thead>
<tr>
<th>References(1)</th>
<th>Hyperlinks(2)</th>
<th>URLs(3)</th>
<th>Seed Paper (cited/citing paper)(4)</th>
</tr>
</thead>
</table>

Citation Links (citing paper)


Hyperlinks in Seed paper


Linked URLs: in seed paper


The analysis of the seed paper has given a body of 43 references, 123 hyperlinks and 12
URLs. The online linkages alone are given above.

This study is part of an on-going research in an effort to build a model for comprehensive reference linking. Our short-term objective in this research is to introduce and test the mechanism to the initial extent and we are improving the mechanism.

10. Summary

The human mind operates rapidly and instantly from one concept structure to other, the web navigational structure can correspond to this view implies creating an effective architecture and implementing it. Thus, it is ideal to construct architecture initially and develop programmes to reflect the architecture.

We emphasise the importance of relevance measurement in concept space as the perfect way in drawing links. Since the architecture proposed rests firmly on the online creators based, the reliance of the linkages tends to be high.

The above presentation is the reflection of an ongoing work on building a citation-assigned parsing as a supplement to concept navigation in web. We are not sure the rate of success in exploring the ultimate link. Hence we are precautious in terming this exercise as supplement rather than the solution.

The online resources linking is at present carried out in different directions by different initiatives. Unlike in print journals reference indexing where a consistent approach is followed, in the projects funded by such as JISC, NSF and CNRI, the linking mechanisms adapted is not uniform. As the online resources are increasing exceedingly, a unified approach is essential.

Content integration will score better than information integration, which would drive the online resource linking process, and it is the real virtual link in the web world.

11. References


[3] Integrating and navigating eprint archives through citation-linking: the open citation linking project NSF / JISC - eLib Collaborative Project: International